Testing Reading Through Summary

Investigating summary completion tasks for assessing reading comprehension ability

For a complete list of titles please visit: http://www.cambridge.org/elt/silt

Also in this series:

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

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

Aligning Tests with the CEFR: Reflections on using the Council of Europe's draft Manual
Edited by Waldemar Martyniuk

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

Examining Speaking: Research and practice in assessing second language speaking
Edited by Lynda Taylor

IELTS Collected Papers 2: Research in reading and listening assessment
Edited by Lynda Taylor and Cyril J. Weir

Examining Listening: Research and practice in assessing second language listening
Edited by Ardeshir Geranpayeh and Lynda Taylor

Exploring Language Frameworks: Proceedings of the ALTE Kraków Conference, July 2011
Edited by Evelina D Galaczi and Cyril J. Weir

Measured Constructs: A history of Cambridge English language examinations 1913–2012
Cyril J. Weir, Ivana Vidaković, Evelina D Galaczi

Cambridge English Exams – The First Hundred Years: A history of English language assessment from the University of Cambridge 1913–2013
Roger Hawkey and Michael Milanovic

Testing Reading Through Summary

Investigating summary completion tasks for assessing reading comprehension ability

Lynda Taylor
Senior Lecturer, Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire
and
Consultant, Cambridge English Language Assessment (part of the University of Cambridge)

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107695702

First published 2013

Printed in the United Kingdom by XXXX

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

Library of Congress Cataloguing in Publication data
Taylor, Lynda B.
Testing reading through summary : investigating summary completion tasks for assessing reading comprehension ability / Lynda Taylor, Senior Lecturer - Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire and Consultant - Cambridge English Language Assessment (part of the University of Cambridge).
pages cm. -- (Studies in Language Testing ; 39)
Includes bibliographical references and index.
ISBN 978-1-107-69570-2
1. Reading comprehension--Ability testing. 2. Reading comprehension--Evaluation. 3. Reading--Ability testing. I. Title.

LB1050.46.T39 2013
372.47--dc23
2013021187

ISBN 9781107695702

For my parents – Frank and Kay,
who introduced me to the joy of reading
and taught me to see a world in the words on a page

Contents

Acknowledgements

The usual advice given to a student on completing their PhD is to publish all or some of it as quickly as possible, either as a monograph or in the form of journal papers. In my own case, professional responsibilities soon took priority and the likelihood of publishing anything from my doctoral dissertation on the testing of reading receded steadily into the distance.

In recent years, however, the focus and outcomes of my earlier research exploring approaches to assessing reading comprehension ability have assumed a fresh relevance and I believe they now represent a timely and useful contribution to our current theory and practice in applied linguistics and language testing. The reason for this is that the past decade has seen a burgeoning interest in the cognitive processing that underpins language use, particularly what distinguishes an expert from a novice user. Increasing attention is also being focused on how far such cognitive processing is adequately represented in approaches to assessing language proficiency. There is growing understanding of the cognitive processes that are typically activated in test and non-test tasks, and of the importance of an appropriate match between the two. This concern for construct validity is particularly relevant when seeking to assess comprehension ability, given that language testers rely heavily upon indirect measures to make visible a process and product located inside the head of the language user.

The nature of comprehension ability and the suitability of approaches for evaluating it are issues with which applied linguists, cognitive psychologists and language testers have grappled for many years. Language testers generally have no difficulty in designing reading tests that activate the lower-level perceptual processes of decoding, lexical access and syntactic parsing, but there remain significant challenges in creating test tasks that successfully activate higher-level conceptual processes of meaning construction and discourse representation in reading, processes which are critical at higher proficiency levels, for example in academic study contexts.

At the University of Cambridge's Research Centre in English and Applied Linguistics (RCEAL), I was fortunate to be able to spend time researching these issues under the supervision of Alastair Pollitt, Senior Lecturer at RCEAL and also Director of the Research and Evaluation Division at the University of Cambridge Local Examinations Syndicate (UCLES). I also benefitted greatly from the research interest and support of Professor Gillian Brown, an eminent authority in text comprehension and discourse analysis,

and Dr John Williams, an experienced cognitive psychologist with expertise in both first and second language acquisition. My overall aim was to investigate what readers actually do when they comprehend a text and to explore the potential for a test task that would reconcile more closely the practice of assessing reading comprehension ability with a sound understanding of the nature of reading comprehension.

Professor Cyril J Weir persuaded me to revisit my PhD research and to remodel the original doctoral thesis in order to make it suitable for publication, by updating the literature in a number of areas and by contextualising the findings for the present day. I am most grateful to him for his encouragement to publish and for his editorial guidance on the manuscript as a whole, as well as to various CRELLA colleagues who assisted with specific chapters: Dr John Field reviewed and advised on Chapters 1 and 2, with regard to the psychological and cognitive dimensions of reading; and Professor Stephen Bax offered input to Chapter 3 on approaches to text and discourse analysis.

It would not have been possible for me to undertake the original research reported in this volume without the support, co-operation and encouragement of various people.

UCLES was extremely generous in providing funding for the doctoral research. In particular I would like to thank Dr Michael Milanovic, Chief Executive of Cambridge English Language Assessment, for the interest and support he showed throughout the study, and indeed throughout my career in language testing over nearly 25 years. Much of my professional life has involved working in close association with Cambridge so it is a privilege and a pleasure to have the opportunity of publishing my work as a volume in the *Studies in Language Testing* series in the year when Cambridge celebrates its 100th anniversary.

Grateful thanks are also due to Professor Charles Alderson whose own work in the field of reading assessment influenced my early interest and enquiry in this area. Professor Alderson kindly served as my external PhD examiner and his encouraging and insightful comments helped inform revisions to the manuscript for this volume.

Special thanks must be recorded to the staff and students at Hills Road Sixth Form College, Cambridge, who were involved with the oral recall study, and to the staff and pupils at Sawston Village College who took part in the written recall study and in the trialling of the two summary completion tasks.

Finally, the publishers are grateful to the copyright holders for permission to use the copyright material reproduced in this book: Macmillan Caribbean for the use of the short story 'Journey by Night' by Undine Giuseppi, and *The Independent* for the use of the editorial piece 'The rights and wrongs of treating anorexia'.

Series Editors' note

Since its inception in 1995, the *Studies in Language Testing* (SiLT) series has published many PhDs of quality. One of the core purposes of this innovative and now well-established series is to support and promote work in the field of language assessment by enabling the language testing community to benefit from research which makes a contribution to the field but which might not otherwise reach publication. PhDs are selected for inclusion in the series in accordance with certain criteria which include:

- being a contribution to knowledge
- being previously unpublished
- having a sound theoretical basis
- being well-referenced to the literature
- being research-based
- being executed with care and rigour
- demonstrating analysis and interpretation which is well-founded
- having the style of an academic monograph.

The first PhD we published was by Anthony John Kunnan on test taker characteristics and test performance (SiLT volume 2). Eight other PhD theses have been published to date. Caroline Clapham documented the development of *IELTS* (*International English Language Testing System*) and looked in particular at the effect of background knowledge on reading comprehension (SiLT volume 4), while Anthony Green investigated the impact of the *IELTS* writing subtest on English for Academic Purposes pedagogy (SiLT volume 25). James Purpura investigated learner strategy use and performance (SiLT volume 8). Kieran O'Loughlin compared direct and semi-direct tests of speaking (SiLT volume 13) and Angela Hasselgreen looked at testing the spoken English of young Norwegians (SiLT volume 20). Dianne Wall and Liying Cheng both investigated aspects of test washback and impact, with Wall studying its effects on the classroom in Sri Lanka (SiLT volume 22) and Cheng carrying out a study on the classroom in Hong Kong (SiLT volume 21). Toshihiko Shiotsu examined the components of L2 reading ability in the context of Japanese learners of English (SiLT volume 32). A number of these theses were also awarded the Jacqueline Ross TOEFL Outstanding Doctoral Dissertation award. SiLT policy is to publish one PhD for every three or four SiLT volumes and in successfully doing this we have enabled high-quality

doctoral research to reach a wider audience than would normally be expected. In this volume we continue this tradition and publish Lynda Taylor's PhD thesis on *Testing Reading Through Summary: Investigating summary completion tasks for assessing reading comprehension ability*.

The publication of a study on the use of summary is timely given it was a testing device in vogue for the first three quarters of the 20th century, disappeared from view in the "communicative revolution" of the 1970s and then re-emerged on the global stage at the start of the 21st century.

At the beginning of the 20th century, précis featured widely in many English for Specific Purposes (ESP) and educationally-oriented tests. Robeson (1913), a Master at Eton, describes its use in Civil Service, Army and Navy qualifying examinations, in the commercial and teacher awards and examinations of the Royal Society of Arts and the London Chamber of Commerce, and by the Oxford and Cambridge Schools Examination Board in its Examination for School or Leaving Certificates. He quotes (1913:9) from the London Chamber of Commerce Regulations:

> [T]he object of the précis is to enable any one reading it to be put into possession, in the smallest space of time, of the essential points of the subject to which the documents refer. The characteristics of a good précis accordingly are (a) the inclusion of all that is important and the exclusion of all that is unimportant in the correspondence; (b) the expression of this in a consecutive story as clearly as possible, and as briefly as is compatible with distinctness.

Summary tasks, which by necessity involved reading comprehension at the global text level, were included as test formats in Cambridge examinations from the 1930s onwards. In 1931 a précis of a passage or a poem had been introduced into the English Literature paper in Part B. Typically, candidates had to choose between summarising a passage, which included defining the meaning of words and phrases in the text, and explaining a poem in detail including a focus on style and diction. In addition, by 1936 there was a further summary task in the English Essay paper.

Summaries were viewed in the school system as valuable, integrated tasks and an appreciation of the validity of the task took precedence over any concern with difficulties of marking. The tasks were intended to test comprehension of a whole passage (careful reading at the global as well as the local level) as well as writing ability and this stands in stark contrast to the emphasis on decoding in many tests of reading and in the research literature in the first half of the 20th century.

Summary was to last as a task in the Certificate of Proficiency in English (CPE) right through to 1975. Given the continued use of CPE for university entrance in the 21st century, the demise of such an authentic academic reading-into-writing task may, with the advantage of hindsight, be regretted

(note, however, its return to favour in 2002 albeit in a reduced intertextual form in the CPE Use of English paper). Cambridge had not been alone in abandoning summary and the well regarded Schools Council Research Studies monograph entitled *The Development of Writing Abilities (11–18)* (Britton, Burgess, Martin, McLeod and Rosen 1975) contains only one brief and fairly disparaging reference to summary on page 47. From a present day perspective, we would argue that (albeit in an integrated format) summary effectively tests the important advanced level reading skill of creating a text level representation, a vital element of academic study, in an authentic manner.

As well as its rebirth in Cambridge examinations, summary has been introduced into the Internet-Based Test of English of a Foreign Language (TOEFL iBT) and the General English Proficiency Test (GEPT) Advanced Reading Test in Taiwan in the 21st century, but relatively little serious research has been carried out on summary as a testing task to date. The literature on the assessment and teaching of reading is immense so it is perhaps surprising to find so few studies available relating to the use of summary as a measurement tool. Their absence is even more surprising given a nascent concern with the construct validity of test formats employed for assessing reading ability in the 21st century.

An overt concern with the constructs being measured in the Cambridge English examinations and their relationship to real-life language use was apparent by the end of the 20th century. The commitment to transparency and the explicit specification of the communicative content of its examinations was further enhanced by Cambridge's adoption of a socio-cognitive approach to language test design and validation in the first decade of the 21st century; such an approach acknowledges that language use constitutes both a socially situated and a cognitively processed phenomenon and that this must be reflected in language assessment theory and practice.

The increased attention paid to cognitive validity at Cambridge came about as a result of a 10-year project (2003–2013) which saw the publication of the 'construct-focused' volumes in the SiLT series (SiLT volumes 26, 29, 30 and 35), guided by Michael Milanovic, Nick Saville, Lynda Taylor, Evelina Galaczi and Cyril J Weir on the editorial steering committee. This ambitious project enabled far greater attention to be paid than previously to the cognitive processing typically activated in test and non-test tasks, and to the importance of an appropriate match between the two. There is now a growing recognition within Cambridge English Language Assessment and its partners, and in the wider international testing community, of the importance for any successful assessment system of seeking and assembling validity evidence on each of the three core aspects of validity: cognitive, context and scoring, which together constitute test construct validity.

Lynda Taylor's PhD thesis was very much ahead of its time when it was

conceived and written in the mid-1990s given that at the time very few language testers and even fewer examination boards paid any serious attention to the cognitive processing underpinning the tasks they employed. Indeed, the cognitive validity of the tasks used in most tests of reading comprehension is a concept that many language testers and examination boards are still struggling to come to terms with. Lynda Taylor tackled the issues involved in addressing this critical component of construct validity for reading tests head on in her thesis at Cambridge University under the supervision of Alastair Pollitt, then Director of the Assessment Division at the University of Cambridge Local Examinations Syndicate (UCLES), and with the support of the other members of her Research Committee, Gillian Brown, an established authority in text comprehension and discourse analysis, and John Williams, an eminent cognitive psychologist. Lynda investigated a test format that reconciles, more closely than any other alternative format, the practice of assessing reading comprehension ability with our current understanding of the nature of reading comprehension. The editors were thus very pleased when she agreed, after much cajoling, to revisit her thesis a number of years later and remodel it as a book for the SiLT series on the use of summary as a language testing task for measuring reading comprehension ability. They felt it would help address an important gap in the research literature on the testing of reading and further ground the need in test design to take account of what learners actually do when they comprehend a text.

Taylor argues that the main experimental aim of the series of studies she carried out was to investigate the key features of readers' mental representation of text and to identify how best to develop a summary completion task which directly addressed those understandings. Developing such a task involved exploring different readers' mental representations of a given text to identify what constituted an adequate verbal summary version of the text in question. Secondly, it required the construction of suitable test items from within the resulting summary which could be used to assess readers' comprehension of the text.

She describes how readers' mental representations of two different texts, one narrative and one expository, were explored through a series of studies and how a text-removed summary completion task was developed to accompany each text. The two summary completion tasks were then trialled on a population of readers and the results from this exercise were compared with an independent measure of reading ability for the same population to determine the effectiveness of the text-removed summary completion format as a measure of reading comprehension ability. An accompanying aim of the investigation was to establish some practical guiding principles for the construction of summary completion tasks.

In Chapter 2 of this volume Taylor reviews the development of different theories of reading and text comprehension over the past century. Particular

attention is paid to the active and constructive nature of the comprehension process, in which meaning is constructed by the reader's cognitive processes interacting with their knowledge base and personal goals. Chapter 3 offers a survey of reading test design, briefly chronicling the historical developments which have led to current practice and it deals in more detail with issues relating to construct validity. In Chapter 4 the rationale for using summary writing tasks as a means of assessing reading comprehension ability is explored along with the problems. An alternative approach – summary completion technique – is considered and the research questions for an empirical study are then presented. Chapter 5 reports on a text recall study designed to investigate readers' mental representations following the reading of two different texts, Text A (*Journey*) and Text B (*Anorexia*). Chapter 6 presents the detailed results of the text recall study for each of these texts, based upon an analysis of readers' mental representations in terms of text-based (micro-) propositions, summarising (macro-) propositions and additional propositions occurring in their oral recalls. Chapter 7 reports on the design of two text-removed summary completion tasks using summaries derived directly from the readers' shared mental representations of Texts A and B. Chapter 8 reports the results from trialling the two summary completion tasks with a population of readers and the concluding Chapter 9 summarises the main research findings of the study, discussing their implications and making suggestions for future areas of research.

This volume offers examining boards as well the teacher in the classroom both practical and theoretical support for developing summary completion tasks to assess reading comprehension. In so doing, it affords them the possibility of employing a task which has potentially greater claims to the mantle of cognitive validity than many other formats in common use for assessing reading comprehension ability.

Cyril J Weir and Michael Milanovic
March 2013

Abbreviations

AEB	Associated Examining Board
BNC	British National Corpus
CEFR	Common European Framework of Reference
CPE	Certificate of Proficiency in English
DIALANG	Diagnostic Language (Assessment)
EAP	English for Academic Purposes
EFL	English as a Foreign Language
ELM	English Language Monitoring (Project)
ELTS	English Language Testing Service
ESOL	English for Speakers of Other Languages
ESP	English for Specific Purposes
GCSE	General Certificate of Secondary Education
GEPT	General English Proficiency Test
HF	High Frequency
IELTS	International English Language Testing System
ISLC	International Study and Language Centre
KR20	Kuder-Richardson 20
KS3	Key Stage 3
L1	First Language
L2	Second Language
MA	Master of Arts
MC	Multiple Choice
MCQ	Multiple Choice Question
MF	Medium Frequency
MR	Macro Rule
PDP	Parallel Distributed Processing
PTE	Pearson Test of English
SP	Summarising Proposition
TEEP	Test of English for Educational Purposes
TOEFL iBT	Internet-Based Test of English as a Foreign Language
TP	Text-based Proposition
TTR	Type Token Ratio
UCLES	University of Cambridge Local Examinations Syndicate
VPA	Verbal Protocol Analysis

1 Introduction

Background

In a recent volume chronicling the historical evolution of assessment constructs and the way these are operationalised through language tests, Weir (2013b) noted how for much of the 20th century the teaching and testing of reading focused on lower-level processing to extract factual meaning at the clause and sentence level, rather than on higher-level processing to combine and integrate text-based and reader-based knowledge sources in order to construct a meaning representation for a text, or across a set of texts. By the 1970s, however, the focus was beginning to shift. Weir highlights an editorial published in issue 15 of *Reading Research Quarterly* (1980:181–182), under the heading 'Why comprehension?', in which the editors noted how the earlier focus in reading was giving way to a new emphasis on comprehension. With greater attention being paid to research into comprehension, i.e. exploration of the cognitive processes involved in meaning construction and the skills and strategies involved, the field of reading began to hold greater interest for language teachers and testers than when the focus had been more narrowly limited to the lower-level processes (Urquhart and Weir 1998).

Against this background, the last 40 years have witnessed significant expansion in the volume of empirical research conducted in the field of reading assessment. Many of the question formats commonly used in reading tests have been the subject of intense scrutiny with regard to issues of validity. Multiple-choice and cloze, in particular, were the focus of considerable attention during the 1980s and 1990s, with large numbers of studies devoted to analysing the efficiency of multiple-choice items or the relative merits of one cloze format over another (Alderson 1980, Bachman 1982). Other research began to explore the role of cultural or background knowledge in a reading test (Clapham 1996), the nature of test taker strategies when assessing reading (Cohen 1984) and the value of reading-into-writing tasks within an academic study context (Hill and Parry 1992), and this continued into the 21st century. More recently, greater attention has focused on systematically investigating the cognitive processes utilised by test takers during a reading test, in particular how these can be affected by the question formats employed (see Khalifa and Weir 2009 for a full discussion of this with an extensive list of references).

Expansion has also taken place more broadly in all areas of both first and

second language reading research. The second half of the 20th century saw advances in the development of theories and models of reading, a trend which continues to this day. In line with greater interest in higher-level processing (as against the lower-level processes of decoding, parsing and extraction of local factual information), considerable attention was directed towards trying to identify and describe the component processes of reading for meaning at the level of discourse construction, as well as towards finding an appropriate model to describe and explain the nature of reading comprehension. Text comprehension models hypothesised the active and constructive nature of the comprehension process in which meaning was generated by the cognitive processes of the reader in association with contextual features of a text. Using text together with pre-existing knowledge, the reader was increasingly perceived as building a personal mental representation which may be modified by the attitudinal characteristics and intentions of the individual.

In light of these developments, it is reasonable to suggest that reading assessment theory and reading comprehension theory must surely overlap and that research in one field is bound to be of direct relevance and value to the other. We might expect there to exist between these two fields a strong reciprocal relationship, through which advances in our understanding of reading processes and products are directly reflected in developments in our reading assessment theory and practice. This has not always been the case, however, and a significant gap has sometimes been perceived to exist between theories of reading comprehension on the one hand and the practice of assessing of reading comprehension ability on the other. One result of such a mismatch is that approaches to reading comprehension assessment risk being undertaken without sufficient regard to latest understanding about the process of reading comprehension based upon empirical research findings.

The aims of the theoretical and empirical research reported in this volume are twofold. First, to examine in greater detail the gap which can exist between theories of reading comprehension on the one hand, and the practice of assessing reading comprehension ability on the other. Secondly, to explore the development of an approach to assessing reading comprehension ability which takes fuller account of how readers actually process and comprehend written text.

The gap between reading comprehension theory and reading test theory

Comments in the reading research literature from the 1980s onwards indicate that various researchers perceived a gap to exist between theories of reading and the theory and practice of reading test design. Farr and Carey (1986) and Anderson, Bachman, Perkins and Cohen (1991) concluded that reading tests had not changed significantly in the previous 50 years and had not therefore

responded to changes in how comprehension was increasingly being understood. Anderson et al (1991:41) commented as follows:

> . . . while models of reading have evolved, changing our thinking about how the printed word is understood, the tests that we use to measure that understanding have not changed significantly. It would thus appear that an examination of the construct validity of current reading tests, vis-a-vis current reading theories, is in order.

In an article calling for a substantial review of approaches to reading assessment, Valencia and Pearson (1987) argued that reading assessment had not kept pace with advances in reading theory, research or practice. The authors suggested at least 11 different features of reading assessment practice which they believed were at direct variance with latest views of the reading process.

Over the following decade, and despite expanding research in the areas of both cognitive and educational psychology, scholars working in the field of language pedagogy and assessment continued to perceive an apparent disconnection between research into the nature of reading and the impact of this upon approaches to assessing reading ability, even if it was beginning to have some influence upon the teaching of reading skills. Grabe (2000:11) commented:

> One strong outcome of this research has been its impact on reading instruction, particularly with respect to greater emphases on word recognition abilities, vocabulary knowledge, strategic processing and awareness of discourse organising principles. It is probably safe to say, however, that there has not been a similar impact on reading assessment.

Alderson (2000:110) also referred to a 'disjunction' between research into reading and research into the testing of reading (though see his more nuanced view on this on page 7).

Explaining the gap between reading comprehension and reading test theory

One reason for a perceived gap between reading comprehension theory and reading test theory and practice may have been the nature of much reading research, particularly its preoccupation with theoretical issues of cognitive processing in reading at the expense of more applied issues in education. During the 1970s and 1980s reading research was primarily the domain of cognitive psychologists and it is possible that some educational theorists and practitioners may have felt marginalised as a consequence.

In a guest editorial for *Reading Research Quarterly*, Vacca and Vacca (1983) complained that, despite advances during the 1970s in research into the basic processes of reading comprehension, applied research issues relating to reading instruction and development remained relatively neglected:

> Applied research questions were dismissed as premature, perhaps even unimportant, as theoreticians and researchers began from the ground up to build and verify theories of the reading process (1983:382).

Taking a similar perspective, Pearson (1979) concluded:

> Too often we have assumed that we must settle issues of basic research before we can tackle issues of applied research . . . such a delay in facing applied research questions may be inadvisable as well as unnecessary (1979:166–167).

Vacca and Vacca (1983) suggested that what was lacking during the 1970s was sufficient bridging between basic and applied research in reading and that an improvement in this situation needed to be a priority for the future.

If it is true that reading research from the late 1960s onwards focused heavily upon modelling the reading process with little reference to applied issues of reading instruction and development or its assessment, then it is perhaps not surprising that applied issues of reading assessment also remained relatively neglected for many years by mainstream reading research. This view was espoused by Valencia and Pearson (1987:727) who suggested that, even though the fruits of reading research were beginning to benefit instructional research, materials and practice in the 1980s, assessment continued to lag behind:

> The advances of the last 15–20 years in our knowledge of basic reading processes have begun to impact instructional research (Pearson, 1985) and are beginning to find a home in instructional materials and classroom practice (Pearson, 1986). Yet the tests used to monitor the abilities of individual students and to make policy decisions have remained remarkably impervious to advances in reading research (Farr and Carey, 1986; Johnston, in press; Pearson and Dunning, 1985).

The suggestion so far has been that much reading research undertaken during the 1970s and 1980s concentrated so heavily upon explaining the basic processes of reading that it had relatively little to say to those involved in applied reading issues as far as instruction and assessment were concerned. It may not be fair, however, to lay blame for limited cross-fertilisation of ideas on reading entirely at the door of cognitive psychologists who were investigating the reading process.

A second and possibly related reason for the perceived gap may have been the way in which some educational theorists and others interested in applied issues of reading (e.g. teachers, syllabus designers and testers) were choosing to deconstruct and analyse the activity of skilled reading. Valencia and Pearson (1987) suggested that the influence of mastery learning during the 1960s, at least in the USA, led to a tendency to conceptualise reading as the mastery of small, separate enabling skills and to regard skilled reading as an aggregation (rather than integration) of these skills. A similar view of reading, i.e. as the aggregation of separate and definable sub-skills, was being developed simultaneously in Britain, particularly with regard to reading in the second language (L2).[1] A taxonomic or hierarchical approach to describing reading sub-skills was becoming increasingly popular (e.g. Davis 1968, Munby 1978), partly because of its potential for ready application in syllabus and course design. A direct legacy of this emphasis upon reading sub-skills was that reading tests were often constructed to test different and specific sub-skills in relative isolation, focusing heavily upon the informational purpose for reading and relying on items testing aspects such as 'literal comprehension' or 'finding the main idea of a paragraph'. It may be important, at this point, to distinguish between the vague notion of sub-skills which covered a multitude of different types of operation from the development of test criteria based upon reader goals. The latter quite closely parallel a movement in listening, where test providers increasingly relied upon *listening for* categories, e.g. *listening for gist*, *listening for information* (see Field 2013 for more discussion). Several published volumes offer comprehensive historical overviews of the sub-skills approach to defining reading ability (see Alderson 2000, Grabe 2009, Grabe and Stoller 2002, and Urquhart and Weir 1998). The practical impact of the reading sub-skills paradigm on task formats in reading tests will be considered more fully in Chapter 3.

A third explanation for the gap may have been that the practice of comprehensive and multi-faceted construct validation of reading tests is a relatively recent development in the language assessment field. Traditional approaches to construct validation tended to be fairly narrow in their focus, typically paying close attention to test content (in terms of representation and relevance), to item and test scores, and to the statistical relationship between these, often through the use of factor analysis. Khalifa and Weir (2009) noted how a *post hoc* factorial approach to defining reading comprehension tended to dominate research into the testing of reading from the 1960s onwards. This approach was based mainly upon a divisibility hypothesis according to which reading ability could be subdivided into various components, each of which could be tested independently and then confirmed by means of quantitative statistical approaches, such as factor analysis. Khalifa and Weir highlighted the limitations of such a psychometrically driven approach due to its heavy focus upon factors that can be shown statistically to contribute to successful

reading test performance while taking little account of the actual components of the reading processes that are necessary for successful comprehension:

> The approach might be described as focusing upon a product in the form of the outcome of a test rather than upon the process that gave rise to it. Thus the data examined is a measure not of successful reading *per se* but of successful performance in the test. The factors underlying the latter do not necessarily hold true for reading activities that take place in the real world (2009:37).

Field (2011) also cautioned against an over-reliance upon seeking to track back from a product or outcome to the process that gave rise to it. He defined criteria for judging test validity as follows:

> The goal is to establish whether the tasks proposed by the test designer elicit mental processes resembling those which a language user would actually employ when undertaking similar tasks in the world beyond the test. The processes in question might relate to the way in which the user assembles or interprets input; or they might reflect the cognitive demands imposed upon the user by facets of the task (2011:67).

It is always possible, of course, that the cognitive processing involved in a reading test only became a significant focus of interest and concern for language testers as suitable methodologies for investigating this emerged during the 1980s. Green (1998) reported how the methodology of verbal protocol analysis (VPA) was being used increasingly through the 1980s in cognitive, educational and social psychology to explore aspects of learning and problem solving as well as differences between expert and novice behaviours. VPA was also used to study both text comprehension (Ericsson 1988, Laszlo, Meutsch and Viehoff 1988) and second language acquisition (Cohen 1986, Faerch and Kasper 1987, Seliger and Shohamy 1989). Its application to the field of language assessment was still quite limited in the 1980s, though see Alderson (1988) and Cohen (1984). The use of VPA to explore tests of second language reading and listening comprehension expanded from 1990 onwards (see, for example, Anderson et al 1991, Buck 1991 and Kobayashi 1995). Green's 1998 volume helped to strengthen the role of VPA in language testing research by presenting and reviewing several empirical studies that specifically used this methodology for construct validation purposes.

It is likely that the traditional preoccupation with issues of psychometric validity and reliability in language testing was also linked with a concern for administrative and economic efficiency in assessment, especially for the large scale testing of reading ability. While Grabe (2000) suggested such an approach was understandable, he also hinted at the potential risk this could pose for construct validity:

> Simple and straightforward measures of main idea and detail comprehension questions on passages, combined with sections on vocabulary, provide strong reliability and at least arguable validity for these testing approaches. The traditional approaches are also popular because they are easy to administer, to score and to scale, and they are economical (2000:35).

The priority, then, in reading assessment has usually been to select reading test tasks that demonstrate psychometric rigour and promise administrative efficiency, rather than design tasks which entail the full range of mental processes typically found in reading activities in the world beyond the test. This tendency led Urquhart and Weir (1998) to express concern that reading tests often failed to sample the full range of real-world reading skills, particularly careful and expeditious reading activity at both local and global level.

Alderson (2000), however, defended the importance for testers of a strong concern for reliability. He also questioned the overall assumption that reading research 'must necessarily impact on research into the assessment of reading' (2000:111), pointing out that the relationship between reading research and research into assessment is inherently two-way, rather than uni-directional, since any research depends upon assessment measures in order to collect the required data.

Despite differing views on the precise nature of the relationship between the outcomes of reading research on the one hand and the theory and practice of reading assessment on the other, the past decade has seen increased efforts to align these two fields more closely for mutual benefit, and to develop new instruments for measuring reading comprehension ability with both pedagogic and research applications. Such efforts have also involved a reappraisal of the theory and practice of construct validation in language assessment, not only for reading tests but also for tests of the other language skills. Weir (2005), for example, was among the first to offer a systematic framework for test development and validation, grounded in the latest theoretical and empirical research, which acknowledged language use as both a cognitively derived and a socially situated phenomenon. The framework can be applied in practice as a methodology for developing language tests and assembling the validation evidence needed to underpin claims about their quality and usefulness. The application of a socio-cognitive approach to developing and validating reading tests was fully articulated in Khalifa and Weir (2009) and has particular relevance for the research reported in this volume.

The constructive and unobservable nature of reading comprehension

Over 25 years ago, Vincent (1985) suggested that the starting point for creative professional initiatives in reading assessment needed to be a thoughtful

and rigorous analysis of what is meant by 'reading'. Thus any new initiative for reading test design requires first of all a detailed appraisal of our current understanding of the nature of reading comprehension.

A recurring feature of attempts to describe the nature of the text comprehension process is the use of terms that reflect *a process of construction*. Gernsbacher's seminal (1990) volume described language comprehension as 'structure building' (see also discussions in Brown, Malmkjaer, Pollitt and Williams (Eds) 1994 and in Kintsch 1998). More recent accounts of how discourse is constructed can be found in Zwaan and Rapp (2006), Long, Johns and Morris (2006) and in Spivey, McRae and Joanisse (Eds) (2012). Extensive research has been undertaken in both cognitive psychology and applied linguistics into the way in which readers integrate the text base with their world knowledge and experience to shape their understanding in both first and second language contexts. As we shall see in Chapter 2, cognitive psychologists and applied linguists alike generally regard the process of text comprehension as active and constructive, according to which meaning is generated by the cognitive processes of the reader using elements of text content, background knowledge and personal goals to construct a mental model which in some way represents their understanding of the text. (For comprehensive overviews and discussion of relevant research in this area, see Alderson 2000, Clapham 1996, Field 2004, Grabe 2009 and Khalifa and Weir 2009).

One possible disadvantage of using a construction metaphor to describe the nature of comprehension is that it suggests a reader's mental representation to be rather fixed or static. It is important to recognise that any mental representation is likely to be quite flexible or fluid, with the potential for being influenced and modified in various ways, both during and after reading, subject to the effect of a wide range of factors, including purpose for reading, integration of existing knowledge, and the processing of unfolding text (Gernsbacher 1990). Even the presence of comprehension questions about the text has been shown to affect the ongoing construction of a test taker's mental representation (Gordon and Hanauer 1993, 1995).

An obvious problem in any attempt to assess reading comprehension ability stems directly from the nature of comprehension itself. Comprehension is essentially an invisible process that takes place inside the head of a reader or listener. It generates an invisible product. Neither process nor product lends itself to external observation. Neville and Pugh (1982) observed that the output of reading is difficult to capture precisely because real-life reading comprehension leads to some modification of the conceptual system. Any attempt at direct assessment of the reading comprehension ability trait is impossible because it is 'a mental operation which is unobservable' (Gordon 1987:5, cited in Anderson et al 1991:44). What is required is some indirect means of making the outcome of comprehension visible to an assessor in a way that is not totally unnatural.

One idea for achieving this has been to ask readers to produce a summary of what they have read as evidence of their comprehension. This approach is attractive in as much as it has an authentic quality to it – what might be termed ecological validity. Within an educational context, for example, readers frequently have to make a summary of a text they have read, although such a summary is likely to be in note form and for their own purpose, rather than in continuous prose for the benefit of someone else to read. A written summary of a reading text (whether in note form or in continuous prose) can nevertheless be considered as the reader's attempt to put into words the mental representation they constructed as a result of reading. It can justifiably be regarded as evidence of the nature and extent of their understanding of a given text and, by extension, of that reader's ability in general to comprehend similar texts.

While this approach presents an intuitively satisfying and convenient format for assessing comprehension, it is also a test format which poses significant problems due to its compositional nature which means that reading skills are conflated with writing skills, or what Weir referred to as 'muddied measurement' (1990:85), i.e. the contamination of the measurement of one skill by the involvement of another or other skills at the same time.

A number of empirical studies have been carried out among both first and second language readers to investigate and describe the processes involved in the activity of summarising. These will be reviewed in Chapter 4 of this volume where the usefulness of summary writing as an appropriate test format for assessing reading comprehension ability will be discussed further, along with considerations of its drawbacks.

Developing a reading comprehension test format that requires a reader to develop a mental representation of a text

Given compositional and other difficulties associated with producing a written (or even an oral) summary of a reading text, one way of resolving this dilemma could be to provide readers with a gapped summary of a text they have read and then ask them to complete the gaps in the summary by inserting missing words or phrases, drawing on their understanding of the original text. With this approach, it would be important for the gapped summary to map *directly* onto the typical mental representation that is generated when reading the source text. Furthermore, all missing information in the gapped summary should ideally correspond to what *most* readers would consider to be salient features of the original text content. Finally, the completion of the gaps in the summary should only be possible based upon the reader's understanding of the source text and not on other types of cue, such as the local co-text or the reader's background knowledge.

A particular advantage of this approach in the assessment context is that it employs an item-based format in which each missing word or phrase within the gapped summary constitutes a single test item that can be objectively scored according to a predetermined mark scheme. This avoids the evaluation problems typically encountered when marking a written summary of a reading text, while test development and equating procedures become much easier to manage, at least in theory.

The gapped summary test format described here – sometimes referred to as *summary completion technique* – has been invented independently several times (Courchene and Ready 1993, Mossenson, Hill and Masters 1987, Pollitt and Hutchinson 1987). Hughes (1989) referred to this reading test format as *summary cloze* and he provided a good example of such a task based upon a newspaper article (1989:122–124). Alderson, Clapham and Wall (1995) commented that, although gapped summary tasks may be difficult to write and need careful pretesting, they can 'work well and are easier to mark' (1995:61). Further examples of gapped summary tasks, taken from the reading test of the *International English Language Testing System* (*IELTS*), are presented by Alderson (2000:240–242). In principle, the technique seeks to interfere as little as possible with the reading process and to make visible the relative strengths and weaknesses of a reader's understanding with as little alteration as possible.

Summary completion formats have been used in both formal and informal reading test contexts (Bensoussan 1983, Courchene and Ready 1993). The format is normally used in the condition where the source reading text remains present throughout the task, i.e. after reading and during completion of the gapped summary. This means that the source text can be re-read and referred to as many times as the reader wishes while they are filling in the gaps in the summary. However, having the source text permanently present may well enable the reader to elaborate their initial mental representation through re-reading of the text while completing the gapped summary. It could be argued that this risks reducing the extent to which the reader is providing evidence of an ecological, or unelaborated, comprehension of the text. What they may actually be providing is evidence of a far more elaborated understanding due to extensive re-reading and task effects than had emerged by the end of their initial reading. It would be unreasonable to suggest that this elaborated understanding is not genuine comprehension for there are many occasions, especially with lengthy or conceptually complex texts, when readers go back and re-read parts of the text several times in order to improve their understanding for a particular purpose. In general, however, much of what we read is read through once in a more or less straightforward way and the understanding we carry away with us as a result of reading, although sometimes quite simple or superficial in terms of processing depth, is nevertheless adequate for our purposes. In this sense, it is possible to conceptualise

different levels or depths of understanding as some writers have indeed done (Gerrig 1988, Spolsky 1994).

An interesting alternative condition for administering summary completion technique (or summary cloze) would be to give readers a text to read through, and then to remove the text from them and ask them to complete a partial summary of the text they have just read. This would remove the opportunity for the reader to elaborate their mental representation through intensive re-reading of the original text. Instead, readers are forced to rely more heavily upon their initial understanding. This approach may offer a useful means of assessing readers' comprehension ability at the sort of level or depth of understanding that is characteristic of most routine reading activity. As a potential reading test format, it stands in marked contrast to most conventional reading comprehension test formats (e.g. multiple-choice questions), which risk stimulating extensive processing and re-processing of a text due to the nature of the task.

In a study of the effects of prior knowledge in reading comprehension tests, Johnston (1984) observed that when the text was permanently available, peripheral questions were very easy to answer because, although peripheral information is not readily stored in memory, it is easily obtained by searching the text. When the text was not available, the task became much more difficult. Johnston suggested that denying the reader access to a text while answering questions about it might offer a valid and complementary approach to assessing understanding:

> . . . if comprehension is defined as the forming of a coherent cognitive model of the text meaning, then interest is most likely to be on the reader storing the central aspects of the text. It seems that the best way to evaluate this is to ask central questions, and possibly to prevent the reader from referring to the text while answering the questions (1984:236).

Summary completion format in the *text-removed* condition thus offers the reading test designer some interesting possibilities. First, it has the advantage that a summary version of the text could in effect be derived directly from the mental representation constructed by a group of readers reading a given text for a specific purpose. Secondly, the summary completion format offers a principled framework within which to identify appropriate reading comprehension test items derived from that summary. Thirdly, the fact that the text is removed following reading means that the focus of the assessment becomes a reader's initial constructed understanding rather than an understanding of text which has been amended or elaborated through extensive re-reading. The type of test question (e.g. multiple-choice prompt) that tends to encourage over-processing of the text or that risks interfering with reading and understanding is entirely avoided, though any cognitive issues associated with memory effects will need to be taken into consideration.

From a practical perspective, the production of a summary completion task raises three important issues for the test developer. The first concerns the selection of an appropriate text and the choice of an appropriate reading purpose and context. The second relates to the generation of a summary version of the text which is derived from the typical mental representation that is likely to be constructed by a group of readers reading for the same purpose. The third concerns the construction of a set of suitable test items. In addition, related issues concerning aspects of text and item difficulty, appropriate response format and test validation will need to be considered.

Aims and scope of the empirical research reported in this volume

This volume reports on a series of empirical studies whose main aim was to investigate readers' mental representation of a text and to develop a text-removed summary completion task which directly addressed their understanding of it. This involved first of all exploring different readers' mental representations of one or more texts to identify what constituted an adequate verbal summary version of each one. Secondly, it required the construction of a set of test items from within such a summary to construct a summary completion task capable of effectively assessing readers' comprehension of the text.

The volume explains how readers' mental representations of two different texts were explored and how a text-removed summary completion task was developed to accompany each of these texts. The two summary completion tasks were then trialled on a population of readers and the results from this exercise were compared with an independent measure of reading comprehension ability for the same population. In this way it was hoped to determine the effectiveness of the text-removed summary completion format as a measure of reading comprehension ability. A secondary aim of this study was to determine whether a preliminary methodology, or at least some practical guiding principles, could be developed for the construction of both text-present and text-removed summary completion tasks.

To contextualise the empirical research, Chapter 2 of this volume reviews in some detail the development of different theories of reading and text comprehension over the past century. Particular attention is paid to theories and models that stress the active and constructive nature of the comprehension process, in which meaning is constructed by the reader's cognitive processes interacting with their background knowledge and personal goals. Chapter 3 then offers an overview of reading test design, describing some of the most widely used formats in the testing of reading comprehension and raising issues of construct validity for much of our current practice. In Chapter 4 the use of summary writing tasks as a means of assessing reading comprehension

ability is explored and the benefits and drawbacks associated with this technique are discussed. An alternative approach – *summary completion technique* – is considered and the research questions for a set of empirical studies are then presented.

Chapters 5 to 8 present the empirical research. Chapter 5 reports on an oral recall study designed to investigate the mental representations constructed by readers as a result of reading of two different texts, Text A (*Journey*) and Text B (*Anorexia*). Chapter 6 presents the detailed analysis and results of the oral recall study for each of these texts and explains how these were used to create a summary for each text that was derived directly from readers' shared mental representations. Chapter 7 reports on how the two summaries were used to form the basis for designing text-removed summary completion tasks, and how a written recall study was undertaken to inform the development of a set of test items. Chapter 8 reports the results from trialling the two summary completion tasks with a population of readers and correlating the outcomes with an independent measure of reading comprehension ability. Finally, Chapter 9 summarises the main findings of the research, discussing its implications and making recommendations for the future.

The research reported in this volume offers a preliminary investigation into the theory and practice of using text-removed summary completion tasks to assess reading comprehension ability. Hopefully, the results of this investigation will stimulate interest in a test format that aims to reconcile more closely the practice of assessing reading comprehension ability with our current understanding of the nature of reading comprehension, leading in turn to further research into summary completion tasks in the future.

Endnote

1. With regard to first language (L1) literacy teaching in the UK in the 1970s, Field (personal communication) notes a parallel movement to the sub-skills approach used for the teaching of L2 reading: '. . . the rise of the whole word and whole language movements which not only tended to the dogmatic but also represented a pendulum swing against traditional sound-spelling instruction. Here, of course, so-called cognitive approaches were partly to blame – not least is Goodman's (1967) assertion that reading was a psycholinguistic guessing game – and the support he received from other luminaries such as Smith.'

2 The theory of reading comprehension

Introduction

This chapter considers the importance of developing a satisfactory theory of reading and comprehension. It reviews how various theories and models of reading and comprehension have developed over the past century or so, paying particular attention to those that stress the active and constructive nature of the process, in which meaning is constructed through the interaction of the reader's cognitive processes with their background knowledge and personal goals, as well as with the text itself. A sound theory of reading comprehension, grounded in empirical research, will be essential for developing a valid test of reading comprehension ability.

Towards a satisfactory theory of reading

Any adequate theory or definition of reading must account for the multiple levels of cognitive processing involved. At the lower levels lie the *perceptual* processes, beginning with the identification of handwritten or printed symbols on a page or on a computer screen. Beyond this initial decoding level are the processes by means of which the reader retrieves lexical entries from the lexicon and parses a string of words to assemble a syntactic structure and add semantic content. (See Nassaji (forthcoming 2014) for a helpful summary of lower-level reading processes in both L1 and L2.) The decoded linguistic material now has to be processed at a higher *conceptual* level. This entails constructing propositional meaning at the clause and sentence level and making the necessary inferences to supply any links that are taken for granted by the writer of the text. The reader then has to integrate this information with their knowledge of the world and with their previous reading experience to support and enrich their mental representation of the text in front of them. The initial process enriches or deepens the propositional information by drawing upon knowledge of the world and the general context. The reader then goes on to draw inferences and resolve anaphoric reference. The highest level of processing entails integrating incoming information (i.e. from the next sentence or paragraph) into the meaning representation of the text constructed thus far as well as monitoring comprehension to check that

the developing discourse representation remains consistent, meaningful and relevant. The resulting product should be an organised representation of a single text or, in some cases, a coherent representation constructed across multiple texts.

Interestingly, just such a theoretical definition of reading was hypothesised by Tinker and McCullough over half a century ago:

> Reading involves the recognition of printed or written symbols which serve as stimuli for the recall of meanings built up through past experience, and the construction of new meanings through the manipulation of concepts already possessed by the reader. The resulting meanings are organized into thought processes according to the purposes adopted by the reader. Such an organization leads to modified thought and/or behavior, or else leads to new behavior which takes its place, either in personal or in social development (1962, cited in Melnik and Merritt (Eds) 1972:38).

This early theoretical definition of reading was confirmed through later empirical research. It is of particular interest for our purposes in that it placed emphasis firmly upon the role of meaning and comprehension in reading activity and, perhaps surprisingly for its time, reflected current views of comprehension as a constructive process.

The reading process has probably been one of the most intensively studied aspects of all human cognitive activity, and Tinker and McCullough stand in a long line of researchers to attempt a plausible and meaningful definition of what actually happens during reading. Determining the detailed nature of the reading process has attracted attention from specialists in a wide range of academic fields, including applied linguists, cognitive psychologists, discourse analysts, literacy researchers and specialists in education. Researchers have dedicated extensive time, energy and resources to investigating both the processes involved in and the products resulting from the activity of reading. These have been investigated for reading in both a first and a second language. Given the broad scope of interest in reading, it would be unrealistic to attempt a comprehensive historical survey of reading research and the many interpretations of reading comprehension that have been proposed over the decades. Fortunately, there exist several recently published accounts offering accessible overviews of reading processes and products, including Field (2004), Grabe (2009), Grabe and Stoller (2002) and Khalifa and Weir (2009). For this reason, attention will focus below primarily upon those theories and research studies which are generally considered central to the way our present understanding of the reading comprehension process has developed, or those which are in some way directly relevant to the research reported in this volume.

Early theories of reading comprehension

Fascination with the nature of the reading process leading to a desire to deconstruct the psychological elements that make up the process of comprehension is not a recent phenomenon. Over a century ago, Huey wrote:

> . . . to completely analyze what we do when we read would be almost the acme of a psychologist's dream for it would be to describe very many of the most intricate workings of the human mind . . . (1908:8).

Huey was one of several early reading researchers who recognised reading to be a *conceptual* as well as a *perceptual* process. A few years later, Thorndike (1917/1972) suggested that reading comprehension called upon the reader's reasoning as well as their perceptual powers. Even at this early stage in reading comprehension research, Thorndike recognised the importance of 'the mental set' which could cover not only the contextual feature of purpose for reading but other features such as background knowledge. Later on, both Fries (1945, 1963) and Gray (1948) also recognised the reader's background knowledge as a potentially significant influence in the reading comprehension process, although empirical investigation of this and similar influences had to wait another two or three decades (see Bernhardt 1991a for a useful list of relevant studies).

Despite some early awareness of the different levels of processing that take place during reading, much reading research in the first half of the 20th century restricted itself to investigating the lower-level processes, using mostly word recognition studies (Venezky 1984). Some authors (Clapham 1996, Samuels and Kamil 1988, Tanenhaus 1988) attributed this to the strong influence of behaviourist thinking on research in psychology and the social sciences, and the resulting emphasis upon the observation of subjects' responses to external stimuli. It was not really until the 1970s that research attention finally began to refocus towards a study of the other mental processes which must be at work during reading and the way in which these mental processes might interact with one another (see Urquhart and Weir 1998:18–21 for discussion of this).

The 1960s onwards saw concerted efforts to describe more clearly components of the reading process, and their relationship to one another, using a more or less formal model of the reading process. Samuels and Kamil (1988) suggested that although theories concerning components of the reading process had existed in the minds of some researchers for several years 'there was simply not a strong tradition of attempting to conceptualize knowledge and theory about the reading process in the form of explicit reading models' (1988:22). The 1960s and 1970s, however, witnessed the development of several different models describing the reading process, some more detailed

than others, and some lending themselves more readily to empirical investigation and validation than others. Vacca and Vacca (1983) referred to the 1970s as 'an era of unprecedented advances in theory and model building leading to a renaissance in research on the reading process' (1983:382).

Early attempts to model reading comprehension

Jenkinson (1972) suggested that the earliest attempt to model the reading process was probably that of Gray (1960) who proposed four different activities: *word perception*, *comprehension*, *reaction to what is read*, and *assimilation of what is read* through the fusion of old and new ideas. Gray seems to have drawn a distinction between 'comprehension', which he related to the literal and implied meaning of the text (described as reading the lines and reading *between* the lines), and 'assimilation' (or reading *beyond* the lines), where readers relate what they have read to their background knowledge.

Gray's model was subsequently revised by Robinson (1966) who observed that the Gray/Robinson model related more to skills and abilities than to processes in reading. It was therefore closer in nature to the taxonomies of reading sub-skills drawn up during the 1960s and 1970s which are discussed more fully in Chapter 3.

In her review of early attempts to model the reading process during the 1960s, Jenkinson (1972) commented that most models of this period sought to cover too many facets of reading and as a result tended to confuse the dynamic activity of the actual reading process by linking it with the techniques and skills which need to be acquired in the process of learning to read. She suggested that quite separate models may be needed to show the interrelationship between skills, techniques, materials and media on the part of the developing reader and the potentially different processing on the part of the mature reader. A clear distinction between reading skills or abilities on the one hand, and reading processes on the other, took quite a long time to emerge.

Reading skills and abilities were variously described as being the products of reading (Strang 1972) or as underlying or contributing to the reading process (Alderson 1990a). However they were defined, it seems to have been the focus on defining reading skills and sub-skills, rather than on defining reading comprehension processes, which shaped and dominated developments in both the teaching and the testing of reading comprehension during the 1970s and 1980s. As Field (personal communication) points out, at some point in the early 1980s writers started using the terms skills and strategies as synonymous, without adequate differentiation, and there were also parallel intuitive attempts to provide taxonomies of sub-skills and taxonomies of strategies. The sub-skills approach can be critiqued on several counts: first, the intuitive nature of sub-skills and the lack of evidence for their psychological reality; secondly, the need to integrate sub-skills into performance; and

thirdly, the miscellaneous nature of what was covered by the term sub-skills, especially in reading. The powerful skills and strategies paradigm of the 1980s may have unhelpfully blurred a proper distinction between the more context-related elements of reading, such as text types and reading purposes, and more cognitive-related elements, such as the levels of processing required of readers. As we shall see later, the work of Weir during the early 1990s (1990, 1993) was instrumental in helping to draw a much clearer distinction between contextual and cognitive parameters in reading (and in other skills), leading eventually to a socio-cognitive validity framework for analysing task features and processing that was initially presented in Weir (2005) and later refined and applied specifically to reading assessment in Khalifa and Weir (2009).

Early attempts to model essential aspects of reading were thus partly motivated by a desire among educationalists to discover a suitable way of organising the instruction of reading. Spache (1964), for example, claimed that a clear definition of reading was essential to planning the goals of the instructional programme. Following the model offered by Bloom (1956) (Ed) in his *Taxonomy of Educational Objectives*, some educational researchers sought to establish a systematic framework for the teaching of reading comprehension through the development of reading skill/sub-skill taxonomies (Barrett 1968, Davis 1968). This taxonomic analysis of reading comprehension (rather than a cognitive processing analysis) came to exert a significant influence on the nature of reading comprehension assessment and this will be discussed in more detail in Chapter 3.

Process and componential models of reading

Some attempts to develop a model of the reading process were shaped by researchers in the field of cognitive psychology interested in the nature of the cognitive processing that occurs during reading. They also drew on work being carried out in the field of artificial intelligence and the computational modelling of language understanding. Three different theories of how readers were believed to arrive at an understanding of text are described in more detail below – *bottom-up*, *top-down* and *interactive* – along with specific examples of their proponents. Urquhart and Weir (1998:39–46) classified all three as 'process models' due to their focus on processing, while Grabe and Stoller (2002:31) referred to them as 'metaphorical models' representing generalisations that arose from comprehension research over the previous three decades.

Bottom-up processing models of reading

So-called bottom-up processing models of reading regarded the flow of information from initial incoming data to final interpretation of meaning

on the part of the reader as uni-directional, passing through a series of discrete stages. Each stage had the function of transforming incoming input (i.e. symbols, letter clusters, single words) and sending the recoded information up to the next stage for further transformation and recoding. Cohen and Upton (2006) and Birch (2007) provide a full explanation of bottom-up models.

One of the earliest bottom-up processing models of reading was that of Gough (1972). His model assumed that a reader first registers all the letters in his visual field and then converts the graphemic cues into phonemic cues before assigning them any meaning. A later example of a bottom-up model was proposed by Carver (1977). The results of a number of reading studies undertaken both prior to and during the 1970s provoked some researchers to criticise bottom-up theories of processing on the basis that they failed to account satisfactorily for the way in which a reader's processing of letters, words and sentences was clearly influenced by syntactic, semantic, lexical and orthographic information from higher up in the processing chain. For detailed criticism of so-called bottom-up processing models, see Rumelhart (1977) and Samuels and Kamil (1988).

The main problem with a narrowly bottom-up view is that evidence tells us that readers do not operate by building smaller units of language into larger ones and then into meanings. Current models recognise that reading takes place at many levels simultaneously. Thus, word recognition considers letter features, letters, letter order, digraphs, recurrent syllables and whole words in parallel – matching the cues against a set of candidate words until one emerges as the obvious match (Rastle 2007). The later part of bottom-up processing is very different; it is what is correctly termed parsing and because reading takes place linearly on a word by word basis, there has to be a degree of forward projection in which the reader anticipates the syntactic structure that is evolving.

Top-down processing models of reading

Proponents of so-called top-down processing models claimed that, instead of moving up from the bottom, readers generate an understanding of text by working in the opposite direction, using their ability to anticipate, hypothesise and confirm or disconfirm meaning on the basis of context.

The most well-known among these is Goodman's model of the late 1960s, which was refined over subsequent years into a relatively formal account of the components and stages of the reading process. Goodman (1967) proposed the concept of the reader as a more active participant in the reading process than had hitherto been recognised. Using insights and methodology drawn from the fields of psycholinguistics (and later, sociolinguistics), he set about designing a model of the reading process which would be 'powerful enough to explain and predict reading behavior and sound enough to be a base on which to build and examine the effectiveness of reading instruction'

(1988:11). Unlike the models of Gough (1972) and Carver (1977), Goodman's model was characterised by its emphasis upon the influence of high-level syntactic and semantic knowledge structures in the process of comprehension, rather than the low-level decoding and recoding of graphic information. Goodman's argument was that reading is a psycholinguistic guessing game in which good readers avoid having to decode text on the page by anticipating upcoming words. However, such an approach was seriously questioned given that it was impossible to predict in the way asserted by Goodman and there were also concerns about Goodman's method and data (see Gough and Wren 1999).

Smith (1971) is often cited as the originator of another top-down theory of reading in which language factors play a more influential role than graphic information. Samuels and Kamil (1988) noted the value of Smith's explanation of the inherent redundancy in language at every level, enabling readers to be highly selective regarding the word and sentence features which they need to identify in order to create meaning for a given text.

Interactive processing models of reading

At this point it may be worth commenting on the various ways in which the term *interactive* is used in discussions of language processing. It is frequently used to describe the nature of the relationship which can exist between a text and a reader, or even between the various elements within a text. In the present review of theories and models of reading, however, this term is used to refer specifically to hierarchical models of reading in which processing at one level or stage (e.g. word perception) is able to interact with processing at another level or stage (e.g. semantic knowledge). Such models are distinguishable from those previously described in which the stages are passed through in a uni-directional, linear fashion. Grabe (1991:384) suggested that while top-down and bottom-up models were 'serial' in nature, interactive models hypothesised a 'parallel' processing approach in reading in which every level or component can interact with any other. Both Rumelhart (1977) and Stanovich (1980) produced examples of interactive processing models of reading.

Dissatisfied with the limitations of uni-directional, linear models to explain a number of occurrences that are known to take place during reading (e.g. the effect of semantic knowledge on word perception and the effect of context on interpretation), Rumelhart (1977) proposed a model in which information from syntactic, semantic, lexical and orthographic sources, together with visual input, converge upon a pattern synthesiser or message centre. This means that each knowledge source is simultaneously able to provide input that interacts with input from other knowledge sources, through the central mechanism of the message centre, in order to generate the most probable interpretation.

Stanovich (1980) claimed that 'interactive models of reading appear to provide a more accurate conceptualization of reading performance than do strictly top-down or bottom-up models' (1980:32) and his model of the reading process was similar to Rumelhart's in as much as he allowed for information from several knowledge sources to be managed simultaneously and interactively. Stanovich, however, added an important extra dimension to his model by introducing the notion that processing deficiency in any one source (e.g. word recognition) on the part of an individual reader could be compensated for by reliance upon another information source (e.g. knowledge of the text topic): 'The compensatory assumption states that a deficit in any knowledge results in a heavier reliance on other knowledge sources, regardless of their level in the processing hierarchy' (1980:63). For this reason Stanovich referred to his model as an interactive-compensatory model. Its importance was that it was among the first models of reading to be able to account satisfactorily for individual differences observed among readers, probably due to factors such as level of language proficiency or extent of background knowledge.

Current models of reading and listening could be said to be interactive in that they assume that both perceptual and conceptual information will guide, for example, word recognition. However, that by no means suggests that they necessarily subscribe to the kind of connectionist modelling favoured by Rumelhart (1977) that is often said to be a prime example of interactionism. Stanovich's model was a very different one, based upon the assumption that decoding was the cue to successful reading but that, where decoding became difficult, readers would rely to a greater degree upon contextual information to compensate for their problems of recognition.

Componential models of reading

Componential models of reading tended to conceptualise reading ability as capable of being deconstructed into constituent skills or types of knowledge, rather than as a set of psychological processes. Since componential models are less relevant to our purposes in this volume, only brief mention will be made of them here (though the issue of background knowledge will be touched upon later in this chapter). Hoover and Tunmer (1993), for example, suggested reading could be understood 'as a set of theoretically distinct and isolable constituents' (1993:4) and they proposed a model of reading with two components: *word recognition* and *linguistic comprehension*. Others chose to accommodate a larger number of constituents within their model. Perfetti (1985, 1994, 1997) and Carver (1997) highlighted the role of factors such as *speed and automaticity of word recognition*, *depth of word representation knowledge* and *fluency in syntactic processing* in a componential model, especially in terms of their ability to explain individual reader differences. Perfetti, in particular, has always been a major proponent

of the critical nature of decoding in successful reading because of the way it releases working memory capacity so that the reader can attend to higher-level issues of meaning. With regard to second language readers, Coady (1979) selected *conceptual abilities*, *process strategies* and *background knowledge*, while Bernhardt (1991a) opted for *language*, *literacy* and *world knowledge*. For a fuller discussion of componential models, see Urquhart and Weir (1998:46–84).

The role and value of models of reading

Given the diverse conceptualisations of reading outlined so far, it may be worth commenting here on the use of the term *model*, and upon the role and value of developing a model of reading. The term itself can carry two different meanings. The first is a processing model as favoured in cognitive psychology, based upon information processing theory, which tracks the way in which a piece of information changes its form during a process. The term is also much more loosely applied to general and unelaborated theories of reading such as the notion that it is mainly bottom-up or mainly top-down. Goldman, Golden and van den Broek (2007) defined the term 'model' as 'a representation of the psychological processes that comprise a component or set of components involved in text comprehension' (2007:27). Samuels and Kamil (1988) suggested that a model fulfils three important functions, relating to past, present, and future. First, it can help us to summarise all the information on a given topic which has been gathered so far. Secondly, it can help us to understand what is likely to be a complex phenomenon by focusing only on the essential for the time being. Thirdly, it can enable us to formulate testable hypotheses which can be used to guide further stages of investigation. Models of the reading process, such as those described above, can only ever be partial and temporary. They help us to organise past observations and current thinking in such a way that we are able to move steadily forwards in improving our understanding of a given topic. Models that are either over-simplistic or extremely complex are unlikely to achieve these objectives (de Jong and Verhoeven 1992, Samuels and Kamil 1988). See Grabe (2009) for a comprehensive overview of the key theories and models to have emerged for reading over the past 30 years together with some discussion of their significance.

Approaches to determining meaning in text

Simultaneous with the development of different psychological theories and models to explain the reading process were attempts to investigate the relationship between text and meaning. Instead of focusing on the processes involved in text comprehension, text analysts preferred to focus on the potential product, i.e. the semantic representation of text.

The propositional representation of discourse

An early and influential approach to text analysis involved analysing the semantic representation of a text's content in terms of its *propositional relationships* (Kintsch 1974, van Dijk 1977). Both writers chose the term *proposition* to denote a conceptual structure that represents the meaning of part or all of a natural language sentence as it is stored in and recalled from memory but which is not necessarily expressed in the exact wording of the original sentence. This view was consistent with evidence from cognitive psychology that the surface level representation, or verbatim memory, for a text or utterance decays very rapidly, presumably because the input is translated into some other, more conceptual form. Jarvella (1971) produced empirical evidence that the surface form of an utterance ceased to be available after a short delay and concluded that language users were under pressure to transform a piece of speech (mainly a clause) into an abstract concept as soon as possible to prevent an overload in working memory. Kintsch stated that 'the memory representation of text is a function of the content of the text, but not of the way in which it is expressed' (1974:107). Van Dijk suggested that the mental representation of a text takes the form of a set of propositions which are themselves hierarchically organised so that the semantic representation of a whole text is its *macro-structure* defining 'the meaning of parts of a discourse and of the whole discourse on the basis of the meanings of individual sentences' (1977:6).

Some issues with a proposition-based approach

Brown and Yule (1983) highlighted some significant issues associated with this type of proposition-based approach to the representation of text content. One is that it too readily assumes meaning to reside primarily in the text and fails to take sufficient account of the importance of interpretation of a speaker or writer's intended meaning as shaped by context. Furthermore, although a proposition-based approach gives the impression of being a highly formal and objective approach to text analysis, the production of a proposition set for most naturally occurring texts, especially in its hierarchical arrangement, remains an essentially arbitrary and subjective affair, determined largely by the text analyst in isolation. Brown and Yule (1983) argued that if the semantic representation of text is regarded as something that occurs in a reader or listener's head (rather than being something encoded in the textual record), then it is probably naive to assume the existence of a single set of hierarchically organised propositions representing the definitive interpretation of the text. In practice (as opposed to theory), more than one interpretation of a text may be possible depending on different readers. Thus a purely proposition-based approach to text analysis will not necessarily provide a principled means for deciding which is the best or 'correct' understanding of a text.

The notion of relativity of interpretation is an important one since it has some clear implications for the testing of reading comprehension. If we take the view that a group of readers will construct multiple and highly personalised interpretations of the same text, then it becomes difficult, if not impossible, to know how to test these, and indeed there are some who oppose reading tests *per se* for this very reason. On the other hand, it is likely that readers of the same text will have a sufficient convergence of point of view or 'reciprocity of perspective' (Schuetz 1953) even despite some potential variability in their mental representations due to personal background, knowledge, experience or purpose. Thus we can take a view of reading which acknowledges the capacity for some measure of individualised interpretation of a text, while at the same time asserting that, for most texts, it is reasonable to expect that a common understanding can be shared across a group of readers and that this understanding can reasonably be assessed. This is consistent with the view articulated by Urquhart and Weir that we can and should be testing comprehensions, but not interpretations (1998:112–120).

It is fair to say that the propositional approach to text analysis exerted a powerful influence on understanding of how text content is processed during comprehension, stored in memory and then recalled at a later stage. Many text recall studies drew, and to some degree still draw, primarily upon the theories of Kintsch and van Dijk (1978) for their methodological approaches to the coding of texts and text recall protocols. Despite the criticism that can be levelled at Kintsch and van Dijk's proposition-based approach to text analysis, Williams (1993) and others, such as Field (2004), suggested that many researchers still regard the propositional level as a pre-requisite for building more complete representations of discourse. In other words, it constitutes the minimum semantic analysis which must be established before context, reference and elaborative inference can play their part. For this reason, the notion of linguistic propositional analysis continues to occupy an essential place in current models of reading (e.g. Field 2004, Khalifa and Weir 2009), despite the fact that the term proposition can be problematic in the testing of receptive skills. Strictly speaking, in semantics (and indeed in much of the cognitive theory), a proposition is a conceptual unit which is independent of context. We therefore cannot talk about a test item as tapping into information at a propositional level, but must instead think in terms of a simple piece of factual information as representing an idea unit (Chafe 1979).

Applications of a proposition-based approach to determining meaning

The application of a proposition-based approach to determining meaning in narrative text was central to producing a *story-grammar* (Mandler and Johnson 1977, Rumelhart 1975, 1977, Thorndyke 1977). Both Rumelhart and Thorndyke adopted a tree-structure approach to describe the steps involved in comprehending a simple narrative text. They developed hierarchically

organised, proposition-based versions of the story content and then compared these with subjects' recall protocols and summaries of the original story. They concluded that it was the components at the top of the hierarchy which were most readily recalled or included in the summaries.

An alternative approach was proposed by de Beaugrande (1980) who avoided the notion of hierarchical organisation and instead represented text-content as a network of relationships between the various elements in a *text-world*. The network could be grammatical in nature, or it could be used to reflect the conceptual relations existing in a text. De Beaugrande identified a problem with this approach in that the text-world represented by the network can be no more than an idealised version of an actual mental representation and is unable to account for the various conceptual relations held in the head of an individual reader.

Van Dijk and Kintsch (1977) adopted a proposition-based analytical approach to research the text structure and processing of expository as well as narrative prose. Kintsch and van Dijk (1978) proposed a model of text comprehension according to which the reader combines propositions to form macro-propositions at a higher level of abstraction. These macro-propositions are in turn combined to form the macro-structure of a text, which represents its gist, or highest level of abstraction. The researchers suggested that certain text types, such as stories or psychological research reports, have a conventional, schematic structure.

Taking a similar approach, Meyer (1975a, 1985) identified five different but typical patterns for expository texts – *collection/sequence* (e.g. lists), *description* (attribution), *causation* (cause and effect), *response* (problem/solution), and *comparison* (compare and contrast). Her belief was that each pattern prompts certain expectations in a reader and thus guides the text comprehension process.

Although investigation extended beyond simple narrative texts to the study of other discourse types, much experimental work continued to rely heavily upon short and artificially constructed texts. This made it difficult to generalise any findings to the comprehension of naturally occurring reading texts and it cannot with certainty tell us anything about how test takers handle the more naturalistic types of text that feature in tests of comprehension.

Some problems associated with a proposition-based approach to analysing the meaning of written text have already been referred to. Clearly, features of text analysis such as presuppositions, implicatures, inferences and other discoursal aspects cannot be adequately accounted for through a purely proposition-based analysis of text. Nor is the potential role of the preceding context (as determined by the writer/reader's purpose and other sociolinguistic constraints) satisfactorily addressed. In seeking to identify how the meaning of a given text can be shared or held in common among a group of readers, it is necessary to go beyond a purely formal propositional analysis.

This will be essential if the text is to be used as the basis for testing reading comprehension and will be explored in later chapters of this volume.

Context and background knowledge in comprehension

As interest in investigating the processes of comprehension grew during the 1970s, it became increasingly clear that understanding a text depended not only upon language knowledge but also upon extra-linguistic knowledge, i.e. knowledge beyond the text. A number of empirical studies explored in more detail how an individual's comprehension might be affected by their background knowledge or by the context for their reading or listening. Some of the most relevant studies and their findings are discussed below.

The role of context and background knowledge

Bransford and Johnson (1973) conducted a series of experiments asking subjects to listen to, comprehend and recall specially constructed texts. The aim was to demonstrate that understanding depended not just on the linguistic content of the text but at least in part upon the *context* in which that text occurred. Experiments showed that comprehension and recall were significantly better when subjects were told the supposed topic before listening to an otherwise virtually incomprehensible text. The researchers concluded that the existence of relevant background knowledge and the ability to activate that knowledge contributed to successful comprehension. In a later study, Bransford and McCarrell (1977) demonstrated a similar result for reading comprehension. They also showed that comprehension was likely to be weaker if subjects were deliberately encouraged to activate the wrong sort of background knowledge, e.g. by being given an inappropriate title for the reading text.

Based partly on this empirical evidence, Brown and Yule (1983) concluded that the title of a text has the potential to function as a powerful thematising element, providing not only a starting point around which the discourse is structured but also certain expectations on the part of the reader or listener which go on to constrain the interpretation they construct.

Several studies confirmed the influence of background culture and related cultural expectations on interpretation (Steffensen, Joag-Dev and Anderson 1979, Tannen 1979). Other research showed that a similar effect can be demonstrated between members of different interest groups (Anderson, Reynolds, Schallert and Goetz 1977, Carrell 1987) or between readers with a high or low-level knowledge of a topic (Freebody and Anderson 1983, Spilich, Vesonder, Chiesi and Voss 1979).

Given the findings described above, it is clear that any attempt to describe and explain the nature of reading comprehension needs to account for how a reader's world knowledge influences the process of interpretation (Brown

and Yule 1983). Various theories have attempted to explain how such knowledge is organised and stored and some of these will be discussed here.

The representation of background knowledge: schema theory

Bartlett (1932) is often considered as the originator of schema theory although Anderson and Pearson (1988) suggested that the underlying principles of schema theory can be identified in earlier work by Gestält psychologists. Bartlett defined a schema as 'an active organization of past reactions or past experience' (1932:201), while Anderson and Pearson defined it as an abstract knowledge structure which 'summarizes what is known about a variety of cases that differ in many particulars' (1988:42). For a helpful illustration of how a schema might be represented, see Anderson and Pearson's discussion of a ship's christening in Carrell, Devine and Eskey (Eds) (1988).

Minsky (1975) proposed that knowledge is stored in memory in the form of information-structures, or *frames*, representing stereotyped situations. He defined a frame as a remembered framework containing labelled slots that can be filled with typical or appropriate features relating to the given frame. For example, in a frame representing a typical room, there will be basic stereotypical expectations concerning the room's shape (probably square or a rectangle), the presence of a door, window, furniture, and so on. If a room is encountered in everyday life (or in some piece of discourse), the room frame is selected from memory and then adapted to fit the current reality by changing the default elements as necessary: for example, that the room has ceiling lights instead of the usual windows, or that all the furniture has been removed. Minsky's frame-system theory offered a useful if limited account of the way in which we hold default concepts in our minds that can be retrieved and utilised to achieve understanding.

Script theory was developed by Schank and Abelson (1977) and Riesbeck and Schank (1978). Although it bears some similarity to Minsky's frame theory, script theory represents an attempt to deal with knowledge relating to a sequence of events (i.e. a script) rather than a more static situation or set of facts about the world (i.e. a frame). Schank and Abelson proposed that much of our understanding of what we read or hear is expectation-based in accordance with our background knowledge and experience, e.g. a simple sentence such as *we had a meal in a restaurant* or *I checked in at the airport* refers indirectly to a whole sequence of operations which we have learned to recognise through repeated exposure.

Efforts continued into the 1980s to develop theories that could satisfactorily account for the role of background knowledge and life experience in text comprehension. Sanford and Garrod (1981) adopted the term *scenario* to describe the interpretative context within which understanding of written text takes place. According to *scenario theory*, a text entitled *Going to a restaurant* or *In court* automatically activates and brings into the representation

certain 'role' slots, such as *a waiter* or *a lawyer*. Sanford and Garrod acknowledged that their concept of the role of pre-existing knowledge representations in text processing had much in common with other studies in which the term *schemata* is more generally used.

While the use of *frame*, *script* and *scenario* suggested a situation-specific approach (e.g. a *restaurant* or *courtroom*), the term *schema* became a more general term for this type of knowledge representation. Frame/script/scenario concepts could be regarded as specific manifestations of a more general schema-theoretic approach to explaining understanding.

Since the literature on schema-theoretic views of comprehension is extensive, no further attempt will be made here to review and discuss them here, other than to consider some specific limitations which have been identified in the relationship between schema theory and the nature of human understanding.

Problems with schema-theoretic approaches

Although the schema-theoretic approaches outlined above offer an attractive and plausible explanation of how background knowledge is organised, stored and retrieved for the process of comprehension, they raise a number of important problems. Brown and Yule (1983) highlighted two particular concerns with the application of schema theory in understanding. First, since a considerable degree of existing knowledge is apparently already assumed on the part of any reader or listener, logic suggests that it would be unnecessary for much of this information to be communicated. A second question concerns the potential for a piece of discourse to activate several different schemata, or even sub-schemata, simultaneously. It is unclear, therefore, how those schemata that are necessary and relevant to interpretation are to be distinguished from those that are not (see Brown and Yule 1983 for a fuller discussion of both these issues).

Furthermore, some researchers expressed reservations about the extent to which schema theory could account adequately for either the level of detail or the general messiness of human general knowledge and background experience (Alba and Hasher 1983). Others questioned whether it could account for the way readers clearly coped with new and incoming information that does not relate to a pre-existing knowledge structure (van Dijk and Kintsch 1983). In addition, it seems that a schema will inevitably be highly personal in nature and how much of it can be guaranteed to be shared by more than one reader remains unclear. Finally, since much of the research investigating schema-theoretic knowledge representations involved very short and artificially constructed texts, there remains a question over the extent to which the theory can be generalised to comprehension of much longer and naturally occurring texts, such as those employed in tests of reading, although researchers varied in the extent to which they perceived this to be a problem (Anderson et al

1977, Grabe 1988). Interestingly, Carver (1992) argued that schema theory was probably more relevant to the reading of longer, more challenging texts, such as those read by college-level students than to most everyday reading (which he referred to, somewhat curiously, as *rauding*).

Despite its limitations, schema theory nevertheless enabled considerable advances to be made in the study and understanding of the comprehension process. By the late 1980s and early 1990s it was widely accepted that schemata do play a role in comprehension even if the process by which this is possible remained relatively unspecified (Anderson and Pearson 1988, Brown and Yule 1983, Clapham 1996, Just and Carpenter 1987, McNamara, Miller and Bransford 1991). Anderson and Pearson (1988) suggested that critical and as yet unresolved issues in developing a theory of language comprehension which incorporates the role of knowledge schemata were: the specification of component parts of a schema and their relationship to one another; the identification of the role of inferencing in schema activation; and some explanation of how people apparently use knowledge of particular cases as well as more abstract and general schemata.

The nature and function of knowledge schemata are clearly of direct relevance to the testing of reading comprehension. Khalifa (1997) demonstrated empirically how significant topic knowledge was in performance on reading tests. Test developers therefore need to have reasonable expectations of how well developed test takers' schemata are likely to be in advance of reading a text in order to avoid unfair bias. There is clearly a bigger issue here – widely addressed in both the reading and the listening literature – relating to familiarity of topic and the extent to which a particular topic may be culturally inaccessible for certain students (see Alderson 2000, Khalifa and Weir 2009 and Urquhart and Weir 1998 for extensive discussion of these issues).

Constructivist theories of comprehension

Dissatisfied with the limitations of describing and explaining human comprehension in terms of surface and propositional representations, some researchers turned their attention to defining a deeper level of representation which could present the content of a text as a state of affairs in a real or hypothetical world. This meant going beyond the information explicitly stated in a text, drawing on background knowledge to construct a wider context in which the text would make sense and establishing relations between elements that are merely implied but not explicitly stated by the text.

A mental models view of text comprehension

Bransford, Barclay and Franks (1972) termed this a *constructivist* view of comprehension and theories of comprehension to adopt this view included the *situation model* (Kintsch 1988), the *mental model* (Garnham 1987,

Johnson-Laird 1983) and *referential representation* (Just and Carpenter 1987). The suggestion was that although linguistic input from a text must first pass through an initial, obligatory stage of propositional encoding, the construction of a deeper level of representation situated in a real or possible world is optional and will depend upon the nature of the text together with the background knowledge and goals of the reader.

Some researchers pointed to the phenomenon of shallow comprehension as evidence that full referential models do not always need to be constructed. For example, when asked to solve the following problem: 'A plane crashed on the border between America and Canada. Where were the survivors buried?' most people simply fail to notice that survivors would of course not need to be buried (i.e. a contradiction in terms). Although the nature of the situation and problem has been identified, only a very shallow semantic analysis is performed upon the word 'survivors' – an analysis which is insufficient to activate the ALIVE property of the word. As we shall see later, the issue of potential variability in the depth and specificity of text representation has important implications for the way in which a test of reading comprehension is designed.

Kintsch and van Dijk's (1978) theory of text comprehension proposed that the content of a text could be represented by a structured set of propositions or *text base*. A number of problems with this approach have already been discussed and the authors eventually revised their earlier theory to take account of some of its shortcomings. The revised version (van Dijk and Kintsch 1983) proposed three levels of representation for a given text: a *surface level*, a *propositional level*, and a *situation model*. They pointed out that situation models are linked to situations in the world as well as to particular texts. Thus a situation model is assembled not only on the basis of the text itself but also on the basis of pragmatic and contextual features relating to the text. In this sense van Dijk and Kintsch subscribed to what is generally known as the mental models view of text comprehension (Garnham and Oakhill 1994).

According to this view of comprehension, readers and listeners construct a mental representation for themselves which satisfactorily accounts for or makes sense of the incoming linguistic input. A mental model is a representation of some real or possible world constructed in accordance with a specific purpose. Various experiments suggested that mental representations are capable of representing both spatial and conceptual situations. Bransford et al (1972), for example, concluded from their study that subjects had set up a situation model with spatial characteristics. Studies of anaphora (Oakhill, Garnham, Gernsbacher and Cain 1992) offered evidence that a situation model can be conceptually as well as spatially oriented.

The original proponent of the mental models approach to comprehension is generally considered to be Johnson-Laird (1983). His mental models framework set out to describe the mental representation of a text in terms

of discourse entities and their relationships rather than a set of propositions derived from a text. Words or expressions in a text functioned as cues used to build a familiar mental model composed of entities arranged in accordance with the relationships expressed in the text. The model is familiar in as much as it represents a real or possible world that is coherent not only with the textual information provided but also with our own experience of such a real or possible (e.g. fictional) world.

An interesting question in this connection is what happens where it is not possible to construct a coherent mental model because the information given in the text is either insufficient, or is somehow idiosyncratic and apparently contradictory. One solution could be the construction of multiple models although this would place heavy demands on working memory. Another solution could be to resort to a purely propositional encoding, or to construct only a partial model and leave indeterminate points at the propositional level. Johnson-Laird advocated the view that, in most communicative situations, the majority of readers are inclined to select a model that is not incompatible with the text. In essence therefore, a mental model is 'a representative sample from the set of possible models satisfying the description' (1983:165). As Field (personal communication) points out, there is always the issue that if a text proves to be more detailed or complex than the reader's goals demand, the reader reserves the right to extract only a partial or indeterminate representation of what is on the page. This might be true of a newspaper article or of a very complex detective novel where the reader knows that the truth will be revealed by the end.

Garnham described a mental model as a structure that is created during the comprehension of a particular text and held in working memory. Importantly, Garnham added that although this structure reflects the structure of the situation in the real or imaginary world that the text is about, it need bear no resemblance to any of the text's linguistic representations (Garnham 1987, Garnham and Oakhill 1992). As we shall see later in this volume, this point has significant implications when considering the activity of summary writing as a means of assessing reading comprehension ability. Even though there must clearly be a strong conceptual match between original text and a summary of it, the linguistic representation of the summary is likely to vary considerably from the linguistic content of the text, making compositional as well as comprehension demands upon the reader. From a testing perspective, this points to summary writing as an integrated assessment construct, raising important issues over how quality of performance should be evaluated and how test scores should be interpreted in terms of reading and writing ability.

The 1980s saw growing interest in *connectionist* views of language processing as a way of explaining how generalised representations (including representations of language) might be constructed on the basis of exposure to

multiple examples. Connectionist systems were originally developed by computer modellers such as Rumelhart and McClelland (1986) in a way that was said to emulate the workings of the human brain, including the nature of cognition in perception, understanding and learning. Field (2004:73) provides the following explanation of the principles underpinning connectionism:

> Like the brain, connectionist models consist of a large number of simple processing units with multiple connections linking them. Activation flows along the connections, just as electrical impulses transmit information through neurons in the brain. The ease with which activation spreads from one unit to another is determined by the *strength* of connections along which it travels. The stronger the connection to a unit, the more readily that unit becomes activated. A connection's strength depends upon how frequently it is used. Thus, over time, connections to a frequent word will become strong, ensuring that the word is activated more rapidly than other less common ones.

In relation to modelling text processing and comprehension, Kintsch incorporated connectionist ideas into the earlier model that he and van Dijk had developed, resulting in Kintsch's (1988) Construction-Integration Model of discourse comprehension, which was essentially a hybrid between the symbolic systems used in the earlier Kintsch models and connectionist systems of researchers like Rumelhart and McClelland (1986). Gernsbacher and Foerstch (1999) provide a helpful description and discussion of Kintsch's Construction-Integration Model for discourse comprehension.

Just and Carpenter (1987) called the reader's representation of *the world* to which a text refers the *referential representation* and they likened it to a situation model embodying the situation described by the text. Thus for the clause 'He flew to Cairo', it may be possible to determine one syntactic analysis and one semantic analysis, but several different referential representations each of which will depend upon context (e.g. did he fly himself to Cairo in a small light aircraft from a local Egyptian airport? Or did he fly on a scheduled airliner from London Heathrow to Cairo?). The authors suggested that the referential representation of a given expression is highly context-sensitive and that the most important factor in determining referential representation is appropriateness to the discourse context. Furthermore, the referential level may in fact be the most important level of analysis since the main function of many texts is to provide information about referents.

In summary then, a particular feature of mental models was their potential for being individual to the reader and context-specific. Several studies demonstrated the varying effects of text, reader and task on the construction of mental models (Perrig and Kintsch 1985, Schmalhofer and Glavanov 1986). If it is the case that mental models are likely to be subject to individual interpretation, then one might question how it is that people can still understand

each other to the extent that they do. Just and Carpenter (1987) explained this phenomenon on the basis that individuals share a great deal of common knowledge about the world.

Garnham and Oakhill (1994) speculated that there was still some way to go to formulate a systematic theory of discourse processing and text representation within the mental models framework. Since then research has continued along a number of different lines of enquiry, including work on the neuroimaging of discourse processing and on the processing of figurative language in discourse. Several useful overviews describe developments in theories of discourse comprehension over the past 10–15 years – see, for example, Long et al (2006) and Zwaan and Rapp (2006). One relatively recent theory that may prove particularly relevant for our purposes is the *Event Indexing Model* (Zwaan and Rapp 2006, Zwaan, Langston and Graesser 1995) which suggests that any event in a narrative is indexed according to five core features – time, protagonists, space, causation, and motivation.

The Structure Building Framework

A particularly influential constructivist view of comprehension, though rather different in type of structure from the mental models approach discussed earlier in this chapter, was provided by Gernsbacher in her *Structure Building Framework* theory. In a series of publications (Gernsbacher 1990, Gernsbacher and Faust 1992, Gernsbacher and Foerstch 1999, Gernsbacher, Varner and Faust 1989), Gernsbacher developed the notion of comprehension as a Structure Building Framework to account for the line of argument expressed through a text and the hierarchy of importance of information points within it. She assumed little, if any, difference between comprehension ability in listening and reading, and she described the functioning of the framework as follows:

> According to the Structure Building Framework, the goal of comprehension is to build cohesive, mental representations or structures. The first process involved in building a structure is laying a foundation. The next process involves developing the structure by mapping on incoming information when that information coheres with the previous information. However, when the incoming information is less coherent, comprehenders employ a different process. They shift to initiate a new substructure. Thus, most representations comprise several branching structures (1990:221).

Gernsbacher's theory proved particularly valuable in that it was one of the first theories of language comprehension which sought not only to account for numerous comprehension phenomena but also to provide an explanation for the differences that can be observed between good and poor comprehenders. (For a fuller description and discussion, see Gernsbacher 1990, 1997, and Gernsbacher and Foerstch 1999). However, it remained unclear

how different substructures relate to one another, and Gernsbacher herself expressed uncertainty as to whether the mental structures and substructures proposed by the Structure Building Framework can be equated to the situational or mental models described above. She did, nevertheless, concede that the cognitive processes and mechanisms involved in structure building also depend upon comprehenders' ability to envision real-world situations.

From the review and discussion of various theories of reading comprehension so far, it becomes clear that by the 1990s cognitive psychologists, text analysts and other reading researchers regarded comprehension as involving processes at a number of different levels, from a basic or superficial perceptual level down to a much deeper level at which some sort of unified mental representation of a text's content is constructed with reference to a world (real or possible) that exists beyond the words on the page. Gerrig (1988) expressed this latter notion in terms of situation models making it possible for readers 'to reason about texts at a level liberated from the actual words and sentences that have comprised the text' (1988:255). With regard to the testing of reading comprehension, therefore, test developers clearly need to be concerned that the reading measures they design are focused on testing this structure building and overall meaning construction.

The importance of anaphora and referential coherence

Psycholinguists and discourse analysts alike acknowledge the importance of the way readers use a variety of linguistic devices to construct meaning, ensuring that propositional content across clauses and sentence is woven together in a coherent manner. In a recent chapter on the comprehension of discourse, Rayner, Pollatsek, Ashby and Clifton (2012) provide a useful overview of how features such as pronominal and nominal anaphora, connectives and thematic coherence contribute to relations between propositions, together with other linguistic devices used by writers to encode the information structure of a text for readers.

An emphasis on anaphoric resolution and referential coherence was central to Sanford and Garrod's Memory Focus Model of discourse comprehension. Like Kintsch, they developed their model over time (Garrod and Sanford 1994, Sanford and Garrod 1981). Gernsbacher and Foerstch (1999) compared Garrod and Sanford's emphasis on anaphoric resolution and referential coherence with Kintsch and van Dijk's (1978) preoccupation with propositional transformations and the representation of meaning, pointing out that these are essentially the same processes. The distinction was helpfully illustrated in the following way:

> For Kintsch and van Dijk, the primary question of interest during processing is "How does this new element change the scenario I am

> constructing?" For Sanford and Garrod the primary question of interest is "Does this new element refer to something mentioned previously in the text, and if so, what?" (Gernsbacher and Foerstch 1999:288).

As we shall see, results of the empirical studies reported in later chapters provide some clear evidence showing how anaphoric resolution and referential coherence, as well as propositional transformation and the representation of meaning, can be key indicators of good and poor comprehension in reading comprehension tests.

The nature and role of inferencing in text comprehension

Important evidence that readers are capable of constructing a mental representation of this depth and richness is their ability to make inferences in relation to a text. For this reason, it is appropriate at this point to consider views on the nature and role of readers' inferencing in text comprehension. The role of inference generation in comprehension has been extensively studied and various classificatory systems for inferences have been proposed, categorising these by content, by function, by logical form and by direction (see Kintsch 1993, Singer 1994, 2007).

Just and Carpenter (1987) classified inferences by direction according to two types. In the first type, when a text does not explicitly indicate how the sentences or clauses of a text are related to one another, the reader must infer the relation so that 'new' information provided by a text can be integrated with 'given' information. This type is termed a *backward* inference by Just and Carpenter, although it has also been referred to as a *bridging*, *integrative*, *connective, necessary* or *linking* inference in the literature. Field (2004:129) described how a bridging inference works:

> To achieve a full understanding of:
> a. *Bill had been murdered. The knife lay by the body.*
> it is necessary to infer that *the body* refers to Bill, that *the knife* was the weapon and that Bill was murdered by stabbing. Only in this way can the reader impose *coherence* upon the text.

A second type of inference proposed by Just and Carpenter (1987) involves embellishing the representation of a text being read. This type is described as a *forward* inference, sometimes also known as a *predictive*, *extrapolative, evaluative* or *elaborative* inference. Field (2004:130) explained elaborative inferences in the following way:

> The reader uses this type of inference to enrich an interpretation, but it is not essential to understanding and can readily be reversed if later information indicates it is incorrect. Cancellation does not cause disruption

> to the representation of the text that has been constructed. . . . Bridging inferences are stored as part of an ongoing mental representation. . . . Elaborative inferences differ from bridging ones in that they do not appear to form part of the mental representation. It has been suggested that, while bridging inferences are made on-line during text processing, elaborative inferences may not be made until later, during recall.

While there is a strong likelihood that *backward* inferences will be made in the comprehension of text in order to achieve coherence, *forward* inferences are optional and are far less likely to occur unless specifically encouraged by some associated task (such as a comprehension question on the reading text or a recall activity). Field (2004) cautioned, however, that the distinction between necessary and elaborative inferences may not always be clear-cut. Nonetheless, as test developers seeking to assess reading comprehension we may well need to draw a clear line between inferences that are necessary to satisfactory comprehension, and thus a legitimate testing focus, and inferences that are more elaborative (and possibly individualised) in nature, and which it would be unfair to try and test (see also Urquhart and Weir's discussion of comprehensions and interpretations (1998:112–120)).

Models of inference generation attempted to explain the evidence arising from experimental studies from differing perspectives. McKoon and Ratcliff (1992) proposed a *minimalist* view, suggesting that inferences are only made when required for local coherence in accordance with the basic principle of parsimony or economy of effort. Williams (1995), however, cited research results indicating that some of the spontaneous inferences made by readers are more closely related to a search for global rather than just local coherence and for this reason a minimalist view of inference generation may be too restrictive.

Graesser, Singer and Trabasso (1994) summarised the various descriptions of inferences occurring in narrative texts and identified as many as 13 different classes which they grouped into four separate categories as follows:

- inferences necessary for local coherence
- inferences necessary for global coherence
- elaborative inferences not necessary for coherence
- inferences which reflect a pragmatic exchange between the author and reader.

Graesser et al proposed a *constructionist* model of inference which is more flexible than McKoon and Ratcliff's minimalist theory and which adopted the view that comprehension is guided by the basic principle of search or effort after meaning. The authors underpinned their theory with three essential assumptions:

- the *reader goal assumption*, according to which a reader constructs a deep-level meaning representation that addresses the reader's goals
- a *coherence assumption*, according to which a reader seeks to construct a meaning representation that is coherent at both global and local levels
- and *an explanation assumption*, according to which a reader attempts to explain why actions, events and states are mentioned in the text because of the desire to achieve coherence in understanding.

Although the constructionist model provided for greater flexibility as far as inference generation is concerned, Williams (1995) pointed out that it still failed to account adequately for those occasions when inferences which are unnecessary for local or global coherence are made as a result of strong contextual constraints or manipulation of the discourse focus, as has been observed with noun instantiations and instruments of action (Garrod and Sanford 1981, O'Brien, Duffy and Myers 1986). Williams concluded that no neat set of assumptions or principles can predict the circumstances under which all types of inference will and will not be made, and that this is hardly surprising given the range of types of inference. He suggested, nevertheless, that there may be a core set of necessary, coherence-maintaining inferences which are essentially retrospective in that they form a bridge between a current element and previous parts of a text. Their function is to support the process of integrating a sentence into the preceding discourse structure, whether at a local or global level. In addition to the core set of necessary inferences, it is possible to conceive of other likely inferences which, while not being strictly necessary to maintain coherence, may nevertheless be generated on the basis of particular activating factors such as the strength of association with explicitly stated information in the text or the reader's motivation and purposes in relation to the text.

In a later chapter this flexible view of inference generation will provide the theoretical context for an analysis and discussion of actual inferences made by readers of two particular texts. Careful consideration will be given to distinguishing those inferences which are optional, and possibly more idiosyncratic in nature, from those inferences which are core or integral to a sound understanding of the text, and which can justifiably be incorporated into a testing format that aims to measure readers' comprehension ability.

The nature of reading comprehension in a second or foreign language

No distinction has been made thus far between the nature of reading comprehension ability in the L1 and reading comprehension ability in an L2. Although the focus of the studies reported in subsequent chapters in this volume is restricted to reading assessment in the L1 context, the issue of L2 reading

assessment remains important given that investigations into L1 reading have been instrumental in informing and guiding the more recently established but equivalent field in L2 reading research and that testing practices for L2 comprehension ability have sometimes drawn heavily upon approaches in the assessment of L1 reading comprehension. Specific issues concerning the nature of L2 reading have been comprehensively reviewed elsewhere so they will not be discussed here. Key references spanning the past 30 years include: Alderson (1984, 2000), Alderson and Urquhart (1984), Bernhardt (1991b), Carrell (1991), Carrell, Devine and Eskey (1988), Clapham (1996), Devine (1988), Enright, Grabe, Koda, Mosenthal, Mulcahy-Ernt and Schedl (2000), Grabe (1991, 2009), Grabe and Stoller (2002), Khalifa and Weir (2009), Koda (2005) and Urquhart and Weir (1998).

More importantly, there is a growing sense that the traditional L1 versus L2 distinction which has often shaped research into reading and other skills is gradually giving way to a different paradigm that draws on the notion of expertise and the novice–expert user continuum, and for which there is a burgeoning research literature (see, for example, Ericsson, Charness, Feltovich and Hoffman (Eds) 2006). Field (2011:68–69) explains this approach as follows:

> . . . the premise is adopted that underlying the four language skills are certain established and shared routines which can be traced by examining and comparing the performance of expert language users. This assumption is supported by two lines of argument:
>
> a) *The universal argument*. All human brains are similarly configured. They can be assumed, at some level of generality, to share processing routines which are broadly similar in that they reflect the strength and limitations of the organ and the means it adopts for transmitting information. These routines might be deemed to contribute not simply to the forms that language takes but also to the ways in which it is processed in performance.
> b) *The expertise argument*. A marked difference between an adult L1 speaker and an L2 learner lies in the fact that the former has had many years of experience during which to develop the most rapid and most effective processing routines for dealing with the vagaries of the target language – and to develop them without competition from deeply ingrained routines associated with another language. An understanding of how such expert users perform should thus assist us in directing the development of novice users. The novice/expert distinction is not, of course, an all-or-nothing one. There exists a continuum of expertise stretching from novice to expert user, which is highly relevant in the context of language proficiency assessment, where gradations of ability need to be distinguished and accredited for teaching and learning, employment or other social purposes.

Thus it is anticipated that results from the empirical research reported in later chapters of this volume will have as much relevance for the design and construction of reading comprehension tests for foreign and second language learners as for those reading in their first language.

Conclusion

What seems clear from the review and discussion in this chapter is that current models for text comprehension, and by extension for reading comprehension, all stress the active and constructive nature of the comprehension process in which the reader's cognitive processes and mechanisms interact to build some sort of representation of text meaning at the discourse level. If we are looking for evidence of successful comprehension on the part of a reader then this will need to be measured through a format that is designed to directly address the high-level discourse representation constructed by the reader in relation to a text. When claiming to test reading comprehension, therefore, test developers clearly need to be concerned that the reading measures they design are capable of tapping into the reader's overall meaning construction.

In light of this, three important questions arise in relation to current approaches to the testing of reading comprehension ability. First, what are the fundamental principles underpinning past and current approaches to the design and construction of tests of reading comprehension ability? Secondly, to what extent do these principles take adequate account of our current understanding of the nature of reading comprehension? Thirdly, how might it be possible to reconcile more closely the latest theory of reading comprehension with the practice of assessing it? As further background to the empirical research reported later in this volume, Chapter 3 will consider these questions in more detail to establish how far conventional reading comprehension tests require the reader to build up a mental representation of text to the discourse level and how far comprehension questions posed on a reading text address the overall understanding that test takers construct from integrating information across the text as a whole.

3 An overview of reading comprehension test design

Introduction

The assessment of reading comprehension ability has for many years been regarded as a key element in determining an individual's level of language proficiency, either in a first or in a second or foreign language. Spolsky claimed that 'the very first modern language test (Handschin, 1919) was a test of silent reading, that is, comprehension', adding that anyone designing a language test today 'would probably start by automatically assuming that the two principal parts will be tests of reading and listening comprehension' (1994:141). A more recent, and fascinating, scholarly contribution on the history of reading assessment in the UK is offered by Weir (2013b), who traces its development from the medieval tests of oral reading in the 14th century up to the silent approaches to testing reading in the 20th century to which Spolsky referred. Weir asserts the value of knowing how and why tests of reading have evolved over time in the way that they have:

> A familiarity with how the reading construct was measured in the past provides us with a valuable perspective when developing 'new' reading tests or critiquing existing ones (2013b:103).

This point is well taken with regard to the research reported later in this volume, the aim of which was to develop a new type of reading comprehension test that seeks to address some of the constraints and limitations inherent in many contemporary reading measures. Weir's comment thus provides a convincing rationale for the brief survey of reading test design offered in this chapter.

Weir's (2013b) account of the development of the Cambridge English examinations illustrates clearly some of the major changes which have taken place over the past century regarding the assessment of reading. Cambridge English Language Assessment (formerly Cambridge ESOL) did not test reading in a dedicated, separate paper until 1975 when a component entitled Reading Comprehension was introduced to the First Certificate in English (FCE) and Certificate of Proficiency in English (CPE) examinations. Previously, the testing of reading in the Cambridge English examinations had

been just one aspect of a number of integrated language use tasks favoured by language testers for much of the 20th century. In the early years of that century, testers were primarily concerned whether students could write a good summary (a shortened version of a written text, containing the main points and omitting minor details), or translate a passage from a work of literature, or read aloud a literary prose passage. Integrated tasks of this nature would clearly *involve* some reading ability but would not *focus* uniquely upon it. In this early period it is unlikely that testers thought of themselves as measuring a reading construct *per se*, or at least as we conceive it today.

Though there is some evidence of more focused reading activities in the 1960s, such tasks were located within the English Language and Use of English sub-tests, and the emergence in 1975 of a dedicated, multiple-choice Reading Comprehension paper in its own right marked something of a watershed. It indicated an increased awareness of, as well as a receptiveness to, new developments in cognitive psychology, applied linguistics and classroom pedagogy. As understanding of comprehension (as opposed to lower-level decoding) developed during the communicative era (Urquhart and Weir 1998) and as classroom reading materials and teaching began to focus on higher-level reading skills and strategies from the 1970s onwards, so test developers at Cambridge, and probably elsewhere, abandoned tasks that had involved reading in an integrated mode (e.g. translation, reading aloud and summary) to focus more specifically on the testing of reading in its own right.

The rest of this chapter will examine the componential skills view which came to shape approaches to testing reading in the 1980s and 1990s leading on to more recent attempts in the 2000s to take account of higher-level cognitive processes in reading test design. The seemingly intractable problem of which reading skills or cognitive processes can be identified or should be tested will be considered and a possible solution will be explored involving a return to a form of summarising task, which is, ironically, the approach that was adopted in the Cambridge exams a century ago.

20th century approaches to assessing reading

Most language proficiency measures today include a test component specifically designed to assess an individual's reading ability. As we shall see below, recent decades have seen rapid expansion in the development of different reading tests for different purposes, some claiming to measure general reading ability and others claiming to assess more specialised reading for academic or professional purposes. Test designers and constructors now have a comprehensive toolbox at their disposal containing different question formats or item types from which to choose when constructing an established reading test or designing a new one. These include task formats such as matching,

true/false, multiple choice, sentence completion, cloze, C-test, short-answer question, information transfer and others.

A distinction is sometimes drawn between *selected response* formats, which require the test taker to choose the correct answer from among a number of options offered (e.g. by shading a lozenge or circling a letter on an answer sheet), and *constructed response* formats, where test takers have to produce a written answer themselves. Several recent publications describe in detail the wide range of question formats or item types used in reading tests and these are helpful in considering the relative strengths and limitations of different test tasks for assessing reading ability – see, for example, Alderson (2000), Hughes (2003), Khalifa and Weir (2009), Urquhart and Weir (1998) and Weir (2013b).

The choice of precisely which task or question format(s) to include in a particular reading test typically depends upon the purpose and context for which the test is designed. For example, a low-level test of decoding skills in reading often contains short items requiring few or no productive skills on the part of a test taker. A higher-level reading test, on the other hand, may require a candidate to summarise part or all of a given text and even evaluate the author's style. Various methodological and pragmatic considerations that apply in reading test design and delivery are clearly important. For example, selected response items, such as 4-option multiple-choice items, tend to be favoured by large-scale test providers because they are easier to score when large test taker populations are involved and because they can contribute to overall test reliability, in terms of both internal consistency and marker reliability.

Of far greater importance, however, is the extent to which any reading test can be assumed to be measuring what it is intended to measure, often referred to as the construct validity of the test. Language testers need to avoid what is generally termed *construct-irrelevant variance*, i.e. the measurement of traits other than the particular trait of interest (Messick 1989). For example, a test which purports to measure a reader's comprehension ability may risk measuring other traits, such as their general knowledge or their writing ability. Thus the test format employed may also be a source of construct-irrelevant variance. If the process of arriving at the correct answers on a reading comprehension test does not adequately reflect the processes involved in reading tasks in the world beyond the test, then the construct validity of the test as a suitable measure of reading comprehension ability is called into question.

Interest in construct validity issues (in terms of the relationships between performance on language tests and the abilities underlying this performance) can be traced back as far as the 1940s (Brereton 1944, Roach 1945). It was the 1960s and 1970s, however, which began to see considerable advances in both theoretical discussion and empirical research though, as Khalifa and Weir

(2009) pointed out, approaches to demonstrating construct validity in the latter decades of the 20th century still drew primarily on *a posteriori* quantitative studies using factor analysis. Only since the turn of this century has there been a sustained focus on exploring the cognitive processing involved in reading tests and in mapping this back to real-world reading activities as part of the activity of construct validation.

The remainder of this chapter reviews and discusses in more detail some of the theoretical developments that strongly influenced approaches to reading assessment in the 20th century and which continue to shape current practice in reading test design in much of the world today, despite our growing understanding of the constructed nature of comprehension. As will become evident, some of the discussion about approaches in reading assessment necessarily echoes the Chapter 2 discussion regarding the evolution of reading theory. Consideration is also given in what follows to the importance of research and construct validity in relation to reading comprehension test design and construction. The aim of this chapter (and also of Chapter 4, which focuses specifically on the use of summarising tasks for testing reading) is to prepare the ground, in terms of both the theory and practice of assessing reading comprehension ability, for the proposal of a radically different form of reading comprehension test.

Early attempts to define the *testable* aspects of reading comprehension ability

Carroll (1972) claimed that it is essential to have a definition of listening or reading comprehension ability if appropriate measures of that ability are to be developed. This view was reiterated by other language testers over subsequent decades – see, for example, the developments in reading assessment described by Alderson (2000), Hughes (1989), Khalifa and Weir (2009), Urquhart and Weir (1998) and Weir (2013b).

According to Carroll (1972) and Andrich and Godfrey (1978/1979), initial attempts to define reading in terms of its testable skills are considered to date from the 1940s. During this period a major debate arose on whether the ability to comprehend what we read or hear should be conceptualised as a uni-dimensional skill or as a set of multi-dimensional skills. Davis (1944) claimed to have confirmed the existence through a factor analytic study of nine independent and testable reading comprehension skills and these were reflected in his Cooperative Reading Comprehension Tests. Using a different factor analytic approach to analyse the same data, Thurstone (1946) disputed Davis' view, claiming that the data could be adequately accounted for by a single factor. This debate was to continue for over half a century and was to have a significant influence over the design and development of later reading proficiency measures.

Taxonomies of reading sub-skills

Davis (1968) reaffirmed the existence of eight separate and testable skills as listed below:

1. Remembering word meanings.
2. Drawing inferences about word meaning from context.
3. Finding answers to questions answered explicitly or in paraphrase in the content.
4. Weaving together ideas in the content.
5. Drawing inferences from the content.
6. Recognising a writer's purpose, attitude, tone or mood.
7. Identifying a writer's techniques.
8. Following the structure of a passage.

This list of eight skills gives the appearance of being hierarchically structured with each skill subsuming the preceding ones. Andrich and Godfrey (1978/1979) commented that although individual test items were not originally constructed with this hierarchy in mind, Davis clearly recognised some sort of hierarchical element when he summarised reading comprehension research. They suggested that Davis mirrored his own set of skills with those described in Bloom's (1956) *Taxonomy of Educational Objectives Handbook 1: Cognitive Domain*.

For each of the eight skills identified, Davis constructed two sets of 12 items each. He then composed two parallel test forms (Experimental Reading Tests – Forms C and D) each consisting of 96 items. Following administration of these two tests, Davis concluded that they succeeded in adequately distinguishing only five out of the eight skills. Carroll (1972) confirmed that it was logical to consider only five of Davis' original eight as being distinct skills: i) remembering word meanings; ii) finding answers to questions answered explicitly or in paraphrase; iii) drawing inferences from the content; iv) recognising a writer's purpose, attitude, tone or mood and v) following the structure of a passage.

Debate over the uni-dimensionality versus multi-dimensionality of reading comprehension ability continued, however, with Thorndike (1973) concluding that although the reliable variance could be accounted for by three factors, 93% of this variance was accounted for by one single factor. This uni-dimensionality versus multi-dimensionality debate can be seen as part of a much larger and ongoing theoretical discussion concerning the nature of language proficiency itself, which found expression in the work and writings of John B Carroll (1972) and John Oller (1972).

Andrich and Godfrey (1978/1979) re-examined one of Davis' reading comprehension tests according to the Rasch uni-dimensional latent trait model, since previous factor analyses of the test items seemed to offer conflicting

interpretations concerning the single-factor versus many-factor issue. They examined two main hypotheses: first, that the items conformed to a uni-dimensional set, and secondly, that the eight specifically constructed groups of items would form an interpretable hierarchical ordering. The results of their study appeared to confirm the uni-dimensional hypothesis quite strongly, although the hierarchical hypothesis was less clearly supported. This study produced two rather interesting results in relation to performance on particular types of item. First, the vocabulary items in the reading test proved to be the fifth most difficult group of items among the eight groups, rather than the easiest group according to the hierarchy. Andrich and Godfrey explained this in terms of the words themselves being difficult but it not being necessary to know them in order to answer the questions in the remaining items. Second, it was observed that seven out of a total of nine items involving humour did not conform to the uni-dimensional hypothesis. The authors suggested that items such as these should probably not be included in comprehension tests, 'since the question of whether a test-taker's sense of humour agrees with the test constructor's and thus affords the choice of a correct response should not be at issue in evaluating comprehension ability' (1978/1979:198). This observation highlights the fact that not everything which makes up our understanding of a text (e.g. appreciation of humour) will necessarily be a suitable candidate for testing purposes. The same applies to elaborative inferences made by reader or listener to enrich their mental representation but which may be highly idiosyncratic and not necessarily mandated by the writer or speaker of the original text.

Based on their examination of Davis' reading test, Andrich and Godfrey concluded that it was quite possible to conceive of reading comprehension as being uni-dimensional with a number of factors representing skills at different levels on the dimension. Subgroup clusters were identified according to mean item difficulty (with vocabulary omitted). So, on the one hand, at what might be termed a macro-level analysis, Davis' test provided a measure of a unitary trait. On the other hand, at what might be termed a micro-level analysis, it appeared to measure four skills, regrouped and renamed as:

1. Understanding the content (easiest).
2. Weaving ideas and meaning from context.
3. Following the structure of the content.
4. Recognising the author's literary methodology (most difficult).

Weir, Hughes and Porter (1990), Weir and Porter (1994) and Alderson (2000) all reviewed the literature discussing the divisibility of reading skills as far as reading assessment was concerned. Alderson concluded there was no consensus on this issue:

> . . . three positions are common:
> • the first is that reading is a unitary skill

- the second is that reading is multidivisible, even though there is no agreement on how many skills might be empirically distinguishable
- the third is that there is a two-way split (2000:95).

See also Urquhart and Weir (1998) for a helpful discussion reviewing the evidence for uni-dimensional and multi-divisible views of reading, and in particular the usefulness of characterising reading according to a 4-component matrix involving global/local and careful/expeditious types of reading.

Davis was not the only researcher whose hierarchical and taxonomic approach to defining reading skills influenced the design and construction of much reading comprehension assessment. Barrett (1968) sought to provide a manageable and understandable means of teaching reading comprehension through his *Taxonomy of the Cognitive and Affective Dimensions of Reading Comprehension*. Like Davis, he broke reading comprehension down into five major skill categories or levels:

1. Literal comprehension (recognition of main idea, details, sequence, comparison and cause and effect).
2. Reorganisation (classifying, outlining, summarising, synthesising).
3. Inferential comprehension (inferring supporting details, main ideas, sequence, comparisons, cause and effect, character traits, outcomes).
4. Evaluation (judgements of fact, opinion, validity, appropriateness, worth).
5. Appreciation (emotional responses to content, characters' language use, imagery).

Although it is difficult to match Barrett's categories precisely to Davis' skills analysis, some measure of overlap is immediately obvious. Once again, the five categories were ordered to move from the easy to the difficult in terms of the apparent demands they make on the reader, although, as Clymer (1968) observed, some tasks both between and within categories may prove more demanding than their theoretical position on the scale would suggest. Barrett's system of categorisation was designed primarily for application in the teaching context, first as a basis for developing teaching purposes and questions to guide children's reading, and secondly as a convenient means for teachers to analyse the types of comprehension questions found in basal reading manuals. It would be unfair to assume that the system was conceived as a methodology for designing and constructing formal reading tests. Nevertheless, it attracted considerable attention for its potential application in developing various tests, including the famous Edinburgh Reading Tests.

One further taxonomy merits attention with regard to the assessment of reading comprehension, especially in the context of testing English as a foreign or second language. Munby's *Communicative Syllabus Design* (1978) offered syllabus designers and materials writers a highly detailed inventory

of multiple aspects of language proficiency. His taxonomy had its origins in two important trends that developed during the 1970s. The first was the growing view of language proficiency as possessing a sociolinguistic dimension. Hymes (1972), for example, conceived of communicative competence as comprising both a linguistic and a sociolinguistic element, thus acknowledging the importance of context to the appropriate use of language and the dynamic interaction that can occur between context and the discourse itself. Around the same time, there was a move within Europe to develop a language teaching programme that would apply across linguistic frontiers, known later as the Council of Europe's *Threshold Level*. The *Threshold Level* had its roots in the Rüschlikon Symposium of 1971, following which Richterich (1972) detailed the language needs and learning needs relevant to the communicative situations in which a learner might have to use a foreign language, especially in relation to mobility across European borders. Howatt (1984) speculated that Richterich's communicative needs analysis provided the starting point for Munby's more elaborate version. Combining a sociolinguistic perspective together with a needs analysis approach, Munby (1978) offered a model for defining the content of purpose-specific language programmes. The model consisted of a vast taxonomy of different but related language skills and sub-skills although these were not presented as being hierarchically organised along the lines of the taxonomies described previously. With regard to comprehension (little distinction was drawn between reading and listening), Munby described at least 16 distinct reading-related skills covering over 60 separate sub-skills. Munby's analysis of the different potential skills and sub-skills involved in English proved to be particularly influential, not only in the field of syllabus and course design but also in the development of tests of English as a foreign/second language, although the approach did attract serious criticism from some quarters (Davies 1981, Mead 1982).

Problems of using a sub-skills taxonomy to define reading ability

Clymer (1968) highlighted several problems associated with a taxonomic approach to defining reading skills such as those described above, whether for teaching or testing purposes. Some issues relate to the nature of taxonomies in general, while others are specific to using a taxonomic approach in reading test design. One general problem is that the orderly presentation of categories in any taxonomy is likely to suggest a much greater precision than the classification actually possesses. Another is that a system of categories fails to take into account the overlap which may exist in certain types of reading test question. Finally, type of comprehension demanded and difficulty of task can result from at least three factors – the text, the questions, and the

reader's background knowledge. Traditional taxonomic approaches, suggested Clymer, can take only the first two into account.

Clymer was not the only writer to criticise the sub-skills approach to analysing reading comprehension ability. Others expressed concern over the level of conjecture involved, issues of impracticality and a lack of empirical evidence (Davies 1981, Mead 1982, Skehan 1984). In a critique of various attempts to validate Bloom's (1956) taxonomy, Seddon (1978) concluded that research had generally failed to provide convincing evidence of either the educational value or the psychological reality of the taxonomy. Lunzer, Waite and Dolan (1979) constructed L1 reading tests for children designed to measure different reading skills but were ultimately unable to find adequate evidence of either an implicational scale among the skills or of the existence of clearly distinguishable skills. Their conclusion was that, in relation to reading tests at least, reading can be defined as a single, unitary skill. Alderson (1990a) also questioned the extent to which it is possible to speak of an implicational scale (or cumulative hierarchy) of reading skills, in which the acquisition and development of lower order skills (such as the ability to understand explicitly stated information) is a necessary prerequisite for the acquisition and development of higher order skills (such as the ability to evaluate). Quinn (1993) condemned any attempt to deconstruct language skill into atomised elements as 'a distorting and trivialising process of reductionism' (1993:71).

Reading sub-skills analyses and the design of reading tests

Concerns about its validity did not prevent the definition of reading in terms of a taxonomy of separate sub-skills from exerting a powerful influence over reading syllabus design and teaching materials produced in the latter part of the 20th century (Grellet 1987, Yalden 1987). This is perhaps understandable given that both Barrett's and Munby's taxonomies were conceived largely to provide direction and specific objectives for the teaching progamme. In addition, however, such taxonomic approaches exerted considerable influence over the design and construction of reading comprehension tests, and other tests of language proficiency, in both L1 and L2.

Following a detailed analysis of three representative American tests of reading, Just and Carpenter (1987) concluded that most standardised reading tests tended to converge on the following four categories:

1. Understanding the important facts in a text.
2. Abstracting the main point of a text.
3. Making the inferences that the author intended.
4. Analysing a text's organisation.

They added two additional categories which they believed were occasionally though not usually addressed through reading tests: first, reading to critically

evaluate the logic of a text, and secondly, recognising the tone and style of a text.

The Test of English for Educational Purposes (TEEP) and the English Language Testing Service (ELTS) were two examples of L2 tests whose reading components were very explicitly linked to a taxonomy of reading sub-skills. The TEEP reading test was based upon a list of 17 enabling skills ordered according to whether they were higher- or lower-level (Weir 1983, 1990) while the ELTS reading test was based on a list of 19 enabling skills (Carroll 1978, 1981, Criper and Davies 1988, Davies 2008). For both tests, explicit claims were made about what each item tested. Less explicit but similar approaches were adopted in the design of other L2 reading tests (see, for example, description and discussion of the Cambridge English reading tests in the 1990s in Weir and Milanovic (Eds) (2003) and Hawkey (2009)).

The strong influence of the reading sub-skills approach during the 1980s and 1990s can be perceived in some of the publications that offered test designers and constructors practical advice on how to develop reading tests. In one of the earliest books offering guidance on writing English language tests, Heaton (1988) wrote of 'reading skills and strategies', drawing directly on Munby's analysis 'to identify some of the specific skills involved in reading' (1988:105). Weir (1990) offered useful advice on the various methods available for testing reading, but at that time wrote little about the nature of the underlying construct of reading. Hughes (1989) discussed the twin notions of broad reading skills or *macro*-skills (e.g. scanning for specific information, identifying stages in an argument), and underlying reading skills or *micro*-skills (e.g. identifying referents for pronouns, using context to guess word meaning, understanding relations between parts of a text). What seems curious from his 1989 analysis is how the micro-skill of *understanding relations between parts of a text* could be distinguished in testing terms from the macro-skill of *identifying stages in an argument*. Interestingly, Hughes did not really address the important issue of the nature of reading comprehension and made almost no reference to comprehension throughout the chapter on the testing of reading in his book. Instead, he offered suggestions about the two broad levels of reading at which he believed testing is most usefully carried out: the micro-skill level, reflecting the underlying reading skills taught as part of a reading course in the belief that they will promote the development of macro-skills; and the macro-level. It seems clear that Hughes saw the micro-skills as being subsumed within the higher-level macro-skills for he stated that, in some tests, there was a case for 'having only items which test macro-skills, since the successful completion of these would imply command of the relevant micro-skills' (1989:117–118). Alderson (1990a) commented that it was common practice for teachers and test writers to assume both a hierarchy and an implicational scale in relation to reading sub-skills.

Weir's (1993) book, entitled *Understanding and Developing Language Tests*, began to refer to some processing aspects of comprehension ability, such as the integration of background knowledge and the effect of reading purpose. He included much discussion of the types of underlying skills and strategies or the specified operations that apparently contribute to the ability to understand text, together with a summary checklist of selected reading operations. Though this continued the earlier tradition of lists of sub-skills, Weir defended the sub-skills approach to reading test design as follows:

> In our present state of knowledge, if we wish to report on students' proficiency in reading, as distinct, say, from writing ability, then we are forced to break reading down into what we conceive are its constituent parts. It is accepted pedagogical practice to break the reading process down in this way and to teach the component skills and strategies to a certain extent separately. So, to that extent, it should be possible to focus on these components for testing purposes. If we can identify skills and operations that appear to make an important contribution to the reading process, it should be possible to test these and use the composite results for reporting on reading proficiency (1993:72–73).

While Weir acknowledged that the reading operations he specified might not be a fully valid model of what reading comprehension actually is, he nonetheless suggested that as soon as individual items are written on a passage we are inevitably engaged in the business of measuring reading skills, either individually or in combination.

It seems clear that, up until the mid-1990s, most authors offering guidance on reading test design and construction either accepted a view of reading as component sub-skills or saw little need to be explicit about the nature of reading comprehension, perhaps because there existed among teachers and testers some intuitive understanding of what was to be tested. Some authors, such as Weir (1993), made limited reference to the processing involved in reading comprehension, but still considered reading assessment largely in terms of the testing of specified reading skills operationalised through individual test items.

Investigating reading test items for construct validity

A more sophisticated and systematic investigation of the relationship between the content of reading test items and their performance gradually became possible towards the close of the 20th century, due partly to the availability of more sophisticated packages for quantitative data analysis, but more importantly to the development and application of qualitative methods such as introspection and retrospection (see Green 1998 for a full discussion with examples). These qualitative techniques enabled researchers in language

testing to actively explore (rather than just speculate on) the cognitive processes employed by test takers when responding to test items. This followed growing criticism of and debate over the traditional use of sub-skills analyses for designing and constructing reading tests.

In the early 1990s language testers were already debating the assumptions made by test constructors about which reading sub-skills were actually tested by which items in a given test (Alderson 1990a, Alderson and Lukmani 1989). In a study of a pilot version of the TEEP test and of the ELTS reading tests, Alderson investigated the relationship between individual items and the reading skills they were supposed to reflect. He reported serious disagreement between expert judges on the nature and difficulty level of skill supposedly being tested by each item. Furthermore, he could establish no stable relationship between (a) the nature and level of skill being tested by the items, and (b) the items' statistical difficulty and discrimination. Both Lumley (1993) and Weir and Porter (1994) challenged Alderson's findings on this point, however. Weir and Porter (1994) criticised Alderson's methodological approach, e.g. the use of untrained Master of Arts (MA) students as expert judges and the lack of adequate guidance or agreed on definitions of skills for informants. They also questioned his assumptions regarding the relative difficulty of test items and the implicational relationships between them. They argued that even if the High/Low distinction was open to question, for practical testing purposes it was nevertheless possible to obtain reliable judgements about likely skill focus from properly selected and trained judges as long as clear definitions of skills were available and there was a shared understanding of what is meant by these. Interestingly, Alderson's later account of the Diagnostic Language (Assessment) (DIALANG) project continued to make reference to targeting skills and sub-skills through the test and implied that expert judges had been capable of evaluating item level and focus for the test (2005:125–131).

A second area for debate concerned the construct validity issue of whether traditional reading comprehension items actually measured comprehension. Alderson (1990a) concluded that introspective and judgemental evidence suggested a reading test item would test more than one skill and that different readers might use different routes to arrive at the same answer. Various studies had identified variables other than comprehension which could contribute to the successful answering of a comprehension test item (Drum, Calfee and Cook 1981, Just and Carpenter 1987, Pollitt, Entwistle, Hutchinson and De Luca 1985). Swaffar, Arens and Byrnes (1991) complained that most reading comprehension items tested only test bottom-up skills, and thus failed to measure the way a reader actually understands text. Bernhardt (1991b) considered the construct validity of cloze testing for the assessment of reading to be profoundly inadequate while West (1991) noted the limitations of multiple-choice (MC) reading items for testing global comprehension. Rupp,

Ferne and Choi (2006) also questioned the value of MC reading tests 'as composite measures of higher order reading comprehension', i.e. their usefulness for assessing comprehension of the macrostructure of a situation model (2006:468). They concluded that the MC format may involve the reader in 'response processes that deviate significantly from those predicted by a model of reading comprehension in a non-testing context' (2006:469) and they hypothesised that '. . . responding to MC reading comprehension questions on many standardised reading comprehension tests is much more a problem-solving process relying heavily on verbal reasoning than a fluid process of integrating propositions to arrive at a connected mental representation of a text' (2006:454). There is also concern that the mental model which is normally created while reading a text is affected if candidates try to incorporate all the options provided to them by a test item into an ongoing text representation. The processing that takes place in working out which option fits, and which does not, would seem to bear little resemblance to the way we typically process texts for information.

Some continued to advocate the use of a sub-skills basis for reading test design and construction on the grounds that, since each test item required some theoretical justification, this was most appropriately supplied by linking it to a specified reading sub-skill or sub-skills even if these sub-skills were still only theoretical constructs with little empirical foundation. Weir (1993) defended this position on the grounds that, by adopting this approach, it was at least possible to investigate from a research perspective whether we were in fact testing what we set out to test. He suggested that over time it might become possible to determine the interactional relationships between the hypothesised skills. In the meantime, however, he concluded:

> The best we can do in the present state of knowledge is to ensure that if we wish to make comments about students' reading at a certain level of proficiency, then we include a range of formats and sufficient items to cover the range of skills that we believe are important and which together equate with our construct of reading (1993:75).

Weir's comments echoed Skehan's pragmatic view that without definitive theories of communicative competence and performance 'testers have to do the best they can with such theories as are available' (1988:211). As we shall see below, Weir's later work during the late 1990s and early 2000s would contribute significantly to the development of a theoretical framework for test development, derived from empirical research, which would assign cognitive processing a place at its heart.

The 1990s saw increasing calls for a reappraisal of traditional approaches to assessing reading comprehension as well as for research into the

development of alternative instruments with greater construct validity as far as cognitive processing was concerned. The International Reading Association, for example, called for measures that engaged and assessed the cognitive processes of reading rather than those that conceptualised reading as being made up of discrete skills (cited in the *Degrees of Literacy Power Program* (Touchstone Applied Science Associates 1991). Alderson (1990b) expressed a similar view:

> What a test of reading tests is not simply what its constructors say it tests, nor what a set of judges considers it to test. It must surely and crucially relate to what happens inside a test-taker's head when he or she responds to an item. Finding out that information, and discovering how generalisable the results are, is a neglected but important research endeavour (1990b:478).

Taylor's doctoral research undertaken during the first half of the 1990s (Taylor 1996) was a direct response to these calls for empirical research into cognitive processing in reading and its implications for the design of reading comprehension tests.

The situation was to change significantly by the turn of the century, not just in response to the sorts of hopes expressed above among language testers working in reading assessment, but also as the outcomes from reading comprehension research increasingly highlighted the potential for cognitive processing issues to explain individual reader differences. Perfetti (1997) argued that research into reader differences was central to understanding the nature of reading. Other researchers took a similar view of the significance of processing efficiencies in comprehension, highlighting the key elements of word recognition, syntactic parsing, proposition integration, text model building, inferencing and monitoring (Carver 1997, Gough, Hoover and Peterson 1996). Enright, Grabe, Koda, Mosenthal, Mulcahy-Ernt and Schedl noted the implications of this view for language testing:

> In language testing contexts, it would seem fairly apparent that vocabulary is a key co-variate with reading, as is, to a lesser extent, some measure of grammar knowledge. The high correlation of listening test scores with reading test scores has been attributed to the importance of general comprehension abilities associated with (a) generating a text model of comprehension, (b) forming an appropriate situation model relating reader knowledge with text information, (c) inferencing of certain types, and (d) monitoring comprehension strategically . . . (2000:3).

In their 1998 volume, Urquhart and Weir observed that a comprehensive model of the processing involved in different types of reading was not yet available and that, until this was forthcoming, testers might need to continue

basing their tests on componential models of reading ability, expressed in terms of skills and strategies. They nevertheless highlighted the importance of closely examining and controlling the textual parameters of any reading passage used for testing purposes as well as of understanding the effect of various performance conditions imposed in a test (e.g. time constraints, text type, response method) on both the process and the product of reading comprehension. Urquhart and Weir presumably looked forward to a day when a more fully developed *processing*, rather than mainly *componential*, model of reading, grounded in empirical research, would become available to language testers to underpin their test development and validation endeavours and to generate convincing evidence (theoretical and empirical) for claims regarding construct validity and test usefulness.

During the early 2000s Weir continued to work on developing just such a framework, seeking to bring together the cognitive and the social dimensions of both text and task within the testing event. The outcomes were a socio-cognitive approach to test development and validation, first published in Weir (2005) and later applied to the assessment of L2 reading in Khalifa and Weir (2009).

Based upon a comprehensive view of the research literature, Khalifa and Weir (2009) identified a number of levels of processing at both the decoding and the meaning construction levels: *word recognition, lexical access, syntactic parsing, establishing propositional meaning at clause and sentence level, inferencing, building a mental model, creating a text level representation*. Though this cognitive processing approach was theoretically better grounded than the earlier skills-based approach, it was not without similar problems in relation to operationalising the approach in an actual reading test. The test developer still has to determine the type and level of cognitive processing an item in a reading test should focus on, e.g. decoding or meaning building, and in what proportion across a set of test items. There is also the possibility that answering a question may entail parallel processing at multiple levels or that candidates may arrive at an answer via different routes. There is an additional decision to be made on how many test items should target each level of processing. So though test items may be based on a more construct-valid processing view of reading than was the case in the skills-based approach, the language tester is still confronted with some difficult decisions.

Conclusion

From the 1970s until the end of the 20th century the dominant paradigm when testing reading remained the skills and sub-skills approach, the legacy of which can still be seen today in many reading tests used around the world. In the 1980s and 1990s, however, testers found it difficult to agree on the sub-skills that could be consensually be identified and tested (Alderson 1990a,

1990b, Weir Hughes and Porter 1990, Weir and Porter 1994). In the early years of the 21st century the general lack of consensus on whether or not reading could be divided into sub-skills, and if so how many, gave way to a focus on the cognitive processes involved in reading in the hope that this approach might offer a satisfactory way forward for the testing of reading. Though much progress has been made towards a cognitive model of reading in recent years, it could be argued that considerable work is still needed to empirically ground the saliency of the various cognitive processes for assessment purposes, especially in terms of knowing exactly which of these processes should be covered in a reading test and by what method (a concern that echoes the pre-2000 worry over which sub-skills to sample in a test, and how).

What may be needed now for improved construct validity in reading comprehension assessment is an approach that will avoid the deconstruction of reading into either skills or processes, and will instead unequivocally account for the way readers naturally process the text as a whole, i.e. constructing a text-level representation. If developing a coherent and meaningful mental representation of the whole text is the highest order processing skill, and one that subsumes all the other levels in its successful completion, then perhaps it is worth focusing on this in our efforts to assess reading comprehension. By transferring our attention to testing the superordinate process in comprehension, i.e. the construction of a mental representation of a text (or even multiple texts), we may be able to escape some of the challenges and limitations encountered to date when designing reading tests. Designing a reading comprehension measure in a format that addresses the reader's holistic mental representation would have the added benefit of removing any need to satisfactorily identify and then adequately sample a range of targeted reading sub-skills or cognitive processes.

Several writers have discussed and advocated the use of recall and summarising tasks as effective measures of reading comprehension since these require readers to demonstrate their understanding of a passage by reproducing its overall content in their own words (Alderson 2000, Bernhardt 1991b, Kobayashi 1995, Riley and Lee 1996, Swaffar et al 1991). There is good reason to suppose, therefore, that recall and summarising tasks may more successfully address the reader's mental representation of the text as a whole and they clearly offer distinct advantages over more commonly used reading test formats, such as multiple-choice, cloze and short-answer questions. Nevertheless, summarising tasks also present a number of significant difficulties as tests of reading comprehension. The advantages and disadvantages of summarising tasks for measuring reading comprehension ability will be discussed fully in Chapter 4 and a possible, and potentially more viable, variant of a summarising task will be considered as an alternative solution – *summary completion technique.*

4 Using summarising tasks to assess reading comprehension ability

Introduction

The review and discussion in Chapter 3 suggested that much reading comprehension test design and development over the decades has been based upon an analysis of reading as a set of clearly distinguishable skills and sub-skills, or more recently as a series of cognitive processes ranging from decoding to various levels of meaning comprehension. In general, test developers construct individual reading test items, e.g. multiple-choice questions or cloze items, in an attempt to target and test discrete sub-skills or processes as a means of assessing a reader's comprehension ability.

Sometimes, however, a reading test will contain a task asking a test taker to *read and summarise* all or part of a written text. Underlying this latter approach is an assumption that satisfactory production of an oral or written summary can provide suitable evidence of the extent to which a given text has been understood by the reader. Summary tasks presumably have a fixed purpose and are premised on a consensus understanding of a text. It seems reasonable to assume that summarising tasks acknowledge, probably more than most other reading test tasks, a view of text comprehension as the construction of a mental representation of the whole text and that they therefore offer an appropriate format for assessing this.

It has been argued that creating a text-level representation constitutes the highest level in a socio-cognitive processing model of reading. Khalifa and Weir described this high-level processing as follows:

> At a final stage of processing, a discourse-level structure is created for the text as a whole. The skilled reader is able to recognise the hierarchical structure of the whole text and determines which items of information are central to the meaning of the text. The skilled reader determines how the different parts of the text fit together and which parts of the text are important to the writer or to reader purpose (2009:52–53).

Khalifa and Weir went on to assert that global, careful reading at the highest level requires the reader to understand the micro- and macro-propositions in a text and how these are interconnected, while integrating new information into a mental model to create a discourse level structure that is appropriate to their purpose (2009:60–61). Creating such a mental model involves most of the other lower-level reading processes and is probably the target reality for many readers in their real-world reading activity, and particularly for students involved in academic study.

This high-level view of reading resonates with the perspective of Enright et al who referred to the 'ability to construct a text model representation of what is read and also the ability to form a relevant situation model' combined with the ability to 'integrate and connect the detailed information provided by the author into a coherent whole' (2000:5–6). In their discussion of the internet-based Test of English as a Foreign Language (TOEFL iBT) reading specifications, Enright et al (2000:6) referred to this type of high-level processing as 'reading to learn' (see also Cohen and Upton 2006):

> From a skills processing perspective, reading to learn requires that a reader form linkages between a more elaborated model of text construction and frames (such as cause/effect, compare/contrast) to organize conceptual information and to understand the author's rhetorical intent. Conceptual knowledge that helps the reader integrate information in a text might include information derived from the text and/or from background knowledge. As such, it might represent an efficient alignment of the text model and the situation model.

Summarising tasks are generally considered to engage readers in precisely this sort of high-level processing because they require readers to identify and organise information that is key to overall meaning, sifting main ideas from supporting details and integrating these into a discourse structure that is consistent with writer/reader purpose.

This chapter considers how summarising tasks have been used in the past for assessing comprehension and then reviews some of the research into cognitive and other factors that are understood to be involved in the process of summarisation. Different types of summarisation are identified from the literature to consider whether and how these might inform our test task design. A range of problems associated with developing and using summary tasks for assessing reading comprehension ability is then explored. The final part of the chapter proposes *summary completion technique* as an alternative approach which has the potential to resolve many of the problems raised by traditional summary tasks while at the same time offering a test format that directly addresses the mental representation of text constructed by a reader. This provides the context and rationale for the series of empirical research

studies into developing and using summary completion techniques that are reported in later chapters of the book.

The use of summary tasks in language tests

A summary is generally defined as a short written or spoken account of text which provides the gist or main points of what has been heard or read, but not the details. Over a century ago, brevity and focus were among the essential qualities of a successful précis of written correspondence highlighted by Robeson (1913):

> The object of the précis is to enable any one reading it to be put into possession, in the smallest space of time, of the essential points of the subject to which the documents refer. The characteristics of a good précis accordingly are (a) the inclusion of all that is important and the exclusion of all that is unimportant in the correspondence; (b) the expression of this in a consecutive story as clearly as possible, and as briefly as is compatible with distinctness (Robeson 1913, cited by Weir 2013b:129).

Robeson's succinct analysis highlights some key principles of a successful summary: first that the key content will be retained and the unnecessary material will be discarded; secondly, that the discourse thread will be clear and accessible to the intended reader, as well as efficient. Both are principles which will apply to the empirical research reported in later chapters.

Summary tasks, in both oral and written formats, have long been used by teachers on an informal or semi-formal basis in classroom teaching and testing (see, for example, Pocock 1917, Robeson 1913). Most reading tests have required a written summary and, for this reason, examples of written summary tasks will be considered here. Examples of more formal, standardised tests that employed written summary tasks for assessing comprehension included English and foreign language examinations used throughout much of the 20th century in both the Scottish and English secondary education systems (e.g. O Grade examinations in English and in French, and Advanced Level examinations in French and other modern languages). The Certificate of Proficiency in English (CPE) (introduced by the University of Cambridge Examinations Syndicate (UCLES) in 1913) included a summary task in its English Literature paper as early as 1931. Weir (2013b) describes how a précis of a passage or a poem was introduced into the English Literature paper in Part B. Typically, candidates had to choose between summarising a passage, which included defining the meaning of words and phrases in the text, and explaining a poem in detail including a focus on style and diction. By 1936 there was a further summary task in the English Essay paper (rebranded in 1938 as the English Composition paper, presumably because candidates by

then had to write more than just an essay). In the 1938 paper candidates were instructed to read a text of 525 words and write a summary not exceeding 185 words.

Whether used with L1 or L2 examinations, and whether on a formal or an informal basis, tasks involving the written summarisation of text enjoyed a natural appeal as communicative approaches to language teaching and testing gained ground during the last quarter of the 20th century (see Chapter 1 of Hawkey 2004), especially where the assessment of language proficiency was linked to the requirements of academic study contexts for tests such as the TEEP. The TEEP was developed in the late 1970s as a university entrance measure by the UK's Associated Examining Board (AEB), though it is now administered by the International Study and Language Centre (ISLC) at the University of Reading. Weir (1990) provides an example of an early TEEP summary task (the first of two writing tasks). This was specifically designed to correspond to the sorts of language activities undertaken by students in higher education. It required test takers to write a summary of about 250 words in 45 minutes, based upon their reading of a written passage in a source booklet but also drawing upon input from the earlier Listening Comprehension part of the test (see Davies 2008:352 for an example of this summary task). A similar summary task, though restricted to reading and writing, was included in another test developed for use in international student admissions, the English Language Testing Service (ELTS) in 1980. ELTS contained a writing task that required test takers to summarise a passage from a domain-specific Source Booklet of reading passages (see Davies 2008).

More recent examples of summary tasks for assessing reading comprehension at a similar proficiency level can be found in the Pearson Test of English (PTE) Academic and in the Advanced General English Proficiency Test (GEPT) in Taiwan. These two tasks helpfully illustrate the continuum along which different variations of a summarising exercise can be conceptualised and constructed. The PTE Academic Writing variant requires test takers to read a short written passage on screen (up to 300 words) and then to summarise it using a single sentence, all within a 10-minute timeframe (see Pearson 2012 and O'Dell, Chandler, da Silva, Cotterill and Hogan 2013 for example tasks). At the other end of the continuum, the Advanced GEPT variant, like the more recent versions of the *Cambridge English: Proficiency* (formerly known as CPE) summary task, involves inter-textual summary, i.e. the summarising of content from more than one text. The written summary must be about 250 words in length and the task completed within 60 minutes (see Weir 2005:167–171 for an example of this task).

Other summary task variants to have appeared in recent years include those used in the new internet-based Test of English as a Foreign Language (TOEFL iBT) and in the British Council Placement Test. The TOEFL iBT task requires test takers to read a text and then complete a prose summary for

which one or two sentences are already provided. The test taker must select three additional sentences to add to this summary from a set of six options expressing the most important ideas in the passage (Cohen and Upton 2006). The summarising task in the British Council Placement Test provides a set of eight statements (labelled A–H), only six of which are true according to the set reading passage. The test taker must choose the six true statements in the set and sequence these in the order in which the information appears in the reading passage (O'Sullivan and Rugea 2011). Both the TOEFL iBT and the British Council Placement tests involve a selection-and-matching exercise (i.e. selected response format), rather than a compositional writing activity (i.e. constructed response format), a design choice on the part of the test developers which was presumably shaped by the computer-based delivery and objective scoring systems that underpin the tests.

The relationship between summarising and higher order reading skills

The rationale for using summary tasks such as those described above to assess reading comprehension lies in the relationship that is perceived to exist between the activity of reading and understanding text, and the activity of summarising. It has been assumed that production of a summary demands the use of so-called higher order reading skills and can thus provide insight into comprehension (Johnston 1981, Kintsch and van Dijk 1978). A summary can indicate, for example, whether a reader has grasped the main ideas, focus and viewpoint of the author of the text, while at the same time avoiding subjective comment and interpretation (Johns 1985). Not surprisingly, such higher order skills are often considered to be critical at higher proficiency levels, e.g. for those engaging in academic study or working in the professions.

In relation to testing, therefore, a summary task is believed, as discussed earlier, to force the test taker to demonstrate their use of higher order reading skills, such as identifying relevant information, distinguishing superordinate from subordinate material and eliminating redundant or trivial information. The premise is that test takers provide evidence of the extent of their comprehension of one or more texts and from this an inference can be drawn about their ability to comprehend texts of a similar nature, or texts in general. It is not surprising that summary tasks have come to enjoy a natural appeal, especially at higher proficiency levels. Not only do they appear to offer a means of testing the more elusive reading skills, such as understanding at a global level, but they do so within the context of a task format which simulates real-world activity. Even in general everyday life (as opposed to academic study or professional work contexts) readers do sometimes need to summarise a text they have read, although this may be more often in note form for personal use, than in continuous prose for the benefit of another reader.

Research into cognitive processes involved in summarising

The probable linguistic and cognitive processes involved in summarising written text (and their implications for reading and writing instruction) were extensively studied from the 1970s through to the 1990s (Armbruster, Anderson and Ostertag 1987, Brown and Day 1983, Chou Hare and Borchardt 1984, Cohen 1993, 1994, Johns 1985, Johns and Mayes 1990, Kintsch and van Dijk 1978, Lehrer 1994, Seidlhofer 1990, Sherrard 1986, van Dijk and Kintsch 1977, Winograd 1984); summarising processes continued to be the subject of review and empirical investigation into the 21st century (Caccamise, Franzke, Eckhoff, Kintsch and Kintsch 2007, Kintsch, Steinhart, Stahl, Matthews and Lamb 2000, Ohno 2005, 2007). Results from extensive research led to the development of theoretical sets of formal rules claiming to explain how readers process and summarise text.

Theories and models of the summarisation process naturally had a direct influence upon pedagogical approaches to reading and text summarisation, and also on approaches to the testing of reading and writing. Grabe and Stoller (2002) suggested it was reasonable to practise summary tasks in teaching and learning because they encourage the reader to build a text model of comprehension which 'amounts to an internal summary of main ideas' (2002:26). They advocated 'practising effective summarising strategies' as an instructional practice that could help students move from learning-to-read to reading-to-learn. Grabe (2009) noted how the development of summary-writing skills could contribute to skills in reading comprehension, citing Trabasso and Bouchard (2002) for evidence of this in both L1 and L2. Others highlighted summary writing as an important type of integrated reading and writing, especially within academic contexts which often require the synthesis of information, sometimes from multiple sources (Grabe 2001, 2003, Hudson 2007, Khalifa and Weir 2009). Alderson (2000), Hughes (2003) and Pollitt and Taylor (2006) all discussed the use of summarising tasks as a technique for assessing reading comprehension. See also Barratt's (2003) report on the summary task that was included in the Use of English paper of the Cambridge Certificate of Proficiency (CPE) from 2002. (Pages 251–260 in Weir and Milanovic (Eds) 2003 provide the rationale for its inclusion in the revised CPE, while pages 313–314 show an example of the task.) Interestingly, from 2013 this task will move from the Use of English to the Writing paper.

If summarising tasks are to be validated as appropriate measures of reading comprehension ability, then it is important we understand and can specify the component cognitive processes of summarisation activity. Some of the more influential theories and models of summarisation are therefore described here in some detail.

Following extensive work with subjects who recalled and summarised narrative stories, van Dijk and Kintsch (1977) proposed a model of summarisation in which the information to be included in any summary is determined by *macrorules* operating on the text propositions of the input text. Such macro-operators, they claimed, enable the information in a text base to be reduced to its gist or theoretical macrostructure. Van Dijk and Kintsch distinguished four macro-rules reflecting the processes of *deletion* (MR-1), *generalisation* (MR-2), *selection* (MR-3) and *construction* (MR-4). They claimed that in rules MR-1 and MR-2 the information is irrecoverably lost, whereas in MR-3 and MR-4 the information is partly (inductively) recoverable 'by general knowledge and frame knowledge concerning normal conditions, components and consequences' (1977:69). Kintsch and van Dijk (1978) tested their theoretical model of text comprehension and production through the generation of recall and summary protocols of a text. They also contended that the four macro-rules of deletion, generalisation, selection and construction were not just specific rules for carrying out a summary writing task, but *general* rules underlying the comprehension of any text. This view was later endorsed by other researchers (Johnston 1984, Winograd 1984). For example, Winograd stated that the ability to identify important elements in a text was a strategic skill underlying both comprehension and summarisation, a view which is still widely held (Enright et al 2000, Grabe 2009).

Brown and Day (1983) combined Kintsch and van Dijk's analysis with an informal consideration of summarisation protocols obtained from children and adults. They identified six basic rules of summarisation: 1) *delete trivial material*; 2) *delete redundant material*; 3) *substitute a superordinate term for a list of items*; 4) *substitute a superordinate action for a list of subcomponents of that action*; 5) *select a topic sentence*; and 6) *invent a topic sentence if one does not already exist*. Brown and Day commented on the similarity between their own six rules and the four macro-rules produced by Kintsch and van Dijk as follows: their deletion rules (1) and (2) were similar to van Dijk and Kintsch's deletion macro-rule (MR-1); their superordination rules (3) and (4) corresponded to van Dijk and Kintsch's generalisation macro-rule (MR-2); and the last two rules (5) and (6) related roughly to Kitsch and van Dijk's selection and construction rules (MR-3 and MR-4). For subsequent studies, Brown and Day conflated rules (3) and (4) into a single superordination rule to generate a final set of five rules: *deletion of trivial material*; *deletion of redundant material*; *superordination*; *selection*; and *invention*.

In a series of three studies involving both children and adults, Brown and Day examined the application of these five rules during the summarisation of expository texts. Results from the first of the three studies appeared to provide evidence of the use of all five rules, with both deletion rules used comfortably by younger children (aged 10). The use of superordination, selection

and invention rules, however, increased only as readers become more mature. In the second study, experts who were interviewed during a summarising task (using a think-aloud protocol) reported using specific strategies for summarising the text. These strategies corresponded closely to the five basic rules identified by the researchers. It was also observed that experts accorded special status to the topic sentences, selecting or inventing these first and then writing their summary around or in support of them. An additional, and quite persistent, rule which kept appearing involved rearranging material across paragraphs, and for this reason a further rule was added to the original set of five. Brown and Day concluded that these six basic rules appeared to capture the essence of the methods of text condensation used by students when formally undertaking a summarising task.

In a study investigating the efficiency of instruction in summarisation skills among high school students, Chou Hare and Borchardt (1984) extended the set of six rules initially drawn up by Brown and Day (1983) to include two more. Like Brown and Day, they had observed that mature summarisers adopted a strategy of combining paragraphs, and for this reason they added a seventh rule – *paragraphs-combining*, along with an eighth rule – *rewriting* – to reflect the 'polishing' strategy apparently adopted by summarisers to produce the finished article. For the Chou Hare and Borchardt study, the final set of eight rules was partially collapsed and redesigned into a rule-sheet containing four general self-management steps, four specific summarisation rules (*collapse lists*, *use topic sentences*, *get rid of unnecessary detail* and *collapse paragraphs*) and one polishing rule. This rule-sheet was later used when training American high-school students in summarisation skills and was found to have positive effects on students' efficiency and use of rules in summary writing.

All these studies suggest that it is possible to identify and describe some of the component cognitive processes that underlie the activity of summarising a text. As we shall see, these findings may well have relevance and implications for the procedures needed to generate a suitable summary of a text which can form the basis of a summary completion task for testing reading comprehension ability.

Research into other factors influencing the summarisation process

The studies described so far all focused on identifying and describing the component cognitive processes underlying summarisation activity. In addition, several studies explored other important factors that may directly influence the process and product of summarisation. These included the *qualities of the text* to be summarised, *procedural aspects* of the summarisation task and the *type of summary* required.

In an analysis of the different textual characteristics that may make it more or less difficult to summarise a text, Hidi and Anderson (1986) highlighted *length*, *genre* and *complexity* of the original piece of writing as three key factors.

With regard to *length*, Hidi and Anderson suggested that the shorter the text, the more likely it was that the ideas were closely related and that they could be expressed in a single topic sentence. Longer texts, however, caused summarising to become more difficult to process since these contained more ideas and, as a result, more evaluations and decisions were required on the part of the reader over which ideas to include or eliminate.

Genre (or type of text to be summarised) was also identified as a factor which may affect difficulty. Some text types were perceived as being essentially easier to summarise than others. For example, it was suggested that narratives are generally easier to summarise than expository texts, particularly for children. Several possible reasons for this were offered:

- readers tend to be more familiar with narrative texts than with expositions
- narratives tend to be organised in a linear structure expressing common temporal-causal relationships which are easier to process and recall than the complex relationships and non-linear structure found in many expository texts (see also Meyer and Freedle 1984)
- expositions tend to contain complex and unfamiliar ideas while narratives are more accessible and straightforward
- in narratives what is interesting may overlap with what is important, while in expositions importance and interestingness are less likely to overlap (Hidi and Baird 1986).

The third characteristic of text that Hidi and Anderson suggested may make a text difficult to summarise was *complexity*. Definitions of text complexity vary considerably and it could be argued that both length and genre are already factors that contribute to text complexity. Other commonly recognised aspects of difficult text, however, include features of linguistic complexity (e.g. low-frequency vocabulary, elaborate sentence structure, poor cohesion) as well as features of conceptual complexity (e.g. degree of abstractness, unfamiliarity of concepts and ideas, lack of coherence). Readability formulae have traditionally been used to evaluate elements of a text's complexity. It is unlikely, however, that readability formulae, of which there are now a significant number, are capable of measuring the degree to which length, genre and complexity interact with one another to make a text more or less difficult to read and summarise (see Davison and Green (Eds) 1988 for a comprehensive review of readability issues).

The literature on potential sources of text complexity, and their interaction with one another, is now so extensive that only brief mention will

be made of them here along with relevant references which the reader is encouraged to consult for more detail. A number of descriptive typologies of text characteristics have been designed specifically for use in test development and validation studies. Typologies include a wide range of textual features, sometimes expressed using different terminology. They include features as diverse as: *length, topic, genre, domain, text source, authenticity, discourse type* and *type of information*; *vocabulary, syntax, cohesion, number of negations, number of functions* and *length of various segments* such as word, sentence, paragraphs; *rhetorical structure and organisation, distribution of new information, concreteness/abstractness, referentials, fronted structure, coherence* and *writer–reader relationships*. For more detail, see Alderson, Figueras, Kuijper, Nold, Takala and Tardieu (2006), Bachman, Davidson, Ryan and Choi (1995), Enright et al (2000), Fortus, Coriat and Fund (1998), Freedle and Kostin (1993), Khalifa and Weir (2009) and (Masi 2002). Weir comments as follows on the wealth of theoretical and empirical research now available concerning aspects of text complexity:

> There appears to be a measure of consensus in the subjective judgements of these different authors on the features to be addressed when considering text complexity. Additionally there is empirical evidence from studies such as Freedle and Kostin (1993) and Fortus et al (1998) that a subset of the listed characteristics do indeed impact on the difficulty of reading comprehension tests for learners (2013a:473).

In recent years it has become increasingly possible to measure some, if not all, contextual features of text automatically, using more sophisticated computer-based tools. Weir continues:

> Until relatively recently we lacked the quantitative tools necessary to compare automatically and accurately the various contextual characteristics of the range of written texts we use in our tests at different levels of ability (Biber, Conrad, Reppen, Byrd, Helt, Clark, Cortes, Csomay and Urzua 2004). However, recent advances in automated textual analysis, computational linguistics and the development of corpora have now made it feasible to provide more quantitative approaches focusing analytically on a wide range of individual characteristics (Crossley, Louwerse, McCarthy and McNamara 2007, Crossley and McNamara 2008, Graesser, McNamara, Louwerse and Cai 2004, Graesser, McNamara, and Kulikowich 2011, Green, Ünaldi, and Weir 2010, Green 2012 and Weir, Bax, Chan, Field, Green and Taylor 2012, Wu 2011). New technologies offer examination boards the potential of a more systematic, efficient way of describing a number of the contextual parameters in the texts they select for their language tests (see Green et al 2010) (Weir 2013a:473).

Graesser et al (2011:223) made a strong case for using a system called Coh-Metrix:

> Recent advances in numerous disciplines have made it possible to computationally investigate various measures of text and language comprehension that supersede surface components of language and instead explore deeper, more global attributes of language. They have allowed the analysis of many deep-level factors of textual coherence and processing to be automated, permitting more accurate and detailed analyses of language to take place. A synthesis of the advances in these areas has been achieved in Coh-Metrix, a computational tool that measures cohesion and text difficulty at various levels of language, discourse, and conceptual analysis.

Despite significant advances in our ability to analyse and quantify important text characteristics in terms of their impact on text complexity and the cognitive demands they make during reading, it is nonetheless important to recognise that such quantitative indices cannot provide a full picture of the complexity of a text for readers. As Graesser et al pointed out, 'successful text comprehension involves much more than an analysis of text characteristics alone because prior knowledge, inference mechanisms, and skills of readers are also critically important' (2011:223). Green et al (2010) and Wu (2011) investigated a number of these qualitative features for texts used in study and for reading tests, and their work attests to the importance of these in establishing text complexity.

Apart from characteristics of the text itself, a number of *procedural aspects* (i.e. the conditions under which a summary is written) may affect the processing of a summary task. Urquhart and Weir (1998) referred to these aspects as 'performance conditions' (1998:119) and Khalifa and Weir (2009) explicitly list these under 'Task setting' within the *context validity* component of their socio-cognitive framework for reading test development and validation (2009:82–104). Such aspects include time constraints imposed upon the task as well parameters such as knowledge of the assessment criteria, but perhaps more importantly in the case of summary these need to be extended to include the presence or absence of the source text during the activity of summarisation. We shall return later to the issues of time constraints and awareness of evaluation criteria when discussing the instructions that accompany summary tasks. Some consideration will be given at this point, however, to the issue of whether the source text remains present during summarisation or whether it has been removed from test takers after reading, since this is the condition that will be operationalised in the experimental studies.

Hidi and Anderson (1986) suggested that the cognitive operations performed by the summary writer under the text-*present* and text-*removed*

conditions may be essentially different. If the text remains *present* throughout the writing of a summary, the role of memory becomes less important since the text can be reviewed over and over again and the summary writer is able to devote attentional resource to evaluating and deciding upon the relative importance of text segments, as well as to monitoring the production of the summary against the original material. On the other hand, this may also encourage the copy and delete strategy observed by a number of researchers (Brown and Day 1983, Johns 1985, Johns and Mayes 1990).

If the text is *removed*, the summary writer is forced to depend heavily on the original mental representation stored in memory in order to retrieve the information to create a summary. Hidi and Anderson suggested that, during construction and storage of their mental representation, the reader may have more actively processed the text content (omitting details, reorganising content, using paraphrases, integrating schemata) than would be the case in the text-present condition. The risk, however, is that the writer may be inclined to leave out parts of the original text as a result of forgetting rather than due to deliberate deletion, condensation and integration of ideas.

Hidi and Anderson reported that the few empirical investigations into this issue produced somewhat ambiguous results. They did, nevertheless, offer some support for the hypothesis that the more active type of processing hypothesised for the text-*removed* condition may result in a more stable memory representation over time.

Although Hidi and Anderson made no mention of this in their 1986 paper, a related issue must surely be the potential difference that might occur in cognitive processing between:

- a reader who is told before reading a text that they will be asked to summarise it after reading but without continuing access to it, and
- a reader who is not told about the need to write a summary of the text until after they have finished reading it and it has been removed from them.

We shall return to this issue when considering the nature of the instructions to accompany a reading and summarisation task.

Different types of summary

A particularly important aspect of summary writing tasks, and yet one which appears to have been relatively neglected in much of the research and pedagogical literature on summarisation, concerns the *type* of summary that a reader is required to produce. It seems only reasonable that the production of a satisfactory summary must in some way be related to a sense of the purpose or function of the summary together with the audience for which

it is intended. It is difficult to find much evidence of attention being given to this issue in the research studies into summarisation or the instructional materials on summarisation. The general assumption appears to be that for any text it is possible to determine a model summary and that any reader who has adequately understood the text should have little difficulty in reproducing such a model.

It seems more likely, however, that there exists not just one summary but several different possible summaries for any given text, each one designed to suit a different purpose and targeted at a different audience. Requiring a reader, especially in a testing context, to generate a summary of a text without providing any indication of why and for whom the summary is being produced seems somewhat unfocused and unfair. Unfortunately, this can be precisely what researchers, course book writers, teachers and test constructors ask readers to do.

Some might be tempted to justify this on the grounds that no satisfactory taxonomy of summary types is available. Although this may be true from a purely academic or pedagogical perspective, Ratteray (1985) provided an extremely useful review (based upon direct experience of real-world summary writing in professional, government and corporate environments) of at least seven summary types which he claimed had evolved over 250 years. He classified them first of all as either *sequential* summaries or *synthesising* summaries. *Sequential* summaries retain the original order in which information was presented (e.g. journal abstracts, the précis, secretarial minutes and the abridging digest). *Synthesising* summaries alter the original sequence of the text to achieve specific objectives (e.g. the locational digest, the restructuring digest and reviews). Specific features of the different summary types specified by Ratteray will not be discussed in detail here although a number of general points will be drawn out of his attempts at summary classification. For certain types (e.g. abstracts), the convention is to retain the words of the original writer, while for others the summary writer extensively paraphrases the original (e.g. précis, see also discussions of précis-writing in Pocock 1917 and Robeson 1913), or even includes their own critical analysis and explicitly stated opinions (e.g. reviews).

Summary length may vary according to type – as little as one tenth of the length of the original text (e.g. abstracts) up to between one third and one quarter (e.g. précis). While some summary types tend to follow a more rigid or stylised pattern (e.g. abstracts, minutes), others may be freer, reorganising the original text content thematically or according to some other criterion (e.g. locational or restructuring digests). Summaries may be of a single text or cover a selection of different texts on a similar theme. They may set out to be analytical and make judgements (e.g. reviews) or they may strive to remain faithful to the original author's style, point of view and intent without any added commentary (e.g. abridging digests).

Ratteray's account of how different summary types are produced and used in everyday life makes it clear that how a summary is written will be significantly affected by both the purpose for which it is designed and the audience for which it is intended. His stress on the importance of purpose and audience for a particular summary is echoed in the wider literature on writing instruction and assessment for both L1 and L2 (see, among others, Grabe and Kaplan 1996, Hamp-Lyons and Kroll 1997, Hyland 2002, Weigle 2002, Weir 2005).

Ratteray surmised that each of the summary types in his categorisation had developed historically in response to challenges facing the professions, government, business and ordinary citizens, all of whom were seeking to absorb increasing quantities of information in an increasingly complex society. This sets the overall process and product of summarisation firmly within a real-world, social context, reminding us that summaries are directly related to specific needs on the part of individuals and society. In one sense, this stands in marked contrast to the world of academic research or pedagogical instruction (and, by extension, testing) in which much of the theory relating to summarisation processes has been derived from the study of summaries that are largely independent of any natural context, i.e. they have no specified social function or audience in a real-world sense. Even specification of length (in terms of number of words or lines) tends to be arbitrary, as we shall see below. The problem of artificiality of context is further compounded when we consider that specially constructed (rather than naturally occurring) texts are often used in research studies into summarisation. There is clearly a risk if theories of the summarisation process are derived mainly from a study of what readers do when asked to summarise a specially constructed text within a largely unspecified context.

Attempts to model the summarisation process artificially in this way inevitably tended to assume the existence of a single summarisation model rather than several potentially different models according to such variables as the individual reader, their level of language proficiency, the purpose for summarising or the intended readership. Features such as purpose, audience and length, as well as time constraints and awareness of criteria for evaluation, can all be found within the specification of context validity parameters articulated by Weir (2005) and Shaw and Weir (2007) together with the socio-cognitive parameters of content knowledge, and lexical, structural and functional features of the text. There is undoubtedly a need for more studies into summarisation where any and all of the variables identified above are carefully specified, controlled and investigated (see also Cohen 1994, Shaw and Weir 2007, Weir 2005).

Ratteray's categorisation focused firmly upon the varying functions of individual summary types and it assumed that summaries were largely written for the benefit of other people (not the summary writer) who had

neither the time, inclination nor opportunity to read the original text(s). Other researchers focused upon the need to distinguish between *reader-oriented* summaries (i.e. produced for others to read) and *writer-oriented* summaries (i.e. produced purely for the personal use of the writer), and the way in which this too could affect the creation of a summary (Cohen 1994, Hidi and Anderson 1986). An example of the former type might be the executive summary placed at the front of a longer project report for the benefit of a management board. An example of the latter type might be a student's notes made on a text for future reference, perhaps to feed into an essay. Such a set of notes might in fact be unsystematic or nothing more than a few main points. In their 1978 study, Kintsch and van Dijk highlighted the potential influence of the reader's goals during the process of comprehension and summarisation but they did not attempt to build this factor into their model. Nor did they seek to account in any detail at that stage for the way in which a reader's background knowledge enables inferences to be made. Later studies in summarisation appeared to take little account of the way in which a reader's purpose could affect both reading and summary production.

The significance of purpose and audience

It seems that many of the empirical studies investigating summarisation processes failed to provide either a clear purpose for summarising or any indication of whom the summary was intended for (other than for the researchers themselves). Brown and Day (1983), for example, reported asking subjects to write two good summaries (one unconstrained and one limited to 60 words) of a text. However, the notion of what constituted a good summary seems to have been left to the judgement of the subjects themselves. Winograd (1984) reported using a similar procedure with a 60-word limit. Chou Hare and Borchardt (1984) explicitly acknowledged Kintsch and van Dijk's point about reader purpose but made no reference in their training rule-sheet to the need to consider either the purpose of the summary being constructed or the intended audience. They too reported asking subjects to write a good summary for their empirical studies with an upper limit of 80 words. Johns (1985) reported asking students and adult fluent readers to write summaries of up to 100 words with no time limit imposed and Sherrard (1986) reported neither time nor length constraints.

Several authors commented upon the role of a reader's sensitivity to what is important in a text (Chou Hare and Borchardt 1984, Johns 1985, Johns and Mayes 1990, Winograd 1984). Few, however, made explicit any relationship between sensitivity to importance and a statement of purpose for summarising. Once again, the general assumption was that for any text it is possible to define a single set of main ideas and their essential supporting

details, irrespective of any stated purpose for reading and summarising, or any variation in the background knowledge and interests of the reader. Once this set of main ideas and supporting details has been defined by a reader it can supposedly be developed into a good summary of the original text.

It is possible that this view directly influenced the nature of any guidance given to learners in instructional materials. Advice on how to summarise did not normally include consideration of the purpose for summarising (see, for example, Davies and Whitney 1984). In a discussion of their findings, Chou Hare and Borchardt conceded that 'the nature of the text and its main idea manifestations influence rule usage, and consequently summary product' (1984:76), but they failed to make any reference to the influence of reader purpose on reading and summarising.

Van Dijk (1979) distinguished two kinds of *relevance* in text comprehension. The first was *textual relevance* that defined important information as that which the author appeared to consider important and chose to signal as such through cues in the text structure at various levels. The second was *contextual relevance* which was assigned to properties of a given text on the basis of their cognitive set (i.e. actual knowledge, beliefs, opinions, wishes, attitudes or tasks). Pichert and Anderson (1977) demonstrated empirically how easily a reader's perception of what was important could be manipulated through the framing of the task. For example, when readers were asked to read the same text and then recall it from the point of view of either a prospective home-buyer or a prospective burglar, readers' perceptions of which textual information was important varied in accordance with the perspective taken. Cohen commented that 'there are undoubtedly differences of perception regarding what a 'main idea' consists of and the appropriate way to write it up. There may also be differing views as to the acceptability of introducing commentary into the summary' (1994:175).

There are clearly multiple factors operating and interacting during the activity of text summarising. Some of these concern the nature of the text itself. Others focus more upon the person of the reader and the nature of the task. It is likely that most or all of these factors affect the nature of the summarisation process and product during reading and summary production. Although much has been learned about the component cognitive processes of summarisation, the influence of contextual variables on summary tasks remains less well-researched and understood. Furthermore, although summary tasks have an understandable appeal as a means of assessing reading comprehension ability and have been recognised as a suitable method for doing just this, they present a number of important problems. The main drawbacks in using summary tasks for assessing reading comprehension will be reviewed and discussed in the following sections.

Problems with using summary tasks to assess reading comprehension ability

Test method

Perhaps the most obvious problem associated with using summary tasks to test reading comprehension ability must be the complex nature of the response involved – the *test method*. As we have seen above, producing an accurate and coherent text summary requires not only appropriate reading skills but also appropriate writing skills and strategies. The process involves deleting trivial and redundant material and identifying relevant information, substituting superordinate terms for lists of terms or sequences of events, and finally, reformulating the selected content so that the final product reads smoothly and coherently (Brown and Day 1983, Chou Hare and Borchardt 1984, Davies and Whitney 1984, Kintsch and van Dijk 1978). The test method inevitably confounds the measurement of reading skills with the measurement of writing skills as was demonstrated by Pollitt et al (1985) in an empirical study of factors affecting the difficulty and quality of comprehension test questions. It is what Weir referred to as 'muddied measurement' (1990:61), i.e. the contamination of the measurement of one skill by the involvement of other skills at the same time, or what Messick (1989) later termed construct-irrelevant variance.

Pollitt et al (1985) examined comprehension tests of English (L1) and French (L2) that included both open-ended test questions and a summary task. Four significant sources of difficulty in reading comprehension were identified. While two of the four sources related to features of the *reading text* (i.e. lexical and syntactic complexity and spread of information), the remaining two sources concerned the nature of the *test task*. First, difficulty in understanding the questions on the text (e.g. because of ambiguity or interaction between sub-questions); and secondly, difficulty in composing or transforming written language in order to produce the response.

Cognitive development

The same study also highlighted an interesting aspect of summarising ability. An analysis of performance by the top 20% of the ability range among the students showed that although they performed very well on the open-ended questions testing their comprehension of the text, their scores on the summary exercise proved surprisingly low by comparison. Other research studies have observed similar findings, i.e. that even good students seem to find the summarising process particularly difficult, not only because it involves more than reading alone, but also because it constitutes the highest, most demanding level of processing in reading (far beyond processing at sentence level which is the focus of so many reading test items).

Brown and Day (1983) examined the way in which the component skills of summarisation in L1 develop between ages 10 and 18, and they concluded that summarising skills are constrained more by general cognitive than by linguistic development. Development is incomplete at any stage before adult maturity, whatever efforts teachers may expend. This may explain why the 16-year-olds in the Pollitt et al (1985) study encountered such difficulty in the summarisation task on their examination paper, even though they seemed to understand the text fully. If summarisation skills are unlikely to be complete until adult maturity is reached, then it would seem unreasonable to use summary tasks in assessing the comprehension ability of younger age-groups.

Some studies commented, however, that even mature adults do not find it easy to produce a summary and that considerable variation can occur across adult summaries of the same text. It may be that summarising skills do indeed develop with age up to a certain level of proficiency but that expert summarisers are most likely to be adults who have received direct instruction in how to summarise. In other words, the ability to summarise represents a novice-expert continuum. It may not be appropriate to regard summary writing skills in the same way as reading skills, i.e. as skills which have naturally developed in full by the time adulthood is reached, though it may also be that even some mature adults are not always capable of a text-level representation.

Garnham (1987) proposed that the mental representation generated by a reader is a model of the world, or of some possible world, and not a model of the text itself. Their model is not necessarily, or even principally, linguistic. The original text has been used to construct a mental representation, which may be visual, or verbal, or both, or indeed involve other kinds of representations than words and pictures. If this is so, then the process of writing a summary – in words, sentences, and discourse structure – involves translating into language a representation that is likely to be visual, auditory, kinaesthetic, emotional or symbolic, or any combination of these, rather than verbal. This may be another reason for some of the difficulties encountered by summarisers, even if they are mature adult readers. The fact that the input to the representation is mostly verbal may not be sufficient to facilitate its retranslation into the words of a summary – especially if the instructions demand that the summary is 'in your own words'.

Extent and efficacy of prior training

If summarising skills depend not only upon the level of cognitive development but also upon explicit training, then a further problem concerns the extent to which test takers may or may not have received such training, together with the effects of any such instruction.

Results of research into the efficacy of instruction in summarising skills have been contradictory. Some studies concluded that direct instruction has

a generally positive effect on the development of summarising skills (Chou Hare and Borchardt 1984, Johns 1985). Instruction may not, however, automatically improve readers' sensitivity to what is important in text or their ability to identify and reproduce main ideas. Bensoussan and Kreindler (1990) found that English as a Foreign Language (EFL) students with a semester's training in summary writing recognised summaries as an important tool for grasping the gist of a text. Nevertheless, the authors still expressed frustration at students' continuing inability to distinguish macro- from micro-propositions. This may have been because no clear purpose for summarising was given to them and, in such cases, identification of main ideas is likely to vary across readers. Irrespective of how successful instruction in summarising skills is, it seems likely that test takers who have received little or no instruction in summarising skills will be disadvantaged by comparison with those who have.

Instructions for a summary task

Several researchers (Brown and Day 1983, Chou Hare and Borchardt 1984, Cohen 1993) referred to the importance of the instructions given to students or research subjects for summarising a text. Cohen (1984) also pointed out, however, that students do not always read the instructions in the test context. As discussed earlier, instructions for most test summary tasks tend to be minimal, with constraints only on the summary's length (often calculated as a ratio of summary length to text length and expressed in terms of the acceptable number of lines or words). Little guidance is offered on style, format, purpose, readership, etc. The important question of the intended audience for a summary and whether this should be explicitly stated in summary task instructions has already been discussed above at some length.

Cohen (1993) highlighted the additional problem of the difference between a real-world summary and a test summary, commenting that test summaries are largely prepared for an assessor 'who has already decided what the text is about and wants to see to what extent the respondents approximate those decisions' (1993:132). Given differing perceptions of what can constitute the main idea or the important or relevant information for a summary, there must be potential for a mismatch between the criteria used for summary production by the test taker and those used for summary evaluation by the assessor. Cohen investigated the effects of different levels of guidance to test takers for completing summary tasks. He observed that although guided or elaborated (as opposed to minimal) instructions can have a positive effect on the summarising of foreign language texts, they may have a mixed (even detrimental) effect on the summarising of native-language texts. This led him to suggest that the nature of summary task instructions may be particularly important in L2 testing. It is worth noting, however, that the specific guidelines drawn

up for Cohen's study concentrated upon a step-by-step analysis of how to read the text and write the summary, similar in content to the instructional rulesheet devised by Chou Hare and Borchardt (1984). No explanation was given regarding the function of the summary, and the intended readership was referred to simply as the reader of the summary. Cohen commented on his own study as follows: 'Although the intention was to indicate for each summary for whom the text was being summarised, this information was inadvertently left out of the instructions for both groups' (1993:143). Presumably a fuller context for reading and summarising should have been given and might have made a difference to subjects' motivation if nothing else.

Test reliability

Cohen's (1993) reference to the potential for a mismatch between a test taker's criteria for summary production and those used for summary evaluation by the assessor raises the important issue of reliability. Several researchers warned that statistical results from summarising tasks are not always consistent with results from other types of reading comprehension tests such as multiple choice, short answer and cloze.

Cohen (1994) cited a study by Shohamy (1984) comparing EFL tests containing summarising tasks with tests containing multiple-choice and open-ended response formats, and in which responses could be in either the native or the foreign language depending upon the test version. Shohamy found the results from the summarising data to be so inconsistent with results on the other tests that she eliminated the findings from the published study. Cohen suggested that a plausible explanation for this was potential unreliability in the ratings of the summaries.

In his own study investigating rater strategies and reliability on summary tasks, Cohen observed significant differences both across individual raters' strategies and across their perceptions of what constituted appropriate rating criteria. Common practice when devising rating keys for summary tasks is to leave the development of the answer key to the test constructor and/or the raters. Given the evidence that even expert raters can disagree on which ideas in a text are essential to the construction of a meaningful summary (Cohen 1994, Sarig 1989), this is unlikely to contribute to high reliability in rating. Cohen concluded that more rigorously developed rating keys needed to be devised (Cohen 1993). Interestingly, Bensoussan and Kreindler (1990) built a rating key based *both* on the suggestions of the respondents as to key ideas *and* on the insights of the raters.

The issue of differing perceptions over what constitutes the main point(s) in a text and therefore what important or relevant information should be included in an adequate summary of it, especially in respect of how that summary is to be evaluated, can be partially resolved through exercises such

as consensual text-mapping. Harri-Augstein and Thomas (1984) suggest a consensual rule-of-thumb approach for distinguishing between main and minor ideas in a text and determining the relationship between them. Sarig (1989:77–94) proposed a 'meaning consensus criterion answer' approach to form a framework of the main ideas and important details a reader might be expected to extract from a text. See also Urquhart and Weir's presentation of a text-mapping technique for use with individuals or a group (1998:306–307) as well as the application of a 'text-mapping' or 'text-diagramming' approach in a specific reading test development project reported in Weir, Yang and Jin (2000: 64–66, 173–178).

Cohen (1993) advocated the development of standardised approaches to rating procedures and a comprehensive programme of rater training. He argued that high reliability in the use of summary tasks was unlikely to be achieved unless several important issues were addressed, i.e. student training in summary writing, the nature of test instructions, the development of scoring keys, the provision of rater training and standardisation in rating procedures.

Test processing

Some doubt has been expressed over the type of processing encouraged by summary tasks in the test context, and particularly the extent to which test takers undertake summary tasks in a way that is consistent with current understanding of the summarising process. This takes us back into the area of cognitive validity as it concerns the task used in the testing process (Weir 2005, Khalifa and Weir 2009, Shaw and Weir 2007).

Cohen (1993, 1994) reported various studies highlighting potential causes of discrepancy between the way test takers are supposed to produce summaries and the way they actually do. He suggested that the testing situation itself imposed constraints on the test taker that are not normally found in the real world, causing test takers to read in an unnatural way and to write summaries that are quite different from what they might normally do.

Despite the generally accepted claim that summary tasks necessarily encourage so-called higher order reading skills, it is quite possible that what they actually encourage is greater processing at the localised word/phrase/sentence level than at the global paragraph/text level, especially among L2 learners. Johns and Mayes (1990) found that neither high-level nor low-level proficiency college-level EFL students used macro-propositions in their summaries and that low-level students in particular did a considerable amount of direct copying from source text into their summaries. Underprepared L1 writers are also inclined to use more reproductions (copying and paraphrase) than macro-propositions (Johns 1985). Other studies observed a similar linear and compartmentalised, rather than global, approach among readers

writing a summary. What is perhaps not clear is the extent to which readers may begin by generating high-level macro-propositions in their reading of a text but subsequently inhibit these and substitute lower-level structures in summary production. The complex demands of producing a summary of a text may have the effect of inhibiting a conceptual transformation of the text content at the highest level thereby causing the comprehension process to be masked in some way and consequently not readily available for assessment purposes.

Implications

There appears to be sufficient evidence from the research reported and discussed above to conclude that, despite their natural appeal, the use of traditional summary tasks for assessing reading comprehension ability raises some serious problems.

Although Cohen (1994) expressed some reservations about the use of summary tasks for assessing reading comprehension, he also suggested that the complexity of the processes involved in summarising and the consequent difficulty in assessing them should not necessarily be taken as grounds for rejecting summary tasks out of hand. Issues of test reliability, he claimed, could be satisfactorily resolved through appropriate test design, through the training of students and raters and through rigorous standardisation procedures. Cohen defended using summary tasks on the grounds that they elicit a wide range of reading strategies, affirming that 'well-constructed summary tests may promote a richer, more interactive approach to reading than do many comprehension tests that focus more on details' (1994:203).

Winograd (1984), however, was less enthusiastic about using summary tasks to assess reading comprehension. In his study of the strategic difficulties involved in summarising texts, Winograd concluded that 'the ability to reduce a passage into a summary through the use of transformations identified in the study did not relate significantly to the ability to comprehend that passage' (1984:421). He gave the following reasons:

> . . . the task of summarising not only requires a reader to construct an internal representation of the author's message, as is required for comprehension but also requires that other, secondary decisions be made about the relative importance of the elements in the internal representation (. . .). Moreover, it seems that these secondary operations require the active control of the reader to a much greater extent than do the comprehension processes which resulted in the internal representation initially. Poor readers run into difficulty with both stages of this task. Not only do they have difficulty in constructing an internal representation of the author's message, but they also have difficulty with the secondary operations required to produce a summary (1984:421).

Winograd drew a distinction between the ability to comprehend a text and the ability to summarise it. In his view 'there is more to summarization than adequate comprehension' (1984:423).

If this is the case, then the major objection to using summary tasks for reading comprehension assessment must be that they combine two dimensions. In assessing reading comprehension ability, we are primarily interested in the first dimension, i.e. the initial internal representation of the text constructed by the reader. The second dimension, i.e. the subsequent transformation of the initial internal representation into a written form, is another matter, which may of course be relevant when assessing writing ability, or when testing integrated reading-into-writing ability, but is not of concern if we are primarily interested in testing comprehension ability. If we incorporate this second dimension into our reading assessment procedures, then we risk gaining a distorted measure of the reading comprehension trait and fall prone to the risk of confounded measurement highlighted by Weir (1990).

It seems, then, that we need to find an alternative test format which will address the mental representation constructed by the reader of a text, enabling them to demonstrate the extent of their comprehension, but which will not require them to translate that mental representation into words.

One way would be to identify predominantly non-verbal summary formats, such as diagrams and flowcharts. By successfully labelling elements or stages in such visual representations of the text content, readers could provide adequate evidence of their comprehension. Some texts (such as those describing linear processes or categorisation) lend themselves relatively easily to this type of visual representation, but the majority of texts are unlikely to do so and this approach may therefore have limited application.

An alternative approach would be to assist readers with the process of producing a verbal summary but do so in such a way that it does not help them in the process of understanding itself. This could be achieved by providing readers with a partial summary framework or scaffolding which they are only able to complete if they have read and understood the text from which the summary framework was derived.

In both these formats, the compositional difficulties of summarising secondary operations referred to by Winograd (1984) are largely avoided. The first method circumvents the problem completely by not asking for a written summary at all, while the second provides direct assistance in helping students write a summary.

Summary completion technique

The second of the two approaches described above has been referred to as *summary completion technique*, and it is this format which provides the focus

for the empirical research reported in the remaining chapters of this volume. Summary completion technique involves first of all selecting a suitable reading text and generating a summary of it. The most important words or phrases of the summary are then deleted to form the test items. The technique has been invented independently several times, for example, by Mossenson, Hill and Masters (1987), Pollitt and Hutchinson (1987), and Courchene and Ready (1993).

Summary completion technique and cloze format

The technique has sometimes been referred to as *summary cloze* (Courchene and Ready 1993), but it is important to recognise that the process of completing such a test has little to do with the test methods properly called cloze. In cloze tests it is intended that successful completion of the gaps in the text should be worked out on the basis of the surrounding co-text at both the local and the global level. In reality, research has shown that most gaps in a cloze test are completed on the basis of the local co-text, i.e. the 4–5 words occurring immediately before and after a gap (Alderson 1978, 1979, 1980, 2000, Kintsch and Yarborough 1982).

From a qualitative and quantitative study of L1 and L2 test takers' responses on a rational cloze test format, Pollitt and Taylor (1993) concluded that cloze technique should not be regarded as a test of reading ability but rather as an analytic language task which makes direct demands upon a language user's analysed knowledge of the language. The study raised significant doubts about the long-established use of cloze test formats to assess reading comprehension.

By contrast, summary completion technique requires a reader to read a text and then complete gaps in a summary of the same text. The knowledge required to fill the gaps correctly can be found only by reading and understanding the original text. Thus the essential difference between summary completion technique and cloze tests is that the answers needed to complete the gaps lie in understanding the original text and the summary of it and in being able to map the one to the other. Correct answers should not be deducible from the gapped passage alone by someone who has not read and understood the original. This can of course be checked and confirmed by trialling the gapped summary in isolation, without giving informants access to the original source text.

The cloze-like format of summary completion technique has the advantage of making the test countable, but the presence of the whole original story makes the processing nature of the task wholly different from that of cloze.

Advantages of summary completion technique

Summary completion technique offers several important advantages over traditional summary tasks. First of all, it makes few demands upon the productive language skills of the reader, especially if a range of options is provided from which the test taker can select the correct answer to fill the gap. This variant has been referred to as a 'banked choice, gapped summary task' and Alderson (2000:242) provides a short example of this selected response format variant taken from an *IELTS* Reading test. Secondly, ability to complete a summary of a text is unlikely to be constrained by cognitive developmental factors in the same way that ability to produce a full summary of a given text appears to be. Summary completion technique does not depend upon readers having been taught how to summarise effectively and the need for complex instructions relating to the type, audience and length of a text summary is avoided. Potential interference from the test method is thus substantially reduced. Finally, the item-based nature of summary completion technique makes it countable, bringing with it the advantage that rating of the test becomes much easier than in traditional summarising tasks involving extensive written production on the part of test takers.

Summary completion technique and the readers' mental representation

The gapped summary given to readers in a summary completion task is designed to scaffold for them the mental representation which they could reasonably be expected to construct as a result of reading a text. The technique makes few demands upon productive language skills so readers are not expected to translate their own mental representation into words.

More importantly, the task avoids asking specific (open-ended or multiple-choice) comprehension questions on the text which might stimulate or interfere with the reader's mental representation. In traditional reading comprehension tests, the text normally remains with the reader throughout. This means that, while answering comprehension questions, the reader draws upon a mental representation which can be continually reinforced and modified through multiple readings of all or parts of the text. It is relatively easy for readers to elaborate their mental representation if prompted to do so by a particular question. Their initial mental representation may have been quite simple, but extra details, whether explicit in the text or inferable from it, may come to be added by rereading the text in response to a question on it.

In listening comprehension tests the text is normally heard once or twice and what remains after listening is just the listener's mental representation constructed on the basis of what was heard, understood and stored in memory. The text itself is no longer available and listening test questions can only be

answered by referring directly to the listener's mental representation. If the aim of reading assessment is to examine and assess a reader's understanding of text, then one way of achieving this would be to make the testing of reading resemble more closely the testing of listening. Summary completion technique offers an opportunity to do this if the reading text is removed immediately after reading and the reader is asked to complete a summary of the text on the basis of their constructed meaning representation, i.e. their comprehension, alone. Removing the written text from readers before asking them questions on it makes the approach more like a listening experience. Potential objections to this approach could be that readers will not know which aspects of the text to focus upon when developing their mental representation of it and that it is unfair to expect them to remember all of it. (Interestingly, similar objections are rarely raised in relation to listening texts, except on the grounds of excessive length.) The nature of the mental representation constructed by a reader of a text could, in part at least, be determined by giving readers a clear reading context and purpose, just as the provision of a clear context and purpose for listening can be used to predispose the listener to attend to certain aspects of the text rather than others and so constrain the type of comprehension which is required. The importance of reading context and purpose was underlined by Garnham and Oakhill: 'mental models are representations of the world constructed for specific purposes, and the model constructed, whether it be from perception, reasoning or language processing, should be the one that is most appropriate for the task in hand' (1992:202).

Readers typically read with some purpose in mind, ranging from the general (e.g. browsing through a magazine for ideas or to pass the time) to the more specific (e.g. following a set of instructions to assemble a piece of furniture). Purpose for reading combines with other factors (e.g. length of time allowed for reading, availability of background knowledge, level of language proficiency) to help shape the mental representation that a reader constructs in order to arrive at an adequate understanding. Purpose for reading is likely to be influential in determining what readers pay attention to or consider relevant. In the test context, setting a purpose or context for reading may be important if the text is to be removed following reading so that the summary is completed on the basis of the reader's mental representation alone.

In other words, readers ideally need to know why they are reading a given text and what sort of understanding they are expected to construct from it. This will enable them to have some idea of what they should pay attention to while reading and enable them to build an appropriate mental representation which they can then draw on to complete the gapped summary. In setting up a purpose for reading, the suggestion is of course not that readers should see the summary before processing the source text. That would be likely to cause a significant test method effect not dissimilar to the artificial processing provoked by multiple-choice questions. The experience of the English

Language Monitoring (ELM) Project (Pollitt, Hutchinson, Napuk, Munro and Dickie 1990) showed that it is possible to control reader and listener purposes by embedding the assessment format within a larger activity-based context (Pollitt 1993). In most test contexts, however, an elaborate contextualisation such as that used in the ELM Project is impracticable, and it is the explicit instructions or rubric for a test task which must convey as fairly and as effectively as possible how test takers should listen or read, and what will and will not be relevant. Pollitt described this as the *teleological* dimension of testing. By providing a teleological purpose for a comprehension task, test designers and constructors also provide themselves with an important criterion for checking the validity of each test item or question. For example, while ideas and information in a reading text which are relevant to the stated reading purpose become valid testing points, irrelevant ideas and information remain invalid. The experimental studies reported below will demonstrate how purpose for reading can be established in a variety of ways.

Summary completion task development

In theory, producing a summary completion task (in either a text-present or a text-removed format) should be relatively straightforward. All that is needed, it would seem, is the selection of a suitable reading text, the construction of a summary of the text and the deletion of key content words and/or phrases to create a set of test items.

In an informal assessment context, constructing a summary completion task may be a relatively simple exercise. In a formal assessment context, however, where issues of test reliability and validity assume far greater importance, a more rigorous and principled approach to test construction is required. The summary of a text must be one which a large sample of different readers can easily match to the original text in terms of a typical mental representation. The words or phrases deleted from the summary to form the test items will need careful analysis to ensure they are appropriate in terms of content and level of difficulty. If a summary completion task is to be used as a means of assessing reading comprehension ability, then it must be possible to demonstrate that it offers a test which is both reliable and valid.

Although some empirical work has been undertaken to validate text-present summary completion tasks, relatively little research has been done on text-removed summary completion tasks. Furthermore, little practical guidance is available to test designers and constructors on how to generate an appropriate summary of a given text and how to identify suitable test items. In light of this, the experimental studies described and discussed in the remaining chapters of this volume set out to explore some of the theoretical and practical issues relating to the design and empirical validation of summary completion tasks for assessing reading comprehension ability.

5 Designing recall studies to explore readers' mental representations of two texts

Introduction

This chapter describes the design, procedures and analytical approaches used in a pair of *oral recall* studies to investigate readers' mental representations of two written texts – a short story (Text A – *Journey*) and a newspaper editorial (Text B – *Anorexia*). The purpose of the studies was twofold. Firstly, to explore the mental representations constructed by readers for two texts of differing genres. Secondly, to identify what might constitute a satisfactory verbal summary of a shared mental representation for each of these texts, with a view to using this as the basis for developing a summary completion task for assessing reading comprehension ability. Outcomes from the oral recall studies will be reported and discussed in Chapter 6, together with an explanation of how the multi-level analyses of the oral recall transcripts were used to construct a satisfactory summary of each text. (Chapter 7 will report further research undertaken to convert the summary into a summary completion task format for assessing reading comprehension ability.)

Methodology

Since both reading process and product are invisible to an observer, a methodology was needed that would permit the elicitation of a reader's mental representation of a given text. Text recall methodology involves comparing subjects' oral or written recalls of a reading text with the researcher's formal analysis of the same textual discourse. This technique has been extensively used over the years as a means of investigating processes in text comprehension.

One of the earliest researchers to use text recall methodology was Bartlett (1932) in his study of memory. Bartlett used text recall to investigate subjects' representation in memory of an American Indian story. His findings from this and other experiments led him to conclude that memory and recall were more than simply a matter of storing and reproducing memories that were

fixed and lifeless. Instead, memory and recall were essentially *constructive*, involving rationalisations of various sorts in conjunction with established knowledge schemata.

Twenty years later, Gomulicki (1956) investigated the text recall process in greater detail, focusing particular attention on *how much* of a text was recalled in relation to its original length, as well as *which parts* of the text were retained and *why* this might be. Gomulicki concluded that the probability of recall of any element in a text was directly related to the contribution of that element to the total meaning of the text. Furthermore, he surmised that selective retention involves a ranking of importance of the various elements in general abstraction from the detailed input. As we shall see, both of these findings are directly relevant to the text recall studies reported in this and subsequent chapters. They are also consistent with the cognitive processing that was highlighted in Chapter 4 as being necessary for successful completion of a summary task.

While both Bartlett and Gomulicki employed recall methodology to investigate memory, later researchers assumed that the approach can be used to study the comprehension of text as well. Van Dijk and Kintsch (1977) used both recall and summary to investigate text comprehension. Like Bartlett and Gomulicki, they asserted that reading recall is essentially a constructive process that draws upon linguistic and semantic information stored in memory:

> When a subject is asked to recall a story he has read, he generates a linguistic output from his memory traces. Whatever remnants of the actual linguistic processing during reading that are still available in memory are used for that purpose, but for the most part the text must be reconstructed from the micro- and macro-structure propositions that represent the reader's memory for the meaning of the text (1977:74).

Following their study of written recalls and summaries produced by readers of a 1,600-word narrative, van Dijk and Kintsch (1977:74) drew the following distinction between the related concepts of recall and summary:

> Recall may be characterized as summary-plus-detail; that is, the statements that subjects make in their summaries tend also to be included in their recall protocols. In addition, recall protocols contain more information about some details of the story, which usually does not appear in summaries.

Riley and Lee (1996) made a similar observation in their comparison of recall and summary protocols as measures of reading comprehension; they found that 'the summaries contained a higher percentage of main ideas than details whereas the recall protocols contained a higher percentage of details

than main ideas' (1996:173). Other researchers recognised the value of recall studies in providing information about comprehension processes and outcomes. For example, Connor (1984) considered recall to be the operational definition of comprehension, while Bernhardt (1991b) suggested that the recall protocol can provide data on the nature of the reading process in terms of how information is analysed, encoded and restructured.

Although comprehension, memory and recall are undoubtedly closely linked to one another, it would be naive to assume that memory and comprehension are one and the same thing, or that text recall is capable of providing full and total access to either memory or comprehension. We might imagine that some processing and comprehension must first of all take place if information is to be stored initially in memory and thus become available for retrieval at a later stage. Although memory enables subsequent retrieval of that information, during recall for example, it may be that some things which are initially comprehended are located only in working memory and are subsequently unavailable for recall. Furthermore, not everything that is stored in long-term memory will necessarily be retrieved during recall. Recall is therefore likely to be able to provide only partial insight into comprehension. For this reason, certain drawbacks have been highlighted in using recall methods and other forms of verbal report (often referred to as verbal protocol analysis, or VPA) as a means of investigating people's understanding (e.g. Brown and Rodgers 2002, Clapham 1996, Ericsson and Simon 1993, McDonough and McDonough 1997).

Nevertheless, as Field (2012) noted, verbal report methods including recall have been widely used in research into expertise generally (Ericsson and Simon 1993) and into cognitive validity specifically (Baxter and Glaser 1998). Recall and verbal report methods have also been extensively researched and exploited since the 1970s as a means of investigating both *first language* (Crain-Thoresen, Lippman and McClendon-Magnuson 1997, Crothers 1972, Frederiksen 1972, 1975, Kintsch 1974, Kintsch and van Dijk 1978, Meyer 1975b, van Dijk and Kintsch 1977) and *second language* processing and comprehension (Bernhardt 1991b, Badger and Yan 2012, Carrell 1983, 1984a, 1984b, Cohen 1998, 2006, 2013, Connor 1984, Connor and McCagg 1983, Faerch and Kasper 1987, Field 2012, Gass and Mackey 2000, Lee 1986, Lund 1991, Mackey and Gass 2005, Riley and Lee 1996, Steffensen 1988, Weir et al 2000, Weir, Hawkey, Green and Devi 2009, Woodfield 2012). Thus despite some limitations (and in the absence of other effective methods) it is reasonable to assume that text recall has the potential to provide us with at least some insights into the reader's mental representation of text and that these will be sufficient to reflect some, if not all, aspects of their comprehension of a given text. For this reason, reading recall methodology involving interviews and stimulated recall was selected as the most effective method currently available of making verbally explicit as many aspects of a reader's mental representation of text as possible.

Design

The hypotheses underlying the design of the study were as follows:

1. Some content features of a text are more likely to be recalled than others.
2. Some content features of a text will consistently be recalled by a group of readers when reading the same text for the same purpose.
3. Content features held in common by a group of readers can be used to construct an adequate verbal summary of the text.

Although van Dijk and Kintsch (1977) suggested that recall can be characterised as 'summary-plus-detail', experience has shown that asking a subject to recall what they can of a text they have just read often results in the reader offering only a short, skeletal account of the text content – sometimes no more than a single sentence – without much of the detail which may characterise their mental representation and which is of interest to researchers. For this reason, it was decided to include in the research design both a *free recall phase* (unprompted) and a *prompted recall phase* (a form of stimulated recall) in which the researcher followed up the participant's free recall with a series of standard probe questions in order to elicit further propositional content. Probe questions were to be asked only after the free recall phase had been completed. It was anticipated this might encourage a fuller recall of general elements of the text already mentioned, as well as recall of additional elements thus far unmentioned.

It was decided to use naturally occurring texts rather than artificially constructed texts which have been the focus of many recall studies. When designing reading comprehension tests, test writers typically select naturally occurring texts, rather than construct their own artificial texts, especially at higher levels of proficiency where issues of contextual validity assume greater importance (see Chapter 4 of Khalifa and Weir 2009 for a full discussion). Grabe (1988) suggested that an important part of reading is the ability to recognise text genres and the distinct text types which are deliberately exploited by writers, while Kobayashi (2002) demonstrated empirically that text type or structure can significantly impact the reading comprehension test performance of test takers at different proficiency levels. With this in mind, two of the most frequently occurring text types (*narrative* and *expository* – see Weigle 2002) were selected: Text A (*Journey*) was a narrative text (a short story), and Text B (*Anorexia*) was an expository text (a newspaper editorial). It was hoped that this might enable a limited comparative analysis across two differing genres. Each text was selected as being reasonably typical of its genre and also representative of the sorts of text that are often selected for reading comprehension tests. (Texts A and B are included as Appendices 1 and 2.) The two texts and their accompanying probe questions for recall were

trialled in a small pilot study. From piloting it became apparent that some of the probe questions used with each text were too directive or suggestive so these were amended to make them less leading in nature.

Reading researchers consistently highlight context and purpose as significant factors that can shape a reader's approach to a text (see, for example, Alderson 2000, Grabe 2009, Khalifa and Weir 2009, Pressley and Afflerbach 1995, Urquhart and Weir 1998). A number of studies demonstrated empirically how context and purpose can influence the interaction of text and reader (Anderson, Pichert and Shirey 1983, Frederiksen 1972, Pichert and Anderson 1977, Schmalhofer and Glavanov 1986). In light of this, it was felt that a clear and plausible reason for reading should be an essential feature of the experimental design. The reading recall task was therefore constructed to incorporate a specific purpose for reading the two texts (see more detail on this below, together with Appendix 3), as well as a partial indication of what would happen after reading. This made it possible to embed the recall phases within a larger and still plausible task. To avoid the recall activity becoming the purpose that might shape their reading, participants were not told in advance of reading that they would be asked to remember or recall the texts. Some recall studies reported giving subjects this information prior to reading (Connor 1984, Gomulicki 1956, Varnhagen 1991). Other recall studies did not always make clear whether participants were told this in advance or not. In either case, it is arguable that such prior knowledge on the part of the reader could significantly influence the nature of their reading process and the resulting output.

In the initial task design, there was also concern that participants might resort heavily to verbatim recall, i.e. the reproduction of exact words immediately after reading or listening to a text (see Field 2004:318–320 for a fuller explanation of this phenomenon). Previous text recall studies have sometimes required participants to complete an additional task immediately after reading but before recall. This additional task acts as an interval 'buffer' to minimise the recall of words positioned late in a text (Freebody and Anderson 1986, Steffensen et al 1979, Varnhagen 1991). For the present study, therefore, it was decided that participants would be given a small buffer task immediately after reading each text and before being invited to engage in free recall. It was hoped that their text recall would thus be drawing upon the mental representation constructed in long-term memory, rather than on the surface linguistic representation maintained in working memory. The two buffer tasks (one introduced after each text reading) were designed to be relevant in a general way to the overall task, and thus directly linked to the purpose for reading given at the outset, but to cause as little additional processing of the textual content as possible. In this way it was believed that they would not seriously influence the reader's mental representation of the text. Even if at the recall stage readers produced some words and phrases appearing in the

written text, it was anticipated that these were unlikely to be linguistic traces from working memory, but would result instead from the informant's active search for linguistic expression, including the selection of words and phrases already activated through the reading experience.

Participants

Participants in the study were 40 students studying for Advanced Level qualifications (in preparation for university entry) at a local sixth form college for 16- to 18-year-olds. All participants were individually invited to take part in a small reading research study and they voluntarily agreed with the option to withdraw at any time. The average age of the group was 17 years 9 months and the sample included 21 girls and 19 boys. The students involved in the study represented a wide range of academic disciplines and can be regarded as reasonably mature readers, given their age, background and level of education.

Materials

Two short texts were selected for the study. Text A was a short story narrative of 526 words entitled *Journey by Night*. Text B was an editorial passage of 389 words entitled *The rights and wrongs of treating anorexia*, taken from a broadsheet newspaper – *The Independent*. (See Appendices 1 and 2.) Text A was a short story with an unexpected twist at the end of it, while Text B focused on a health issue of direct relevance to teenagers. Both texts were selected for their interest value and for their accessibility to the intended readership in terms of topic and its treatment, as well as for their proven suitability and use as text types in reading tests.

Procedures

Two different conditions were employed for the reading recall exercise. Half the participants were interviewed individually, while the other half of the group were interviewed in pairs. The same procedures were adopted in each condition. The reason for using some paired interviews was to gauge whether a paired recall approach would prove as effective as singleton interviews (see Haastrup 1987 for a similar paired approach to explore learners' lexical inferencing procedures). It was felt that a paired approach might be more appropriate if the recall design was later extended to interviews with younger participants, since a one-on-one, adult–pupil interview can be particularly daunting for younger teenagers and children. In the event, no obvious differences between the paired and individual interviews were observed, so for the purposes of analysis all paired interviews were later treated as single recalls.

Participants were first told that they were taking part in a small-scale study into aspects of reading text selection for testing at the Key Stage 3 (KS3) level within the English secondary education system (i.e. the national testing of 13/14-year-olds). It was explained that two different reading texts had been

identified for possible use with KS3 pupils. Before finally deciding whether to use these two texts or not, perceptions relating to the texts' interest value and perceived difficulty were being sought from students reasonably close in age to KS3 pupils. Participants were told that they would be asked to read Texts A and B and to consider how accessible they considered each text to be for KS3 pupils, in terms of its ideas and language. This approach enabled a clear and plausible context and purpose for the participants' reading activity to be established at the outset.

A protocol was used to guide the reading recall interview in the following way (see Appendix 3). Participants were invited to read silently and at their own speed the first of the two texts. The first reading text was then removed from the reader(s) who were asked to complete a brief questionnaire gathering personal details such as name, gender, age in years/months, Advanced Level exam subjects, general interests and future plans. This brief questionnaire constituted the first of the buffer tasks positioned between reading and recall. After completing the questionnaire, participants were invited to freely recall as much as they could of the first text. Following this free recall phase, participants were asked a selection of probe questions in a prompted recall phase to elicit any additional details (see Appendix 4).

At the end of the prompted recall phase participants were invited to give their views on the suitability of the first text's subject matter for the KS3 age group as well as the accessibility of the language in it. After sharing their views on these points, participants were invited to read the second text at their own pace. Following reading, the second text was removed and participants were given a list of suggested topic areas relating to the KS3 curriculum. They were invited to tick one or more of the topic areas into which they felt the second text might reasonably fit. This constituted the second buffer task. The remainder of the interview followed exactly the procedures used for the first text, i.e. free recall phase, prompted recall phase and views on text suitability and accessibility for a KS3 cohort. The order of presentation of the two texts was counterbalanced across the group to eliminate any possible practice effect.

All interviews were recorded onto cassette and a full orthographic transcription was made of the free recall and prompted recall phases for the purpose of detailed propositional analysis (as described below). In addition, brief field notes were made of each participant's reactions to the texts for future reference. (A sample transcript for one of the interviews can be found in Appendix 5.)

Approaches to analysing the reading texts and the recall protocols

In the 1970s several formal models of text analysis were developed to describe text content at both sentence and discourse level (see some discussion of this in Chapter 2). Much early text analysis concentrated on narrative prose

(Mandler and Johnson 1977, Rumelhart 1975, Thorndyke 1977), but attention later turned to the development of systems for analysing the content and structure of expository prose (Kintsch and van Dijk 1978, Meyer 1975a). Some researchers asserted that these theoretical models can be used to analyse readers' recall protocols of a reading passage in order to reveal the processes in text comprehension and to compare these with the researcher's abstract (propositional) representation of the same text.

When devising appropriate research tools for the analysis of text and text recalls, researchers have frequently based their approach on the Kintsch and van Dijk (1978) theory of micro- and macro-structures and/or on the hierarchical content-structure theory proposed by Meyer (1975a). Research tools devised for analysing text and recall protocols have generally made use of an analytical unit derived directly from the ideas or *propositions* contained in the reading text and in the corresponding recall or summary. Individual or simple propositions are first identified through a clausal analysis of the text. These are then combined to generate complex propositions which, it is claimed, represent hierarchical levels of the text structure.

Some discourse analysts commented that a proposition-based representation of discourse content may be less useful than it first appears (Brown and Yule 1983). The notion of *proposition* (which stems originally from formal logic) was adopted by linguists and can now be found throughout the literature on text analysis. Brown and Yule (1983) pointed out, however, that writers in this field were inclined to interpret the term in different ways: some used it to refer to sentences or statements; others used it to represent conceptual structures; and yet others used it to mean the representations in which all knowledge is stored. This confusion surrounding the definition of a *proposition* was discussed further in Lyons (1977). Brown and Yule also warned that a proposition-based analysis of a given text cannot of itself constitute a representation of a reader's understanding of that text since computing the intended meaning of a speaker or writer depends upon knowledge of many details 'over and above those to be found in the textual record of the speaker's/writer's linguistic production' (1983:116). While they cautioned that a propositionally based analysis of a reader's recall protocol cannot necessarily constitute a full representation of that reader's understanding of the text, Brown and Yule nonetheless acknowledged that such an analysis can provide us with some meaningful insights into the reader's mental representation of the text held in long-term memory.

With this in mind, the research tools devised in the present study for analysing and coding both the reading texts and the recall protocols draw only partially on the approaches to text analysis originally developed by Meyer (1975a) and by Kintsch and van Dijk (1978). Consideration was also given to the contribution of other theoretical and practical work undertaken in the fields of text and discourse analysis, and this will be discussed later in this chapter.

Analysis of text-based propositions occurring in the recall protocols

It was decided in principle that all recalls of Text A and Text B should first be analysed using a set of propositional units or *text-based propositions* derived directly from the original reading texts. In practice this meant that in some cases a short sentence in the text provided a single idea or text-based proposition. In other cases a compound or complex sentence containing several ideas needed to be broken down into a number of separate text-based propositions. This approach also fits neatly with the psycholinguistic definition of a proposition as 'an abstract representation of a single unit of meaning' (Field 2004:225).

A set of 89 text-based propositions was drawn up for Text A (*Journey*) and a set of 61 text-based propositions for Text B (*Anorexia*). The two sets of propositions were checked with an independent judge (experienced discourse analyst – Professor Gillian Brown) and amended following discussion. Text-based propositions maintained the same wording and ordering as the original reading texts, except in cases where it seemed advisable to insert a full noun phrase as the subject in order to avoid any ambiguity of reference. (Note that the concept of text-based proposition developed for this analysis is different from Kintsch and van Dijk's (1978) notion of text proposition.) The two finalised sets of text-based propositions (TP) are shown in Tables 5.1 and 5.2 below.

The relevant set of text-based propositions was used to analyse the 30 recalls that had been generated for each of the two texts. The fact that half the participants were interviewed individually and the other half were interviewed in pairs means there was a total of 30 (rather than 40) recalls.

Whenever a proposition from the set of text-based propositions was present in a recall transcript it was coded 2, 1 or 0. A proposition occurring in the free recall phase was coded 2. A proposition which did not occur until the prompted recall phase was coded 1. If a proposition failed to occur within either the free recall or the prompted recall phase, then it was coded 0. Using the approach described above, 10% of the recalls for each of the two texts were analysed by an independent rater. The degree of consensus between first and second rater was calculated to be 84% which was considered to be an acceptable level of inter-coder consistency (Miles and Huberman 1994).

After manual analysis of the recalls and manual coding of the text-based proposition units for each text, the resulting datasets were analysed using a Rasch analysis program. This provided frequencies of answer codes together with an estimate of the relative *ease* with which the propositions for each text were remembered by the readers in both the free and prompted recalls. Results of these analyses are reported in Chapter 6.

Although a propositionally based analysis at the clause and sentence level proved effective in coding much of the content of participants' recalls

Table 5.1 Set of text-based propositions for Text A (*Journey*)

TP01	the man stood alone
TP02	the man was leaning against a post
TP03	the man was shifting his weight from one foot to the other
TP04	the hour was late
TP05	the taxi-stand was empty
TP06	the street was silent
TP07	the man looked up and down
TP08	the man was hoping that
TP09	some vehicle would come in sight
TP10	the man wanted to get home
TP11	no vehicle came
TP12	the silence began to pall
TP13	the man started to whistle
TP14	there was no mirth in his whistling
TP15	the man soon stopped whistling
TP16	it was midnight
TP17	the man was ten miles away from home
TP18	what was the man to do?
TP19	it was out of the question to begin to walk that distance
TP20	a dark cloud passed across the sky
TP21	the dark cloud hid the few pale stars
TP22	the few pale stars that had been there
TP23	the noise of a falling dustbin reached the man's ear
TP24	some dog must have been scattering the dustbin's contents
TP25	instinctively the man's hand felt for his wallet
TP26	the wallet was still there
TP27	if only the man had a stick
TP28	the man had nothing with which
TP29	the man might protect himself
TP30	the man began to walk up and down, up and down
TP31	what was that in the distance?
TP32	at last two headlights were drawing near
TP33	the man stepped into the middle of the street
TP34	the man held up his hand
TP35	the car stopped
TP36	"Are you a taxi?"
TP37	the man asked
TP38	"Will you take me to Valencia?"
TP39	"Get in"
TP40	said the driver
TP41	the driver was opening the door
TP42	the man sat beside the driver
TP43	the man was glad to be on his way home at last
TP44	the man had felt so lonely while
TP45	the man had been waiting
TP46	if only someone would say something
TP47	in the semi-darkness of the car the man turned to look at the other passengers
TP48	no one else was there
TP49	the driver said nothing to the man
TP50	the car sped along
TP51	the man mustn't allow himself to think of that
TP52	the man glanced at the driver
TP53	again the man's hand went to his wallet
TP54	the man had heard of passengers being attacked at night

Table 5.1 (continued)

TP55	the man had heard of passengers being robbed
TP56	that couldn't happen to the man
TP57	if only the man could see the other man's face clearly
TP58	the man had no idea who
TP59	the driver was
TP60	the man kept his eye intently on the driver during the seemingly interminable journey
TP61	now the two men were approaching a spot where
TP62	the road branched off in another direction
TP63	there were tall, dark bushes around
TP64	the car slowed down
TP65	the driver was looking at the man
TP66	the driver took something short and black from the side-pocket of the car
TP67	the thing looked like an iron tool
TP68	would the driver attack him with that?
TP69	"Stop (the car)!"
TP70	the man heard himself screaming
TP71	the man's heart beat so fast with fear that
TP72	the man could hardly breathe
TP73	the car did not stop
TP74	instead the car went faster and faster
TP75	now the two men were nearing the man's destination
TP76	did the driver intend to take the man past?
TP77	"Put me down here"
TP78	the man cried out
TP79	the man still had his eyes on the driver
TP80	the man quickly stepped from the car
TP81	the car came to a standstill
TP82	the man fumbled in his wallet for the fare
TP83	the taxi was no longer there
TP84	"There'll be no more night passengers for me again"
TP85	exclaimed the driver
TP86	with a sigh of relief the driver hurriedly moved off
TP87	the driver's hand tenderly caressed the heavy spanner with which
TP88	the driver had meant to defend himself
TP89	had that queer passenger attacked him

for both texts, this approach nevertheless revealed certain limitations. Inevitably, participants' recalls were composed of more than just propositions which could be matched directly to original propositions contained in the text. In some cases, for example, participants would subsume several text-based propositions under a more general superordinate or *summarising* proposition that was not explicitly expressed in the reading text. At other times, participants appeared to offer *inferences* which might reasonably be provoked by the text. There were also occasions when participants offered propositions which were clearly linked to the text but which were inaccurate in some sense. Interestingly, evidence of the latter two categories of propositions was also reported by Lehrer (1994) in a study of lecture summaries.

Table 5.2 Set of text-based propositions for Text B (*Anorexia*)

TP01	the case of Samantha Kendall has highlighted a confusion
TP02	there is a confusion in public thinking
TP03	Samantha Kendall is an anorexia nervosa sufferer
TP04	Samantha Kendall discharged herself from hospital
TP05	doctors feared for her life
TP06	anorexia nervosa is a disturbing disease
TP07	anorexia nervosa is a perplexing disease
TP08	ten years ago anorexia was still dismissed as nothing more than slimming gone too far
TP09	today anorexia is recognised as a medical condition
TP10	a medical condition which can be treated
TP11	the degree to which has become a topic of debate
TP12	whether treatment should be carried out without a patient's consent
TP13	researchers have suggested two psychiatric explanations behind the onset of anorexia
TP14	one explanation is that
TP15	the patient is faced with an unacceptably stressful adult life
TP16	the patient is faced with an unacceptably difficult adult life
TP17	the patient is trying to retreat into childhood
TP18	the patient is trying to avoid leaving childhood
TP19	another explanation is that
TP20	choosing what to eat is often an attempt to exert control by people
TP21	specifically choosing what not to eat is often an attempt to exert control by people
TP22	people feel their lives are too constrained in other ways
TP23	the truth is that
TP24	the syndrome remains imperfectly understood
TP25	a great many resources have been devoted to the study of anorexia
TP26	it is beyond doubt however that
TP27	anorexia is a severe psychiatric disorder
TP28	there is no other way to describe an illness that
TP29	the illness allows a patient to look in the mirror at their own emaciated body
TP30	the illness allows a patient to look in the mirror at their own starved body
TP31	the illness allows a patient to see someone
TP32	someone obese staring back
TP33	severe sufferers often deny that
TP34	severe sufferers are trying to kill themselves
TP35	the diet severe sufferers are pursuing
TP36	the diet is all too likely to make death inevitable
TP37	the 1983 Mental Health Act provides for sufferers from severe psychiatric disorders
TP38	sufferers can be held in hospital for treatment against their will
TP39	there is a danger that
TP40	sufferers will do harm to themselves
TP41	sufferers will do harm to others
TP42	one in ten anorexia sufferers dies
TP43	doctors are sometimes reluctant
TP44	doctors use their powers under the law
TP45	this is often because of a fear that
TP46	treatment by compulsion is self-defeating
TP47	force-fed victims of anorexia often return to starvation diets when
TP48	they get home
TP49	there is clearly work to be done
TP50	done in making the treatment of extreme anorexia more humane
TP51	the treatment often involves leaving patients in isolation

Table 5.2 (continued)

TP52	the treatment often involves leaving patients without their clothes
TP53	the treatment often involves watching patients
TP54	patients eat
TP55	patients go to the lavatory
TP56	there are shortcomings of the available treatment
TP57	should not obscure the fact that
TP58	there is an alternative to treatment
TP59	which can sometimes be death
TP60	if doctors made more use of the powers available to them
TP61	lives could be saved

Analysis of summarising propositions occurring in the recall protocols

Given the presence in the recalls of propositions that could not be categorised as text-based, a second level of analysis and coding of all recall transcripts was conducted using a set of higher-level summarising propositions for each text. It was hoped that this approach would accommodate the various additional features of the text recalls observed and reported above.

While the generation of text-based propositions representing semantic information contained in clauses (i.e. at the lowest level in the discourse) is relatively straightforward, the identification of suitable higher-level, macropropositions presents more of a challenge. As Connor (1984:244) pointed out:

> The number of possible argument slots in a simple sentence is finite and may be listed. It is more difficult, however, to list a set of all the possible relations for a text, to show how clauses (simple sentences) fill text-level case or rhetorical roles.

Meyer (1975a, 1975b), van Dijk (1980) and van Dijk and Kintsch (1983) all discussed the notion of 'levels' in relation to propositions in text, suggesting that propositions may differ according to their degree of generality. Where several propositions occur in a sequence, these can be integrated into a higher-level macrostructure which provides a global meaning for the more specific or localised sequences within a text. Such macrostructures can themselves be grouped into more abstract or general propositions. In theory, therefore, it becomes possible to build up a number of levels representing the propositional content of the text. Multi-level approaches to text recall or summary analysis and coding, often based upon the theoretical frameworks of Meyer and van Dijk and Kintsch, have been widely used and reported (Connor 1984, Lehrer 1994, Meyer 1987, Varnhagen 1991).

Meyer (1975b) proposed that a hierarchical diagram of text structure can be created by identifying the rhetorical predicates within the text. Rhetorical

predicates, which may or may not be lexicalised (e.g. *but*), function as labels for relationships between content words in the text. Although such an approach has sometimes been used to generate a theoretical hierarchy of the superordinate ideas contained in a text, it restricts itself to using the wording of the original text. For this reason, it seemed unable to account directly for the type of summarising propositions that were observed in the participants' recalls of Texts A and B.

An alternative approach to describing higher-level structures of text was proposed by van Dijk and Kintsch (1983) who suggested that semantic mapping rules (macro-rules) function to reduce and organise the more detailed information of a text's propositional content. The macro-rules of *deletion*, *generalisation* and *construction* operate on sequences of individual propositions, and even macro-propositions, to generate several levels of macro-structure which in turn represent the more global meaning of the text. In the present study, this approach was felt to be unsuitable as a means of analysing the texts and their recalls since it assumes that for any text there exists a single theoretical semantic representation which can be objectively described. It seems likely, however, that attempts by different individuals to produce a single summary sentence relating to any set of text-based propositions will result in a variety of different, though probably related, interpretations of what should be included in the summarising sentence. In effect, then, it is likely to be a researcher's (or a reader's) subjectivity, rather than anything inherent in the text itself, which primarily determines how deletion, generalisation and construction operate in relation to text content. (One is tempted to speculate on how far this might constitute a problem for contemporary tests of reading that demand a one-sentence summary of a text, such as the Pearson Test of English (Academic).)

Interestingly, this effect was confirmed in a small-scale pilot study undertaken in the very early stages of the research project reported here. Four adult readers were asked to read the same text and then write a sentence summarising the content of each of seven short paragraphs making up the text. Considerable variation was observed across the four readers' summary sentences for each of the seven paragraphs. The variation was evident in several ways, including the extent of propositional encoding, the role of thematisation and the use of stylistic features. In a follow-up study, the four summary sentences produced by the four readers for each paragraph were collated to create a set of four multiple-choice options for each of the seven paragraphs making up the text. The text and accompanying summary sentence options for each paragraph were then given to a sample of 31 mature readers with the instruction to read each paragraph of the text in turn and to decide which of the four summary sentence options offered best summarised that paragraph in their view. Among the 31 participants substantial consensus on the 'best' summary sentence was achieved for only two out of the seven paragraphs

(68% and 61% respectively). For the remaining five paragraphs, responses were distributed fairly evenly across three or even four of the summary sentence options, suggesting considerable variation in participants' perceptions of what constituted 'a good summary'. In practice, the summary sentences were conceptually different in terms of the macro-propositions they encoded; it was more than just a question of the same proposition being expressed through alternative lexical choices or stylistic preferences.

These observations, although from a relatively small exploratory study, are consistent with the findings of other researchers. For example, Zuck and Zuck (1984) asked professional readers from different backgrounds to extract a main idea from a text and found that readers failed to achieve agreement on what the main point of the text was. Instead, each reader constructed a text meaning compatible with their own field of expertise. Sarig (1989) asked seven 'model' readers to read the same English text and to extract or create propositions relating to the main ideas contained in the text in accordance with principles proposed by Kintsch and van Dijk (1978) and van Dijk and Kintsch (1983). Sarig reported that, out of a total of 12 propositions delineated as main ideas by the seven readers, seven of these 12 achieved majority level (over 50%) agreement among the seven readers, and only four out of these seven achieved high consensus (over 85%). Given the level of consensus achieved for seven and four of the 12 propositions respectively, these might be useful thresholds to bear in mind when determining those propositions that should be included in a summary of a text, and this approach will be explored further in Chapter 6. Chou Hare and Borchardt (1984) reported that, even after intensive training in the use of summarisation rules, native speaker high school students failed to improve in their ability to identify or invent topic sentences relating to main ideas. Chou Hare and Borchardt explained this as the result of insensitivity to importance in text, though this view may fail to acknowledge the subtle difference between the comprehension and the interpretation of a text (see Urquhart and Weir 1998).

All these studies seem to suggest that different mature skilled readers will not necessarily produce similar results when asked to summarise or create a macrostructure (to use van Dijk and Kintsch's term) of the content of a short paragraph or text. Furthermore, different readers will not necessarily share a common view on what constitutes the best summary sentence for the content of a given paragraph. They are likely to be influenced in their construction or selection of a summary sentence by a variety of different factors, including features of propositional content, style, emphasis, thematisation and referential interpretation. A superordinate macro-proposition (i.e. at the highest level of the text) may be equivalent to what is sometimes called the 'gist' of a text. In the practical testing context, gist questions such as '*give a title to this text*' are rarely presented in an open-ended format because experience shows that too many potentially acceptable answers are offered which

makes it difficult to achieve reliable scoring. Once again, one wonders how this dilemma is resolved for the summary sentence task in the current Pearson Test of English (Academic).

In light of the challenges outlined above, it was decided not to attempt a supposedly objective higher-level text analysis along the lines suggested by either Meyer or van Dijk and Kintsch, but instead to rely on a more intuitive yet principled approach to identifying summarising propositions for analysing the recalls for Texts A and B. Summarising propositions for the two texts were constructed, therefore, partly in accordance with the orthographic paragraph structuring for each text, and partly in accordance with other identifiable points of topic shift.

The terms *paragraph* and *topic*, and the relationship between the two, have been the subject of extensive debate in the literature. Some authors have suggested that the orthographic paragraph may not necessarily indicate a change in the writer's topic, but may be dictated instead by eye appeal or by printing conventions (Hinds 1977, Longacre 1979). Hinds commented that paragraphs in journalistic texts in particular are often determined on the basis of appearance. Both writers propose the existence of formal linguistic markers (e.g. pro-nominalisation, sentence-initial adverbial expressions) for signalling the beginning and end of paragraphs. However, as Brown and Yule (1983) observed, these are often genre-specific and for this reason it may be difficult to identify formal paragraph boundary markers which can be generalised across a range of naturally occurring written or printed discourse. Using a text in which the orthographic boundaries as they appeared on the printed page had been removed, the authors illustrated the difficulty of identifying suitable formal markers that signal with any degree of certainty where the original orthographic paragraph divisions occurred. Brown and Yule concluded, therefore, that indentation in text functions not simply as a cosmetic device but instead as a primary indicator of topic shift, helping the writer in their task of structuring blocks of information and staging the development of the discourse:

> Rather than treat the indenting of the first line of a paragraph as simply some cosmetic device, as Longacre (1979) does, we might look upon it as an indication by a writer of what he intends us to treat as the beginning of a new part of his text (1983:99).

Brown and Yule also suggested that a reader of a given text may be capable of suggesting other divisions between sets of sentences within the text which are perfectly reasonable, even though these may be in addition to the orthographic paragraph boundaries indicated by the text's original writer.

In light of the above discussion, therefore, summarising propositions for Texts A and B were generated according to two guiding principles:

1. A summarising proposition was generated from a sentence or set of sentences that constituted an orthographic paragraph.
2. A summarising proposition was generated from a sentence or set of sentences that formed only part of a larger orthographic paragraph, but which could nevertheless be identified as marking a topic shift in relation to the preceding text.

It is interesting that for Text B the boundaries for summarising propositions which resulted from an initially intuitive sub-division of orthographic paragraphs into smaller segments coincided, in all cases but one, with generally accepted formal markers of possible topic shift in text. These included an initial adverbial expression indicating temporal sequence (identified by Longacre 1979) – *Ten years ago* – and several connectives (described as macrostructure connectives by van Dijk 1977) – *But*, *Yet*.

In the case of Text A, boundaries for summarising propositions were identified intuitively to begin with. However, on closer analysis it proved possible to justify the suggested boundaries on the basis of thematisation within the discourse. Brown and Yule suggested that an analysis of thematic structure can be helpful in identifying topic area and the organisation of text structure, and that thematisation in discourse is an essential method by which a complete text can be subdivided into smaller chunks.

> Thematic organisation appears to be exploited by speakers/writers to provide a structural framework for their discourse which relates back to their main intention and provides a perspective on what follows (1983:143).

While it is recognised that the development of this second research tool (i.e. sets of summarising propositions for Texts A and B) is somewhat subjective in nature, it is nevertheless an approach that is both principled and systematic. More importantly, test developers require procedures that are accessible and practical. It was anticipated that the qualitative information provided through an analysis using summarising propositions might usefully supplement the quantitative information provided by the initial text-based proposition analysis. Furthermore, once each summarising proposition had been identified and defined, it was also possible to match it directly to a particular set of text propositions from the first analysis.

The two finalised sets of summarising propositions (SP) can be seen in Tables 5.3 and 5.4. The same coding scheme was used as in the first analysis.

An analysis using first of all text-based propositions and secondly summarising propositions successfully accounted for most of the content of readers' recall protocols for both Texts A and B. It did not appear capable, however, of explaining some additional propositions produced by readers in their recall protocols but which could not easily be matched to the sets of text-based and

Table 5.3 Set of summarising propositions for Text A (*Journey*)

SP01	a man was present
SP02	the place was dark and lonely
SP03	the man wanted to get home
SP04	the man tried to keep his spirits up
SP05	the man was anxious about how to get home
SP06	the man was aware of the darkness intensifying
SP07	the man was aware of the noises of the night
SP08	the man was concerned for his personal safety
SP09	a car came along and stopped
SP10	the man and the driver briefly exchanged words
SP11	the man initially felt relieved
SP12	the man started to feel uncomfortable
SP13	the man started to think about being attacked
SP14	the man grew increasingly suspicious of the driver
SP15	the car drove into a country area
SP16	the driver reached for something
SP17	the man panicked
SP18	the car failed to stop
SP19	the man got out of the car
SP20	the man tried unsuccessfully to pay
SP21	the driver drove away fast
SP22	the driver was afraid of being attacked

Table 5.4 Set of summarising propositions for Text B (*Anorexia*)

SP01	a recent news case illustrates confused thinking about anorexia
SP02	there have been changes in recent years in the way anorexia is regarded
SP03	treating anorexia patients without their consent is a matter for debate
SP04	two psychiatric explanations have been offered for anorexia
SP05	extensive research has not provided us with a full understanding
SP06	anorexia is certainly a severe psychiatric disorder
SP07	severe anorexia is likely to result in death
SP08	the law makes it possible to force treatment on anorexia sufferers
SP09	doctors are reluctant to use legal force because it is ineffective in the long run
SP10	there is a need for more humane approaches to treating extreme anorexia
SP11	treatment by force is justified if the alternative is death

summarising propositions. It appeared that while some of these additional propositions were shared by several readers, others were highly idiosyncratic. Interestingly, other recall studies reported having found it necessary to modify or extend their initial analytical systems so as to accommodate particular features of the text(s) under scrutiny (Connor and McCagg 1983, Varnhagen 1991). Furthermore, Lehrer's (1994) study of students' comprehension of classroom lectures commented that different analysis systems may be better for recall of some texts than for recall of others. For this reason, a suitable approach for analysing additional propositions in the recall protocols for Texts A and B was sought.

Analysis of additional propositions occurring in the recall protocols

In their discussion of a possible theoretical model of text comprehension and production, Kintsch and van Dijk (1978) suggested that recall and summarisation protocols are likely to contain

> . . . information not based on what the subject remembers from the original text, but consisting of reconstructively added details, explanations and various features that are the result of output constraints characterizing production in general (1978:374).

In one of a series of studies investigating the comprehension and retention of linguistic information by readers, Stein and Bransford (1979) used the term *elaboration* to describe additional propositions generated by the reader of a text during recall which were not explicitly stated in the original text. They demonstrated that by activating relevant and existing knowledge some readers are capable of generating their own elaborations in relation to a text. Bransford, Stein and Shelton (1984) suggested that such elaborations not only make the interpretation of text more meaningful for the reader but also make the information easier to remember and learn.

Steffensen et al (1979) used the term elaboration in a similar way and introduced the term *distortion* to describe (a) alterations of the explicit text, and (b) the addition to the text of incorrect information. In reviewing the work of Bransford et al (1984) and Steffensen and Joag-Dev (1984), Alderson and Urquhart (1984) concluded that all readers (irrespective of whether they are reading in their native or foreign language) experience the need to make sense of the texts they read, and that elaborations and distortions probably form an integral part of the normal comprehension process. If this is the case, it would seem reasonable to assume that any elaborations and distortions produced in the recall protocols of readers of a text accurately reflect aspects of their mental representation. If they reflect features shared in common by several readers, then they may be relevant to any attempt to construct a summary of the text based upon readers' mental representations.

Distortions are likely to reflect aspects of readers' mental representations in direct conflict with the content of the original text, i.e. offering evidence of misunderstanding. For this reason they may provide valuable insight into points within the text where readers' comprehension breaks down and thus offer some guidance on key elements of a correct understanding of the text which could become a focus for testing, e.g. misunderstanding over whether the passenger paid the taxi driver or not.

Elaborations, on the other hand, present some difficulty. It is not immediately clear to what extent readers' elaborations of a text are highly individual and personalised, or, alternatively, may be held in common by several different

readers when reading for a common purpose. While some elaborations may be generated and shared by a majority of readers, others may be more idiosyncratic. Elaborations shared by a sufficiently large number of readers could perhaps be considered an important part of an adequate mental representation of the text, and it might thus be reasonable to include such elaborations in any summary of that text. Identifying such elaborations is unlikely to be a straightforward process, however, and will require some principled approach in order to distinguish elaborations which can in some sense be considered as 'encouraged' or 'authorised' by the writer of the text from those which are purely a matter of individual interpretation. The literature on types of inferencing is likely to help us in this regard.

Clearly, the additional propositions that occur in recall or summarisation protocols present a challenge to the analyst since it is not immediately obvious how best to categorise and code them. Previous recall studies have adopted a variety of approaches, e.g. by focusing upon the accuracy, completeness and/or cultural appropriateness of any additional material. For example, in their study of cross-cultural factors affecting reading comprehension, Steffensen et al (1979:17) coded recalls not only for text propositional content but also for: (i) *culturally appropriate elaborations* (i.e. information intruded from appropriate schema); (ii) *distortions* (i.e. information attributable to a lack of knowledge); and (iii) *overt errors* (i.e. not obviously related to cultural background). Studies investigating the process and product of summarisation (as opposed to recall) have not generally coded summary protocols for elaborations but have instead focused on scoring protocols for the use of summarisation rules (Brown and Day 1983, Chou Hare and Borchardt 1984, Sherrard 1986, Winograd 1984). Since summaries are considerably more constrained by the original text than reading recalls, elaborations and distortions generated by the reader are perhaps less likely to occur and are therefore considered less worthy of attention. Even so, in an analysis of summary protocols produced by native speaker university students, Johns (1985) found it necessary to score not only correct replications at the propositional and macro-propositional levels, but also distortions of idea units, combinations, macro-propositions and meta-comments. Johns and Mayes (1990) adopted a similar approach in a later study with non-native speaker students.

It is clear from the different studies referred to above that interpretations of what actually constitutes an elaboration and a distortion can vary. Alderson and Urquhart (1984) commented that what Steffensen and Joag-Dev (1984) sometimes described as a distortion could equally be described as another form of elaboration when viewed from another perspective. In conclusion therefore, and given the potential ambiguity surrounding use of the terms elaboration and distortion, there seems little point in seeking to develop a detailed specification or taxonomy of types of elaboration and

distortion generated by readers of text. The issue of practicality also militates against this approach. For the purposes of analysing the recall protocols for Texts A and B, therefore, the elaboration and distortion distinction adopted by previous recall studies will not be adopted.

In her analysis of students' summaries of lectures, Lehrer (1994) also found it necessary to code the protocols for propositions which did not correspond to any in the original lectures. She described the additional propositions in terms of *inferencing*, i.e. they were either correct inferences, or unreasonable inferences, or propositions which were too vague to be coded as either correct or unreasonable inferences. This appeal to inferencing as a way of describing the additional propositions generated by readers in their recall protocols may have been avoided by researchers in other studies because of the extensive and long-standing debate over the nature and role of inferences in text comprehension. Nevertheless, the field of psycholinguistics readily accepts inferencing as an essential process utilised by a reader to impose coherence on a text. Field offers the following definition of inferencing:

> The process of adding information which is not linguistically present in a text. This is often because a speaker/writer has recognised that certain details and logical connections do not need to be specifically expressed because the recipient will co-operate in supplying them (2004:129).

Current understanding of the nature of inferencing offers a much more integrated and flexible account than previously of how inferencing functions in text comprehension and meaning construction (see Chapter 2).

Instead of taking a largely taxonomic or all-or-none approach to inference description, some researchers proposed a continuum of inference activation (Gerrig 1993) that is affected by variation in reader abilities, reader goals, text characteristics, samples of inferences or experimental contexts (Graesser and Kreuz 1994). Several researchers acknowledged that an inference may in fact be activated to a greater or lesser degree, rather than on an all-or-none basis (Gernsbacher 1990, Sharkey and Sharkey 1992).

In light of this, the additional propositions occurring in readers' recall protocols for Texts A and B will be described and explained in terms of inferencing, using insights gained from relevant theories of inferencing in text comprehension (Brown and Yule 1983, Oakhill and Garnham 1988, Singer 1990, 1994). The discussion will be contextualised within Brown and Yule's (1983) pragmatic definition of inferences as the 'connections people make when attempting to reach an interpretation of what they read or hear' (1983:265). Inferencing is therefore perceived as a process which is 'context-dependent, text-specific and located in the individual reader (hearer)' (1983:266).

Chapter 5 has described in detail the experimental design, data collection and analytical procedures developed for studying readers' mental representation of two texts. Chapter 6 will report the multi-level analysis of readers' recall protocols for both Texts A and B.

6 Using readers' mental representations to construct summaries of two texts

Introduction

This chapter reports and discusses results from a multi-level analysis of reading recalls for Text A (*Journey*) and Text B (*Anorexia*). The oral recalls were gathered, transcribed and analysed using the data collection and analytical procedures previously described in Chapter 5 (see Appendix 5 on pages 225–226 for a sample transcript of an oral recall for Text A). Chapter 6 describes how a detailed and systematic analysis of the text-based, summarising and additional propositions contained in the recall transcripts opened a window into how readers appeared to construct their mental representations of the short story and the newspaper editorial, thus providing valuable insights into the nature of readers' comprehension of these texts. It demonstrates how readers integrated the text-based content with their own background knowledge and experience to build meaning, making necessary inferences to support local and global coherence in their comprehension as well as other inferences that enriched their interpretation in various ways. The latter section of Chapter 6 explains how outcomes from this multi-level textual analysis of oral recalls were employed to construct a satisfactory summary of each text which could be used for reading test development purposes. Chapter 7 then reports how these two summaries were developed into item-based summary completion tasks for assessing reading comprehension ability.

Analysis of text-based propositions in recall transcripts of Text A (*Journey*)

Identifying sets of text-based propositions recalled by readers

A preliminary analysis of Text A recall transcripts determined the frequency with which individual *text-based propositions* from the narrative were recalled by readers. (See pages 92–93 in Chapter 5 for the definition of a text-based proposition, together with the set of 89 text-based propositions that were generated from Text A.) Further analysis set out to define sets of individual

text-based propositions according to different selection criteria. By applying different criteria it was hoped a principled approach might emerge that could be used in the future for devising a summary of any similar text. General principles for defining possible sets of text-based propositions centred upon:

a) the number of readers (i.e. proportion of the total group) who recalled each individual text-based proposition, and
b) whether the text-based proposition was recalled only in the free recall phase, or across both recall phases (i.e. free and prompted).

It was anticipated that exploratory analyses along these lines might help to reveal the text content that was held in memory by most readers as part of their mental representation of the short story, including how far some of this content might be considered *core* (as revealed during free recall) or more *peripheral* (revealed only after prompting). Six tentative criteria for defining the sets of text-based propositions were devised as follows:

i) propositions included by 20 out of 30 readers in FREE recall
ii) propositions included by 20 out of 30 readers in ANY recall
iii) propositions included by 15 out of 30 readers in FREE recall
iv) propositions included by 15 out of 30 readers in ANY recall
v) the top third most frequently recalled propositions in FREE recall (i.e. high-frequency – top 33%)
vi) the top third most frequently recalled propositions in ANY recall (i.e. high-frequency – top 33%).

Criteria (i) and (ii) established a relatively demanding threshold of 66% consensus among readers, while criteria (iii) and (iv) set a less demanding threshold of 50% agreement, taking account of previous findings in the research literature on levels of consensus in recalling main and subsidiary ideas (e.g. Sarig 1989). Taking a slightly different approach, analysis of the top third most frequently recalled text-based propositions was included (criteria (v) and (vi)) to see whether this might shed further light on the nature of core, peripheral and superfluous material in readers' mental representation of text.

Comparing content across the sets of text-based (TP) propositions

In accordance with the criteria outlined above, six sets of text-based (TP) propositions were identified across the recall transcript and these are shown in Tables 6.1 to 6.6. The codes and wording for the text-based propositions come from Table 5.1 on pages 92–93. The content of each set is discussed in some detail and differences between the sets are highlighted.

Set T(i), comprising just five (6%) of the 89 text-based propositions recalled by at least two-thirds of the readers, offers a highly skeletal account

of the short story narrative. It incorporates no more than the bare minimum of information: *a man is waiting late at night*; *a car stops*; *the man fears attack by the driver*. This is reminiscent of the brief, one-sentence synopsis often used to describe the plot of a book or film, although, interestingly, it fails to include the crucial twist at the end of the tale.

Table 6.1 Text-based propositions included by 20 out of 30 readers in FREE recall – Set T(i)

TP04	the hour was late
TP35	the car stopped
TP45	the man had been waiting
TP68	would the driver attack him with that?
TP71	the man's heart beat so fast with fear that

Set T(ii) includes nine (almost 10%) of the 89 text-based propositions and it adds a small amount of extra detail to the skeletal storyline provided by Set T(i). In particular, it adds the critical feature of *the driver reaching for some sort of weapon* which contributes to *the man's panic*. It also adds more from the end of the narrative and from the driver's perspective which we might argue is critical to the short story.

Table 6.2 Text-based propositions included by 20 out of 30 readers in ANY recall – Set T(ii)*

TP04	
TP35	
TP45	
TP66	the driver took something short and black from the side-pocket of the car
TP68	
TP71	
TP81	the car came to a standstill
TP85	exclaimed the driver
TP86	with a sigh of relief the driver hurriedly moved off

** text-based proposition codes from T(i) are included here but without their wording in order to highlight the information that is being added as a result of probing*

Sets T(i) and T(ii) (as shown in Tables 6.1 and 6.2) comprise text-based propositions recalled by 20 out of the 30 readers. Clearly, a two-thirds threshold sets quite a high bar and it resulted in a fairly minimal account of the story, probably not enough to form the basis of an adequate summary. Setting a lower threshold of 15 out of 30 of readers (i.e. 50%) is likely to produce a larger set of text-based propositions recalled by readers and to result in a fuller account of the narrative.

In line with this, Set T(iii) in Table 6.3 shows how the number of text-based propositions has increased from the original five in Set T(i) to a total of 13

Table 6.3 Text-based propositions included by 15 out of 30 readers in FREE recall – Set T(iii)

TP04	the hour was late
TP10	the man wanted to get home
TP35	the car stopped
TP42	the man sat beside the driver
TP45	the man had been waiting
TP66	the driver took something short and black from the side-pocket of the car
TP68	would the driver attack him with that?
TP71	the man's heart beat so fast with fear that
TP80	the man quickly stepped from the car
TP81	the car came to a standstill
TP85	exclaimed the driver
TP86	with a sigh of relief the driver hurriedly moved off
TP88	the driver had meant to defend himself

(14%). This set contains the basic thread of the story in terms of its action sequence but it also includes more detail relating to the intentions and actions of the man (his *wanting to get home*, his *getting into and out of the car*) as well as those of the driver (his *reaching for a weapon for the purposes of self-defence*). Set T(iv) in Table 6.4 adds 12 more propositions that emerged as a result of probing, almost doubling the total to 25 (28%) of the 89 possible text-based propositions.

This expanded set offers a much more coherent summary of the events in the short story than any of the previous sets. In this account, it is noticeable that propositions frequently occur in clusters of two or three, linked to specific elements in the story: 01/04/10, 23/25, 35/38/42, 54/57, 66/67/68, 77/78, 80/81/82, 84/85/86, 88/89.

Table 6.4 Text-based propositions included by 15 out of 30 readers in ANY recall – Set T(iv)*

TP01	the man stood alone
TP04	
TP10	
TP23	the noise of a falling dustbin reached the man's ear
TP25	instinctively the man's hand felt for his wallet
TP35	
TP38	"Will you take me to Valencia?"
TP42	
TP45	
TP54	the man had heard of passengers being attacked at night
TP57	if only the man could see the other man's face clearly
TP66	
TP67	the thing looked like an iron tool
TP68	
TP71	
TP77	"Put me down here"

Table 6.4 (continued)

TP78	the man cried out
TP80	
TP81	
TP82	the man fumbled in his wallet for the fare
TP84	"There'll be no more night passengers for me again"
TP85	
TP86	
TP88	
TP89	had that queer passenger attacked him

** text-based proposition codes from T(iii) are included here but without their wording in order to highlight the information that is being added as a result of probing*

Sets T(i) to T(iv) (as illustrated in Tables 6.1 to 6.4) were selected by analysing the consensual recall of text-based propositions by 66% and 50% of the readers respectively.

Taking a slightly different approach, the text-based propositions present in the recall transcripts were categorised and analysed by *frequency of occurrence* (i.e. high/mid/low-frequency) to see whether this might shed further light on the nature of core, peripheral and superfluous content in mental representations of the story. The top third most frequently recalled text-based propositions (i.e. the top 33%) were analysed, according to whether these occurred during free recall or as a result of probing (see Tables 6.5 and 6.6).

Sets T(v) and T(vi) show the high-frequency (top 33%) text-based propositions that occurred in free and prompted recall phrases and these add considerably more details of the story relating to:

the setting – *midnight, ten miles away from home*
the problem – *it was out of the question to walk that distance*
the car's approach – *headlights drawing near*
the man's thoughts/actions – *stepping into the middle of the road, touching the wallet again, shouting stop, stories of passengers being attacked*
the driver's thoughts/actions – *caressing the spanner, the taxi's disappearance, no more night passengers.*

When comparing and contrasting the sets of text-based propositions shown in Tables 6.1 to 6.6, we can see that the setting of different criterial parameters, i.e. (i) to (vi), generates very different summary versions of the short story in the form of text-based propositions as far as the narrative's content and structure are concerned. In terms of reader consensus, it seems 66% agreement may be too high an expectation, while 50% agreement at least begins to generate a reasonable summary of the text, though still accounting for only 14% of the original content. The additional content that emerged from the prompted phase fleshed out this version quite significantly, resulting

in a summary version which accounted for 28% of the original content – much closer to generally accepted views of what constitutes a summary (see the earlier discussion in Chapter 4). The implications of the effect of prompting on recall will be considered further below.

Table 6.5 High-frequency text-based propositions occurring in FREE recall – Set T(v)

TP01	the man stood alone
TP04	the hour was late
TP10	the man wanted to get home
TP16	it was midnight
TP17	the man was ten miles away from home
TP19	it was out of the question to begin to walk that distance
TP23	the noise of a falling dustbin reached the man's ear
TP25	instinctively the man's hand felt for his wallet
TP32	at last two headlights were drawing near
TP33	the man stepped into the middle of the street
TP35	the car stopped
TP38	"Will you take me to Valencia?"
TP42	the man sat beside the driver
TP45	the man had been waiting
TP53	again the man's hand went to his wallet
TP57	if only the man could see the other man's face clearly
TP66	the driver took something short and black from the side-pocket of the car
TP67	the thing looked like an iron tool
TP68	would the driver attack him with that
TP69	"Stop (the car)!"
TP71	the man's heart beat so fast with fear that
TP77	"Put me down here"
TP78	the man cried out
TP80	the man quickly stepped from the car
TP81	the car came to a standstill
TP82	the man fumbled in his wallet for the fare
TP86	with a sigh of relief the driver hurriedly moved off
TP87	the driver's hand tenderly caressed the heavy spanner with which
TP88	the driver had meant to defend himself
TP89	had that queer passenger attacked him

Table 6.6 High-frequency text-based propositions occurring in ANY recall – Set T(vi)*

TP01
TP04
TP10
TP23
TP25
TP32
TP33
TP35
TP38
TP42

Table 6.6 (continued)

TP45	
TP53	
TP54	the man had heard of passengers being attacked at night
TP57	
TP66	
TP67	
TP68	
TP69	
TP71	
TP77	
TP78	
TP80	
TP81	
TP82	
TP83	the taxi was no longer there
TP84	"There'll be no more night passengers for me again"
TP85	exclaimed the driver
TP86	
TP88	
TP89	

* *text-based proposition codes from T(v) are included here but without their wording in order to highlight the information that is being added as a result of probing*

Analysing low-frequency text-based propositions

Sets T(i)–(iv) suggest that certain text-based propositions figured very strongly in readers' mental representations as evidenced by the recall transcripts. Other propositions seemed to figure very little for whatever reason. For example, although 50% or more of readers were able to recall *the noise of the falling dustbin* at the start of the story, only a few readers (16%) recalled the reference to the *dark cloud hiding the pale stars*. Low-frequency text-based propositions (i.e. the bottom 33%) occurring in either free or prompted mode are listed in Table 6.7.

This set of low-frequency text-based propositions includes some details from the start of the story – *the post, shifting from foot to foot, whistling, the cloud, the stars* – together with a considerable number of references to the state of the driver and his passenger during the journey. The interesting question arises of why these particular propositions should have been recalled so much less frequently than others. Were they less memorable and if so, why? It may be that certain details of the story simply did not appear important (e.g. TP02, TP03). Alternatively, this information may have been inhibited by other factors, such as the adjacency of something else which did seem important. A number of these low-frequency propositions have a repetitive quality to them (i.e. they are linked to recurring features in the story – TP52, TP60 and TP79). So it may be that readers were less likely to recall and repeat some

Table 6.7 Low-frequency text-based propositions in ANY recall (bottom 33%)

TP02	the man was leaning against a post
TP03	the man was shifting his weight from one foot to the other
TP07	the man looked up and down
TP08	he was hoping that
TP09	some vehicle would come in sight
TP12	the silence began to pall
TP14	there was no mirth in his whistling
TP15	the man soon stopped whistling
TP18	what was the man to do?
TP20	a dark cloud passed across the sky
TP21	the dark cloud hid the few pale stars
TP30	the man began to walk up and down, up and down
TP31	what was that in the distance?
TP41	the driver was opening the door
TP43	the man was glad to be on his way home at last
TP44	the man had felt so lonely while
TP46	if only someone would say something
TP49	the driver said nothing to the man
TP51	the man mustn't allow himself to think of that
TP52	the man glanced at the driver
TP55	the man had heard of passengers being robbed
TP56	that couldn't happen to the man
TP58	the man had no idea who
TP59	the driver was
TP60	the man kept his eye intently on the driver during the seemingly interminable journey
TP61	now the two men were approaching a spot where
TP64	the car slowed down
TP65	the driver was looking at the man
TP72	the man could hardly breathe
TP79	the man still had his eyes on the driver

feature which they had already mentioned. Some of the low-frequency text-based propositions might reflect content features of the story that readers find more difficult to recall, either because they do not figure in the reader's mental representation at all, or because they are not sufficiently activated during the recall process. For this particular narrative text, such propositions include reported thoughts and feelings (rather than actions or urges or strong emotions) and these may be details of the narrative which could reasonably be omitted when attempting to construct a summary of the text.

Effect of prompting on recall of text-based propositions

The percentage of readers recalling each individual text-based proposition was plotted to try and evaluate the possible effect on prompting on recall. Figure 6.1 shows ease of remembering for the 89 text propositions before prompting (i.e. during the free recall phase), while Figure 6.2 shows ease of remembering after probing (i.e. during the prompted recall phase).

Figure 6.1 Ease of remembering before probing (Text A)

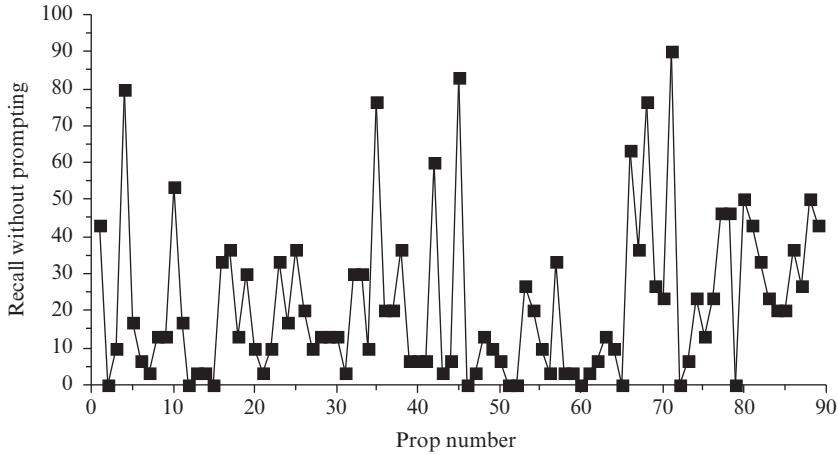


Figure 6.2 Ease of remembering after probing (Text A)

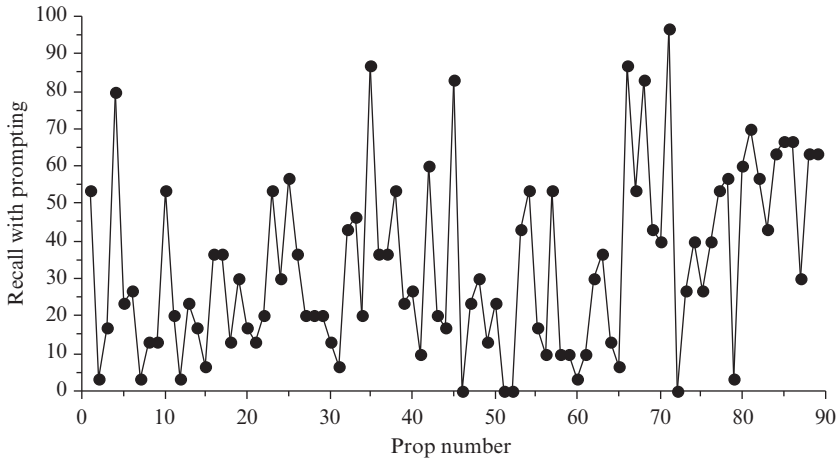


Figure 6.1 shows that the text-based propositions most easily recalled by readers in the free phase appear to cluster in three particular locations within the story. The first of these three clusters occurs close to the start of the story, focusing on the initial setting of the scene: TP04 – *the hour was late*, TP10 – *the man wanted to get home*. The second cluster occurs in the middle of the story at the point where there is partial resolution of the dilemma facing the man: TP35 – *the car stopped*, TP42 – *the man sat beside the driver*. The inclusion of proposition TP45 here (*the man had been waiting*) is probably a direct result of the coding system used. A large

number of the Text A recalls mentioned 'a man waiting' so it seemed appropriate to use proposition TP45 for this since no other proposition was available in the set; as a result, it has been somewhat displaced from the first cluster. The third cluster occurs towards the climax of the story, the point at which the man's fear is reaching its peak: TP66 – *the driver took something short and black from the side-pocket of the car*, TP68 – *would the driver attack him with that?*, and TP71 – *the man's heart beat so fast with fear that*. Similar peaks at similar points in a story have been observed by other researchers when analysing readers' recalls of artificially constructed short stories (Malmkjaer, personal communication).

Figure 6.2 shows the ease with which text-based propositions were remembered after probing (i.e. during the prompted recall phase). The overall pattern remains largely the same as in Figure 6.1 – the three clusters are still clearly visible. Figure 6.2 suggests that probing increased ease of remembering by an average of approximately 10%. It also shows that while probing made a significant difference to the remembering of certain propositions by readers, there were others where it made no difference at all. For this reason, closer attention was focused on how far prompting affected ease of remembering. Figure 6.3 below illustrates the increase in percentage terms in remembering of propositions as a result of prompting.

Figure 6.3 Percentage increase in ease of remembering after probing (Text A)

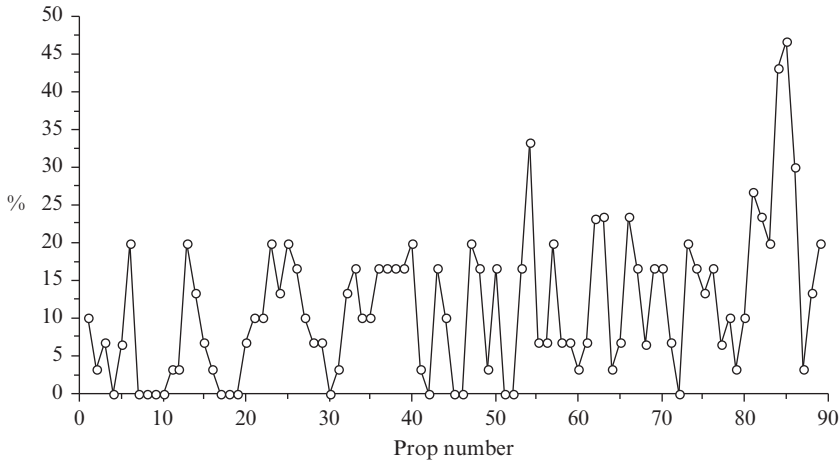


Figure 6.3 suggests that probing increased ease of remembering for 19 (21%) of the 89 text-based propositions by least 20%. These 19 propositions are listed in Table 6.8.

Many of the text-based propositions that became easier to remember after probing seem to relate directly to peripheral details of the narrative.

Table 6.8 Propositions more easily remembered after probing (Text A)

TP06	the street was silent
TP13	the man started to whistle
TP23	the noise of a falling dustbin reached the man's ear
TP25	instinctively the man's hand felt for his wallet
TP40	said the driver
TP47	in the semi-darkness of the car the man turned to look at the other passengers
TP54	the man had heard of passengers being attacked at night
TP57	if only the man could see the other man's face clearly
TP62	the road branched off in another direction
TP63	there were tall, dark bushes around
TP66	the driver took something short and black from the side-pocket of the car
TP73	the car did not stop
TP81	the car came to a standstill
TP82	the man fumbled in his wallet for the fare
TP83	the taxi was no longer there
TP84	"There'll be no more night passengers for me again"
TP85	exclaimed the driver
TP86	with a sigh of relief the driver hurriedly moved off
TP89	had that queer passenger attacked him

One reason may be that this level of detail is not volunteered by readers in their free recall of the text, but is more likely to appear when asked whether they can recall any *additional* content from the text. Text content mentioned by readers in a free recall is presumably what they consider to be the most important information, or information at a higher level of generalisation. Less important information relating to peripheral or supplementary details may well figure in their mental representation of a text, but may be called up through probing rather than volunteered in free recall. Evidence that prompting seems to call up peripheral details which are not central to the text may militate against its use as a method for identifying the content for a summary.

Interestingly, Figure 6.3 clearly illustrates that the more easily remembered propositions for the narrative include set TP84–89 which relate to the ending of the story. Since the story ending could be described as central rather than peripheral, it seems surprising that it should have been mentioned so little in free recall. Perhaps the unexpected shift of perspective at the end of the story (from the passenger's to the driver's point of view) made this feature difficult for some readers to integrate into their mental representation. Increased ease of remembering after probing may also indicate the strong influence of probe question 6, i.e. *Can you recall anything more of the driver's feelings at the end of the story?*

A study was also made of any text-based propositions where prompting seemed to have no effect, i.e. no increase in ease of remembering after probing took place. The list of these propositions is given in Table 6.9.

Three reasons may explain why no increase in ease of remembering after probing was observed for these particular text-based propositions:

1. Some propositions may have been so memorable that probe questions had no additional effect, e.g. TP04, TP10, TP17, TP42, TP45. Four out of these five already occurred in the skeletal account provided by the Set T(iv) propositions.
2. Some propositions may not have been any easier to recall because the probe questions asked were not sufficiently sensitive, e.g. TP08, TP09, TP18, TP19, TP30.
3. Some propositions may have been so hard to recall that even sensitive probe questions did not improve their chances, e.g. TP07, TP46, TP51, TP52, TP72.

Table 6.9 Propositions for which probing seemed to have no effect

TP04	the hour was late
TP07	the man looked up and down
TP08	he was hoping that
TP09	some vehicle would come in sight
TP10	the man wanted to get home
TP17	the man was ten miles away from home
TP18	what was the man to do?
TP19	it was out of the question to begin to walk that distance
TP30	the man began to walk up and down, up and down
TP42	the man sat beside the driver
TP45	the man had been waiting
TP46	if only someone would say something
TP51	the man mustn't allow himself to think that
TP52	the man glanced at the driver
TP72	the man could hardly breathe

It is noticeable that the reported thoughts and feelings that suggest a generalised uneasiness in the text (rather than strong urges and emotions) are among those propositions which were difficult to recall even with probing (e.g. TP07, TP08/09, TP18/19, TP30, TP46, TP51, TP52, TP72). It may be that the emotional colour in a text does not need to be included in a summary, unless it is germane to the action or plot. It is also possible that the very nature of the coding system resulted in some propositions appearing artificially difficult to remember. For example, each complex sentence from the story had to be separated out into multiple text-based propositions and this may have meant that where one proposition in the sentence dominated, those around it simply faded into the background, e.g. TP02 and TP03 (low-frequency) were overshadowed by TP01 (high-frequency).

Insights into readers' structuring of information from Text A

Analysis of text-based propositions occurring in readers' recalls of Text A may indicate something about the way in which readers structured the

information from the text. Some propositions offered during recall appear to reflect nodes for the structuring of information in a reader's mind, while other information units or propositions were perhaps integrated into the overall structure not only by clustering them around a particular node but also by linking them to other nodes at the same time. This may constitute evidence of propositional or semantic networks (see Field 2004) and supports the presence of spreading activation in accordance with recent connectionist approaches to language processing (see below). It would provide a plausible explanation for the many occasions in the recall transcripts when a reader's recall of one proposition from the text apparently triggered their recall of a second or a third. The following short extracts from the recall transcripts illustrate this phenomenon, according to which certain pieces of information from the story appear strongly linked to one another (numbers in brackets refer to individual recall protocols). For example, the mention of *money* sometimes seemed to trigger the mention of *wallet* and *checking his wallet*:

> *he couldn't find his money + well + he was looking for his wallet because + oh he kept going for his wallet all the way through the journey to make sure it was still there* (#010)
> *the taxi's moved off before he can pay any money + and also he's checking the wallet the whole time + and the taxi moves off* (#013)
> *and + er + worried about stories he'd heard of of people being attacked by the taxi drivers + they were about people taking money from his wallet + he kept checking his wallet all the time* (#001)

In a similar way, *checking the wallet* and *looking for something to protect himself* appeared to be linked in readers' minds:

> *he checks for his wallet + he wanted a stick earlier so he + um + to defend himself + keeps checking that he's got his wallet for money* (#009)
> *he checks to see if his wallet is in his pocket and he thinks oh the wallet's there + then he thinks have I got anything to protect myself with + you know* (#018)
> *he was thinking about + um + whether he had anything to protect himself with but he didn't + and he kept putting his hand down to check that he still had his wallet* (#030)

Finally, *being attacked* and *checking for his wallet* also appeared to trigger one another:

> *he started having sort of paranoid delusions about the taxi driver and thinking that the taxi driver was going to steal all his money + and he kept on checking to see if his wallet was in his pocket* (#016)
> *thought he was going to get robbed + he kept feeling for his wallet* (#021)
> *he keeps touching his wallet + checking it was there + he thinks + he thinks he's going to be beaten and robbed* (#023)

that's right because he looks in his wallet + he'd been getting out his + he was handling his wallet while they were in the car + and could be he was going to get mugged (#027)
he thought he was going to be mugged beaten up or have his money taken or whatever and he kept reaching into his pocket to check his wallet (#016)
he's a bit paranoid that um someone's going to attack him and steal his wallet (#003)

If a node is activated, either as a result of unprovoked retrieval or in response to a probe question, then perhaps it activates in turn the information units clustered around it or linked to it in some way making them equally available for recall. This may also explain why propositions occurring close to one another in the text all seemed to be more easily recalled after prompting and by more or less the same degree, e.g. TP23/25, TP62/63/66, TP81/82/83, TP84/85/86. Some degree of interdependence between elements within the information structure is implied. This effect is consistent with a schema-theoretic view of processing in reading comprehension in which a reader's schemata (or knowledge already stored in memory) interact with new and incoming information, thereby enabling its integration into a coherent mental representation.

It is also consistent with the spreading activation theory of semantic processing discussed by Collins and Loftus (1978), developed to explain the way related knowledge structures become activated in memory. Field defines spreading activation as a 'process which speeds up the retrieval of lexical items that are associated with one that has just been seen or heard. The earlier word is said to *prime* the later one' (2004:288). Spreading activation was a central feature of parallel distributed processing (PDP) models of language processing (McClelland, Rumelhart and Hinton 1986). In the case of Text A (*Journey*) it seems likely that elements such as the *wallet*, the *money* and the *taxi fare* are all linked coherently within a familiar 'taxi ride' schema that extends further to incorporate a 'potential personal attack' schema. This links those elements relating to being *beaten*/*robbed*/*mugged* and wishing for some form of *protection* in the form of a *stick*.

Particularly interesting in this regard is the scope for individual differences in the way schemata may affect the emerging mental representation. Although the readers in this group appeared to hold many aspects of their schemata in common, there are clear cases of the effect of individual differences in existing knowledge structures, e.g. in the way readers integrated the mention of 'Valencia' into their mental representation. In general, it seems that a substantial number of readers clearly understood the word as the name of a place to which the man wished to go:

he explains he wants to go to valencia (#002)
he asked to go to valencia or something like that I think (#007)

he asks the driver to take him to valencia (#008)
he wants to go to valencia + asks to go to valencia (#009)
hoped it was a taxi + it was + asked to go to somewhere + valencia (#015)
he said taxi and then valencia + or wherever he was going + valency or whatever + unpronounceable word (#021)
are they + is it in a foreign country + the man asks to go to valencia or something + isn't that in spain (#022)
he wants to go to valencia (#023)
he gets in + asks to go to valencia (#028)
he asked if it would take him to valencia (#029)

Presumably, the relatively mature and experienced readers in this study arrived at this interpretation because some statement of destination would be quite normal in their own personal 'taxi ride' schema. Some readers, however, apparently encountered difficulty incorporating this word into their emerging mental representation. Perhaps this was because, despite some personal experience of taxis, the word did not conform to their expectations of a place-name. It is, after all, recognisable to many as the name of a Spanish rather than an English town and some readers clearly drew on their background knowledge of European geography to infer this. Ironically, what none of the readers realised was that this particular tale came from a published collection of Caribbean short stories. Thus the story was presumably set in the Caribbean and the mention of Valencia is more likely to have referred to the town of that name in Trinidad and Tobago!

Alternatively, perhaps the word 'Valencia' corresponded more closely to another slot in their taxi-ride schema, such as the name of the taxi company. Illustrative examples from the recalls are given below:

he opened the door and said taxi + and then he said something that I didn't really understand + valencia or something + I don't know quite what (#010)
stuck his hand out and said taxi valencia for some strange reason (#016)
he sort of runs out into the road and shouts taxi valencia + I don't know why he shouts valencia (#018)
the person whose taxi it is says taxi and then their name + I think it's valencia (#019)
then he said a name + taxi + valencia taxis + he just said + er + valencia taxis (#027)

Given the apparent confusion for some readers regarding the Valencia mention, it seems reasonable to consider making this explicit in any summary, and especially to avoid making it the focus of a test item.

The evidence provided above suggests that the recalls offer plausible insights into the information structuring of readers' mental representations of the short story, and in particular the effect of individual differences. What is clear for our purposes is that a reader's background or existing knowledge structures can

significantly affect their construction of a coherent mental representation of a text. The influence of readers' schemata in terms of how they contribute to the process of making inferences will be explored later in this chapter.

Some conclusions from analysing text-based propositions in Text A recalls

The analysis of text-based propositions occurring in the 30 recall transcripts for Text A confirms that there are indeed certain propositions of a text which readers recall more readily than others. It also suggests that while some propositions in a text will consistently be recalled by a group of readers when reading the same text for the same purpose, others will not. It seems reasonable to suggest, therefore, that the relative frequency of occurrence of text-based propositions in readers' recall protocols of the text should be one of the things to consider when identifying propositions to be included in an adequate summary of the text.

Analysis of summarising propositions in recall transcripts of Text A (*Journey*)

Identifying sets of summarising propositions recalled by readers

A second-level analysis of the recalls for Text A was focused at a higher level of the text structure to explore whether some summarising propositions were more frequently offered by readers than others. (See pages 95–100 in Chapter 5 for the definition of a summarising proposition together with the set of 22 possible summarising propositions created for Text A.) Once again, it was decided to investigate the extent to which different sets of summarising propositions might assist in constructing a summary of Text A that could form the basis for developing a summary completion task. General principles for identifying sets of summarising propositions were based upon:

> a) the number of readers (i.e. proportion of the total) who offered an individual summarising proposition, and
> b) whether the summarising proposition occurred only in the free recall phase or across the two recall phases (i.e. free and prompted).

It was anticipated that this analysis might help to reveal how readers employed macro-propositions (rather than micro-propositions) when constructing their mental representation of the short story. Eight tentative criteria for identifying sets of summarising propositions were devised as follows:

i) propositions offered by 20 out of 30 readers in FREE recall
ii) propositions offered by 20 out of 30 readers in ANY recall
iii) propositions offered by 15 out of 30 readers in FREE recall

iv) propositions offered by 15 out of 30 readers in ANY recall
v) the top third most frequently occurring propositions in FREE recall (i.e. high-frequency – top 33%)
vi) the top third most frequently occurring propositions in ANY recall (i.e. high-frequency – top 33%)
vii) the top two-thirds most frequently occurring propositions in FREE recall (i.e. high and medium-frequency – top 66%)
viii) the top two-thirds most frequently occurring propositions in ANY recall (i.e. high- and medium-frequency – top 66%)

As with the text-based propositional analysis, criteria (i) and (ii) established a relatively demanding threshold of 66% consensus among readers, while criteria (iii) and (iv) set a less demanding threshold of 50% agreement. Analysis of high-frequency (top 33%) summarising propositions was included (criteria (v) and (vi)) and high/medium-frequency criteria (vii) and (viii) were added to these in case these proved to be needed, given the relatively small number (22) of summarising propositions available.

Comparing content across the sets of summarising propositions

Using the criteria shown above, eight sets of summarising propositions were identified. The codes and wording for the summarising propositions come from Table 5.3 on page 100. Since there proved to be no difference between the propositional content of Sets (v) and (vi), the FREE/ANY recall distinction was dropped for this pair. The same was true for Sets (vii) and (viii). Contents of most sets are shown in Tables 6.10 to 6.15.

Table 6.10 Summarising propositions offered in FREE recall by 20 out of 30 readers – Set S(i)

SP01	a man was present
SP09	a car came along and stopped
SP13	the man started to think about being attacked
SP14	the man grew increasingly suspicious of the driver
SP16	the driver reached for something
SP19	the man got out of the car
SP22	the driver was afraid of being attacked

Set S(i), the seven (31%) out of 22 summarising propositions offered by at least two-thirds of the readers, gives a simple account of the short story: *the presence of a man*, *the arrival of a car*, *the man's growing fear for his safety*, *the driver reaching for something*, *the man getting out of the car* and *an explanation of the driver's action*. This set of summarising propositions provides an accurate and at least partially coherent account of the events in the story. Interestingly, it contains no description of the setting at the start of the story which, it could be argued, is a major element contributing to the story's plot development.

Table 6.11 Summarising propositions offered in ANY recall by 20 out of 30 readers – Set S(ii)*

SP01	
SP02	the place was dark and lonely
SP09	
SP13	
SP14	
SP16	
SP19	
SP21	the driver drove away fast
SP22	

** summarising proposition codes from S(i) are included here but without their wording in order to highlight the information that is being added as a result of probing*

Set S(ii) includes nine (40%) of the 22 summarising propositions and it elaborates on the storyline given in set S(i). Extra information is now provided about *the setting at the start of the story*, as well as about *the driver's speedy getaway at the end* – a direct result of the information contained in SP22.

Setting a lower threshold of 15 out of 30 (i.e. 50%) of readers generated a Set S(iii) comprising 11 (50%) of the 22 summarising propositions.

Table 6.12 Summarising propositions offered in FREE recall by 15 out of 30 readers – Set S(iii)

SP01	a man was present
SP02	the place was dark and lonely
SP03	the man wanted to get home
SP08	the man was concerned for his personal safety
SP09	a car came along and stopped
SP13	the man started to think about being attacked
SP14	the man grew increasingly suspicious of the driver
SP16	the driver reached for something
SP19	the man got out of the car
SP21	the driver drove away fast
SP22	the driver was afraid of being attacked

Set S(iii) contains specific references to the man's *psychological state* at the start of the story: the fact that *he wanted to get home* and that *he was concerned about his personal safety*. Both of these might be considered important elements in the story.

Set S(iv) contains 14 (63%) of the summarising propositions and adds information about *the brief exchange between man* and driver when the car stops, as well as about *the point during the journey at which the man panicked on seeing the driver reach for something*, and the fact that *he did not succeed in paying the fare* after getting out of the car.

Table 6.13 Summarising propositions offered in ANY recall by 15 out of 30 readers – Set S(iv)

SP01	
SP02	
SP03	
SP08	
SP09	
SP10	the man and the driver briefly exchanged words
SP13	
SP14	
SP16	
SP17	the man panicked
SP19	
SP20	the man tried unsuccessfully to pay
SP21	
SP22	

Sets S(i) to S(iv) (as shown in Tables 6.10 to 6.13) were selected by analysing the consensual use of summarising propositions by 66% and 50% of the readers respectively. Once again, it was decided to analyse the summarising propositions by *frequency of occurrence* (i.e. high/mid/low-frequency) to see whether this might shed further light on the nature of core, peripheral and superfluous content in the mental representation of text. High-frequency summarising propositions (the top 33%) offered across both recall phases are shown in Table 6.14, while high- and medium-frequency summarising propositions (the top 66%) are shown in Table 6.15.

Table 6.14 High-frequency summarising propositions offered in FREE/ANY recall – Set S(v/vi)

SP01	a man was present
SP02	the place was dark and lonely
SP09	a car came along and stopped
SP13	the man started to think about being attacked
SP14	the man grew increasingly suspicious of the driver
SP16	the driver reached for something
SP19	the man got out of the car
SP22	the driver was afraid of being attacked

Set S(v/vi) includes eight (36%) of the 22 summarising propositions and is similar to set S(ii) except that it does not include mention of the driver's speedy getaway.

Set S(vii/viii) contains 15 (68%) of the 22 summarising propositions and is similar to set S(iv), except that it has the added reference to *the man being anxious about how to get home*.

Table 6.15 High- and medium-frequency summarising propositions offered in FREE/ANY recall – Set S(vii/viii)

SP01	
SP02	
SP03	the man wanted to get home
SP05	the man was anxious about how to get home
SP08	the man was concerned for his personal safety
SP09	
SP10	the man and the driver briefly exchanged words
SP13	
SP14	
SP16	
SP17	the man panicked
SP19	
SP20	the man tried unsuccessfully to pay
SP21	the driver drove away fast
SP22	

Analysing low-frequency summarising propositions

As with the text-based propositions, it is perhaps the low-frequency summarising propositions (shown in Table 6.16) which are of particular interest.

Table 6.16 Low-frequency summarising propositions in ANY recall

SP04	the man tried to keep his spirits up
SP06	the man was aware of the darkness intensifying
SP07	the man was aware of the noises of the night
SP11	the man initially felt relieved
SP12	the man started to feel uncomfortable
SP15	the car drove into a country area
SP18	the car failed to stop

Once again the question of interest is why these particular propositions should have been offered so infrequently by readers in their recalls. Several different factors may have been instrumental.

One possible explanation for the low frequency of SP04, SP11 and SP12 is that readers found it relatively difficult to integrate the subtler *emotional* aspects of the story into their mental representation after only a single reading. The recalls seem to show that although readers found the early anxiety, growing fear and ultimate panic of the man in the story memorable (i.e. all these are emotional elements which could be regarded as central to understanding the story), they were apparently much less sensitive to those points in the story where the man's emotional state swings, albeit temporarily, in a different direction, e.g. the man's attempt to keep his spirits up at the start of the story, and his initial relief when he gets into the car. This explanation

is supported by the fact that of the 13 text-based propositions which correspond to SP04, SP11 and SP12, seven (53%) are categorised as low-frequency and 10 (76%) occur in the bottom 50% of text-based propositions when ordered by frequency of occurrence. Only two of the corresponding propositions are categorised as high-frequency, probably because they refer to a significant development or feature of the story: TP42 – *the man sat beside the driver* (i.e. the man got into the car), and TP45 – *the man had been waiting* (see the earlier reference to the influence of the coding system on the frequency of this proposition). Furthermore, SP04, SP06 and SP07 all reflect the generalised (as opposed to specific) uneasiness referred to by the writer throughout the text. It is also interesting to note that the other three summarising propositions which make reference to the man's emotional state – SP05, SP08 and SP17 – still occur no higher than the medium-frequency set of summarising propositions. This finding replicates the observation made regarding emotional colour in the text-based propositional analysis above.

We might conclude that high-frequency positions seem to be reserved for actions in the story. This may reflect Gomulicki's finding that the most frequently recalled elements of both narrative and descriptive texts related to their action content (i.e. actions normally proved to be more effective stimuli than static conditions):

> In short, the strongest reaction was to elements bearing the action content of a passage. Agents had the next best representation, followed by the effects or recipients of the action. The weakest reaction was to items which, while serving to fill in the general picture, only retarded the action (1956:91).

Interestingly, Gomulicki also observed that in descriptive passages subjects took advantage of real and potential narrative elements, often expressing the content in terms of *doing* rather than *being* something.

If it is the case that readers' mental representations are likely to reflect more of the agents and actions in a story and less of its emotional shading (unless that emotional shading is strongly tied into an event or action in the narrative), then this could have implications not only for what is included (or not) in a summary but also for the testing focus in any comprehension test based upon the text. Test items focusing on a reader's awareness of the more subtle emotional aspects of a narrative text are likely to be particularly difficult unless the reader has been specifically prompted in advance to focus on this aspect of the story (or is able to go back over the text and process it again in greater depth). On the other hand, it may be that the careful testing of such aspects could offer one means of distinguishing *average* from *good* comprehension.

Another factor influencing the low frequency of certain summarising propositions may have been the presence within the same segment of text of a word, phrase or proposition that exerted some sort of inhibiting effect.

On several occasions, readers themselves reported that a particular word or phrase had caught their attention, either because they didn't understand it or because it seemed curious to them in some way. SP04 offers a good example of this phenomenon. Within the text-based propositions corresponding to SP04 are two of the words remarked upon by a number of readers: TP12 – *the silence began to pall*, and TP14 – *there was no mirth in his whistling*. On several occasions, different readers commented upon these two words as being 'difficult', 'slightly formal or strange', 'curious', 'unusual' or 'old-fashioned'.

An analysis of these two words using the VocabProfile tool in Compleat Lexical Tutor (Cobb 2006) shows them to be among the least frequent in the short story when compared against the lexical content of the British National Corpus (BNC): *pall* falls into the BNC/K8 band, while *mirth* falls into the BNC/K11 band. Given that both lexical items are thus relatively low-frequency in everyday English and occur close to one another in the text (only eight words apart), it may be that at these points readers found it difficult to integrate the more subtle emotional aspects of the narrative into their mental representation precisely because they were distracted by these less familiar words. It seems reasonable to conclude that a reader's mental representation of text will be partly influenced by the extent to which they encounter and process unusual or unfamiliar words, phrases or concepts in a text. It could be argued that such words should not be included in the text and certainly should not be made the focus for testing.

Closer examination of low-frequency summarising proposition SP06 showed that all three text-based propositions which it encapsulated (TP20, TP21 and TP22) were also low- to medium-frequency, indicating that this part of the story (i.e. *the dark cloud hiding the few pale stars*) was relatively unmemorable perhaps for the reasons already discussed. By contrast, in SP07 the text-based propositions (TP23 – *the noise of the falling dustbin* and TP24 – the reference to *some dog scattering its contents*) were high- and medium-frequency respectively. It is possible that these last two details were memorable or few enough to be integrated into a summarising proposition. The low-frequency of SP015 is perhaps similar in this respect since, although it incorporates two medium-frequency text propositions (TP62 – *the road branched off in another direction* and TP63 – *there were tall, dark bushes around*), it also covers three other low-frequency text propositions.

The low frequency of SP018 is particularly curious and may reflect a measure of ambiguity within the text itself. The ambiguity concerns whether or not the car in the story initially failed to stop at the man's request. The text reads as follows: *"Stop!" he heard himself screaming, and his heart beat so fast with fear that he could hardly breathe. But the car did not stop. Faster and faster instead it went*. In general, most readers appeared to conflate the two references to asking for the car to stop, reporting only a single request. However, some recall transcripts appear to indicate two different possible

interpretations for this segment of text. For example, some readers clearly believed the man did actually scream aloud 'Stop!' and that the driver failed to stop the car in response:

> *and he shouted stop but then it got faster* (#007)
> *the boy asked the taxi driver to stop but he carried on faster* (#014)
> *(the man) screams to the driver stop + um + for some reason it doesn't + I'm not sure why + um + and then then he asks the driver to put him down there* (#015)
> *when he said stop it got faster and faster + it seemed to him to get faster and faster and wouldn't stop* (#027)
> *he says he's aware of screaming stop stop* (#028)

An alternative interpretation could be that the man screamed only in his imagination and that the driver's apparent failure to stop was simply because his passenger had not actually screamed it aloud at this stage. Although no recall transcript included an explicit example of this second interpretation, it is clear that at least two readers were aware of this potential confusion or ambiguity of interpretation, as shown below:

> *it was confusing + he screamed stop and it didn't + and then it did* (#009)
> *he asked him to stop once + I don't think he did + or it said he kept going + it wasn't quite clear + didn't he ask him to stop and he didn't but then he finally stopped* (#010)

In effect, it seems that either interpretation could be potentially valid in understanding the story, thus a decision would need to be made on whether this feature should be included in a summary and, if so, with which interpretation. One possibility might be to include the more typical interpretation of the two (i.e. that the man screamed aloud, rather than just in his imagination, the first time). The preferred, and safer, option may be to avoid inclusion of this ambiguous element altogether. Whichever approach is taken, it is an interpretation feature of the story which is unlikely to make a satisfactory testing focus.

Some conclusions from analysing summarising propositions in Text A recalls

Analysis of the summarising propositions occurring in the 30 recall transcripts for Text A appears to confirm that certain summarising propositions of a text are likely to be constructed and deployed in reading recall more readily than others. It also suggests that while some of these propositions will consistently be used by a group of readers when reading the same text for the same purpose, others will not. It seems reasonable to suggest, therefore, that the relative frequency of occurrence of summarising propositions in readers' recall protocols of Text A should be one of the things to consider when identifying propositions

to be included in any summary of the text, and that a further consideration will be the extent to which these form a coherent account of the text.

Analysing additional propositions occurring in recall transcripts of Text A

The presence of additional propositions in readers' recalls of Text A

One of the most interesting features of readers' recall transcripts for Text A is the extent to which they contain more than just material that can be matched to corresponding text-based or summarising propositions. Recalls frequently include additional propositions such as the underlined clauses in the extracts below:

> *he didn't want to walk home because it was too dangerous*
> *there was a dustbin so it's probably an urban setting*
> *he got in the back of the taxi*

The 30 recall transcripts for Text A were analysed to examine these additional propositions in detail. It was anticipated that these additional propositions would provide some evidence of reader inferences and could therefore be accounted for by referring to theories of inferencing.

Chikalanga (1992) distinguished between *propositional* inferences, which are text-based (i.e. within the text), and *pragmatic* inferences, which are based upon information from outside the text. Others have categorised inferences in a similar way including Crothers (1978), Farr, Carey and Tone (1986), Hughes (1993, 2003) Pearson and Johnson (1978) and Weaver and Kintsch (1991). Graesser et al's 1994 review of psychological research into inferencing distinguished inferences that are *necessary* to support local and global coherence (e.g. referential, case structure role assignment, causal antecedent, superordinate goal, thematic and character emotional reaction) from other inferences that are more *elaborative* and less likely to be necessary for coherence (e.g. causal consequence, instantiation of noun category, instrument, subordinate goal action and state). The former, suggested Graesser et al (1994), are generated online (i.e. spontaneously at the time of first reading), while the latter are generated offline (i.e. after reading, perhaps in response to comprehension questions). A few research studies suggested that some elaborative inferences may be made spontaneously during reading, in cases for example where the background knowledge activated by the text is sufficiently constraining (Garrod and Sanford 1981, O'Brien et al 1986).

Examples of necessary and elaborative inferences offered by readers of Text A will be discussed in turn in the next section. The overall aim was to determine the types and relative frequency of inferences made by readers and

to consider their relevance for any summary of the short story. While the generally recognised categories of *necessary* and *elaborative* inferences will help to organise the discussion below, it may be that these categories are not always discrete and are prone to "leakage". It is also difficult to determine whether the Text A inferences under discussion were made online while reading or offline during free recall or in response to probe questions. However, it is not the timing of inference generation which is the central issue here. Of greater importance is distinguishing those inferences that are *essential* to readers' mental representation of Text A and which can thus legitimately be included in a summary of the text. Necessary or propositional inferences are more likely to be shared across a group of readers and thus feature as part of a summary, while pragmatic or elaborative inferences are likely to be more idiosyncratic, reflecting individual reader differences in background knowledge and experience. As such, they are less likely to be candidates for inclusion in a summary.

Inferences necessary to support local and global coherence for Text A

Readers of Text A undoubtedly generated some online inferences which maintained coordinates of time, space and topic during initial reading of the short story. These more retrospective (or bridging) inferences were presumably necessary for readers to construct (and later report) an accurate and coherent mental representation. If these necessary inferences had not been made, then one might have expected extensive evidence of readers' inability to make accurate and coherent referential connections between successive mentions of people, places and events in the narrative. Interestingly, although there was very little evidence of referential inferencing breakdown in the oral recalls of mature readers in this study, we shall see in Chapter 7 that substantial evidence of this was found in a study of *written* recalls of Text A by *younger* readers.

Other examples of necessary inferences generated by readers of Text A include those relating to superordinate goals and those of a thematic nature concerning the main point or moral of the text.

i) Inferences relating to superordinate goals

Several readers inferred a reason for the passenger touching his wallet several times during the short story:

> *keeps checking that he's got his wallet for money* (#009)
> *he kept going for his wallet all the way through the journey to make sure it was still there* (#010)
> *he kept on checking to see if his wallet was in his pocket* (#016)
> *he checks to see if his wallet is in his pocket and he thinks oh the wallet's there* (#018)

he keeps touching his wallet + checking it was there (#023)
he kept putting his hand down to check that he still had his wallet (#030)

It may be that the author's repeated mention of the same action was influential in provoking readers' inferencing on this point, whereas a single mention would not have had the same effect. Similarly, at least one reader inferred a reason for the passenger wanting a stick at one point:

he wanted a stick earlier so he + um + to defend himself (#009)

ii) Thematic inferences

Most readers seemed to have little difficulty in inferring one of the main points of the text (i.e. that the passenger thought he was in danger of being attacked or robbed):

he's a bit paranoid that um someone's going to attack him (#003)
he started having sort of paranoid delusions about the taxi driver and thinking that the taxi driver was going to steal all his money (#016)
he thought he was going to be mugged beaten up or have his money taken or whatever (#016)
thought he was going to get robbed (#021)
he thinks he's going to be beaten and robbed (#023)
could be he was going to get mugged (#027)

Examples of elaborative inferences associated with Text A

Most of the additional propositions in the Text A recall protocols can probably be regarded as evidence of elaborative (or pragmatic), rather than necessary, inferencing.

i) Inferences about the main character

It is clear from the protocols that readers varied in their interpretation of exactly who the participant in the story was. Although the text specifies the *maleness* of the character, his age is not specified, and readers varied in the inferences they made concerning how old he was. Variation in interpretation was as follows: 21 out of 30 readers referred to the participant as an adult male – *a man*, *a bloke*, *a guy*, and for two of these readers he was clearly *a young man* and *a middle-aged man*. Interestingly, three readers referred to him as *a boy* (in one instance *a little boy*) and one reader was unsure whether he was *a man or a boy*. The remaining five readers used indeterminate terms such as *someone, somebody*, and *a person*. Clearly the vast majority of readers instantiated an adult male, but it is interesting that at least three readers seem to have pictured him as younger, perhaps someone of their own age (i.e. mid- to late teens), or even younger. Although one could argue that the participant

is most likely to have been an adult male (since he was alone, late at night, 10 miles from home, looking for a taxi, in possession of a wallet), the grounds for doing so are experiential or cultural rather than text-based.

ii) Inferences about the location

The text does not tell us whether the story begins in a large town, a small village or in open country. Nevertheless, five readers reported a clear picture of where they thought the protagonist was at the start of the story:

> *it's somewhere in civilisation because of the dustbin* (#008)
> *there was a dustbin so it's probably an urban setting* (#012)
> *it's a quiet deserted sort of backstreet type of location* (#017)
> *a dark lane + like an alleyway + um + like you see in a cartoon* (#024)
> *waiting to get in another town + maybe a village* (#024)

For some readers the mention of a dustbin seemed to trigger (or prime) an inference about an urban or back-street location. This is presumably an example of personal experience or culture dictating that dustbins are generally found in built-up, inhabited areas rather than in open country.

iii) Inferences about the setting/atmosphere

Some readers clearly made elaborative or pragmatic inferences about the setting at the start of the story and about how the man felt:

> *I have the impression that it was misty + I don't remember for certain but I have the impression it was + it gave across that impression + um + fairly quiet + noone else around + probably very late at night + early morning* (#015)
> *it's dark + he's shifting from foot to foot + he's nervous + for some reason I got the impression it was cold but that's probably because there's a star + it was a clear sky and then the clouds passed over* (#018)
> *it goes on about how dark it is + how cold he is* (#016)
> *he was + I think he was cold + well he was sort of standing swaying to and fro + stepping from side to side and he was tired* (#002)

Text-based references to the time of night, to the clear sky and to the man's movements evidently prompted some readers to infer that it was cold, even though this is not implied, and is probably unlikely given that the story location was actually in the Caribbean!

iv) Inferences about why the main character was there

No explanation of the events which led the character to this point is offered in the story, but at least four readers drew their own pragmatic conclusions:

has he missed the bus (#011)
this person who's obviously just been out for the night (#019)
he'd obviously been out at night and he was waiting at a taxi-stand (#030)
a man obviously waiting for his taxi after a night out (#027)

These readers have sought to provide for themselves a satisfactory explanation as to why the character in the story should be outside, alone, very late at night. From the following examples, at least two readers believed the participant to have already ordered a taxi and to be waiting for it to arrive:

he was getting nervous because the taxi was late (#027)
he'd called the taxi + I think he'd called the taxi so he was just waiting for it (#021)

Two readers seemed to infer that the man's reason for not walking home was more to do with potential danger than with the issue of how far it was:

he didn't want to walk home because it was too dangerous (#007)
it was safer to take a taxi than to walk (#020)

v) Inferences about Valencia

Inferences relating to the mention of Valencia were considered at some length earlier in this chapter (see pages 118–119) so will not be discussed again here.

vi) Inferences about where the character sat inside the car

Although Text A makes it clear that the passenger *sat beside the driver* (i.e. in the front of the car), at least three readers recalled the passenger as being in the back of the car. This may be because background knowledge dictates that, in Britain at least, a passenger normally sits in the back of a taxi. In some parts of the world, however, it is common practice for a passenger to sit beside the driver, especially if the taxi is an unofficial one:

the taxi stops and he gets in the back + and he notices that there're no other passengers there + and he just sits in the back (#016)
he couldn't see the taxi driver and he was in the back and so he got a bit frightened (#028)
he got in the back of the car (. . .) he jumped out from the back of the taxi (. . .)
he gets in the back he just sits in the back (#019)

vii) Inferences about whether the car was a taxi or not

The nature of the car which picked up the passenger is one of the aspects of the story which seems to have provoked a large amount of reader

inferencing. It is clear from the protocols that readers shared a number of inferences about how the man viewed the car that stopped to give him a lift. At least four readers inferred that the man at first believed the car to be a genuine taxi:

> *he assumed the car was a taxi* (#004)
> *what he thinks is a taxi eventually turns up* (#022)
> *a car pulls up which he assumes is a taxi* (#013)
> *a vehicle which he presumed was a taxi* (#030)

Another four readers inferred that the man later became less convinced of the taxi's genuineness:

> *he wasn't even sure that it was a taxi* (#009)
> *he's not sure whether he's in a taxi* (#020)
> *he didn't actually know it was a taxi I think* (#017)
> *he starts to wonder about whether it's really a taxi or whether the taxi driver is really a taxi driver* (#008)

Two readers expressed their own uncertainty as to whether it was a genuine taxi:

> *I don't think the car was a taxi 1 it was just someone picking him up out of goodwill (#004)*
> *it doesn't actually say whether the car's a taxi or not* (#018)

This may have been a critical element of the story in terms of overall plot and structure, since the passenger's uncertainty about the status of both the car and driver was probably a key factor in his increasing anxiety and ultimate panic. Interestingly, the reader who gave one of the fullest and most accurate recalls commented as follows:

> *it doesn't actually say whether the car's a taxi or not + I was + I thought the story was going to turn on that* (#018)

viii) Inferences about where the spanner was

Text A is explicit about the fact that the driver took the spanner from *the side-pocket of the car*, yet it is interesting that one reader 'translated' this into *the glove compartment*:

> *they see the taxi driver reach for something + an object in the glove compartment (. . .) they realise that the spanner in the glove compartment* (#019)

ix) Inferences about whether the passenger paid the driver
Although the story implies that the driver of the car moved off before the passenger could pay his fare, at least one reader recalled the opposite:

I think he pays (#009)

Another reader recalled the passenger being told not to pay:

the taxi driver tells them not to bother to pay (#019)

Some conclusions from analysing the inferences in Text A recalls

The intention in the previous section was not to list and categorise *all* the inferences made by readers of Text A. Some inferences may have been made but not reported by readers. Others may not have taken the form of additional propositions. Instead, the analysis of inferences for which there was evidence in the recalls in the form of additional propositions aimed to gain insight into the mental representations of the story constructed by readers. This analysis shows that readers of Text A generated both necessary (or propositional) inferences to support their understanding and inferences of a more elaborative (or pragmatic nature). The former can be considered essential for ensuring local and global coherence. The latter are not strictly necessary for coherence but they do enrich an individual's mental representation in a variety of ways, perhaps to make it more memorable.

The range of pragmatic inferences presented above suggests that readers took the information provided by the text and, by integrating it with their existing knowledge and experience of the world, elaborated *beyond* the text to generate an enriched mental representation. It is likely, for example, that personal experience of the world prompted some readers to infer that it was *too dangerous* for the man to walk home and others to infer that it was *too far*. These more elaborative inferences cannot be confirmed or refuted by the information explicitly stated in the text and both interpretations are valid. Given the variable and idiosyncratic nature of inferences such as these, we should be cautious about incorporating them into a summary of the text. If such inferences are built into the summary, perhaps to enrich or flesh it out in some way, then it would be unwise to base test questions on this content. Urquhart and Weir's (1998) warning about testing comprehensions and not interpretations is relevant here (see further discussion on this point below).

Some inferences made by readers are clearly inconsistent with information that is explicitly stated in the short story. For example, the passenger did not sit in the back of the car, but *beside the driver*; and the driver did not take the spanner from the glove compartment, but from *the side-pocket of the car*. Both examples suggest that readers are capable of miscasting certain elements of the story in their mental representation after a single reading and

these examples are similar to what Steffensen and Joag-Dev (1984) described as distortions. In both cases, however, the reader has substituted a plausible alternative which makes sense in a local context if not in the context of the story as a whole, and which may have been provoked by personal "taxi ride" or "car interior" schemas. In a similar way, one reader recalled not a taxi stand but a *bus shelter*. One might argue that examples of incorrect inferences could offer useful insights into how and where comprehension can break down, and thus provide guidance for constructing appropriate test items in the summary completion task. However, as test developers we would need to be sure that such content points were central and salient to overall comprehension of the text rather than just peripheral details of little importance.

The analysis of inferences in the Text A recall transcripts enables consideration of the types of inference which could justifiably be included in an adequate summary of the text. It is relatively easy to justify the inclusion of necessary inferences (e.g. those relating to anaphoric reference, superordinate goals, theme and so on) since these are generally regarded as essential to coherent understanding. Elaborative inferences are more difficult to justify. Clearly, elaborative inferences activated by knowledge or experience of a highly personal nature, such as *it's a black taxi* or *that was the end of the driver's shift*, are less likely to be appropriate for inclusion than more generally shared inferences such as the idea that the man had been *out for the night* or that the initial setting for the story was a built-up area. It was suggested above that a certain amount of leakage between the *necessary* and *elaborative* categories is possible, due to constraining factors such as purpose for reading and background knowledge, and the test designer may need to exercise some judgement in determining whether inclusion of a more elaborative-type inference can be justified in a summary.

Graesser and Kreuz suggested a constructionist theory of inference generation proposing that inferences in narrative are not generated promiscuously, but rather in accordance with the reader's need to explain *why* characters perform actions and *why* involuntary events occur:

> . . . the reader wants to know (a) the goals, motives, or reasons that explain an agent's actions; and (b) the events, actions and states that cause or enable involuntary events. Comprehenders are particularly sensitive to actions and events in the world rather than to constancies, because such changes frequently convey interesting information: new information in the situation model, violations of normative standards, danger, obstacles to an agent's goals, goal conflicts between agents, methods of repairing planning failures, emotions that are triggered by goal failures and/or goal satisfaction, and other interesting occurrences that have adaptive significance to organisms. The comprehenders' attention is captured whenever there is a deviation from homeostatic balance in a physical, social or psychological system . . . (1994:153–154).

The study of inferences generated by readers of Text A suggests that they are generally consistent with the theory of narrative inferencing proposed above. Several of the inferences clearly seem motivated by readers' desire to arrive at a coherent understanding of why actions, events and states occur in the narrative (resonating with Zwaan and Rapp's (2006) discussion of the Event Indexing Model). This may serve as a guiding principle when making a judgement about which inferences to include in a summary task (see below) but given that such inferences are not mandated by the text they should not be the focus of testing points.

Constructing experimental summary variants of Text A

Following the multi-level analysis of readers' recalls of Text A in terms of their text-based and summarising propositions, together with a study of reader inferences in the transcripts, consideration was given to how to use these findings to construct various experimental summaries of the two texts. Three different methods for constructing the experimental summaries were explored:

i) Summaries comprising only text-based propositions
An initial attempt was made to construct and compare summaries of Text A using *only text-based propositions* in accordance with the levels of reader consensus and frequency of occurrence identified in the recall transcripts (see pages 106–112 in this chapter). This generated multiple summary variants (derived from Sets T(i)–T(vi)), containing differing degrees of detail and ranging in length from 27 to 209 words.

ii) Summaries comprising only summarising propositions
A similar approach was adopted using *only summarising propositions*, in line with the criterial categories used for the analysis (see pages 120–124 in this chapter). The summary variants derived from Sets S(i), S(ii), S(iii), S(iv), S(v/vi) and S(vii/viii) ranged in length from 45 to 89 words.

iii) Summaries combining text-based and summarising propositions
The third method for constructing possible summaries of Text A was to *combine corresponding sets of text-based and summarising propositions*, i.e. each summary variant comprised a set of summarising propositions and all text-based propositions associated with these. This method initially generated multiple drafts in line with the criterial categories analysis (see pages 106–124 in this chapter). However, it quickly became clear that some high-frequency text-based propositions that were candidates for inclusion in a summary were actually associated with summarising propositions that were *medium*-frequency (the middle 33%), rather than high-frequency (the top 33%). For this reason, two additional summary variants were created combining high-frequency (HF) text-based propositions and high/medium-frequency (HF/MF) summarising propositions recalled in ANY recall.

In all the experimentally constructed summaries described above, the order and wording of the text-based and summarising propositions was strictly maintained.

Rationale for developing multiple summary variants

The overall aim in developing multiple summary variants using the three methods described above (i.e. derived from text-based propositions, or summarising propositions, or a combination of the two) was to work towards determining what an accurate and coherent summary of Text A might look like which would match well to most readers' mental representations of the text. Furthermore, the optimum summary would need to be adequate in terms of length and degree of detail if it was to form the basis for developing a summary completion task for testing reading comprehension ability.

Several general principles were kept in mind at this point. First, the optimum summary would need to *accurately reflect the content of the original* text. Secondly, the summary would need to be *a coherent and standalone text in its own right*. Thirdly, the summary would ideally need to *include elements of the original text which most readers would find salien*t as well as *some elements which might cause weaker readers difficulty*, i.e. where comprehension might reasonably be expected to break down. An example of this might be a sufficiently developed referential chain across the summary. Finally, the summary would need to be *long enough to be capable of generating a large enough number of locally independent test items* to meet the requirements for test reliability.

In light of these factors, it was expected that the most promising summary of Text A for subsequent development into a summary completion task would be one containing several levels of detail since this would offer maximum scope for constructing a large enough set of test items (probably 30–40 items to meet standards of test reliability) that focused beyond the most basic level of detail and were capable of engaging readers' cognitive processing at a number of different levels. Table 6.17 shows the multiple draft summaries that were constructed according to methods (i), (ii) and (iii) above, with the reasons why each of these variants was either accepted or rejected for future development as a potential summary completion task.

The draft summary that was finally selected for further exploitation was the variant derived by combining summarising and text-based propositions recalled by 50% of the readers. It amounted to approximately 45% of the original text word count and is shown in bold italic in Table 6.17. At this point the selected draft was edited to ensure that it read smoothly and coherently. For the most part, this involved the appropriate use of pro-nominalisation, the introduction of appropriate conjunctive expressions and a small amount of reordering of propositions to ensure text cohesion and fluency.

The earlier discussion on inferencing highlighted the variability of

Table 6.17 Multiple summaries constructed using different criteria

Criteria used to construct draft summary	Length (words)	Accept/ reject	Reason for acceptance/ rejection of draft summary
T(i)	27	Reject	Single level of detail + too short
T(ii)	60	Reject	Single level of detail + too short
T(iii)	86	Reject	Single level of detail + too short
T(iv)	173	Reject	Single level of detail + too short
T(v)	206	Reject	Single level of detail
T(vi)	209	Reject	Single level of detail
S(i)	45	Reject	Single level of detail + too short
S(ii)	56	Reject	Single level of detail + too short
S(iii)	70	Reject	Single level of detail + too short
S(iv)	87	Reject	Single level of detail + too short
S(v/vi)	51	Reject	Single level of detail + too short
S(vii/viii)	89	Reject	Single level of detail + too short
S/T(i/ii) – ANY recall by 20/30 readers	100	Reject	Too short
S/T(iii/iv) – ANY recall by 15/30 readers	***235***	***Accept***	***Content and length about right***
S/T(v/vi) – HF propositions (all) in ANY recall	260	Reject	Too detailed + too long
S/T(vii/viii) – HF/MF summarising + HF text-based propositions in ANY recall	297	Reject	Too detailed + too long

inferences made by readers of Text A as well as the difficulty of determining which necessary and/or elaborative inferences should be included in any summary of the text. In the end, it was decided to integrate only those inferences which could be considered necessary to maintain local or global coherence. This meant confirming that logical or necessary inferences relating to referential chains, superordinate goals and theme of the narrative were represented within the summary (i.e. pronominalisation as part of a referential chain, and mention of a growing sense of panic on the part of the passenger) and thus could be exploited for item construction if so desired. No elaborative inferences were introduced relating to any of the types (i) to (ix) discussed on pages 130-134 for the reasons outlined above.

The final outcome, therefore, was a 235-word summary of Text A that combined summarising and associated text-based propositions, and that incorporated a limited number of authorised inferences, as shown in Appendix 7 on page 229.

Analysis of text-based propositions in recall transcripts of Text B (*Anorexia*)

The detailed and systematic approach described above for analysing the recall transcripts for Text A was implemented in exactly the same way to analyse

the Text B recall transcripts. Readers are reminded that while Text A was a 526-word narrative short story, Text B was a 389-word expository newspaper editorial, i.e. a text of a different genre. (The two texts can be found in Appendices 1 and 2 on pages 220–221.)

Identifying sets of text-based propositions recalled by readers

A preliminary analysis of Text B recall transcripts determined the frequency with which individual text-based propositions were recalled by readers. (See pages 91–95 in Chapter 5 for the definition of a text-based proposition, together with a list of the 61 text-based propositions that were generated from Text B.) Further analysis set out to define sets of individual text-based propositions according to different selection criteria. Several general principles were kept in mind at this point. First, the optimum summary would need to *accurately reflect the content of the original* text. Secondly, the summary would need to be *a coherent and standalone text in its own right*. Thirdly, the summary would ideally need to *include elements of the original text which most readers would find salient* as well as *some elements which might cause weaker readers difficulty*, i.e. where comprehension might reasonably be expected to break down. An example of this might be a sufficiently developed referential chain across the summary. Finally, the summary would need to be *long enough to be capable of generating a large enough number of locally independent test items* to meet the requirements for test reliability.

The selection criteria used for defining sets of text-based propositions were identical to those used for Text A, i.e.:

> a) the number of readers (i.e. proportion of the total group) who recalled each individual text-based proposition, and
> b) whether the text-based proposition was recalled only in the free recall phase, or across both recall phases (i.e. free and prompted).

It was anticipated that exploratory analyses along these lines would reveal the text content that was held in memory by most readers as part of their mental representation of the newspaper editorial, including how far some of this content might be considered *core* (as revealed during free recall) or more *peripheral* (revealed only as a result of prompting).

Comparing content across the sets of text-based (TP) propositions

Six sets of text-based (TP) propositions were identified across the Text B recall transcripts in accordance with the following criteria:

i) propositions included by 20 out of 30 readers in FREE recall
ii) propositions included by 20 out of 30 readers in ANY recall

iii) propositions included by 15 out of 30 readers in FREE recall
iv) propositions included by 15 out of 30 readers in ANY recall
v) the top third most frequently recalled propositions in FREE recall (i.e. high-frequency – top 33%)
vi) the top third most frequently recalled propositions in ANY recall (i.e. high-frequency – top 33%).

As with Text A, criteria (i) and (ii) established a relatively demanding threshold of 66% consensus among readers, while criteria (iii) and (iv) set a less demanding threshold of 50% agreement, in line with previous research on consensus levels when recalling main and subsidiary ideas. Analysis of the top third most frequently recalled text-based propositions was included (criteria (v) and (vi)) to see whether this might shed further light on the nature of core, peripheral and superfluous material in readers' mental representation of text. The content of each set is shown in Tables 6.18 to 6.23 in some detail and differences between the sets are highlighted. The codes and wording for the text-based propositions come from Table 5.2 on pages 94–95.

Table 6.18 Text-based propositions included by 20 out of 30 readers in FREE recall – Set T(i)

None

Not one of the 61 text-based propositions was offered in FREE recall by at least 20 out of 30 subjects. This was clearly a high threshold and, when compared with the short story (see page 107), it may indicate the level of reading comprehension challenge posed by this type of text (see the earlier discussion on text difficulty in Chapter 4, as well as some further reflection on the comparative difficulty of Texts A and B in Chapter 8).

Table 6.19 Text-based propositions included by 20 out of 30 readers in ANY recall – Set T(ii)

TP04	Samantha Kendall discharged herself from hospital
TP09	today anorexia is recognised as a medical condition
TP27	anorexia is a severe psychiatric disorder
TP29	the illness allows a patient to look in the mirror at their own emaciated body
TP31	the illness allows a patient to see someone
TP32	someone obese staring back
TP53	the treatment often involves watching patients

The prompting phase clearly proved quite stimulating and Set T(ii) includes seven (11%) of the 61 text-based propositions offered in ANY recall by at least

20 out of the 30 subjects. These include references to *anorexia as a medical condition and psychiatric disorder* as well as a cluster of propositions relating to *how anorexia sufferers tend to see themselves* (TP29, TP31 and TP32). Mention is also made of the *type of treatment* used, and of *Samantha Kendall*. Of these seven propositions, six refer to *what anorexia is*, *what it does* and *how it is treated*. In other words, they provide the basic conceptual framework necessary to any discussion of the rights and wrongs of different treatments for the condition.

Table 6.20 Text-based propositions included by 15 out of 30 readers in FREE recall – Set T(iii)

None

Once again, none of the 61 text-based propositions was offered in FREE recall by at least 15 out of 30 subjects, in marked contrast to the results for the short story (see pages 117–118).

Table 6.21 Text-based propositions included by 15 out of 30 readers in ANY recall – Set T(iv)

TP03	Samantha Kendall is an anorexia nervosa sufferer
TP04	Samantha Kendall discharged herself from hospital
TP08	ten years ago anorexia was still dismissed as nothing more than slimming-gone-too-far
TP09	today anorexia is recognised as a medical condition
TP13	researchers have suggested two psychiatric explanations behind the onset of anorexia
TP14	one explanation is that
TP18	the patient is trying to avoid leaving childhood
TP20	choosing what to eat is often an attempt to exert control by people
TP27	anorexia is a severe psychiatric disorder
TP29	the illness allows a patient to look in the mirror at their own emaciated body
TP31	the illness allows a patient to see someone
TP32	someone obese staring back
TP38	sufferers can be held in hospital for treatment against their will
TP52	the treatment often involves leaving patients without their clothes
TP53	the treatment often involves watching patients
TP54	patients eat
TP55	patients go to the lavatory

As with Text A, prompting proved effective and Set T(iv) now contains 17 (28%) of the 61 text-based propositions and adds a substantial amount of detail to Set T(ii). The *contrast between modern and previous attitudes* towards anorexia is referred to, together with the *current psychiatric explanations* offered. The *possibility of forced treatment* is mentioned as well as more detail

about *the nature of such treatment*. *Samantha Kendall's involvement* is also made clear – she is an anorexia sufferer. In this set, several clusters of propositions for related information can be identified: 03/04, 08/09, 13/14, 18/20, 27/29/31/32, 52/53/54/55. This replicates a finding in the comparable study with Text A (see page 109).

Sets T(i) to T(iv) (as illustrated in Tables 6.18 to 6.21) were identified by analysing the consensus of text-based propositions recalled by 66% and 50% of the readers respectively. The text-based propositions present in the recall transcripts were also categorised and analysed by frequency (i.e. high/mid/low-frequency) to see what light this might shed on the nature of core, peripheral and superfluous content in readers' mental representations of the newspaper editorial. As with Text A, the top third most frequently recalled text-based propositions (i.e. the top 33%) were analysed, according to whether these occurred during free recall or as a result of probing (see Tables 6.22 and 6.23).

Table 6.22 High-frequency text-based propositions occurring in FREE recall – Set T(v)

TP03	Samantha Kendall is an anorexia nervosa sufferer
TP04	Samantha Kendall discharged herself from hospital
TP09	today anorexia is recognised as a medical condition
TP11	the degree to which has become a topic of debate
TP12	whether treatment should be carried out without a patient's consent
TP13	researchers have suggested two psychiatric explanations behind the onset of anorexia
TP14	one explanation is that
TP18	the patient is trying to avoid leaving childhood
TP20	choosing what to eat is often an attempt to exert control by people
TP24	the syndrome remains imperfectly understood
TP27	anorexia is a severe psychiatric disorder
TP29	the illness allows a patient to look in the mirror at their own emaciated body
TP31	the illness allows a patient to see someone
TP32	someone obese staring back
TP37	the 1983 Mental Health Act provides for sufferers from severe psychiatric disorders
TP38	sufferers can be held in hospital for treatment against their will
TP42	one in ten anorexia sufferers dies
TP44	doctors use their powers under the law
TP49	there is clearly work to be done
TP50	done in making the treatment of extreme anorexia more humane

Set T(v) contains 20 (32%) of the 61 text-based propositions and adds references relating to: the *debate over whether treatment should be carried out with or without a patient's consent*; *our limited understanding of the anorexia syndrome*; *the provision of the 1983 Mental Health Act* and *its use by doctors*; *the death rate among anorexia sufferers*; and *the need for more humane approaches to treatment*.

Table 6.23 High-frequency text-based propositions occurring in ANY recall – Set T(vi)*

TP03	
TP04	
TP08	ten years ago anorexia was still dismissed as nothing more than slimming-gone-too-far
TP09	
TP11	
TP12	
TP13	
TP14	
TP17	the patient is trying to retreat into childhood
TP18	
TP20	
TP22	people feel their lives are constrained in other ways
TP24	
TP27	
TP29	
TP31	
TP32	
TP37	
TP38	
TP42	
TP44	
TP49	
TP50	
TP52	the treatment often involves leaving patients without their clothes
TP53	the treatment often involves watching patients
TP54	patients eat
TP55	patients go to the lavatory

* *text-based proposition codes from T(v) are included here but without their wording in order to highlight the information that is being added as a result of probing*

Set T(vi) adds seven propositions which did not appear in Set T(v). These include references to: *the attitude towards anorexia ten years ago*, *more detail of the explanations given for anorexia*, and *more details of the treatment used*. It is likely that the relative increase in frequency reflected in the seven 'new' text-based propositions contained in Set T(vi) (08, 17, 22, 52, 53, 54, 55) was a direct result of probe questions 2, 3 and 6, i.e.:

> *2. Can you recall anything in the editorial about how attitudes to anorexia have changed over the years?*
> *3. Can you recall any explanations for anorexia given in the editorial?*
> *6. Can you recall any methods of treatment for extreme anorexia which were mentioned?*

When comparing the sets of text-based propositions shown in Tables 6.18 to 6.23, we can see that, just as it did for the short story, the setting of different criterial parameters, i.e. (i) to (vi), generates very different summary versions

of the newspaper editorial as far as its content and structure are concerned. In terms of reader consensus, it seems 66% agreement may be too high an expectation, while 50% agreement at least begins to generate a reasonable summary of the text (if only after some prompting), accounting for 28% of the original content. The implications of the effect of prompting on recall will be considered further below.

Analysing low-frequency text-based propositions

While certain propositions from the text seem to figure very strongly in readers' mental representations as captured in sets T(i)–(iv), other propositions figured very little for some reason. For example, although over 15 (50%) readers recalled that *sufferers can be held in hospital for treatment against their will*, even after prompting only one reader (3%) recalled that *a great many resources have been devoted to the study of anorexia*. The low-frequency propositions (i.e. the bottom 33%) occurring in free and prompted mode are shown in Table 6.24 and discussed more fully below.

Table 6.24 Low-frequency text-based propositions in ANY recall

TP01	the case of Samantha Kendall has highlighted a confusion
TP02	there is a confusion in public thinking
TP05	doctors feared for her life
TP06	anorexia nervosa is a disturbing disease
TP07	anorexia nervosa is a perplexing disease
TP15	the patient is faced with an unacceptably stressful adult life
TP16	the patient is faced with an unacceptably difficult adult life
TP23	the truth is that
TP25	a great many resources have been devoted to the study of anorexia
TP26	it is beyond doubt however that
TP28	there is no other way to describe an illness that
TP33	severe sufferers often deny that
TP34	severe sufferers are trying to kill themselves
TP35	the diet severe sufferers are pursuing
TP36	the diet is all too likely to make death inevitable
TP41	sufferers will do harm to others
TP56	there are shortcomings of the available treatment
TP57	should not obscure the fact that
TP58	there is an alternative to treatment
TP59	which can sometimes be death

Few readers made an explicit connection in their oral recalls between the case of Samantha Kendall and the public debate relating to varying perceptions of the disease (TP01 and TP02). This is somewhat curious given that these two propositions constitute the opening of Text B and the opening section is generally considered to be one of the most memorable sections of a text.

Other clusters of poorly recalled propositions are TP33/34/35/36 and TP56/57/58/59. In the first of these, the cluster relates to a complete sentence at the end of a paragraph and it concerns the fact that severe sufferers often deny trying to kill themselves although the diet they are pursuing is all too likely to make death inevitable. It is possible that the details given in this cluster did not appear important or were regarded as only a subsidiary example of how self-deluding anorexia sufferers can be. However, it may be worth noting that this sentence follows on immediately from a graphic description of how anorexia sufferers can look at themselves in a mirror and yet see someone obese staring back. It may be that the memorability of the graphic description had an inhibiting effect upon memory and recall for the following sentence.

Another inhibiting effect on memory and recall for this set of propositions may have been the presence of two pseudo-technical words in the preceding part of the paragraph – 'emaciated' and 'obese'. An analysis of these two words using the VocabProfile tool in Compleat Lexical Tutor (Cobb 2006) shows them to be among the least frequent in the newspaper editorial when compared against the lexical content of the British National Corpus (BNC): *emaciated* falls into the BNC/K12 band, while *obese* falls into the BNC/K8 band. Given that both lexical items are thus relatively low-frequency in everyday English and occur close to one another in the text (only six words apart), it may be that readers were distracted by these lexical items, just as they had been by similarly infrequent words in the Text A short story (see page 126). Once again, it seems reasonable to conclude that a reader's mental representation of text will be partly influenced by the extent to which they encounter and have to process unusual or unfamiliar words, phrases or concepts in a text. It could be argued that careful consideration should be given to whether such words are included in the summary; if they are, it would be unwise to make them a focus for testing.

The second cluster – TP56/57/58/59 – occurs towards the end of the final paragraph of the text and concerns the need to recognise that, despite its shortcomings, available treatment for severe anorexia does at least offer an alternative to death. Once again, this cluster occurs immediately after a graphic description of the way in which sufferers may be treated and this too may have exerted some sort of inhibiting effect.

The least frequently recalled text-based propositions include those which constitute rhetorical statements on the part of the author or possess an epistemic status (e.g. TP23 – *the truth is that*, TP26 – *it is beyond doubt however that*, TP28 – *there is no other way to describe an illness that*). Interestingly, it seems that readers tend not to recall this type of proposition as readily as those which carry more obviously factual information relating to the speaker's topic.

There remain only two other text-based propositions which carry a low-frequency status – TP05 – *doctors feared for her life* and TP25 – *a great many resources have been devoted to the study of anorexia*. This may have been because these propositions carried information considered of relatively little salience. They may also have been more difficult to integrate into the mental representation as a whole, perhaps because they were less well connected to other units of information.

One technical reason for the low-frequency status of some text-based propositions in Table 6.24 could be that frequency of recall has been artificially depressed by the way in which the coding scheme operated. For example, in the original text there were some cases where a predicate carried more than one argument (e.g. *this disturbing and perplexing disease*; *a treatable, medical condition*; *faced with an unacceptably stressful or difficult adult life; at her own emaciated, starved body*). In such cases, multiple arguments were normally separated out into different text-based propositions (i.e. TP06 – *anorexia nervosa is a disturbing disease*, and TP07 – *anorexia nervosa is a perplexing disease*; TP15 – *the patient is faced with an unacceptably stressful adult life*, and TP16 – *the patient is faced with an unacceptably difficult adult life*). As a result of this, it is possible that frequency of recall has become artificially split between the two propositions. One way to resolve this might be to combine the frequencies for each part and to adjust their ranking accordingly. In the cases of TP06 and TP07 this would still make no difference to their position as low-frequency. TP15 and TP16 would, however, move a little way up the ranking but would still rate as only medium-frequency. Only the text proposition pair TP40 and TP41 would move up to become high-frequency.

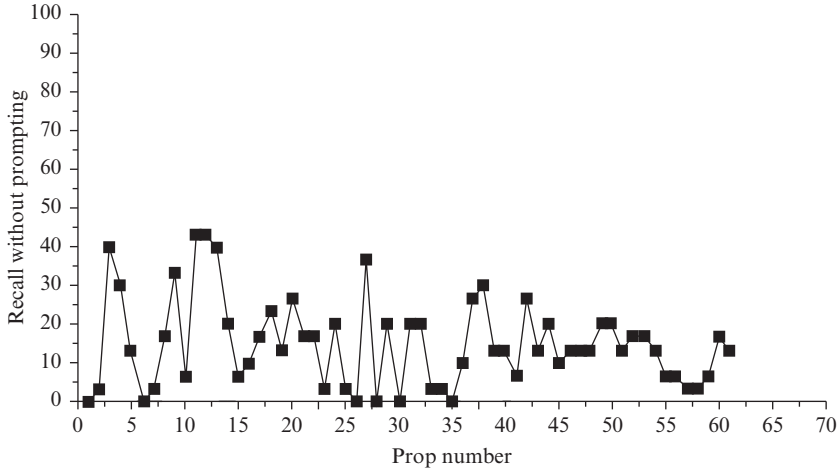
In terms of the psychological theory of summarisation, one might expect readers during recall to use a superordination rule in relation to multiple arguments. However, in cases where an appropriate superordinate term is not immediately obvious, it may be that readers use one of the two terms, perhaps the first or the most general, to stand for both.

Effect of prompting on recall of text-based propositions

As was done for Text A, the percentage of subjects recalling each text-based proposition was plotted. Figure 6.4 shows ease of remembering for the 61 text-based propositions without prompting.

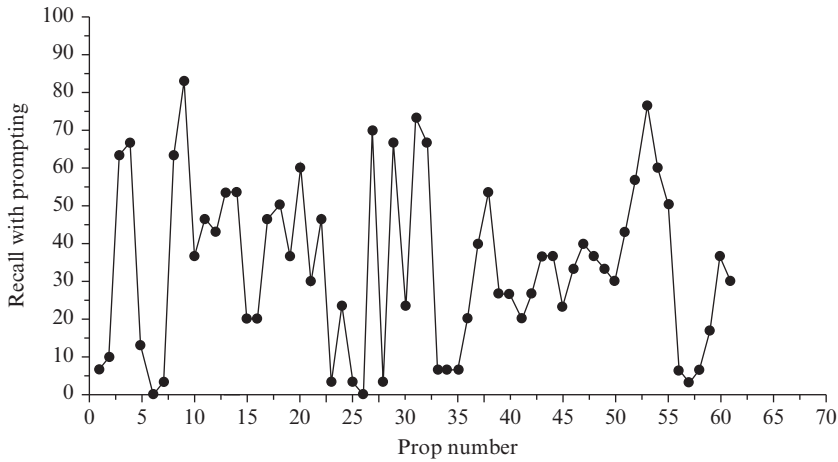
The text-based propositions most easily remembered from Text B are clearly visible, apparently clustering in four particular locations within the text. The first of these occurs close to the start of the text and focuses on the individual case which sparked the debate: TP03 – *Samantha Kendall is an anorexia nervosa sufferer* and TP 04 – *Samantha Kendall discharged herself from hospital*. The second cluster relates to the debate surrounding treatment of patients without their consent (TP11 and TP12). The third peak occurs

Figure 6.4 Ease of remembering before probing (Text B)



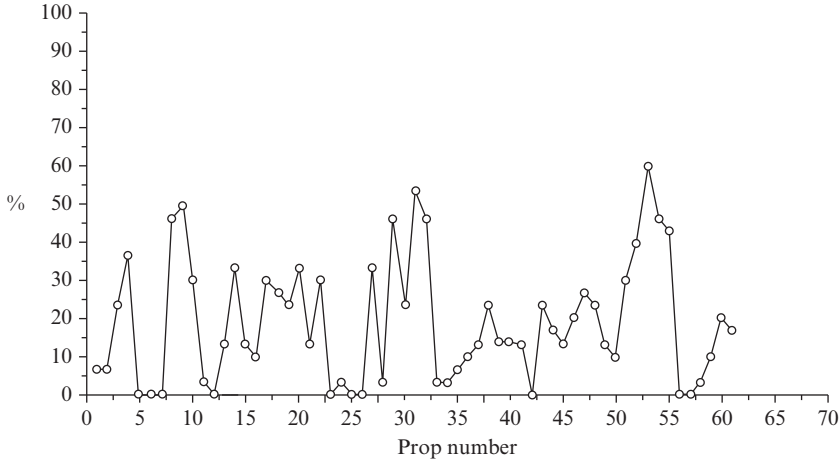
almost halfway through the text and reflects the general statement in TP27 that anorexia is a severe psychiatric disorder. A fourth cluster can be discerned two-thirds of the way through the text (TP37 and TP38), relating to mention of the Mental Health Act which allows for psychiatric sufferers to be forcibly treated in hospital. Also noticeable is a gentle but relatively steady decline in ease of remembering from beginning to end, with only a slight upturn at the very end of the text. Figure 6.5 shows the ease with which text-based propositions were remembered after probing.

Figure 6.5 Ease of remembering after probing (Text B)



While the overall shape of the graph remains largely the same, it appears that prompting increases ease of remembering by an average of almost 20% (twice the value for Text A), and by more than 50% for some propositions. It is also clear that there were instances where it made no difference at all. For this reason, closer attention was paid to the extent to which prompting did or did not have an effect on ease of remembering. Figure 6.6 shows the increase in percentage terms in remembering of individual text-based propositions after prompting.

Figure 6.6 Percentage increase in ease of remembering after probing (Text B)



Prompting produced an increase of at least 20% in ease of remembering for 27 (44%) of the 61 text-based propositions and these propositions are listed in Table 6.25.

It is highly likely that the substantial increase in ease of remembering for these propositions resulted from asking the probe questions during the prompted phase of the recall exercise. For example, improved recall of TP03/04 may be a direct response to probe question 1 – *Can you recall anything in the editorial about a person called Samantha Kendall?* and other single or clustered text propositions could be linked to the probe questions as follows:

Probe question 2: TP08/09/10
[Can you recall anything in the editorial about how attitudes to anorexia have changed over the years?]
Probe question 3: TP14 and TP17/18/19/20/22
[Can you recall any explanations for anorexia given in the editorial?]
Probe question 4: TP27/29/30/31/32

Table 6.25 Propositions more easily remembered after prompting (Text B)

TP03	Samantha Kendall is an anorexia nervosa sufferer
TP04	Samantha Kendall discharged herself from hospital
TP08	ten years ago anorexia was still dismissed as nothing more than slimming-gone-too-far
TP09	today anorexia is recognised as a medical condition
TP10	a medical condition which can be treated
TP14	one explanation is that
TP17	the patient is trying to retreat into childhood
TP18	the patient is trying to avoid leaving childhood
TP19	another explanation is that
TP20	choosing what to eat is often an attempt to exert control by people
TP22	people feel their lives are too constrained in other ways
TP27	anorexia is a severe psychiatric disorder
TP29	the illness allows a patient to look in the mirror at their own emaciated body
TP30	the illness allows a patient to look in the mirror at their own starved body
TP31	the illness allows a patient to see someone
TP32	someone obese staring back
TP38	sufferers can be held in hospital for treatment against their will
TP43	doctors are sometimes reluctant
TP46	treatment by compulsion is self-defeating
TP47	force-fed victims of anorexia often return to starvation diets when
TP48	they get home
TP51	the treatment often involves leaving patients in isolation
TP52	the treatment often involves leaving patients without their clothes
TP53	the treatment often involves watching patients
TP54	patients eat
TP55	patients go to the lavatory
TP60	if doctors made more use of the powers available to them

[Can you recall any effects of the illness on the patient which were described?]
Probe question 5: TP38
[Can you recall anything mentioned about the current legal position on treating anorexia sufferers?]
Probe question 6: TP51/52/53/54/55
[Can you recall any methods of treatment for extreme anorexia which were mentioned?]
Probe question 7: TP60
[Can you recall the writer's point of view on treating extreme anorexia sufferers?]

A study was also made of those text-based propositions where prompting seemed to have no effect. These are listed in Table 6.26.

It is notable that eight out of these 10 propositions are located in the low-frequency range (see Table 6.24 on page 144). Unlike Text A, Text B appeared to contain no text-based propositions which were already so memorable they remained unaffected by prompting. Only TP12 and TP42 gained

Table 6.26 Propositions for which prompting seemed to have no effect

TP05	doctors feared for her life
TP06	anorexia nervosa is a disturbing disease
TP07	anorexia nervosa is a perplexing disease
TP12	whether treatment should be carried out without a patient's consent
TP23	the truth is that
TP25	a great many resources have been devoted to the study of anorexia
TP26	it is beyond doubt however that
TP42	one in ten anorexia sufferers dies
TP56	there are shortcomings of the available treatment
TP57	should not obscure the fact that

high-frequency status in FREE recall and medium-frequency status in ANY recall. There may be several reasons why no increase in ease of remembering after probing was observed for some propositions:

> i) some propositions may not have become easier to recall because the probe questions asked were not sufficiently sensitive (e.g. TP05, TP25, TP42)
> ii) some propositions may have been so hard to recall that even sensitive probe questions did not improve their chances (e.g. TP23, TP26, TP56, TP57); these include rhetorical statements by the article's author
> iii) as discussed above, it is possible that the coding system may have resulted in some propositions appearing artificially difficult to remember (e.g. TP06, TP07, TP12).

Insights into readers' structuring of information from Text B

The analysis of readers' recall protocols for Text A (*Journey*) earlier in this chapter appeared to provide evidence of a triggering effect, in which recall of one proposition from the text seemed to provoke the recall of a second or third (see pages 117–120) and it was suggested that this might reflect the way in which readers have structured the information units or propositions in the text, by clustering them around a particular node or by linking them simultaneously to other nodes.

A similar effect in readers' recalls of Text B (*Anorexia*) was much harder to discern. One reason for this may be that Text B propositions tend to be long and rather dense in nature and, as a result, they may have been more difficult to "network" than the shorter and simpler descriptive and narrative details of Text A. It is also possible that because certain elements in Text A reappear at different points throughout the narrative (e.g. references to *money*/*wallet*, *being robbed*), these need to be reattached to the thread of the discourse and thus are prone to multiple activation. For Text B, however, the content of each paragraph tends to be fairly unified and discrete. Spreading activation during semantic processing at the local level

(i.e. within a paragraph) must undoubtedly occur and probably shows up in recall as sequentially linked propositions. However, the potential for spreading activation to occur across paragraphs in Text B may be considerably less than it is in Text A.

There were some occasions when a reader's existing knowledge structures clearly intruded into the process of comprehension and probably influenced their construction of a coherent mental representation of the text. For example, several readers demonstrated an existing knowledge of anorexia, and also of a similar eating disorder – 'bulimia':

> *they're the people who starve themselves really + they don't eat enough* (#001)
> *the condition where + mainly girls + they sort of starve themselves to make themselves ill* (#002)
> *bulimia is similar to this where they force + where they eat and then binge and then make themselves sick + but this is basically where they just don't eat at all* (#016)

Some readers even appeared to be familiar with effects of the disease that were not mentioned explicitly in the text:

> *apart from obviously losing a tremendous amount of weight you also become ill* (#001)
> *obvious loss of weight* (#008)
> *it reduces their metabolism so it stops their body working properly + so they begin to have like even more psychiatric problems + er + fitting into life* (#011)
> *depression + feeling lonely and cut off* (#021)

Some readers suggested additional explanations for anorexia to those mentioned explicitly in the text:

> *some kind of wanting to rid themselves of something or other + a kind of cleansing operation where their body couldn't be dirty if it didn't have anything left in it therefore they didn't eat* (#004)
> *I think it could be linked to advertising because of the prominence of superthin waifs everywhere* (#005)
> *people trying to look better + being more like stars and models* (#021)
> *it's peer pressure* (#021)

When recalling methods of treatment for anorexia, it was striking how many readers seemed to draw very easily upon an existing knowledge of the most commonly known treatment for anorexia – *force-feeding* – even though there is only one, very brief explicit reference to this in the text (*force-fed victims of anorexia often return to starvation diets when they get home*):

like forcefeeding you + it's really by forcefeeding people into eating (#001)
and forcefeeding them as well (#002)
another one was forcefeeding (#007)
force-feeding + dripfeeding (#009, #027)
force-feeding (#011, #012, #018, #020)
forcefeed + to make the people eat the food so that they don't kill themselves (#016)

One reader demonstrated knowledge of other approaches to treating anorexia:

a lot of the treatment's not actually feeding + it's talking about the image about themselves + and a lot of counselling and support (#027)

At least two readers were clearly aware during the recall itself of the difficulty involved in maintaining a distinction between their relevant background knowledge and the information provided by the text:

they say that + um + they + I think they can be force-fed but I don't think it said that in the article (#004)
I don't want to say anything about the article because I can remember the actual case and how she was allowed to dismiss herself + so I don't want to read anything into the article because of having seen it in the news (#006)

The examples given above demonstrate that readers' background or existing knowledge structures were an integral part of their construction of a coherent mental representation of Text B, just as they were for the short story in Text A discussed earlier in this chapter.

Some conclusions from analysing text-based propositions in Text B recalls

The analysis of text-based propositions occurring in the 30 recall protocols for Text B once again confirmed that certain propositions tend to be recalled more easily by readers than others, and that while some propositions in a text will consistently be recalled by a group of readers when reading the same text for the same purpose, other propositions will not. Though the degree of consensus among readers over core information in the expository Text B seems to have been somewhat lower than for the narrative Text A (to judge from the smaller sets of text-based propositions identified), the findings of these oral recall studies appear to suggest that readers can reach a significant level of agreement on certain key ideas in a text. This finding appears to be in line with earlier findings in this area (e.g. Sarig 1989). It also confirms that identifying the propositions that should be included in an adequate summary of a text could justifiably take account of the relative frequency of occurrence of text-based propositions in readers' recall protocols of the text in question.

Analysis of summarising propositions in recall transcripts of Text B (*Anorexia*)

Identifying sets of summarising propositions recalled by readers

A second-level analysis of the recalls for Text B was focused at a higher level of the text structure to explore whether some summarising propositions were more frequently offered by readers than others. (See pages 95–100 in Chapter 5 for a definition of summarising proposition together with a list of 11 possible summarising propositions created for Text B in Table 5.4 on page 100.) Once again, it was decided to investigate the extent to which different sets of summarising propositions might assist in constructing a summary of Text B that could form the basis for developing a summary completion task. It was anticipated that this analysis might help to reveal how readers employed macro-propositions (rather than micro-propositions) when constructing their mental representation of the newspaper editorial. The selection criteria used for defining sets of summarising propositions were identical to those used for Text A, i.e.:

i) propositions offered by 20 out of 30 readers in FREE recall
ii) propositions offered by 20 out of 30 readers in ANY recall
iii) propositions offered by 15 out of 30 readers in FREE recall
iv) propositions offered by 15 out of 30 readers in ANY recall
v) the top third most frequently occurring propositions in FREE recall (i.e. high-frequency – top 33%)
vi) the top third most frequently occurring propositions in ANY recall (i.e. high-frequency – top 33%)
vii) the top two-thirds most frequently occurring propositions in FREE recall (i.e. high- and medium-frequency – top 66%)
viii) the top two-thirds most frequently occurring propositions in ANY recall (i.e. high- and medium-frequency – top 66%)

As with the text-based propositional analysis, criteria (i) and (ii) established a relatively demanding threshold of 66% consensus among readers, while criteria (iii) and (iv) set a less demanding threshold of 50% agreement. Analysis of high-frequency (top 33%) summarising propositions was included (criteria (v) and (vi)). Given the relatively small number of summarising propositions available (only 11), and drawing on prior experience with Text A (see page 121), high- and medium-frequency criteria (vii) and (viii) were added just in case these proved to be useful.

Comparing content of sets of summarising propositions

Using the criteria outlined above, multiple sets of summarising propositions were initially identified. Contents of all sets are shown in Tables 6.27 to 6.33. (The codes and wording of the summarising propositions come from Table 5.4 on page 100.)

Table 6.27 Summarising propositions offered in FREE recall by 20 out of 30 readers – Set S(i)

None

Table 6.28 Summarising propositions offered in ANY recall by 20 out of 30 readers – Set S(ii)

SP02	there have been changes in recent years in the way anorexia is regarded
SP06	anorexia is certainly a severe psychiatric disorder

Set S(ii) includes two (18%) of the 11 possible summarising propositions focusing on two specific aspects of the text content: the fact that *attitudes to anorexia have changed over the years* and that *anorexia is undoubtedly a severe psychiatric disorder*. While these two propositions provide key themes from the text, they cannot really be seen as constituting a coherent summary of it.

Table 6.29 Summarising propositions in FREE recall by 15 out of 30 readers – Set S(iii)

None

Table 6.30 Summarising propositions in ANY recall by 15 out of 30 readers – Set S(iv)

SP02	there have been changes in recent years in the way anorexia is regarded
SP03	treating anorexia patients without their consent is a matter for debate
SP06	anorexia is certainly a severe psychiatric disorder
SP08	the law makes it possible to force treatment on anorexia sufferers

Set S(iv) extends the number of summarising propositions to four (36%) out of the 11 and adds two references which are in some sense closely linked to one another: that *anorexia sufferers can by law be forcibly treated* but that *such treatment is a subject for debate*. Set S(iv) offers a more comprehensive overview of the Text B content but fails to include some of the important elements, particularly the Samantha Kendall case study.

Sets S(ii) and S(iv) (as shown in Tables 6.28 and 6.30) were selected by analysing the consensual use of summarising propositions by 66% and 50% of the readers respectively. As with Text A, it was once again decided to analyse the summarising propositions by frequency (i.e. high-, medium- and low-frequency) to see whether this might shed further light on the nature of core, peripheral and superfluous content in readers' mental representations of

the text. High-frequency (the top 33% most frequently occurring) and high/medium-frequency (the top 66% most frequently occurring) summarising propositions were analysed according to whether these occurred during free recall or as a result of probing (see Tables 6.31 to 6.32).

Table 6.31 High-frequency summarising propositions offered in FREE recall – Set S(v)

SP03	treating anorexia patients without their consent is a matter for debate
SP04	two psychiatric explanations have been offered for anorexia
SP06	anorexia is certainly a severe psychiatric disorder
SP08	the law makes it possible to force treatment on anorexia sufferers

Set S(v) (Table 6.31) also contains four (36%) of the 11 summarising propositions. It shares three of these with Set S(iv) and adds a reference to the *two psychiatric explanations which have been offered for anorexia*. Set (vi) (in Table 6.32) adds mention of *changes in recent years in the way anorexia is regarded*, a summarising proposition which appeared in both Sets S(ii) and S(iv).

Table 6.32 High-frequency summarising propositions offered in ANY recall – Set S(vi)*

SP02	there have been changes in recent years in the way anorexia is regarded
SP03	
SP04	
SP06	
SP08	

** summarising proposition codes from S(v) are included here but not their wording in order to highlight the information that is being added*

Sets S(vii) and S(viii) (shown in Table 6.33) share both HF and MF summarising propositions, extending the total number to eight (73%) out of 11.

In these sets the debate is contextualised by referring to *a recent news case*, and they also incorporate references to *the result of severe anorexia* and the reason *why doctors are reluctant to forcibly treat sufferers*.

Analysing low-frequency summarising propositions

The analysis of the frequency of occurrence of summarising propositions in readers' recalls of Text B was undertaken to gain further insight into readers' mental representations of the text. As with the earlier analysis of text-based propositions, it is the low-frequency summarising propositions (shown in Tables 6.34 and 6.35) that are of particular interest.

Table 6.33 High/medium-frequency summarising propositions offered in FREE/ANY recall – Sets S(vii)/(viii)

SP01	a recent news case illustrates confused thinking about anorexia
SP02	there have been changes in recent years in the way anorexia is regarded
SP03	treating anorexia patients without their consent is a matter for debate
SP04	two psychiatric explanations have been offered for anorexia
SP06	anorexia is certainly a severe psychiatric disorder
SP07	severe anorexia is likely to result in death
SP08	the law makes it possible to force treatment on anorexia sufferers
SP09	doctors are reluctant to use legal force because it is ineffective in the long run

Table 6.34 Low-frequency summarising propositions in FREE recall

SP02	there have been changes in recent years in the way anorexia is regarded
SP05	extensive research has not provided us with a full understanding
SP10	there is a need for more humane approaches to treating extreme anorexia
SP11	treatment by force is justified if the alternative is death

Table 6.35 Low-frequency summarising propositions in ANY recall

SP01	a recent new case illustrates confused thinking about anorexia
SP05	extensive research has not provided us with a full understanding
SP10	there is a need for more humane approaches to treating extreme anorexia
SP11	treatment by force is justified if the alternative is death

An analysis of text-based propositions to which these low-frequency summarising propositions relate may help to explain why they should have been offered so infrequently by readers in their recalls.

SP01 covers seven text-based propositions, of which five were categorised in the earlier analysis as low-frequency (TP1, TP2, TP5, TP6 and TP7) and only two as high-frequency (TP3 and TP4). These last two propositions may have been high-frequency because they were in direct response to the opening probe question about the person of Samantha Kendall.

SP02 relates to one medium-frequency (TP10) and two high-frequency text-based propositions (TP8 and TP9). Once again the high frequency of TP8 and TP9 propositions was probably a function of one of the probe questions. Both SP01 and SP02 have been included in some earlier sets of summarising propositions and are therefore not the lowest frequency.

SP05 covers two low-frequency (TP23 and TP25) and one MF (TP24) text-based propositions, while SP11 covers four low-frequency (TP56–59) and two MF propositions (TP60–61).

A general conclusion from this analysis could be that the low-frequency summarising propositions tend to relate to parts of the text which readers found less memorable overall. Not only is the text-based propositional

information not integrated into a higher-level representation (i.e. summarising proposition) but it may not even register as being memorable at the basic propositional level.

SP10, however, is very different in that it relates to three medium-frequency text-based propositions (TP49–51) and four high-frequency propositions (TP52–55). One explanation for this might be that all seven propositions, and especially the last four, convey information which is somewhat shocking (i.e. the graphic description of some treatments) and which is therefore likely to be more memorable. This may also explain why the same information is not readily integrated into a higher-level summarising proposition but instead remains encoded at the level of the text. It is also possible that the high-frequency status of TP52–55 is a function of one of the probe questions.

Some conclusions from analysing summarising propositions in Text B recalls

The analysis of summarising propositions occurring in the 30 recall transcripts for Text B appears to confirm, as it did for Text A, that readers constructed and recalled certain summarising propositions more readily than others. It also suggests that while some of these were consistently recalled by a group of readers when reading the same text for the same purpose, other summarising propositions were not. It seems reasonable to suggest, therefore, that the identification of summarising propositions for inclusion in a summary of Text B could take account of the relative frequency of occurrence of such propositions in readers' recall of the text.

Analysing additional propositions occurring in recall transcripts of Text B

The presence of additional propositions in readers' recalls of Text B

In the same way as with Text A (*Journey*), readers' recall transcripts for Text B (*Anorexia*) included more than just material which could be matched to corresponding text-based or summarising propositions. Examples of some additional propositions appearing in the transcripts are shown in the underlined clauses below:

> *there's a certain girl + um +* <u>*that died recently from anorexia*</u> (#010)
> <u>*they don't want to become women*</u> *+ they want to revert back to childhood* <u>*and*</u> <u>*stay a children's size*</u> (#026)
> *they watch people eat (. . .)* <u>*to make sure they didn't make themselves sick,*</u> <u>*to make sure they're eating*</u> (#016)

This type of additional proposition is treated here as a form of inference generated by the reader. The 30 recall transcripts for Text B were analysed for evidence of readers' inferencing in the form of additional propositions and to determine the relative frequency with which such inferencing occurred.

It is difficult to determine whether the Text B inferences currently under discussion were made online or offline (i.e. during recall and/or in response to probe questions). Once again, however, it is not the timing of inference generation which is the central issue here but the content of such inferences and the extent to which they reflect aspects of readers' mental representation of Text B which might validly be included in a satisfactory summary of the text. Examples of both necessary and more elaborative inferences offered by readers of Text B will be discussed in turn. (Earlier discussion of theories of inferencing can be found on pages 128–136.)

Inferences necessary to support local and global coherence for Text B

Readers of Text B presumably did generate online inferences during their initial reading of the newspaper editorial and these helped to maintain co-ordinates of time and space, and especially topic or theme. Such retrospective inferences (otherwise referred to as propositional or bridging inferences) enable readers to construct (and later report) a coherent mental representation though it is difficult to isolate specific examples relating to time and space from the recall transcripts of Text B. It may be that evidence of such inferences is more likely to be found in recalls of a story narrative, such as that in Text A, where the chronological sequencing of time, place, characters and actions is more salient and critical. Inferences which seem to contribute to the maintenance of coherence and which can be relatively easily identified from the Text B recall transcripts include, firstly, inferences about the person of Samantha Kendall, and secondly, those which express the main theme or point of the text.

i) Inferences about the person of Samantha Kendall

One obvious inference which seems to have been made quite readily by a large number of readers was that Samantha Kendall was female. Since no other clue is present in the text, we must assume that it is probably readers' background knowledge of English male and female names which enabled them to make such an inference. In most cases, readers described Samantha as a 'girl', perhaps implying (again on the basis of general knowledge) that she had not yet or had only recently reached adulthood:

a/the girl (#001, #004, #006, #007, #013, #015, #020, #029)
one particular girl (#014)
she (#023, #024, #027)
I don't remember the name of the girl + I think it was a girl (#008)

Only one reader referred to her as a 'woman':

> *there was this woman in the news* (#019)

As we shall see later, a number of readers elaborated upon the necessary inference that Samantha was female to infer how old she was.

At least one reader inferred that *she had been admitted to hospital* (#004), although this might be regarded as more of a presupposition than an inference.

ii) Thematic inferences

Other necessary inferences generated by readers of Text B include thematic inferences relating to the main point or moral of the text. These are perhaps closely linked to readers' interpretation of the views of the writer. At least one reader inferred the overall theme or point of the text as an expression of the need to change the law regarding compulsory treatment of anorexia sufferers:

> *it's about whether they should change the laws on anorexia + make it so we can compulsorily treat anorexia sufferers* (#018)
> *it does cite the 1983 act which is what some years ago + um + so it seems that there's a problem of laws being old and by implication out of date through the article + and that's something which is a problem* (#005)

This inference was initially shared by another reader who subsequently shifted to a different main point.

> *they thought the laws should be changed + or not the laws should be changed but the attitudes of doctors should be changed to make sure that patients are treated better* (#027)

This view of the text's theme as being the need to review and possibly change current approaches to treating anorexia sufferers was shared by several other readers:

> *it's saying if treatment of anorexia and anorexia sufferers should change or whether it's ok as it is* (#019)
> *it's about whether treatments of anorexia should be changed* (#030)
> *they thought the treatment should be looked into and made more humane* (#012)
> *it offers more humane ways of treating the disease + people are trying to find more humane ways of treating it* (#003)
> *they were trying to make that sort of more + more humane* (#017)
> *they're trying to improve ways of treating anorexia* (#023)

A significant number of readers seem to have regarded the main theme as the issue of whether doctors should impose treatment upon anorexia sufferers more than they do at present in order to save lives:

whether doctors should be allowed + um + you know + keep someone in hospital who doesn't particularly want to stay in hospital + but if it means saving their life then they should do it + or whether they shouldn't (#023)*do doctors have a right to + um + treat someone and prevent them + um (. . .) do doctors have the right to stop them* (#030)
if doctors use their powers more under the law then more lives would be saved (#028)
they think that doctors should do more about it (#021)
the opinion of the person who wrote it was that they should keep them in so lives could be saved (#011)
maybe it was + you know + that it shouldn't be personal choice + they should + um + have to stay in hospital (#013)
it argued would it be more humane to let them die and I think he thought it wouldn't be + so more people should be treated (#010)
he tended to think they should be kept in care (#024)

Two readers seemed to feel that the main point or argument of the text was not entirely clear:

I don't think they were really sure what it was best to do (#007)
it says that doctors should hold these laws whereas it's just told us that + um + a deeper understanding is needed + it's a very bad argument (#005)

Examples of elaborative inferences associated with Text B

i) Inferences about the person of Samantha Kendall

Several readers clearly recalled or believed the text to have contained information about Samantha Kendall's age:

a thirteen year old girl (#009)
an eighteen year old (#018)
she was fourteen I think (#003)
was she nineteen (#010)
I can't remember how old she was (#007)

These inferences may have been provoked as a result of readers' prior knowledge of the news case from other sources. Alternatively, since newspaper reports frequently report the age of the characters in their stories, perhaps readers expected to have seen a reference to her age in the text itself.

Some readers made similar inferences about the outcome of Samantha Kendall's condition and even that of a twin sister:

I'm not sure – maybe she died (#008)
there's a certain girl + um + that died recently from anorexia (#010)

she died recently (#028)
her twin sister died (#006)

Once again, since there is no information about any of this in the text, we must assume that these inferences were provoked through the activation of knowledge from other sources (either about Samantha Kendall or about similar cases), probably to enrich their mental representation and make it more coherent and memorable.

Some readers, wishing perhaps to find an explanation for Samantha Kendall's action, speculated upon why she had discharged herself from hospital:

the girl had walked out of the doctors saying she must have help (#001)
perhaps she didn't believe that any more could be done for her (#013)
did her doctor refuse to treat her for some reason (#018)
she discharged herself from hospital presumably because she didn't want to carry on being treated (#021)

At least one reader confessed to being unsure of the reason:

she discharged herself from hospital + I don't know why actually (#029)

And one reader speculated upon the consequences of her action:

it was very dangerous for her to be away from medication because of her disease (#015)

ii) Inferences about the causes of anorexia

Several readers elaborated quite extensively upon the causes of anorexia offered in the text, perhaps as a direct result of prior knowledge or experience. Some inferred that the 'unacceptably stressful and difficult adult life' was caused by pressure from other people:

they might be pressured by people and it's a way of retreating from pressure (#003)
I also think it's a lot of pressure by family and parents + friends (#027)
it's peer pressure (#021)
people think that they're ugly (#016)

Others appeared to infer a link between retreating into childhood and a desire to remain small or undeveloped:

they're trying to starve themselves to get them back to childhood + to being small (#011)
they don't want to become women + they want to revert back to childhood and stay a children's size (#026)

by sort of starving yourself you're + um + stopping your body developing into a woman's body (#017)

One reader inferred that anorexia sufferers wish to constrain their own lives rather than allow them to be constrained by others:

they're putting a constraint on their own lives + they set their own lifestyles (#025)

Another suggested that through their diet sufferers are seeking to control a life that is chaotic:

they're trying to exert control over a sort of chaotic life + and they're sort of rationalising their diet (#028)

And one reader seems to have drawn the strong inference that *more or less no-one knows why anorexia occurs* (#027).

iii) Inferences about attitudes to anorexia

At least one reader inferred that anorexia was not previously regarded as easily treatable:

they said it's treatable now + it didn't used to be as treatable (#026)

Another reader inferred that anorexia was not curable:

it's actually a psychiatric problem therefore not necessarily curable (#015)

Two readers drew inferences about changes in public attitudes to anorexia:

it's gaining a lot more sympathy (#019)
it's not seen as people trying to fool themselves any more + it's seen as something they can't help (#020)

Some readers made similar inferences about the extent and success of research into anorexia, even though the text actually offers very little information on this point:

the whole issue of anorexia has been difficult to research and draw any conclusions about + it mentions some types of experiments they do (#017)
they should be researching more into the effects + that there's not enough being done (#022)
the worry is not enough research is being done + more research needs to be done into helping these patients + a lot more can be done for this illness + it's calling for more research to be done (#018)

iv) Inferences about doctors' attitudes and powers in treating anorexia sufferers

Readers varied in the inferences they drew concerning the attitudes and powers of doctors faced with treating anorexia sufferers. Several readers inferred that doctors are legally empowered to force treatment upon anorexia sufferers but are reluctant to use such powers, an inference which could be described as strongly authorised by the text:

> *a health act in 1983 which said they could force treatment on people with anorexia* (#007)
> *anorexia is a psychiatric problem as opposed to a just a sort of medical + um + medical one + um + which is why it's governed under the mental health act* (#015)
> *the 1983 + I think + mental health act + um + allows doctors to force feed the patient* (#020)
> *doctors have got it in their power to detain them as not in control of their mental health + but most are unsure about this because of the detrimental effect it might have* (#009)
> *doctors haven't been wanting to keep patients in hospital and out of danger in a sort of temporary way* (#017)
> *they're reluctant to do it because of the controversy when they get back home* (#028)
> *doctors aren't sure how to treat it* (#009)
> *failure to integrate the 1983 mental health act* (#029)
> *the doctors should be using the mental health act to + er + help these patients* (#020)

Other readers inferred that doctors could not force treatment upon anorexia sufferers without their consent:

> *it can only be done with their consent + patient's consent* (#014)
> *you can't compulsorily treat someone + can you* (#018)
> *if they want help they can get it + if they don't + it's up to them* (#019)
> *there's no way the hospital can keep you in if you don't want to be treated* (#019)
> *treatment's not compulsory and sufferers can take it or leave it* (#021)
> *doctors cannot make them stay if they discharge themselves + they tend to be lenient on that* (#024)

One reader reported being confused about the extent to which it was possible to force treatment upon an anorexia sufferer:

> *was it that doctors can actually treat the patients without their consent or was it the other way round + I think there was something about that that confused me when I was reading it* (#026)

v) Inferences about approaches to treating anorexia sufferers

Some readers inferred a specific reason for the approaches to treatment described towards the end of the text:

they were watched in everything they did (#013)
they're watched (. . .) to see their behaviour patterns I guess (#015)
they watch people eat (. . .) to make sure they didn't make themselves sick, to make sure they're eating (#016)

iv) Inferences reflecting emotional or critical reactions to the text

On a few occasions readers appeared to offer inferences (or metalinguistic comments) that reflected their own reactions either to the content of the text itself or to the writer's position. Some readers clearly reacted negatively to the approaches to treating anorexia:

treating them like a little kid really + they should be not so strict with the way they are treating (#025)
degradation + humiliating (#006)
another way was humiliation (#028)
um + I can't remember what the next one was because the last one was so shocking (#029)
you can't just + um + put someone on a force-feed diet (#003)
they should be treated humanely + um + shouldn't be forcefed (#020)

Some readers differed in how far they sympathised with the writer's stance:

it seems to present a sort of fairly balanced view of it (#006)
the author's view is correct (#017)
this journalist's extremely biased one-sided view (#021)

Some conclusions from analysing the inferences in Text B recalls

As with Text A, the aim was not to try and categorise *all* the inferences made by readers of Text B but instead to gain insights into readers' mental representation of the newspaper editorial by analysing some of the inferences for which there is evidence in the form of the additional propositions. The analysis shows that readers of Text B generated both bridging inferences necessary to understanding and also inferences of a more elaborative or pragmatic nature which help to enrich the individual's mental representation.

What is perhaps striking about the Text B recall transcripts is the number and diversity of elaborative inferences made by readers of the newspaper editorial – far more, it would appear, than for the short story narrative in Text A. This might be due to a more personal or real-world connection with the theme and its treatment in the text, than was the case with Text A.

Clearly, readers took the information provided in Text B and, by integrating it with their existing knowledge and experience of the world, and of the subject matter in particular, elaborated beyond the text to generate a mental representation which was not just sufficiently accurate and coherent for their purpose, but which also represented a world that was very familiar for them. It is quite likely, for example, that general knowledge enabled readers to identify Samantha Kendall as a female, probably in her teens. But it may well have been some personal or specialised knowledge of anorexia itself (either as a sufferer oneself, or as someone who knows a sufferer, or simply by virtue of being in the age-bracket which is most affected by the condition and thus targeted with information and advice through health professionals and the media) that enabled readers to build some of the more elaborative inferences.

In reviewing the differences in interpretation which Text B provoked, it is necessary to consider the types of inference which could justifiably be included in an adequate summary of the text. It is relatively easy to justify the inclusion of necessary inferences (e.g. referential and thematic) since these are generally regarded as essential to coherent understanding. The potential inclusion of elaborative inferences is more risky, however. It has already been suggested that a certain amount of leakage between necessary and elaborative categories is possible, due to constraining factors such as purpose for reading and background knowledge. Clearly, elaborative inferences activated by knowledge or experience of a highly personal nature, such as *she was fourteen* or *it's a lot of pressure by family and parents*, are much less likely to be appropriate for inclusion than more generally shared inferences such as the idea that anorexia is gaining a lot more sympathy and that because the condition is imperfectly understood, more research needs to be done. Given that this type of text appears to provoke a wide range of elaborative inferences in readers, testers will need to be extremely careful in building a consensus of precisely which elaborative inferences can be justified for inclusion and whether these might form the basis of test items. An individual test writer working alone to develop a text summary and test items is always at risk of allowing their own individual and idiosyncratic elaborative (i.e. non-consensus based) inferences to become items in the test. Even if such inferences are valid for an individual's interpretation of the text, they cannot be justified as part of a shared comprehension and they should really be avoided in any summary and as a focus for testing.

The study of inferences generated by readers of Text B suggests that they were generally motivated by the reader's desire to integrate incoming information from the text with their existing knowledge structures so as to arrive at a coherent understanding of actions, events and states. In the light of this, a useful guiding principle might be to include in a summary of Text B those necessary inferences and consensus elaborative inferences which provide the explanation for actions, events and states.

Constructing experimental summary variants of Text B

Following close analysis of the text-based and summarising propositions occurring in reader recalls of Text B, together with study of readers' inferences in the transcripts, consideration was given to how to construct various experimental summaries of the newspaper editorial – the most suitable of which could be selected at a later stage to serve as the basis for a summary completion task. It was hoped that insights from this exercise, combined with insights from the same approach with Text A, might enable a practical and principled methodology to emerge for constructing text summaries for test development. The same three methods were used as for Text A:

i) Summaries comprising only text-based propositions
An initial attempt was made to construct and compare summaries of Text B using *only text-based propositions* in accordance with the levels of reader consensus and frequency of occurrence identified in the recall transcripts (see pages 139-143 in this chapter). This generated multiple summary variants (derived from Sets T(ii), T(iv), T(v) and T(vi)), containing differing degrees of detail and ranging in length from 53 to 165 words.

ii) Summaries comprising only summarising propositions
A similar approach was adopted using *only summarising propositions*, in line with the criterial categories used for the analysis (see pages 153–156 in this chapter). The summary variants derived from Sets S(ii), S(iv), S(v), S(vi) and S(vi/viii) ranged in length from 20 to 73 words.

iii) Summaries combining text-based and summarising propositions
The third method for constructing possible summaries of Text B was to *combine corresponding sets of text-based and summarising propositions*, i.e. each summary variant comprised a set of summarising propositions and all text-based propositions associated with these. This method initially generated multiple variants in line with the criterial categories analysis (see pages 139–155 in this chapter). However, as in the case of Text A, it quickly became clear that some high-frequency text-based propositions that were candidates for inclusion in a summary were actually associated with summarising propositions that were *medium*-frequency (the middle 33%), rather than high-frequency (the top 33%). For this reason, two additional summary drafts were created combining high-frequency (HF) text-based propositions and high/medium-frequency (HF/MF) summarising propositions recalled in ANY recall.

In all the experimentally constructed summaries described above, the order and wording of the text-based and summarising propositions was strictly maintained.

The rationale for developing multiple summary variants was discussed earlier in this chapter with regard to Text A (see page 137). The reader is reminded of the four key principles underpinning decisions about which summary to select from the available drafts. The optimum summary would need to:

1. Accurately reflect the content of the original text.
2. Be a coherent and standalone piece of discourse in its own right.
3. Include elements of the original text which most readers would find salient as well as some elements which might cause weaker readers difficulty.
4. Be long enough to be capable of generating a large enough number of locally independent test items to meet the demand for test reliability.

As with Text A, it was expected that the most promising summary of Text B for subsequent development into a summary completion task would be one containing several levels of detail since this would offer maximum scope for constructing a large enough set of test items focusing beyond the most basic level of detail and capable of engaging readers' cognitive processing at a number of different levels.

Table 6.36 shows the multiple draft summaries that were constructed

Table 6.36 Multiple summaries of Text B constructed using different criteria

Criteria used to construct draft summary of Text B	Length (words)	Accept/ reject	Reason for acceptance/ rejection of draft summary
T(i)			
T(ii)	53	Reject	Single level of detail + too short
T(iii)			
T(iv)	138	Reject	Single level of detail + too short
T(v)	168	Reject	Single level of detail + too short
T(vi)	165	Reject	Single level of detail + too short
S(i)			
S(ii)	20	Reject	Single level of detail + too short
S(iii)			
S(iv)	42	Reject	Single level of detail + too short
S(v)	37	Reject	Single level of detail + too short
S(vi)	42	Reject	Single level of detail + too short
S(vii/viii)	73	Reject	Single level of detail + too short
S/T(i/ii) – ANY recall by 20/30 readers	73	Reject	Too short
S/T(iii/iv) – ANY recall by 15/30 readers	180	Reject	Length OK but lacking detail
S/T(v/vi) – HF propositions (all) in ANY recall	***272***	***Accept***	***Content + length about right***
S/T(vii/viii) – HF/MF summarising + HF text-based propositions in ANY recall	294	Reject	Too detailed + too long

according to methods (i), (ii) and (iii) above, with the reasons why each of these variants was either accepted or rejected for future development as a summary completion task.

Since the nine summary variants comprising only text-based or only summarising propositions (i.e. derived by methods (i) and (ii) above) limited themselves to a single level of detail, these were all rejected for this reason.

The summary that was finally selected for further exploitation was the draft derived by combining the high-frequency (top 33%) summarising and text-based propositions recalled by readers, as shown in bold italic in Table 6.36. In terms of number of words, this draft summary amounted to approximately 69% of the original text (i.e. 272 compared with 389 words). This percentage constitutes a very high proportion of the original word count and it falls well beyond the normal boundary for what is generally considered a summary. The literature (e.g. Ratteray 1985) suggests that length of summary can vary according to summary type – usually anything between one-tenth of the original text (e.g. an abstract) to one-third (e.g. a précis). Although the 180-word alternative draft (based on combined propositions recalled by 50% of the readers) offered a better length, its quantity of content and level of detail was considered insufficient to support the number of test items (i.e. 30–40) that were thought to be needed for a reliable summary completion task. Since the aim here is to generate a reliable test task with a sufficient number of test items, adequate length becomes an important issue. It is also important to bear in mind that once key words or phrases are deleted from the summary to create the test items, then the remaining text contains even fewer words, so it needs to be a good match to the original even with the deletions. One issue that test developers may need to take into account when developing summary completion tasks, therefore, is whether certain text genres are more difficult to summarise economically than others, especially texts that are lexically dense and conceptually complex (see more discussion of this below and in Chapter 9).

At this point the selected summary was edited to ensure that it read smoothly and coherently. For the most part, this involved the appropriate use of pro-nominalisation, the introduction of appropriate conjunctive expressions and a small amount of reordering of propositions. This process also helped to significantly lower the word count bringing it closer to 60% of the original text word count.

The earlier discussion on inferencing highlighted the variability of inferences made by readers of Text B as well as the difficulty of determining which necessary and/or elaborative inferences should be included in any summary of the text. In the end, it was decided to integrate only those inferences which could be considered necessary to maintain local or global coherence. This meant confirming that any logical or necessary inferences relating to referential chains and thematic content were already represented within the

summary (i.e. pronominalisation as part of a referential chain concerning Samantha Kendall, and mention of the thematic thrust of the test, i.e. the need for further progress in how the condition is treated) and thus could be exploited for item construction if so desired. No elaborative inferences were introduced relating to any of the types (i) to (vi) discussed on pages 158–164 above.

The final outcome, therefore, was a 234-word summary of Text B that combined summarising and associated text-based propositions and that incorporated a limited number of authorised inferences, as shown in Appendix 7 on pages 229–230.

Some reflections on a comparison of oral recalls for Texts A and B

Comparing Texts A and B in terms of free and prompted recall

A major difference was observed between readers' recalls for Text A and their recalls for Text B as far as their overall structure was concerned. This difference is probably a function of the two different text types and their individual rhetorical structures.

Text A (*Journey*) is a narrative short story with a clear chronological ordering from beginning to end. This undoubtedly imposed strong constraints upon the structure of readers' recalls, providing them with a clear linear pattern for the events of the narrative. Only occasionally did readers recall details or incidents outside their natural place within the sequence of events in the story. When this did happen, it seemed to be because mention of one detail or event spontaneously triggered mention of something related but not necessarily contiguous at that point in the sequence.

Text B (*Anorexia*), on the other hand, is an expository newspaper editorial with none of the clear chronological sequencing that is a central thread in Text A. While the Text A recalls followed a fairly rigid, linear pattern, the Text B recalls tended to be much less constrained or clearly defined in terms of their structure, and demonstrated many more individual idiosyncracies. Since no constraining chronological thread or sequence of events was available for organising recall of the text content, readers sought some other principled approach. Some readers took as their point of departure the mini-narrative relating to Samantha Kendall, and this may reflect the journalistic device used in the text of using a specific case-study to introduce a more abstract discussion (i.e. working from the particular to the general). Some Text B readers even seemed to amplify the mini-narrative here by introducing inferences relating to the age of the girl, reasons why she left hospital, and her eventual death (perhaps seeking for themselves a coherent and satisfactory closure for the narrative).

An alternative approach observed in some of the reader recalls for Text B was to use the historical perspective as a point of departure, i.e. the fact that there have been changes over time in public attitudes to anorexia. This is consistent with the observation from other studies that informants may take advantage of both real and potential narrative elements in their recalls (Gomulicki 1956). Impressionistic evidence suggests that readers preferred both the mini-narrative and the historical perspective approaches to structuring their recalls over a purely factually based description of anorexia and its treatment. In addition, readers of Text B frequently recalled propositions in a different order from the one in which they appeared in the text (although this was rarely the case for Text A recalls). In summary, therefore, it seems that recalling the expository newspaper editorial proved to be a much less simple affair than recalling the short story narrative. The inherent structure of a text is likely to be a critical factor when selecting texts for developing summary completion tasks, as well as its difficulty level in terms of its lexical, syntactic, propositional and rhetorical characteristics.

A further difference between recalls for Texts A and B concerns the extent to which readers apparently established associative networks of propositions relating to the text content. For Text A (*Journey*), for example, the mention of one proposition appeared to stimulate recall of two or three other propositions and probe questions provoked recall of whole clusters of propositions. This was far less noticeable for the recalls of Text B (*Anorexia*) perhaps because the dense, information-loaded propositions were more difficult to connect up through such networks of association.

It seems clear, then, that Texts A and B are likely to provoke two different types of mental representation in readers. An attempt to describe the structure of the mental representation for Text A (*Journey*) in diagrammatic form would probably show a central, linear sequence of five or six main stages in the story discourse (e.g. man alone, man afraid, car arrives, and so on), with clusters of associated propositions linked to the central sequence as well as loops backwards and forwards. Text B (*Anorexia*), on the other hand, would probably be characterised by a much less defined structure with single and clusters of propositions being much more loosely attached to one another. Instead of a linear sequence for Text B, we might envisage a wheel-like diagram, with a central node to represent the Samantha Kendall case study and spokes radiating out from it to small circles containing the various sub-themes that are addressed in the text through individual paragraphs, e.g. the change in attitudes to anorexia, psychiatric explanations for the disorder, the legal situation, etc.

Comparing Texts A and B in terms of summary construction

One significant difference between the experimental summaries constructed from Texts A and B proved to be length. Summaries constructed from Text

A tended to be considerably shorter than those for Text B. This may relate directly to the density of propositions in Text B and the fact that it is difficult to condense discursive text and yet still retain a meaningful summary. A narrative text such as Text A, however, often contains a substantial amount of material that is included for stylistic purposes, e.g. to create atmosphere. As a result, it becomes easier to trim the text down much further and still retain an accurate and coherent summary of events.

A further point for comparison between summaries of Texts A and B concerns the extent and nature of any inferences which can be incorporated. The most helpful approach in this case seemed to be to include in the summaries of the two texts only necessary inferences and those essential to an understanding of what was going on (if this needed to be made more explicit). As a general rule, care needs to be taken to avoid including any inferences that are not authorised by the text and which might be in conflict with the valid interpretation of some readers (e.g. the likely age of the main character, or the potential cause of an event). For example, for Texts A and B it seems that readers readily made inferences about the people referred to. Though these may have been only minimal inferences, concerning features such as gender/age or the reason why something happened, they suggest that, when constructing a mental representation of a text, readers may feel the need to instantiate a prototypical person which satisfies their own understanding. A similar phenomenon was observed in an earlier study of cloze task processing (Taylor 1991) where, even though the text gave no indication of gender, all 300 readers in the sample clearly interpreted the taxi driver mentioned in the text to be male rather than female (presumably in line with general experience of the world).

Trying to make potentially ambiguous inferences explicit through the summary of a text may well conflict directly with the mental representation constructed by some readers. For this reason, when constructing a summary of any text, it is much safer to limit the summary content to matters of comprehension rather than interpretation. It is also clear that readers varied somewhat in the evaluative inferences they made regarding interpretation of the writer's stance or point of view. Again it may be wise to avoid including an explicit statement of the writer's point of view if this is somewhat nuanced.

Conclusion

Chapter 6 has presented a detailed and systematic analysis of readers' oral recalls of two texts – a short story (Text A – *Journey*) and a newspaper editorial (Text B – *Anorexia*). This multi-level analysis offered valuable insights into the way readers appear to structure text-based and other information in order to construct a mental representation of narrative and expository text. It demonstrates that while many aspects of meaning construction are shared in

common by readers, certain aspects are far more personal and idiosyncratic in nature. Results of the analyses were used to construct several summary variants of each using a variety of different methods and according to different sets of parameters. The resulting drafts were compared in terms of their potential for future development and a single summary, comprising a combination of summarising and text-based propositions, was selected for each text for the purposes of developing two summary completion tasks for assessing reading comprehension ability.

One aim of the research reported in this chapter was to identify a suitable methodology for constructing a summary of a text that could be developed into a summary completion task. Results in Chapter 6 suggest that to construct an adequate summary of a 350–550-word text, both macro-propositions and a sufficient level of detail will need to be included. This means identifying important summarising macro-propositions and important text-based micro-propositions, and then combining these to form the final summary text, using appropriate linguistic features to ensure surface cohesion and stylistic fluency.

Chapter 7 will describe how a set of *written* recalls of Texts A and B was collected from a broader population of readers including both strong and weak readers. Analysis of these written recalls facilitated the development of a set of test items for the Text A and B summaries and the resulting summary completion tasks were then trialled to assess their effectiveness as measures of reading comprehension.

7 Developing summary completion tasks for Texts A and B

Introduction

The previous chapter explained how a detailed propositional analysis of oral recalls of Text A (*Anorexia*) and Text B (*Journey*) by a cohort of mature readers led to the construction of a summary of each text that combined summarising and text-based propositions and included authorised inferences (see Appendix 7). The two summaries were intended to form the basis for developing summary completion tasks for testing reading comprehension ability. Chapter 7 explains how the summary completion tasks were constructed by deleting words and phrases from the Text A and Text B summaries to form a set of test items for each one.

A set of *written* recalls of Texts A and B were first of all collected to explore reading comprehension performance across a much larger and broader sample of readers than previously, including both competent and less able readers. This was done to identify differing levels of quality of text comprehension and recall for each text. It was anticipated that analysis of these differing levels, combined with insights from the study reported in Chapters 5 and 6, would help to confirm the adequacy of the Text A and B summaries (in terms of their content coverage and length) and also to inform the development of a set of comprehension test items for each task formed by gaps in the summary. The aim was to confirm both summaries in terms of their propositional match to evidence from another cohort of readers regarding what is salient within their constructed mental representation, and particularly any critical points where comprehension is at risk of breaking down among less able readers. It was anticipated that identifying instances where comprehension broke down would help in the targeting of word and phrase deletions for constructing test items. Chapter 7 goes on to describe how the resulting summary completion tasks were trialled on a different population of readers to establish their effectiveness as a test of reading comprehension ability.

Collecting a set of written recalls for Texts A and B

Participants

Participants involved in the written recall study were 82 Year 10 students (14/15-year-olds) studying General Certificate of Secondary Education (GCSE) English at a state secondary school in England. The sample population was selected to ensure that the full ability range for the year group was represented. English teaching for this age group was organised by the school into 10 sets streamed according to ability, with sets paired across the two halves of the school, referred to as the North and South Bands. Thus Sets N1 (higher ability) to N5 (lower ability) in the North Band are comparable to Sets S1 to S5 in the South Band. Since the Head of English reported Sets 1 and 2 to be of more or less comparable ability, only four out of the five English sets from each Band were used. Participant numbers involved in the study are shown in Table 7.1.

Table 7.1 Number of participants in the written recall study

Text A (*Journey*)	**Text B (*Anorexia*)**
Set N1 = 29	Set S1 = 30
[Set N2 not used]	[Set S2 not used]
Set N3 = 25	Set S3 = 24
Set N4 = 16	Set S4 = 20
Set N5 = 8	Set S5 = 8
Total = 78	**Total = 82**

Materials

The two reading texts used were Text A (*Journey*) and Text B (*Anorexia*) as previously described in Chapter 5 (see Appendices 1 and 2 on pages 220–221). A pair of tasks was produced to accompany and contextualise each reading text: Tasks A1 and A2 for Text A, and Tasks B1 and B2 for Text B. The A1/B1 and A2/B2 tasks were designed to provide readers with a clear and plausible context and purpose for their reading activity. Task A1/B1 (on the first handout) asked participants to read the text and consider whether it would be suitable for use as a reading passage in a Key Stage 3 (KS3) (Year 9) English test, taking into account whether the ideas or the language it contained might cause any difficulty for 13/14-year-olds. (This task was directly relevant to the trialling population for the summary completion task at a later stage.) The task instructions explained that they would be asked a few more questions about the text after reading it. Task A2/B2 on a second handout asked participants to respond to three 3-option multiple-choice questions (MCQs) relating to the perceived interest value, complexity of ideas and complexity

of language of the text, and then to write down as much of the reading text as they could recall. It was made clear that errors of grammar, spelling and punctuation in their written performance would be disregarded so they did not need to worry about these. A final question asked participants to sum up in one or two sentences what they thought the writer was trying to say in the text. Written instructions, spacing and layout for the tasks were all checked in consultation with an experienced secondary English teacher and test-writer. (All the tasks are included as Appendix 6 on pages 227–228.)

Procedures

The tasks were administered by class teachers during normal English lessons in accordance with clear written instructions on the procedures to be followed (see Appendix 6). To minimise disruption to the school's normal timetable, the two texts (A and B) and their accompanying tasks were divided between the class sets: Sets N1, N3, N4 and N5 received Tasks A1 and A2, while Sets S1, S3, S4 and S5 received Tasks B1 and B2. Participants were given 3 to 4 minutes to read the Task A1 or B1 instructions and its accompanying text. The Task A1/B1 handouts were then collected in by the class teacher and participants were given a separate handout for Task A2 or B2. They were allowed a maximum of 15 minutes to complete the questions and written recall on this sheet (undertaken without access to the original text as this had already been collected in on the first handout). At the end all Task A2 or B2 handouts were collected in by the class teacher. Time allowances required for the tasks had been calculated and confirmed as suitable following a small pilot study.

Analysis

The Task A2 and Task B2 sheets containing participants' multiple-choice responses and their written recalls were analysed in three stages. First, frequency of responses to the MCQs (i) to (iii) was calculated (see Tables 7.2–7.4).

Results suggested that a good proportion of the candidates found the texts to be 'quite interesting', 'generally easy to understand' in terms of the ideas and 'generally easy to understand' in relation to its linguistic complexity. This was an encouraging outcome since it generally confirmed the suitability of both texts for their intended audience and purpose in terms of interest value and accessibility. Significantly more candidates reported finding the ideas and language of Text B 'sometimes difficult to understand' than those of Text A, suggesting that the content and language made Text B a more challenging passage for readers. This was subsequently confirmed in the analysis of performance on the two summary completion tasks.

Table 7.2 Perceived interest value of Texts A and B

(i) Would 13/14-year-olds find them:	Text A (*Journey*) (%)	Text B (*Anorexia*) (%)
A very interesting?	10 (12.8)	12 (14.6)
B quite interesting?	63 (80.8)	60 (73.2)
C not at all interesting?	3 (3.8)	9 (11.0)
no response	2 (2.6)	1 (1.2)
	78 (100)	82 (100)

Table 7.3 Perceived ideational complexity of Texts A and B

(ii) Were the ideas:	Text A (*Journey*) (%)	Text B (*Anorexia*) (%)
A generally difficult to understand?	1 (1.3)	3 (3.7)
B sometimes difficult to understand?	13 (16.6)	38 (46.3)
C generally easy to understand?	62 (79.5)	40 (48.8)
no response	2 (2.6)	1 (1.2)
	78 (100)	82 (100)

Table 7.4 Perceived linguistic complexity of Texts A and B

(iii) Was the language:	Text A (*Journey*) (%)	Text B (*Anorexia*) (%)
A generally difficult to understand?	1 (1.3)	3 (3.7)
B sometimes difficult to understand?	29 (37.2)	46 (56.1)
C generally easy to understand?	46 (58.9)	32 (39.0)
no response	2 (2.6)	1 (1.2)
	78 (100)	82 (100)

Secondly, the content of the written recalls was analysed using procedures followed in the English Language Monitoring (ELM) Project (Pollitt et al 1990) for developing holistic scales for writing assessment. (The approach used in the ELM Project was derived from the Primary Trait method used in United States' monitoring surveys of writing – see Mullis 1980.) All the written recalls for Texts A and B were first sorted by the researcher into a rough rank order according to how accurate and comprehensive the recall appeared to be. The rank ordered set for each text was then subdivided into as many stable categories as could reliably be distinguished. A brief description of each distinguishable category was drawn up which concentrated on defining its essential character. This description also included, where possible, specific examples to illustrate relevant characteristics of the category, especially regarding degree of successful comprehension or where comprehension appeared to break down. In this way, holistic, empirically based levels of performance quality for the written recalls of Texts A and B were distinguished and described. Five levels of quality of comprehension and recall were distinguished overall for each of the two texts, as shown in Tables 7.5 and 7.6.

Table 7.5 Levels of quality of comprehension and recall for Text A (*Journey*)

Level	Description	No of recalls (%)
Level 0	No evidence of understanding.	0 (0.0)
Level 1	A very general statement of content in one or two sentences and including a collection of discrete facts from one or two parts of the text often in verbatim form. A noticeable lack of coherence.	7 (9.0)
Level 2	A partial account with some level of detail often in verbatim form. Chronological ordering sometimes distorted and several sections of the text omitted. Generally characterised by at least one major comprehension error, e.g. *he paid the driver, the driver attacked the man, the passenger reached for the spanner, the wallet was missing*.	31 (39.7)
Level 3	Accurate account of text content. Good chronological ordering. Good level of detail and coherence, but often characterised by the omission of a particular section of the text, e.g. *the man's unsuccessful attempt to pay, the driver's perspective at the end*. Evidence of paraphrasing.	25 (32.1)
Level 4	Full and accurate account of the text content. Good chronological ordering. Substantial level of detail. High level of coherence. Extensive evidence of paraphrasing.	15 (19.2)

Table 7.6 Levels of quality of comprehension and recall for Text B (*Anorexia*)

	Description	No of recalls (%)
Level 0	No evidence of understanding.	2 (2.4)
Level 1	A collection of discrete facts from one or two parts of the text often in the form of verbatim recall and sometimes recalled incorrectly, e.g. *doctors can't help unless they get permission, doctors give drugs but most don't survive, one treatment is to make sufferers stand in front of a mirror and look at themselves*. A noticeable lack of coherence. Sometimes includes non-text-based information.	37 (45.1)
Level 2	A partial account with some level of coherence and of verbatim detail. Several sections of the text omitted. Sometimes characterised by a major comprehension error, e.g. *anorexia has been diagnosed as a problem for 10 years, doctors think it is best for patients to have no choice about treatment*, etc. Sometimes includes non-text-based information.	19 (23.2)
Level 3	Accurate account of text content. Good level of detail and coherence, but often characterised by the omission of one or two sections of the text, e.g. *the case of Samantha Kendall, changes in our view of anorexia, types of treatment, the need to use force if necessary*. Evidence of paraphrasing.	16 (19.5)
Level 4	Full and accurate account of the text content. Substantial level of detail. High level of coherence. Ordering of the recall tends to correspond to paragraph ordering of the original text. Extensive evidence of paraphrasing.	8 (9.8)

The final stage of this analysis was to list the participants' attempts at producing a one- or two-sentence overall summary of what the writer was trying to say in the text. These responses were also grouped and recorded according to the five levels of quality of performance described above.

The value of analysing the written recalls in this way lay in the insights it provided over what constitutes *full and successful comprehension*, and what might be regarded as only *partial or inadequate comprehension* of Texts A and B, in particular those *points where comprehension risks breaking down* among less able readers for some reason. Given that most reading comprehension tests aim to spread readers out along a continuum of some sort, ranging from poor to good, it will presumably be important for a summary completion task to achieve a similar outcome. Thus any summary that forms the basis for a summary completion task will need to reflect what might be regarded as an accurate and sufficient mental representation of the text. In addition, the test items based upon that summary will need to focus on elements in the representation that are associated with full and successful comprehension but also partial and inadequate comprehension.

The characteristics of the recalls assigned to Level 4 (see Tables 7.5 and 7.6) helped to confirm what an optimum summary should look like for each text, while the features that characterised Levels 3, 2 and 1 provided valuable task design guidance in two areas: first, in confirming any key content to be retained in the summary (e.g. for Text A – *the man's unsuccessful attempt to pay, the driver's perspective at the end*); and secondly, with regard to points within the text where comprehension clearly broke down for weaker readers and thus where a test item might reasonably be located (e.g. *he paid the driver, the driver attacked the man, the passenger reached for the spanner, the wallet was missing*). All these examples testify to poor, or failing, comprehension and might thus be areas of the summary where test items could be located to discriminate between successful and weaker readers.

The two summaries constructed at the end of Chapter 6 had been edited to ensure the text was accurate and sufficiently comprehensive in content and cohesive in its surface structure. Where appropriate, logical relationships and necessary inferences that were implied between actions, states or ideas within each text (e.g. cause and effect, or contrast) had been made explicit by introducing suitable linking adjuncts and complex syntactic structures. Specific examples of co-ordinating and subordinating conjunctions used to achieve cohesion and fluent expression included words such as *however, but, because, although, despite, when* and *before*. It was anticipated that some of these words might later be candidates for deletion to form test items in the gapped summary given that they had the potential to test key aspects of successful comprehension, e.g. the order in which salient events happened in the story, or the salient contrast between past and present attitudes expressed in the newspaper editorial.

Constructing a set of test items from each summary

As explained above, the analysis of the written recalls described above in Tables 7.5 and 7.6, and especially the Level 1, 2 and 3 performance descriptors, was used to help target key areas in the summary where individual test items might be located. For example, specific attention was paid in Text A to those parts of the summary relating to:

- whether payment was made (or not)
- who was attacked (or not)
- who reached for a weapon (or not)
- whether the wallet was missing (or not)
- whose perspective was highlighted at the end of the story.

For Text B, for example, attention focused on whether test items could be located in those parts of the summary which addressed:

- doctors' freedom in how to treat anorexia sufferers (or not)
- types of treatment and survival rates for the condition
- how long anorexia has been a medical diagnosis.

Since variation across levels of comprehension and recall performance appeared to centre on the degree of fullness, accuracy and coherence of understanding, single words and short phrases were sought for potential deletion to generate test items reflecting these aspects of understanding.

Higher-level performances for Text A (*Journey*) (see Table 7.5) generally demonstrated a clear understanding of how individual actions in the story were attributed to the two characters, while lower-level performances showed signs of confusion in following referential chains (e.g. some written recalls reported the passenger rather than the driver as the person who pulled out the metal object, and some recalls failed to grasp the important shift to the driver's perspective at the end of the story). For this reason, some test items for Text A focused on understanding the *reference chains* in the story since the ability to infer referential links in a text can be considered a crucial element of comprehension. The following examples illustrate this type of item (the word/s in bold constitute the deletion that formed the test item):

Text A: (Item 27) *suddenly* ***the driver*** *pulled out what looked like an iron tool;*
(Item 36) ***the driver*** *hurried away with a sense of relief.*

Another group of items for Text A focused on understanding the different *emotional states* of characters in the story. Written recalls at most levels referred to the emotions of fear and panic that were central to the story's action, but only higher-level performances integrated the more subtle or

peripheral emotions of nervousness and relief that were expressed at different points in the story. The following examples illustrate this type of item:

Text A: (Item 17) *at first he felt* ***relieved***
(Item 18) *then he began to feel* ***nervous*** *again*
(Item 25) *a feeling of* ***terror/panic/horror*** *overwhelmed the man*
(Item 36) *the driver hurried away with a sense of* ***relief***

Referential chains and emotional states proved to be an important focus for constructing test items for Text A, but far less so for Text B given its more factual presentation and more objective tone.

A further group of items for both texts focused upon what might be termed *coherence markers or logical connectors*. These had been built into the summaries in order to make explicit logical relationships through linking adjuncts, such as co-ordinating and subordinating conjunctions. Examples of such items include:

Text A: (Item 19) *he began to feel nervous again* ***because*** *he couldn't see the driver's face*
(Item 34) *he got out his wallet to pay the driver* ***but*** *the car drove away*

Text B: (Item 13) *sufferers may want to retreat to childhood* ***because*** *they cannot face*
(Item 18) *our understanding of the disease,* ***despite*** *extensive research, is limited*
(Item 28) ***although*** *one in ten sufferers dies, doctors are reluctant to*
(Item 37) *it may be the case,* ***however****, that such treatment is justified*

Items such as these tested readers' understanding of *essential relationships between items of propositional content* within the text, such as *cause and effect* or *compare and contrast*.

Not surprisingly, the majority of test items constructed within each gapped summary focused on salient details judged to be essential to adequate comprehension of the text. They included key persons, objects, actions, conditions, values and reasons, as illustrated by the examples below. In some cases, these details were explicitly expressed in the original text, as in the following examples:

Text A: (Item 12) *anxious for his own safety, he reached for* ***his wallet***
(Item 28) *suddenly the driver pulled out what looked like* ***an iron tool***
(Item 38) *vowing not to take any more* ***night passengers***

Text B: (Item 1) *a girl suffering from anorexia* ***left*** *hospital*
(Item 29) *one in* ***ten*** *sufferers dies*
(Item 34) *such as leaving patients* ***in isolation/without their clothes***

In other cases, items focused on information that was implied by the text or inferable from it. Necessary inferences made by readers in both oral and written recall studies were used to guide the design of this type of item within each gapped summary. Examples include:

Text A: (Item 2) *he was keen to reach* ***his home/destination***
(Item 3) *it was too* ***late/dark/dangerous/far*** *to think about walking*

Text B: (Item 33) *they think the treatment will be* ***ineffective/pointless*** *in the long run*
(Item 38) *such treatment is* ***justified/acceptable*** *if the alternative is death*

In some cases details contained in text-level propositions acted as prompts for test items constructed out of higher-level summarising propositions (or vice versa) as in the following example:

Text A: (Item 5) *he began* ***whistling*** *in order to keep his spirits up*

In summary, therefore, the construction of test items (and thus the deletion of key words and short phrases from the summary) centred mainly upon the following aspects of the texts:

1. Salient details of the story/article.
2. Affective elements (i.e. emotional colouring) where this was key to the mental representation of text.
3. Logical connectors.
4. Referential chains.

Once a provisional set of at least 30 test items had been constructed in the form of numbered gaps within each summary, both sets were trialled on a small number of teenage readers. Some informants in this opportunistic sample were first of all invited to complete each gapped summary task without reading the original text beforehand. They were observed while doing this and were briefly interviewed about their experience immediately afterwards. As a result of this exercise, it quickly became clear that each task contained a substantial amount of internal scaffolding and support which made it possible for informants to complete many of the items successfully

but without any reference whatsoever to the source text. They managed to achieve this either by exploiting the local co-text within the summary itself, or by referring to a more general context (e.g. their own experience and expectations about how a short story develops, or their own background knowledge of a topic and how it is typically treated in a newspaper article). Thus principles of local interpretation and analogy caused the summary task to more closely resemble a traditional cloze test than a test of the mental representation that results from reading the text.

For this reason, both gapped summaries were carefully revised to reduce the level of internal support supplied by the summary text. This was done by removing, for example, some of the internal cues provided by lexical substitution or cohesive markers. The removal of some internal support cues appeared to introduce a risk that subjects might be confused about how stages of the summary related to stages of the original text. To avoid this, the paragraphing structure of each gapped summary was carefully controlled to make it clear how the summary mapped onto chunks of the original source text. It was hoped this would provide appropriate additional support for the test taker to be able to match the summary readily to the original text. For example, the Text A (*Journey*) summary was divided into three paragraphs relating clearly to the main chunks of the story: *the man's situation before the car arrived*; *the events of the car journey*; and *the driver's perspective at the end*.

Unlike the narrative text, the Text B (*Anorexia*) summary did not possess the supportive scaffolding provided by the chronological constraints of the story. For this reason, the Text B summary was subdivided into six paragraphs that corresponded directly to recognisable chunks of the original source text. In addition, five of these paragraphs were given section titles or sub-headings. It was believed that this would provide subjects with a form of signposting as they drew on their understanding to complete the gaps. Interestingly, recalls had shown that readers of the newspaper editorial sometimes varied in the starting point and trajectory for their recall activity (see further discussion of this on pages 169–170). Thus it was felt desirable to clearly map the gapped summary to the source text, giving it a clear internal content sequence which matched the original starting point and trajectory of the editorial. Each section heading was carefully constructed so as not to provide inappropriate clues as to what to insert in the gaps which followed it. The inclusion of section headings was also justified on the grounds that such markers commonly occur in expository texts, including newspaper reports. However, given that the original newspaper editorial did not include subheadings, one issue for later consideration will be whether the provision of section headings in the Text B summary risked preventing some of the important propositions and provided a mental frame that would not be there in normal processing.

Further cycles of small-scale trialling were done following each successive revision of the items to check how many could now only be completed after

reading and understanding the original text. Care was also taken to ensure that for each gap in the summary there existed plausible distractors and that any obvious interdependence between gaps was minimised. Once again observation and retrospection techniques were used to investigate this. After several cycles of small-scale trialling and revision, the two summary completion tasks were considered ready for trialling with a larger population. A comparison of the gapped summary tasks is shown in Table 7.7.

Table 7.7 Comparison of Text A (*Journey*) and Text B (*Anorexia*) summary completion tasks

	Text A (*Journey*)	Text B (*Anorexia*)
Number of words in original reading text	526	389
Number of words in final constructed summary	214	250
Length of summary as % of original text	40.7	64.7
Number of constructed test items	40	39
Number of words deleted to form test items	55	54
Number of deleted words as % of summary	25.7	21.6
Number of words remaining in summary after deletion of test items	159	196
Number of items testing salient details	24	32
Number of items testing logical connectors	5	7
Number of items testing referential links	4	
Number of items testing emotional elements	7	

Trialling the summary completion tasks

Participants

The participants involved in trialling were 170 KS3 pupils (i.e. Year 9, 13/14 years old) at the same local state secondary school which took part in the written recall study. This sample population was selected because an independent measure of reading comprehension ability for the same year group was due to be available in the form of KS3 teacher assessment, thus enabling a correlational study to be undertaken later. The use of almost the entire year-group ensured that the full ability range was represented.

Materials

The two reading texts were those previously described in Chapter 5: Text A – the short story (*Journey by Night*), and Text B – the newspaper editorial (*The rights and wrongs of treating anorexia*). As in the written recall study, each reading text was accompanied by a pair of tasks. Tasks A1 and B1 were identical to those described earlier in this chapter (see Appendix 6

on pages 227–228). Tasks A2 and B2 asked participants to do the following activities: first, to answer the same three 3-option MCQs as in the written recall study (relating to the interest value, complexity of ideas and complexity of language); and second, to complete a gapped summary of the text they had just read by writing *up to three words* in the spaces provided on the sheet (see Appendices 8 and 9 on pages 231–236).

All written instructions for the tasks had been checked in consultation with an experienced secondary English teacher and test writer. Careful attention had also been given to layout issues relating specifically to each summary completion task, including typeface and size, clarity of punctuation, positioning of item numbers, and the provision of a title. The general principle was that maximum support and clarity should be balanced against minimum interference. Both the summary tasks, together with their accompanying instructions, are included as Appendices 8 and 9 on pages 231-236.

Procedures

The tasks were administered by class teachers during normal English lessons in accordance with clear written instructions (see Appendices 8 and 9). All participants completed both tasks (i.e. both Tasks A1/A2 and B1/B2) so that their reading comprehension performance across both texts could later be compared. They were given Task A1 (*Journey*) first since this was believed to be the easier of the two tasks (narratives are generally considered to be easier than expository texts (Brown 1994)). Participants were allowed 3–4 minutes to read the instructions and text. Task A1 sheets were then collected in by the class teacher and participants were given Task A2. Participants were allowed a maximum of 15 minutes to complete the questions and the summary completion task. Task A2 sheets were then collected in. Participants were next given Task B1 (*Anorexia*) and Task B2 to complete under the same conditions as for Tasks A1/A2. Appropriate time allowances required for the tasks had been estimated and confirmed through a small pilot study.

Analysis

The responses of the 170 participants to both summary completion tasks were retyped exactly as they had been written on the answer sheets in order to produce two data files. No changes were made to spelling, punctuation or grammar. The raw data files were then checked using SumCom0 (1987), one of a set of four computer programs designed specifically to help with the analysis of data from cloze or summary completion tests involving written responses of one or two words. SumCom0 was run several times to detect any data entry errors and corrections were made to the two data files where necessary.

The checked data files were then analysed using SumCom1 to produce two output files for each task: first, a *treeFile* listing in alphabetical order the various responses made to each test item, together with the frequency of each response; second, a *codeFile* identifying each different response to a single test item using a different code number. The treeFiles were then used to determine the acceptable and unacceptable answers for each test item. Since the summary completion tasks are intended to be a test of comprehension ability (rather than of grammatical knowledge or use), acceptability of answers was determined solely on the basis of appropriate meaning. No account was taken of errors of spelling, punctuation or grammatical form (of which there were many). For a small number of items only one answer was possible (e.g. *wallet*, *driver*), while for most items multiple responses, all of which could be considered as semantically comparable within the context, were judged equally acceptable for marking purposes (e.g. *spanner/tool/weapon/metal or iron tool/bar/object/pole/rod, explanations/reasons/theories*). Any queries relating to the acceptability or otherwise of responses were discussed with another experienced examiner and test constructor, though in practice there were very few of these. Once a list of acceptable answers (or answer key) had been finalised for each task, the codes for the answer key were inserted into the relevant codeFile. (The issue of multiple answer keys for summary completion tasks, and the implications of this for the take-up of this task format, will be considered in Chapter 9.)

The completed codeFile for each task was then used as input to the SumCom2 program which scored the data 0 or 1 in accordance with the acceptable codes listed in the SumCom1 codeFile. The resulting *markedFile* was analysed using Tradanal, a classical test analysis program, to produce a raw score distribution and summary test statistics. A Rasch program using the dichotomous model was also used to produce a more detailed analysis of both item and person performance. In the light of this analysis, particularly the item fit statistics, some minor adjustments were made to the key for Task A and both the classical test analysis and the Rasch analysis were redone. No changes proved to be necessary for the Task B key. Results and discussion of the analyses for both summary completion tasks are reported and discussed in Chapter 8.

Validating the summary completion tasks against another measure of reading comprehension ability

To validate the summary completion tasks against an independent reading comprehension measure, results from the teacher assessments for reading (which formed part of the national KS3 English testing regime in secondary education in England) were gathered for each of the 170 KS3 students who took part in the summary completion task study. This meant that raw scores

for the two summary completion tasks could be correlated against KS3 English teacher assessment results to investigate the relationship between scores on the summary completion tasks and on an independent measure of reading comprehension ability. The results of this analysis are also presented and discussed in Chapter 8.

8 Trialling of summary completion Tasks A and B

Introduction

Chapter 8 reports and discusses findings from the large-scale trialling of the two summary completion tasks – Task A (*Journey*) and Task B (*Anorexia*). This chapter presents score distributions, item facility values, test reliability coefficients, misfit statistics and item sequence/item misfit correlations relating to both tasks. Some of the problematic items are discussed, suggesting possible reasons for their seemingly poor performance along with ideas for potentially improving them through revision. The chapter concludes by reporting and discussing outcomes from an attempt to validate the two summary completion tasks against an independent measure of reading comprehension ability used within the national English test system in secondary schools.

Results and discussion of analysis for Task A (based on Text A – *Journey*)

Distribution of test scores

The mean score in the raw score distribution for 170 subjects was 21.14 out of a possible total score of 40 with a standard deviation of 9.15. Following minor adjustments to the key, as discussed in Chapter 7 on pages 184–185, the mean score was revised slightly to 21.70 with a standard deviation of 9.27 indicating an even healthier distribution of scores.

Item facility values

The mean facility value for the 40 test items was estimated to be 52.8 with values ranging from 12.9 (Item 26) to 80.0 (Item 32). Following minor changes to the key, the mean item facility was revised upwards slightly to 54.3, with items ranging from 12.9 (Item 26) to 81.8 (Item 3). The content and nature of individual test items can be checked by referring to Appendix 7 on page 229 and Appendix 8 on pages 231–233.

Test reliability

Rasch analysis estimated the internal consistency of the test to be 0.899 (Kuder-Richardson (KR) 20 = 0.917). Following minor changes to the key, reliability improved slightly to 0.903 (KR20 = 0.921) indicating that the test items were working in the same direction to measure the same trait.

Misfitting test takers

Rasch analysis reported only one misfitting person, Candidate 031. On closer inspection, this proved to be a test taker who had left blank a series of eight gaps in the middle of the test (Items 28–35). A possible explanation for this could be that, at this point in the story, this candidate completely lost track of the reference chains relating to the driver and passenger and then found it difficult to recover until the start of the final paragraph (Item 36).

Misfitting items

Item fit statistics are shown in Table 8.1 in accordance with the commonly accepted threshold of greater than −/+2 (McNamara 1996).

The significant negative misfit (weighted z) for Items 40, 36, 31, 30 and 33 (shaded at the top of Table 8.1) may well have resulted from a certain lack of local independence in these items. All five items occur fairly close together in the last quarter of the test. In addition, Items 31 and 33 are related in meaning (*let out*/*dropped off* and *got out*/*jumped out*). Items 36 and 40 could be linked together in as much as faulty understanding may have led to the correct answers (*the driver* and *the passenger*) being confused or interchanged. It is possible, therefore, that candidate responses to one or more of these items may have affected their performance on the others. A manual study of candidate scripts for this task also showed that in a large number of cases candidates simply inserted *He* for Item 36, thereby failing to distinguish between the two male characters in the story (*the driver* and *the passenger*) and forfeiting the mark for that item. This response was less common among the higher-scoring candidates, but frequent among lower-scoring candidates who may have been trying to mask a lack of clear understanding concerning the referential chain at this point and hedge their bets.

Significantly positively misfitting items include Items 28, 9, 1, 3 and 23 (shaded at the bottom of Table 8.1). An analysis of several high-scoring candidates suggested that these individuals had failed on three of these items through the inflexibility of the marking key, rather than through any failure of comprehension on their part. For this reason the key was extended for Items 3, 23 and 28 and the item analysis was redone. Revised item fit statistics following changes to the key for Task A for these items are shown in Table 8.2.

A comparison of the changes to the most misfitting items following amendments to the key is reported in Table 8.3. In each analysis the

Table 8.1 Summary of item statistics (Task A) – items in misfit order

Output scale is set at 1.000 * logits + 0.000				
Question name	**Threshold difficulty**	**Standard Error**	**Weighted MS**	**Weighted z**
Item 40	0.597	0.179	0.73	−4.12
Item 36	1.047	0.185	0.77	−3.15
Item 31	−0.082	0.179	0.84	−2.34
Item 30	−0.877	0.193	0.80	−2.26
Item 33	−0.632	0.187	0.83	−2.08
Item 37	1.114	0.186	0.85	−1.93
Item 21	0.850	0.182	0.87	−1.82
Item 16	0.011	0.178	0.88	−1.77
Item 22	0.980	0.184	0.88	−1.54
Item 29	−0.020	0.178	0.89	−1.50
Item 39	0.073	0.178	0.90	−1.38
Item 35	−0.335	0.182	0.91	−1.21
Item 18	−0.950	0.195	0.89	−1.16
Item 11	−0.735	0.189	0.91	−1.05
Item 20	0.754	0.180	0.94	−0.83
Item 38	−0.950	0.195	0.94	−0.65
Item 10	0.196	0.177	0.95	−0.63
Item 32	−1.661	0.222	0.92	−0.56
Item 26	2.522	0.248	0.94	−0.37
Item 13	−0.841	0.192	0.96	−0.37
Item 15	−1.520	0.216	0.97	−0.18
Item 34	−1.025	0.197	0.98	−0.16
Item 17	0.196	0.177	0.99	−0.07
Item 14	−1.566	0.218	0.99	−0.02
Item 8	−0.303	0.181	1.01	0.16
Item 27	1.114	0.186	1.01	0.17
Item 5	0.042	0.178	1.02	0.28
Item 7	−0.564	0.185	1.02	0.29
Item 19	0.442	0.178	1.02	0.33
Item 2	−0.735	0.189	1.09	1.04
Item 4	0.011	0.178	1.08	1.08
Item 25	1.182	0.188	1.10	1.22
Item 6	−0.598	0.186	1.12	1.45
Item 12	−0.335	0.182	1.15	1.87
Item 28	0.227	0.177	1.16	2.15
Item 9	1.394	0.193	1.23	2.39
Item 1	1.431	0.194	1.24	2.44
Item 3	−1.345	0.208	1.38	3.02
Item 23	0.135	0.178	1.30	3.90
		Mean	0.99	−0.25
		S.D.	0.14	1.68

difficulty scale is re-centred such that the mean difficulty is kept at zero. Since three items (Items 3, 23 and 28) were made easier through accepting more responses, most others increase in difficulty by about 0.1. This changes the test construct slightly and so the fit of every item to the construct may change slightly. Positive misfit for Items 23 and 28 has been substantially reduced and, despite being easier, these have now become better items. Interestingly, there is almost no change in the positive misfit for Item 3.

Table 8.2 Summary of revised item statistics (Task A) – items in misfit order

Output scale is set at 1.000 * logits + 0.000				
Question name	**Threshold difficulty**	**Standard Error**	**Weighted MS**	**Weighted z**
Item 40	0.691	0.180	0.74	−3.82
Item 36	1.148	0.187	0.77	−3.06
Item 30	−0.806	0.194	0.80	−2.26
Item 33	−0.557	0.188	0.83	−2.07
Item 31	0.001	0.180	0.85	−2.05
Item 37	1.216	0.188	0.85	−1.86
Item 16	0.096	0.179	0.87	−1.82
Item 29	0.064	0.180	0.90	−1.33
Item 22	1.080	0.185	0.90	−1.28
Item 23	−1.324	0.212	0.88	−1.08
Item 39	0.158	0.179	0.92	−1.08
Item 35	−0.255	0.183	0.92	−0.99
Item 18	−0.880	0.196	0.91	−0.93
Item 11	−0.662	0.190	0.92	−0.85
Item 21	0.754	0.181	0.94	−0.83
Item 10	0.221	0.179	0.95	−0.69
Item 26	2.650	0.250	0.91	−0.56
Item 38	0.850	0.182	0.96	−0.55
Item 20	0.850	0.182	0.96	−0.52
Item 24	−0.880	0.196	0.95	−0.51
Item 32	−1.601	0.224	0.94	−0.44
Item 13	−0.769	0.193	0.97	−0.26
Item 15	−1.459	0.218	0.99	−0.06
Item 34	−0.956	0.199	0.99	−0.06
Item 27	1.216	0.188	1.00	0.04
Item 14	−1.505	0.220	1.00	0.06
Item 28	−0.769	0.193	1.01	0.17
Item 17	0.283	0.179	1.01	0.19
Item 8	−0.223	0.182	1.02	0.29
Item 7	−0.488	0.187	1.04	0.54
Item 19	0.533	0.179	1.05	0.74
Item 5	0.127	0.179	1.06	0.81
Item 2	−0.662	0.190	1.10	1.15
Item 4	0.096	0.179	1.09	1.18
Item 25	1.250	0.189	1.11	1.25
Item 6	−0.522	0.187	1.14	1.64
Item 12	−0.255	0.183	1.17	2.06
Item 1	1.539	0.196	1.25	2.47
Item 9	1.501	0.195	1.28	2.81
Item 3	−1.753	0.232	1.43	2.82
		Mean	0.98	−0.27
		S.D.	0.14	1.48

Correlating item sequence with item difficulty and item misfit

A lack of correlation between item difficulty and item misfit (−0.034, subsequently revised to −0.054) suggests that no spurious effects were caused by guessing.

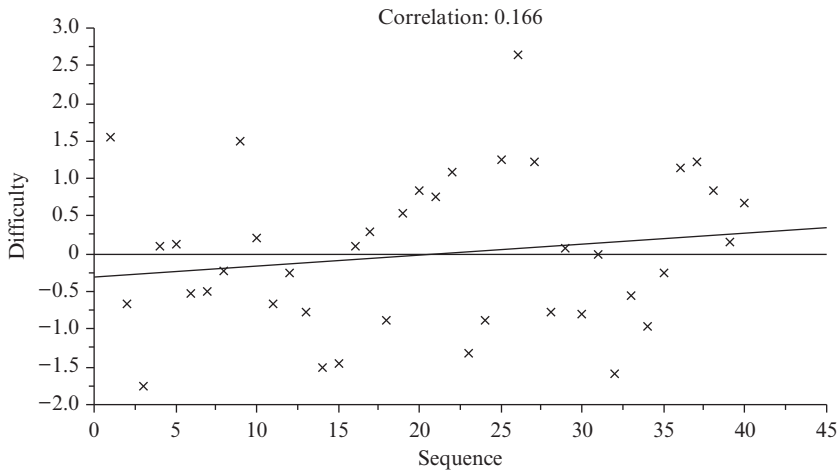
Table 8.3 Comparison of changes to misfitting items (Task A)

Output scale is set at 1.000 * logits + 0.000				
Question name	**Threshold difficulty**	**Revised Diff**	**Weighted z**	**Revised weighted z**
Item 40	0.597	0.691	−4.12	−3.82
Item 36	1.047	1.148	−3.15	−3.06
Item 31	−0.082	0.001	−2.34	−2.05
Item 30	−0.877	−0.806	−2.26	−2.26
Item 33	−0.632	−0.557	−2.08	−2.07
Item 6	−0.598	−0.522	1.45	1.64
Item 12	−0.335	−0.255	1.87	2.06
Item 28	0.227	−0.769	2.15	0.17
Item 9	1.394	1.501	2.39	2.81
Item 1	1.431	1.539	2.44	2.47
Item 3	−1.345	−1.753	3.02	2.82
Item 23	0.135	−1.324	3.90	−1.08

The correlation between item sequence and item difficulty was initially calculated to be 0.18, indicating only a slight tendency for items to get harder as the test progressed. Following changes to the key, this coefficient was reduced slightly to 0.166 as shown in Figure 8.1, probably as a direct result of Items 3, 23 and 28 having become easier.

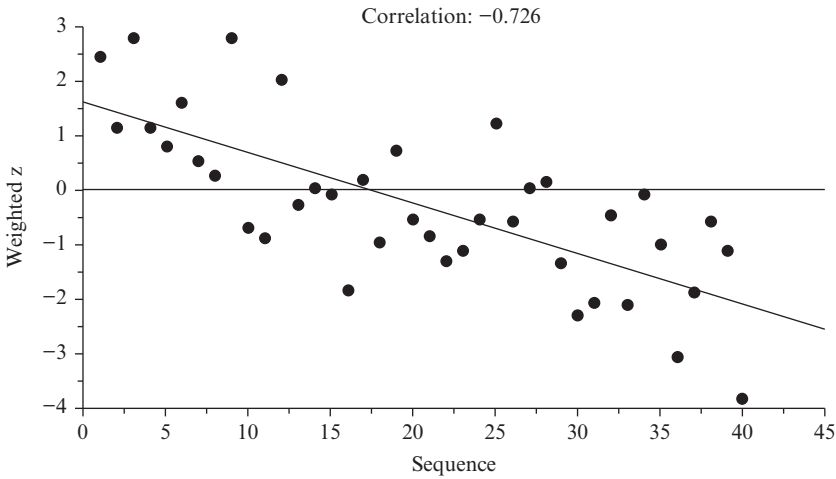
The correlation between item sequence and item misfit, however, was calculated to be −0.60, indicating higher misfit closer to the beginning of the test and a tendency towards overfit as the test progresses. On average, the items in this test are overfitting. This result suggests that weaker candidates found

Figure 8.1 Correlation of item sequence and item difficulty (Task A)



the later items in the test more difficult to complete than the earlier items. An explanation for this could be a type of jigsaw effect. In other words, just as in a jigsaw puzzle the later pieces of the puzzle are usually easier to insert than earlier pieces because the picture is already partially assembled, so it may be that as more and more of the gaps in a text are cumulatively completed, the easier it will become to complete the remainder. It could perhaps be argued that this corresponds precisely to the nature of comprehension, i.e. initial construction of meaning by readers needs to be secure enough if further building is to take place on top of this. If stronger candidates complete the earlier gaps successfully (as they are likely to do), then they will invariably have a better chance of completing the remainder successfully. Weaker candidates, on the other hand, may struggle to complete the early gaps successfully and as a result find it increasingly difficult to complete the remaining gaps. This effect may be exacerbated if deletions are too numerous and too close together and this may contribute to item interdependence. It is possible that, as a result of being unable to complete successive gaps, demotivation will increase among weaker candidates and thus contribute to depressing the results still further. Changes to the key caused the correlation to increase to −0.73 as shown in Figure 8.2 and this may have been due to a substantial reduction in misfit for Items 23 and 28.

Figure 8.2 Correlation of item sequence and item misfit (Task A)



Conclusions

Test reliability for Task A can be considered fairly good for a 40-item test (0.92) and, in general, item difficulty for Task A was well matched to the ability of the student group. The mean ability of the sample group was estimated at

0.145 (increasing to 0.233 after key changes), suggesting the sample group to have been slightly more able than the average item.

Most of the test items in Task A functioned well in terms of fit with only a small number showing significant misfit or overfit. Changes to the key clearly improved fit for Items 3, 23 and 28, and even though Items 1, 3, 9 and 12 are still positively misfitting, the degree of misfit is not extreme. The negatively misfitting items (Items 30, 31, 33, 36 and 40) suggest excessive interdependence between these items and one solution might be to remove some or all of these items from the task in the hope of improving it. However, Items 30, 31 and 33 register a misfit value of slightly above 2 which is not extreme and could probably be tolerated. Items 36 and 40 present more of a problem since both have a more serious misfit value of above–3. One possible improvement to these two items would be to insert the article *The* before Item 36, thus discouraging the response *He* which may have contributed to misfit for this item. The item would then read as follows: *The* (36) . . . . . *hurried away with a sense of* (37) . . . . . Unfortunately, lack of time and resources did not allow for further trialling and analysis of Task A within the context of this study. The lack of correlation between item difficulty and item misfit suggests that guessing had little effect on performance for Task A. Furthermore, the very small correlation between item sequence and item difficulty suggests that factors such as lack of time and onset of fatigue are unlikely to have significantly influenced performance. However, the high correlation between item sequence and item misfit raises an important issue of the extent to which test items across a task such as this can be considered as locally independent. This issue and its implications for summary completion tasks and other text-based sets of reading test items will be discussed further in Chapter 9.

Results and discussion of analysis for Task B (based on Text B – *Anorexia*)

Distribution of test scores

The mean score in the raw score distribution for 170 subjects was 12.87 out of a possible total score of 39 with a standard deviation of 8.25, suggesting that subjects found Task B considerably more difficult than Task A. Possible reasons for this differential performance across the two tasks are discussed below.

Item facility values

The mean facility value for the 39 test items was estimated to be 33.01, with values ranging from 4.7 (Item 4) to 66.5 (Item 26).

Test reliability

Rasch analysis estimated the internal consistency of the test to be 0.881 (KR20 = 0.915), once again indicating that the items were working together as a coherent set.

Misfitting candidates

The analysis reported only two misfitting test takers – Candidates 049 and 050. Closer inspection of the responses by these two candidates suggested that in both cases they had incorrectly answered several groups of consecutive items. Candidate 049 provided incorrect responses to the following item sequences: 7–11, 13–14, 18–19, 27–30, 30–33 and 37–38. Candidate 050 performed in a similar way but on different item sequences: 4–6, 8–9, 15–19, 25–30 and 36–38. Just as a series of unanswered items may have caused one candidate to misfit on Task A, so it may have been this same pattern of response, i.e. sequences of incorrectly answered items, which caused these two individuals to register as misfitting on Task B. This suggests a degree of item interdependence (see further discussion of this below).

Misfitting items

Item fit statistics for all 39 items are shown in Table 8.4.

The significant negative misfit (weighted z) for Items 21, 22, 23 and 24 (shaded at the top of Table 8.4) is most likely to have been caused by a lack of local independence in these items. All the items are mid-range in difficulty, but all four occur within the same sentence and are closely linked in terms of the propositional information they represent: *Otherwise extreme sufferers would not be able to* (21) ***look at*** *themselves in* (22) ***the mirror*** *and* (23) ***see*** *someone* (24) ***obese***. A study of some of the incorrect responses to these four items, as well as a study of individual candidate scripts, indicated that in several instances candidates left all four of these gaps blank. On other occasions, however, subjects chose to complete this set of four items using sets of alternative answers as follows:

> *Otherwise extreme sufferers would not be able to*
> (21) ***care for/control/cope with/cure/fend for/feed/help (etc.)***
> *themselves in*
> (22) ***bed/college/confined spaces/everyday life/home/hospital (etc.)***
> *and*
> (23) ***around/at home/help/need/hurt/cause/think/tell/that (etc.)***
> *someone*
> (24) ***can die/care for them/else/has to help (etc.)***

These two patterns of behaviour (i.e. leaving all four gaps blank or providing a set of alternative incorrect answers) were observed for both higher and

Table 8.4 Summary of item statistics (Task B) – items in misfit order

Output scale is set at 1.000 * logits + 0.000				
Question name	**Threshold difficulty**	**Standard Error**	**Weighted MS**	**Weighted z**
Item 21	0.305	0.203	0.75	−2.75
Item 23	0.784	0.222	0.70	−2.68
Item 24	0.784	0.222	0.72	−2.51
Item 22	0.470	0.209	0.79	−2.09
Item 9	0.737	0.220	0.79	−1.87
Item 12	−1.213	0.183	0.86	−1.83
Item 29	−1.820	0.190	0.85	−1.80
Item 27	−0.822	0.183	0.91	−1.28
Item 30	−1.820	0.190	0.90	−1.22
Item 15	1.666	0.277	0.79	−1.21
Item 10	−0.658	0.184	0.91	−1.19
Item 32	0.882	0.227	0.87	−1.01
Item 14	−1.377	0.184	0.93	−0.83
Item 37	1.522	0.266	0.86	−0.83
Item 31	0.386	0.206	0.92	−0.78
Item 18	2.543	0.369	0.81	−0.68
Item 34	−1.927	0.192	0.94	−0.65
Item 19	1.997	0.307	0.89	−0.46
Item 28	0.985	0.232	0.94	−0.43
Item 11	0.556	0.212	0.95	−0.41
Item 35	−1.049	0.183	0.97	−0.35
Item 36	0.225	0.200	0.98	−0.19
Item 26	−2.074	0.196	0.99	−0.04
Item 39	−0.252	0.189	1.00	−0.03
Item 4	2.682	0.388	0.97	0.01
Item 16	−1.750	0.189	1.00	0.06
Item 20	−0.287	0.189	1.03	0.34
Item 5	0.345	0.204	1.03	0.37
Item 6	0.882	0.227	1.05	0.44
Item 25	−0.356	0.187	1.04	0.49
Item 2	0.187	0.199	1.05	0.53
Item 38	0.645	0.216	1.06	0.54
Item 33	0.784	0.222	1.10	0.82
Item 3	−0.356	0.187	1.09	1.17
Item 17	−0.356	0.187	1.11	1.32
Item 1	−1.477	0.185	1.12	1.53
Item 13	0.225	0.200	1.26	2.53
Item 8	−0.217	0.190	1.33	3.65
Item 7	−1.785	0.190	1.44	4.53
		Mean	0.97	−0.23
		S.D.	0.15	1.53

lower scoring subjects. What seems clear from the range of alternative but incorrect answers for these four items is that in many cases candidates sought to relate and make sense of the local co-text through their insertions into gaps 21 to 24. It seems likely that if candidates succeeded in answering Item 21 correctly, they would probably have activated the correct concept and so have

little difficulty in providing correct answers for the remaining three items (i.e. *looking in* ***the mirror*** *and* ***seeing*** *someone* ***obese***). If, on the other hand, candidates supplied an incorrect answer for Item 21 at the outset, then they would probably be activating a different concept and would be unlikely to arrive at suitable answers for the remaining three items. Thus it could be argued that Items 21 to 24 are actually functioning as a single item. Once again, this raises the issue of deletions being potentially too numerous and too close together.

Significantly positively misfitting items include Items 7, 8 and 13 (shaded at the bottom of Table 8.4). Items 7 and 8 are seriously overfitting and there may be several reasons for this. In general, surprisingly few candidates succeeded in expressing the change in attitudes to anorexia described through the sentence containing Items 4 to 8. What they were expected to do was to draw out the contrast between past and present attitudes to anorexia as follows:

> *Changing attitudes*
> (4) ***Although*** *anorexia* (5) ***used to be*** *regarded as a* (6) ***slimming problem****, it*
> (7) ***is now*** *generally accepted as a* (8) ***medical condition****.*

However, on many occasions subjects simply reiterated in Items 7 and 8 the information they had already provided in Items 5 and 6. Interestingly, the first item in this group, Item 4, is by far the most difficult item in the test with a facility value of only 4.7 and an estimated threshold difficulty of 2.682. This item – *Although*/*Though* – was one of the logical connectors introduced into the summary to compare and contrast past and present attitudes to anorexia expressed in the original text. The subsequent Items 5 and 6 are significantly easier, with facility values (difficulty estimates) of 25.3 (0.345) and 18.2 (0.882) respectively. It is possible that if candidates did not immediately grasp the contrastive nature of this sentence as expressed through the sentence-initial *Although*, then this made successful completion of Items 7 and 8 more difficult. A total of 64 different responses to Item 4 were offered across 170 candidates and incorrect answers offered by several subjects included such responses as: *Before, Once, Ten years ago, Generally, Normally, The* and *Her*. Clearly, the nature of the word required in this gap was not immediately obvious to candidates so it is possible that Item 4 may have significantly affected performance on Items 7 and 8. Another example of a sentence-initial *Although* occurs in Item 28. The estimated difficulty of this item (0.985) and its misfit value (–0.33) are well within acceptable limits, but Item 28 is still the sixth most difficult item in Task B, with a facility value of only 17.1. It is possible that test items which target sentence-initial words or phrases may carry with them an additional difficulty for that reason and should be avoided. (Interestingly, Taylor (1991) observed a similar finding in a small study of cloze completion.)

Even if candidates failed to insert *Although* at Item 4, some subjects may have succeeded in identifying the references to past and present attitudes, but

chose to express the present attitude through Items 5 and 6, and the past attitude through Items 7 and 8 (i.e. their responses were in the opposite order to the order in which these ideas appeared in the original text and were thus designed to appear in the summary). Although it could be argued that the comprehension of these subjects is accurate, the design of the marking key was unable to accommodate and credit this pattern of response behaviour. This highlights the need for test constructors to anticipate and predict, as far as they are able to, those occasions when test takers may provide correct responses but in reverse order to that which is expected in the key. Such instances will require a more flexible marking key if such behaviour is to be properly credited.

Item 13 is only slightly overfitting and has a facility value (difficulty estimate) well within normally acceptable limits 27.1 (0.225). An analysis of responses to this item indicated that a large number of candidates (including high-scoring test takers) had inserted *and*, *or* or *so* into this gap instead of the required *because*. The original text clearly states that *an unacceptably stressful or difficult adult life* is a cause of patients *trying to retreat into childhood or avoid leaving it*. For this reason, no extension was made to the key.

In the case of Task A, simple changes to the scoring procedure immediately improved the quality of several misfitting items. Necessary amendments to the summary also appeared relatively straightforward. For Task B, however, the problems relating to misfitting items were more complex and it was clear they required more than just re-marking and re-analysis of subjects' performance. Lack of time and resources prevented further investigation (i.e. through revision, re-trialling and re-analysis) within the context of this particular study, but factors contributing to the misfit undoubtedly merit further exploration. One useful way of exploring some of the issues raised above would be to undertake a protocol analysis of test takers as they complete the summary completion task. This might offer valuable insights into specific problems when processing the test items as well as provide further evidence to support claims of construct validity for the task.

Correlating the sequence of items with item difficulty and item misfit

A lack of correlation between item difficulty and item misfit (0.058) suggests that no spurious effects were caused by guessing. The correlation between item sequence and item difficulty was calculated to be −0.042 suggesting that the test does not become more difficult as it progresses (see Figure 8.3).

The correlation between item sequence and item misfit, however, was calculated to be −0.338, once again indicating higher misfit towards the beginning of the test and a tendency towards overfit as the test progresses, probably as a result of excessive interdependence in later items (see Figure 8.4).

Although this correlation is not as high as it was for Task A, it nevertheless

Figure 8.3 Correlation of item sequence and item difficulty (Task B)

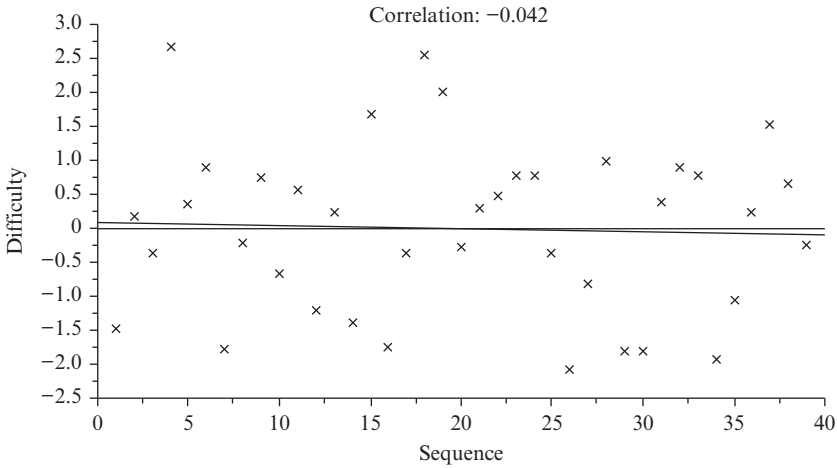
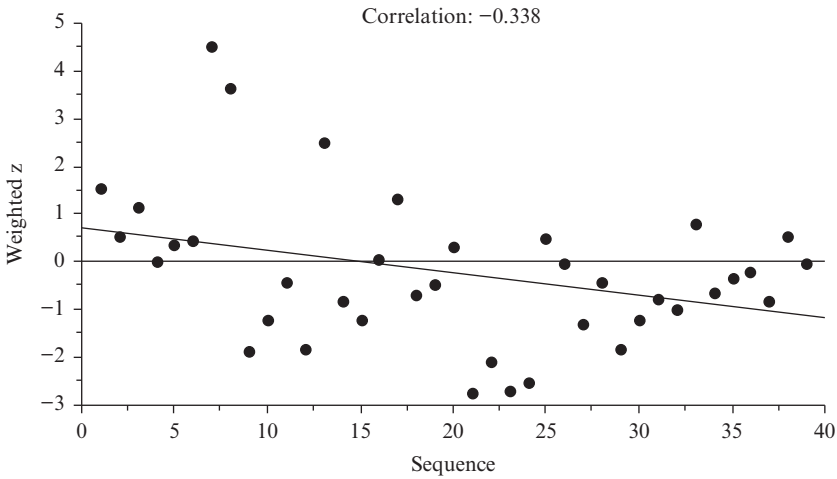


Figure 8.4 Correlation of item sequence and item misfit (Task B)



suggests that weaker candidates once again tended to find the later items in Task B more difficult to complete than the earlier items. Recalculating the correlation after excluding the highly misfitting Items 7 and 8 reduced the coefficient only slightly to −0.231.

Conclusions

Test reliability for Task B is generally acceptable. However, the mean ability of the sample group was estimated at −1.029, suggesting the sample group to

have been a logit below the average difficulty of the items. Thus Task B was less well matched to the ability of the student group than Task A.

Most of the test items in Task B functioned well in terms of fit with only a small number showing significant misfit (i.e. above +/−2). The significantly negatively misfitting items (Items 21, 22, 23 and 24) clearly suggest a lack of local independence between these items and one approach would be to remove some or all of these items completely. However, none of these items is registering extreme misfit and an alternative solution might be to retrial Task B having reinstated Item 21 in the summary to provide an unambiguous cue for the remaining three items. This might have the effect of reducing negative misfit for Items 22, 23 and 24. The degree of positive misfit for Item 13 is not extreme (2.35) and this item could either be retained as it is or reinstated in the summary. Items 7 and 8 present more of a challenge since both have an extreme misfit value of well over 4. One possible solution would be to reinstate Item 4 in the summary, i.e. insert *Although* at the start of the sentence, so that the contrastive nature of the sentence becomes immediately clear. Another possibility would be to add appropriate temporal markers (such as *ten years ago*/*nowadays*) to the relevant clauses in this sentence in order to cue the past and present attitudes to anorexia in the correct position. Some attempt has been made in this section to suggest why the items behaved as they did and how they might be improved, but more adequate data is required (possibly by means of verbal protocol analysis) for any firm solutions to be determined.

The lack of correlation between item difficulty and item misfit suggests that guessing had little effect on performance on Task B. Furthermore, the very small correlation between item sequence and item difficulty suggests that factors such as lack of time and onset of fatigue are unlikely to have significantly influenced performance. However, the negative correlation between item sequence and item misfit partially replicates the finding for Task A and once again raises the issue of the extent of interdependence of test items across an entire task, which will be discussed in Chapter 9.

Comparing the relative difficulty of Tasks A and B

Trialling results indicated that test takers found Task B rather more difficult than Task A. Several reasons for this can be offered. Narrative texts are generally recognised as being easier to understand than other types, e.g. expository texts (Brown 1994). This is probably because the latter type often deals with a topic and content that make heavy propositional demands on the reader, in terms of lexical, syntactic and discoursal complexity.

The relative difficulty or 'readability' of any reading text can be measured in various ways by examining its lexical, syntactic and discourse features (Alderson 2000, Carrell 1987, Khalifa and Weir 2009, Weir 2013a). Readability measures typically take account of a variety of textual

characteristics such as word and sentence length, lexical frequency and density, and clausal complexity. An analysis of lexical content, using software packages such as Compleat Lexical Tutor/Web VocabProfile (Cobb 2006) can provide some indication of text difficulty as occasioned by lexical complexity in terms of frequency; in general the more frequent the word in written texts (BNC) the easier it is to process. Table 8.5 compares Texts A and B across a selection of readability measures.

Table 8.5 Readability measures for Texts A and B

	Text A (*Journey*)	Text B (*Anorexia*)
Number of words	526	389
Number of paragraphs	18	5
Number of sentences	53	16
Sentences per paragraph (mean)	2.9	3.2
Words per sentence (mean)	9.9	24.4
Characters per word (mean)	4.1	4.9
Type-token ratio	0.46	0.56
Lexical density	0.37	0.41

According to these measures at least, the two texts clearly present readers with differing levels of challenge in terms of their respective lexical and syntactic demands. Text A lies towards the easier end of the continuum, while Text B shows itself to be much more demanding and difficult to understand. In Text A sentence length is much shorter and there are many more of them, while Text B is characterised by fewer and much longer sentences which will be more demanding in terms of processing. The type-token ratio (TTR) indicates the level of lexical variety or diversity within a text, on a scale of 0–100. A higher TTR indicates a larger amount of lexical variation and a lower TTR indicates less lexical variation. There is a 10% difference in the TTR across the two texts, suggesting that the Text B difficulty probably lies in its vocabulary as well as its syntactic complexity and conceptual organisation. This small-scale comparative analysis makes the case for the importance of establishing the contextual as well as the cognitive comparability of texts if tests are to be targeted at the same level.

Validation of Tasks A and B against an independent measure of reading comprehension ability

Correlation of Task A with Task B

Raw scores for Tasks A and B were first of all correlated to produce a Pearson product moment coefficient of 0.694, suggesting a moderately high correlation between the two tasks. If both tasks are testing reading ability, then it is

tempting to ask why the correlation between them was not higher. One explanation may be that there was a significant difference in how the two tasks performed, perhaps partly because the difficulty of Text B and its items caused some candidates to become distracted or to give up. On the other hand, evidence from other studies suggests that correlations between tasks based on *different* texts are not necessarily as high as might be hoped. For the TEEP study, for example, Weir (1983) reported correlations mostly in the 0.6s with nothing above 0.75 (1983:1,127–1,129). The implication of this is that it suggests a strong text effect and confirms the importance of paying attention to contextual validity parameters if we seek to have parallel texts and tasks at the same level (see Chapter 4 in Khalifa and Weir 2009 for an extensive discussion of this). In relation to comparability of cognitive demands made on the reader, it would be interesting to explore further, perhaps through a retrospective study, whether more summary propositions were tested in B than in A. If so, it is likely that a higher processing ability was involved for Task B and this threatened the comparability of the two tasks in cognitive validity terms.

Correlation of individual Tasks A and B with KS3 English teacher assessments

Raw scores for Tasks A and B were individually correlated with the KS3 English teacher assessment levels for Reading to produce a Pearson product moment coefficient. The results are reported in Table 8.6.

Table 8.6 Simple correlation with KS3 English teacher assessments for Reading

KS3 English teacher assessments for Reading correlated with:	
(i) Task A raw scores	0.734
(ii) Task B raw scores	0.757

Coefficients (i) and (ii) suggest a reasonably strong positive relationship to exist between KS3 teacher assessments for Reading and each summary completion task. This relationship appears to be marginally stronger for Task B. These results provide evidence for summary being a test of reading insofar as we can rely on teacher assessments being a measure of the reading construct (see more on this below). Interestingly, the correlations reported here are higher than those reported between the TEEP and external indicators by Weir (1983:497).

Figures 8.6 and 8.7 plot the nature of the relationship for (i) and (ii). These show a slight curve in the case of both Task A (Figure 8.5) and Task B (Figure 8.6). This is more marked in the case of Task B where there is some evidence of a floor effect.

Figure 8.5 Simple correlation of Task A scores with KS3 English teacher assessments for Reading

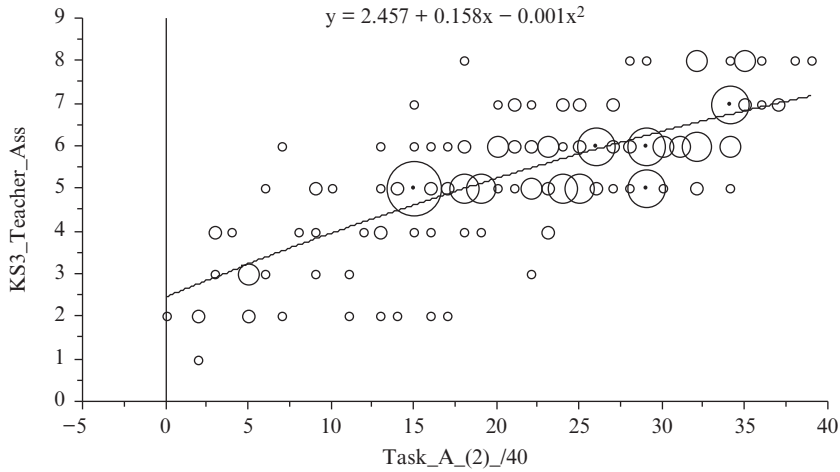
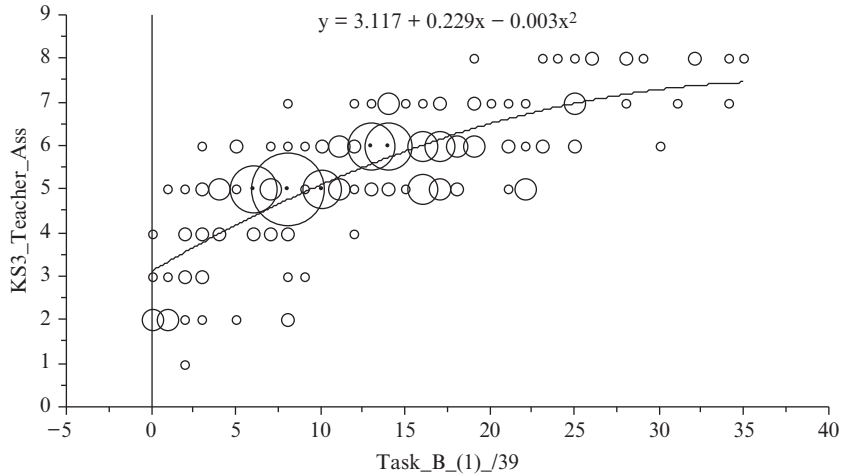


Figure 8.6 Simple correlation of Task B scores with KS3 English teacher assessments for Reading



Since a simple correlation coefficient is known to underestimate a relationship if it is at all curvilinear, a binomial (second order) regression technique was used to correct for attenuation and to describe the relationships more accurately – see Table 8.7 which shows an improvement in fit for Task B.

Table 8.7 Second order correlation with KS3 English teacher assessments for Reading

KS3 English teacher assessments for Reading correlated with:	
(i) Task A raw scores	0.737
(ii) Task B raw scores	0.774

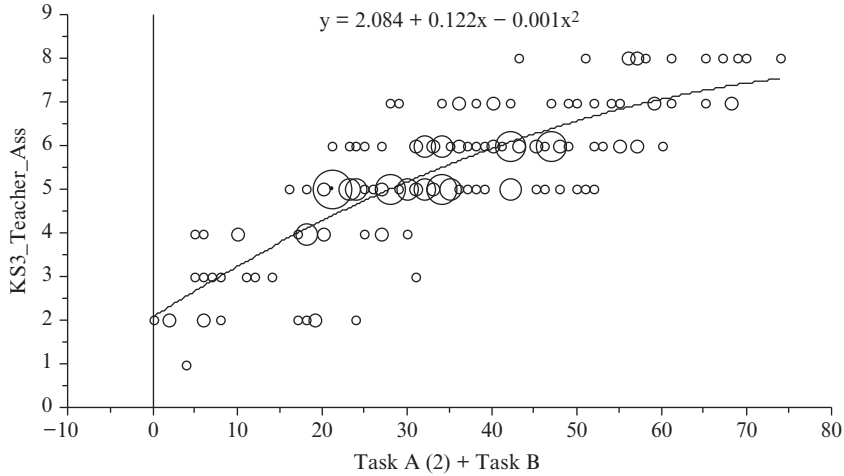
Correlation of combined Tasks A and B with KS3 English teacher assessments

Since an increase in the number of observations contributes to increased reliability and validity, raw scores for Task A (40 items) and Task B (39 items) were aggregated to produce a set of new combined scores out of 79.

Separate internal consistency reliabilities for Tasks A and B (0.90 and 0.88 respectively) generated a composite test reliability of 0.947 using the Spearman-Brown formula. Using the KR20 reliability indices for Tasks A and B (0.921 and 0.915), the reliability estimate for the composite test improved further to 0.957. This of course assumes both tasks to be measuring the same trait, though we should not be naive about potentially significant differences across the two tasks in terms of various characteristics. One such characteristic is their linguistic features, e.g. sentence length, lexical range, syntactic complexity. A more detailed computational analysis of the two tasks, e.g. using Coh-Metrix, might well highlight other lexical, syntactic and conceptual features which differentiate them. Tasks are well-known to vary in difficulty according to their genre and rhetorical function, and subject knowledge and topic familiarity play their part. It would also be interesting to explore to what extent the two summary completion tasks differ in their processing levels. For example, are they comparable in terms of their balance of macro- and micro-propositional content? Is there more of the latter in B? Did the insertion of some elements to aid cohesion actually make processing more difficult? There is undoubtedly scope for further research into multiple linguistic and conceptual aspects of the two gapped summary tasks, as well as the source texts, using more recent tools developed for more sophisticated text and task analysis.

Detailed reliability figures for KS3 assessment were not available at the time this study was conducted, but were generally believed to be in the region of 0.9. If this is so, then a correlation coefficient of 0.77 or above can be considered very high. Disattenuated for the effects of unreliability, it is estimated to be about 0.84. Such results could be seen as particularly impressive for a test of no more than 40 minutes in length.

Figure 8.7 Simple correlation of combined Task A and B scores with KS3 English teacher assessments for Reading



The new combined scores for Tasks A and B were correlated with the KS3 English teacher assessments for Reading to produce the following results, as shown in Figure 8.7:

Pearson product moment coefficient:	0.809
Second order coefficient:	0.821

Conclusions

Whether used individually or together, summary completion Tasks A and B appear to offer a reliable instrument for assessing reading comprehension ability. Teacher assessment is frequently used as a criterion for test validation and for this reason the KS3 English teacher assessment for Reading was selected as an appropriate validation criterion for this study. KS3 English teacher assessments for Reading were based upon a broad range of reading behaviour by candidates throughout the school year, including (though not limited to) the reading of literary classics such as Shakespeare. Summary completion Tasks A and B are of course not capturing candidates' ability to read and understand Shakespeare and it is therefore hardly surprising that the correlation is slightly lower. The fact that it is as high as 0.821 suggests that, taken together, the pair of summary completion tasks represents a valid approach to assessing reading comprehension ability.

The data in Table 8.8 suggests that both summary completion tasks are extremely comparable to the KS3 English teacher assessments.

Table 8.8 Correlation with KS3 English teacher assessment levels for Reading

KS3 English teacher assessment levels for Reading correlated with:	
(i) Summary Completion Tasks A and B	0.821

Conclusion

Chapter 8 has reported and discussed the results of trialling the two summary completion tasks – Task A (*Journey*) and Task B (*Anorexia*). Chapter 9 will draw together some conclusions from the theoretical and empirical research reported in this volume. Drawing upon the experience described in Chapters 5 to 8, in particular, the concluding chapter will seek to offer some practical guidance for test developers on a methodology for constructing summary completion tasks to assess reading comprehension ability. The final part of the chapter will highlight some possible directions for future research in this area.

9 Conclusions and recommendations

Introduction

The motivation for the research reported in this volume was a desire to explore the viability and construct validity of a reading test format which seeks to address directly the reader's understanding in the form of the mental representation generated by a text. The background to the study was a concern that traditional reading test formats too often fail to take adequate account of what we now understand about the constructed and cumulative nature of comprehension across a text as a whole. There remains a tendency in many reading comprehension tests to focus test items on information at the factual level, or at best at the level of interpreting the writer's intentions, rarely tapping into processing at the level of discourse representation. Decisions about choice of test format and test item focus are generally based upon the individual connoisseurship of the test writer, without much reference to how most readers would read a particular text (or set of texts) or what discourse level representation of the text(s) they would be likely to construct. Furthermore, many widely used reading test formats have the potential for interfering substantially with the reader's natural processing of text. Having to cope with a complex comprehension test format such as MCQ, for example, can seriously impact on the comprehension process itself. As a result, the construct validity of many traditional reading tests is called into question.

Chapters 2, 3 and 4 chronicled the evolution of reading and text comprehension theory and reviewed the theoretical basis for much past and current practice in reading test design. Text-removed summary completion technique was proposed as a test format which might reconcile more closely the practice of assessing reading comprehension ability with current theory about the nature of text comprehension. The empirical research reported in Chapters 5, 6, 7 and 8 explored multiple aspects of readers' mental representations of two different texts, drawing on these to develop two text-removed summary completion tasks which were trialled to establish their validity as appropriate measures of reading comprehension ability.

The research findings reported and discussed above suggest that it is indeed possible to design text-removed summary completion tasks which

can function as effective measures of reading comprehension. Summary completion tasks A and B both achieved high levels of test reliability as well as an acceptable level of correlation with a recognised independent reading measure, indicating them to be a valid approach to assessing reading comprehension ability.

It is clear, however, that, just as the production of a well-functioning cloze task is not as straightforward as many suppose it to be, so the generation of a good summary completion task involves taking account of various complex and interrelated factors. It was intended that one outcome of the empirical investigations reported above would be some practical guidance for test designers on how to set about constructing a well-functioning text-removed summary completion task. Thus Chapter 9 will seek to draw together some guidelines and recommendations for test developers on the key issues of text selection, summary construction and test item construction. The final part of this concluding chapter will consider some possible directions for future research in relation to using summary completion tasks to assess reading comprehension ability.

Selecting a suitable text for designing a summary completion task

A major consideration when designing a text-removed summary completion task must be the *selection of a suitable text*. General principles of text selection are already well-established for the design of traditional reading test formats such as cloze and multiple choice, and issues of linguistic complexity (lexical and syntactic), conceptual familiarity, cultural appropriateness and interest value are obviously considered as a matter of course when choosing reading passages for test purposes. There are two additional issues, however, which assume considerable importance when selecting texts for text-removed summary completion tasks. The first of these is the *internal structure and length of* the original source text, and the second is the likely *context and purpose for reading and understanding* the text.

With regard to the first of these issues, i.e. *internal structure and length of text*, experience gained during the present study suggests that where a text contains within it a clear linear and/or chronological thread, e.g. in a predominantly narrative text type such as Text A (*Journey*), readers will find it considerably easier to construct a full and coherent mental representation of the content. It seems more difficult, however, for readers to construct a well-ordered and coherent mental representation of a text that is more loosely constrained, i.e. a descriptive, discursive or expository text such as Text B (*Anorexia*). This may suggest that a narrative-type text is more suitable for testing at lower proficiency levels, while an argumentative text is appropriate for higher levels. It resonates with the distinction that Khalifa and Weir (2009)

drew between the testing of reading up to B1 level of the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001), and testing of reading from B2 and beyond.

From the test developer's perspective, it also seems easier to construct a satisfactory summary of a text that possesses a clear linear or chronological thread, as compared with a text which may have a logical organisation within it but which has a much less rigidly constrained internal framework (whether in terms of chronology, or a set of sequential processes, or even stages of argumentation). The preferred solution, one imagines, will be to choose a text with a clear structure in the first place, or one with as clear a structure as possible. If it proves necessary, the test designer can choose to impose a more explicit external framework upon the constructed summary by introducing section titles or paragraph headings. These can act as organisational markers to help readers match the summary to the original source text (as was done in the case of Task B). While it could be argued that this approach (i.e. the introduction of subheadings) may not necessarily align with the ordering of content as mentally represented by some readers, this risk needs to be balanced against the need to ensure that all key content is accounted for somewhere in the summary, irrespective of the order in which it might be structured in readers' minds, and that all readers can access this content relatively easily so as to map their own understanding onto it.

In addition to internal structure there is the question of the *length* of the selected text. It is accepted that both Texts A and B were relatively short (526 and 389 words respectively). Although text length (and thus content) was clearly sufficient for readers to construct a mental representation of each text and for this to be accessed for comprehension assessment purposes, it would be interesting to see how this might change with the use of *longer* texts. This is an important consideration given that, at higher proficiency levels, and particularly in academic study contexts, both L1 and L2 readers are likely to be encountering much longer texts as a rule, possibly texts containing several thousand words (Engineer 1977, Weir 1983). We might imagine that for such lengthy texts (e.g. research papers or book chapters), a reader's typical mental representation is likely to be at a much higher level of abstraction and generalisation, with far less detail at the micro-propositional level, and that this will need to be reflected in any summary of that text devised for comprehension assessment purposes. Similar considerations will apply if a summary is constructed to represent content from more than one textual source, i.e. an inter-textual summary, though the presence of micro-propositions from each source text may be salient here if they are critical for distinguishing the contribution of each source text to the summary as a whole, e.g. if the summary addresses a theme or problem that is shared across two or more texts but also combines differing scenarios or solutions that are proposed by different texts.

Awareness of the implications of text type and internal structure may

therefore be an important factor when selecting a text, or texts, for developing a summary completion task.

A second consideration must be the simultaneous choice of an appropriate and plausible context and purpose for reading and understanding the text. At the very least, careful thought needs to be given by test developers to why a reader would most probably read a given text in the first place, what sort of mental representation they would typically construct as a result of reading and what purpose(s) their understanding of this text would be made to serve (e.g. extracting evidence to support an argument, or critically reviewing a piece of writing). The aim of the test designer will be to communicate this sense of context and purpose explicitly through the instructions or rubric for the test task (or some other form of task setting). This will hopefully provide readers with the necessary contextualisation to guide their reading activity and constrain their mental representation. A clear statement of the purpose for which readers are being asked to read and understand the text will assist the test designer not only in guiding the construction of the summary but also in determining an appropriate focus for constructing valid test items according to what is salient and consistent with the intended purpose for reading. For English for Academic Purposes (EAP) students, for example, the purpose will most likely be to read to extract salient information for completing an assignment. Readers will need to extract the main ideas from the text and understand the relationships between these.

In the trialling study reported in Chapter 7, readers' mental representations of the two texts were constrained by asking them to consider how suitable they considered Texts A and B for use in KS3 tests for 13/14-year-olds. In high-stakes tests, it will be important to consider even more carefully the design of an appropriate reading purpose and context since this is likely to impact directly on the nature of the task processing and consequently the validity of the test.

Constructing a suitable summary for developing a summary completion task

Once a suitable text and a compatible reading context and purpose have been determined, the next stage of design will be to construct an appropriate summary of the chosen text. The nature of the summary will be partly shaped by the stated purpose for reading the text, but more importantly by the way in which readers actually process and understand the text in question.

Several possible methods for constructing a summary were identified and explored empirically in this volume, all drawing directly upon a detailed analysis of the mental representations constructed by readers, as revealed by their oral and written recalls (see Chapters 6 and 7). Recall studies of two different texts demonstrated that both summarising *macro-propositions* and text-based

micro-propositions are likely to be needed to construct an adequate summary and that these will need to be identified and selected from the pool of all the summarising and text-based propositions available for a given text. Readers typically recall some propositions more readily than others and evidence suggests that they can achieve a good level of agreement on which are the most salient propositions for a text (and thus which are peripheral or superfluous). Some propositions are consistently recalled by a group of readers when reading the same text for the same purpose, while others are not. Relative frequency of occurrence of propositions in readers' recalls of a text should be one of the **key** things to consider when identifying those propositions to be included in an adequate summary of it. For Text A the most promising summary resulted from combining summarising and text-based propositions recalled by 50% of readers. For Text B the most promising summary resulted from combining the high-frequency summarising and text-based propositions (i.e. the top 33% most frequently occurring summarising and text-based propositions). We might speculate here on why these two different approaches emerged from the exercise and what this may imply about the likely potential of different text types and their treatment by the test developer. Narratives, such as Text A, are likely to be more readily understood thanks to a more familiar rhetorical structure which is easier to process. In addition, the language is likely to be syntactically and semantically simpler. Thus a higher level of agreement might be expected in readers' recalls. Expository texts, on the other hand, are generally considered harder to process and their lexical and syntactic complexity is likely to be greater, presenting a greater cognitive load to the reader. Additionally, they tend to be propositionally more demanding. Thus a 33% cut-off point for high-frequency summary and text-based propositions is probably more realistic. (As Khalifa and Weir (2009) observed in their analysis of the Cambridge English Reading tests, narrative reading passages feature largely at the A2 and B1 levels while expository texts are reserved for B2 level and above.)

It would clearly be impractical (and probably unnecessary) to attempt the sort of sample sizes used for the oral and written recalls in the research reported above, but it would be quite feasible to gather recalls from a small group of well-chosen individuals, as well as from the test designer. A practical and effective initial approach to summary construction could involve asking several independent readers to read the selected text once through in accordance with the agreed context and purpose (and under the same time constraints that would be faced by the candidate in the test) and then to recall it, either orally or in written form. The recall could take the form of a list of main ideas with some key supporting details, maybe in the form of a diagrammatic mapping of the salient content (as discussed on page 76). This set of parallel recalls from several different readers, all of whom were reading for the same reason and under comparable conditions, could then

be used to identify the most salient points in the text which must be included in a summary, as well as those parts of the original source text which are less memorable and can therefore justifiably be discarded. By way of example, Weir et al (2000:173–178) reported the successful use of text-mapping as a practical way of achieving this in the context of high stakes language testing for academic purposes.

The next design stage involves the test developers in ordering the selected summarising propositions according to the paragraphing and topic-shift structure of the original source text. Summarising propositions can then be amplified by inserting the selected text-based propositions into the appropriate places. The addition of text-based propositions provides explanation and exemplification, and helps to link together the summarising propositions more cohesively. It also helps to ensure the necessary length and level of detail within a summary so that a sufficient number of test items can be generated.

At this point, the inclusion of authorised inferences in the summary needs to be considered. Given the variable and often idiosyncratic nature of inferences, test developers should be cautious when incorporating inferences into a summary, limiting themselves to those necessary inferences which are mandated by the text (e.g. linked to referential chains, superordinate goals, or thematic content such as the point or moral of the piece). It is clear that readers make a wide variety of elaborative inferences to enrich their mental representation, thus making it more coherent or more memorable. Recalls of Texts A and B showed that such inferences can include material relating to a main character, a location/setting, an atmosphere, a cause or reason for something, as well as other types. However, any attempts by the test developer to make potentially ambiguous inferences explicit through the summary of a text may well conflict directly with the mental representation constructed by some readers, and this may be akin to the test method effect that is created by some distractors in multiple-choice comprehension questions. For this reason, it is safer for test constructors to limit any inferences to matters of comprehension rather than interpretation. It is also clear that readers can vary somewhat in the evaluative inferences they make regarding interpretation of the writer's stance or point of view. Again it may be wise to avoid including an explicit statement of the writer's point of view unless this is unambiguous in the original.

Once an appropriate set of summarising and text-based propositions (plus authorised inferences) has been decided upon, consideration will need to be given to editing these to generate a cohesive and fluent text. A summary containing a large proportion of words and phrases from the original text may be easier to match to its source, but this may also mean that the test items can be completed too easily from verbatim recall. The use of extensive paraphrasing may make the summary less easy to match to its source, but this may also guarantee test items that depend upon understanding rather than verbatim

recall of the original. If extensive paraphrasing is used, it seems sensible to consider dividing and organising the summary clearly into paragraphs or sections that correspond obviously to chunks or topic shifts in the original text. It also seems advisable to check that the language used in the summary is syntactically and lexically less complex than that used in the original text, thus ensuring that it is not the language of the summary that is being tested. Fortunately, automated text analysis tools nowadays make this relatively easy to do (see Weir et al 2012 and Weir 2013a for a full listing and discussion of these).

One significant difference between the summaries constructed for Texts A and B was their length relative to the original source texts: the Text A summary was proportionately shorter than the Text B summary. This may relate directly to the density of propositions in Text B and the fact that it is difficult to condense discursive text and still retain a meaningful summary. A narrative text such as Text A, however, often contains a substantial amount of material that is included for stylistic purposes, e.g. to create atmosphere, and which can be excised without much impact. It is easier to trim the text down and still retain an accurate and coherent summary of events. This finding is worth bearing in mind when selecting texts from different genres and constructing summaries of them.

Throughout the process of summary construction test developers are advised to keep in mind the following four key principles:

1. The summary must *accurately reflect the content of the original* text.
2. The summary must be *a coherent and standalone text in its own right*.
3. The summary must *include elements/propositions from the original text which most readers would find salient* as well as *some elements/propositions which might cause weaker readers difficulty*.
4. The summary must be *long enough to be capable of generating a large enough number of locally independent test items* to meet the demand of test reliability.

The critical issue remains the extent to which the resulting summary is compatible with the stated or likely purpose for reading and understanding the original text. If it is not compatible, then test takers may encounter difficulty in mapping the summary onto the mental representation they have constructed for themselves through reading and the validity of the summary completion task as a whole will be compromised.

Constructing a set of test items for a summary completion task

Once an accurate and coherent summary of the text has been generated as described above, the final stage of the test task design process will be to construct a set of suitable test items, formed by deleting single words and short

phrases from the constructed summary. The number of items to be constructed may depend partly upon the length of the summary and how many items it can reasonably support, and partly on whether the task is intended to stand alone as a single reading test or will form part of a larger battery of test items.

In the empirical study reported in this volume, the tasks were designed with 40 and 39 items respectively and the large number of items in each task undoubtedly contributed to the high reliability of the tasks. However, it was estimated that it would have been possible to reduce the number of items in each task to around 30 (or even 25) and still achieve acceptable reliability figures. This is encouraging since it may mean that after trialling or pretesting of the initial task some of the less well-functioning test items can be discarded.

The words or phrases selected for deletion to form the test items should relate to the salient content features of the text and the selection of these can be guided both by the content of the summary itself (i.e. key words or phrases that are present in the summary text) and by the experience of investigating readers' mental representations preceding the summary construction. As suggested above, this investigation could be achieved by means of collecting a number of oral or written recalls (as demonstrated in Chapters 6 and 7), or through mind- or text-mapping exercises such as those described by Weir et al (2000). Identification of salient points or information may need to vary according to text type. When testing understanding of a narrative text, the test items may be best targeted at key aspects such as referential chains for characters, objects and actions, superordinate goals and actions in an unfolding drama, the thematic content or point of the story, together with some of the stronger emotional colouring where this is fundamental to the action in the narrative. In an expository text, the test items may focus more appropriately on the implied relationships between ideas, e.g. cause/effect, compare/contrast, as well as on salient people, objects, events, actions or concepts.

Appropriate item selection underpins the validity of the task in terms of the cognitive processing involved, since the gaps in the summary should test only those aspects of the text that are salient when the text is read for the specified purpose. Test items should not be targeted at convenient trivia or at peripheral features that are irrelevant to the intended reading purpose. Ideally, the test task as a whole should cover all of the important aspects of the text consistent with the stated reading purpose and, normally, every salient point should be made the subject of some gap in the summary. In effect, this means that gaps are most likely to centre upon nominal or verb phrases, but it is also possible to target core logical connectors that reflect the rhetorical structure of the text or the relationship between ideas within it. Some gaps may be matched directly to words and short phrases in the source text, while others may reflect a degree of paraphrasing or necessary inferencing.

Oral or written recalls of a text gathered from less able as well as proficient readers can usefully reveal points where comprehension breaks down and this may also help to highlight the location for or the nature of appropriate deletions to create test items. Examples of incorrect or faulty inferences can offer useful insights into how and where comprehension can break down, and thus provide guidance for constructing appropriate test items in the summary completion task. However, test developers still need to check that such content points are central and salient to overall comprehension of the text rather than just peripheral details of little importance.

A key consideration in designing the test items is that it should never be possible to complete a gap correctly on the basis of the surrounding co-text or by drawing on background knowledge or prior expectations of discourse (and the measures suggested below should be employed to confirm this). The correct response to each item should ideally only be possible as a result of reading and understanding the original source text. Furthermore, it should be possible to complete each gap with at least one or two plausible distractors. The syntactic structure of the gap and its surrounding co-text needs to be transparent to test takers, but since the emphasis is upon *meaning* rather than form, it should be clear to test takers that acceptable responses may take the form of more than one word (unlike most cloze tests) and that there may be multiple correct responses which can be used to complete any gap.

The possibility of multiple correct responses invariably makes the process of marking the completed summary tasks considerably more challenging, although it would presumably be possible to develop a comprehensive mark scheme by analysing a proportion of the candidate responses before confirming the final answer key. If constraining the response format is a priority or a necessity for reasons of administrative ease and convenience, then consideration can always be given to converting the summary completion task to a multiple-matching format, sometimes referred to as a 'banked choice, gapped summary task' (Alderson 2000:242).

When the test items have been constructed, they will need to be pretested or, at the very least, trialled by asking some informants to complete the task without reading the original text. It will soon become clear if there are items which can be correctly answered on the basis of common sense or cues from the local co-text. The potential for acceptable alternative responses may also be highlighted in this way. Of particular concern will be any test items which are interdependent in some sense, i.e. where completing one gap correctly or incorrectly automatically makes another item easier or more difficult to answer. Following trialling, amendments may be required to individual items and to the answer key and items which prove to be highly interdependent may need to be removed. Alternatively, a cluster of such items may need to be treated and scored as a 'testlet'. In the case of large-scale testing (e.g. for public examinations), this initial trialling phase will normally be followed up

with a formal pretesting phase to establish the difficulty of individual test items and of the task as a whole.

Directions for future research

The theoretical and empirical research reported in this volume has been able to offer only a preliminary investigation into the rationale for and practical issues surrounding the use of text-removed summary completion tasks for assessing reading comprehension ability. While the initial findings are encouraging, there are clearly several important issues which require or merit further investigation. Some of these are discussed below.

Investigating variations in summary task format

For practical reasons, the scope of this study was restricted to the design of textual summaries in continuous prose. It is important to be aware, however, that there exist alternative formats for summarising a text. A text could be summarised as a set of notes, for example, and some texts (e.g. those describing linear processes or sets of characteristics) lend themselves to summarisation in the form of a diagram, table or flowchart. A major advantage of these latter formats is that the verbal content of the summarising process is reduced to a minimum and the summary may therefore become easier to construct and to mark. If readers are asked to summarise a text as a picture or series of pictures, then a word-based summary could be avoided altogether. A disadvantage with such approaches, however, could be that it will be more difficult for test takers to match the summary format to the original text and the summary framework may therefore need careful structuring with headings and numbering. Since it is possible for a single text to be summarised in a number of different ways, one interesting avenue for research would be to investigate the potential variation in test takers' performance on different summary completion task formats derived from the same text (e.g. a verbal summary, a set of pictures or illustrations, a summary diagram or flowchart) especially to explore whether certain text genres lend themselves more to one summary format than another.

Another issue worthy of further investigation is the potential difference in performance on a summary completion task depending upon whether the text remains present throughout the task or is removed immediately after reading. The purpose of removing the text after reading is to force readers to build a discourse representation which they then have to draw upon to complete the gapped summary. The alternative condition according to which the text remains present after reading and during completion of the gapped summary may well result in a much easier gapped completion task, involving mainly matching techniques. While this may be appropriate at lower

proficiency levels, it is unlikely to offer useful insights into discourse comprehension ability at the higher levels.

Similarly, it may be instructive to compare the difference in summary completion task performance when the text is presented as a reading activity and as a listening activity. Brown and Yule (1983) commented on the possibility that memory is modality-specific, i.e. that different memory representations for the same text result depending upon whether the text is encountered in the spoken or written mode. The use of text-removed summary completion tasks may be one way of exploring this notion further.

Where computer-based test administration is possible, the text-removed condition for summary completion may be especially appropriate since it simplifies the process of removing the reading text before administering the gapped summary task. In addition, computer delivery of the test may enable the task format to be explored in a variety of other ways. For example, the delivery of the reading text via computer screen makes it possible to manipulate the reader's access to the reading text in terms of speed and form of presentation. It would be interesting to investigate at what point time pressure exerted upon the reader begins to cause a breakdown in comprehension and then to explore the nature of any such breakdown. Many reading tests focus strongly on assessing 'careful' reading but pay only lip service to the testing of 'expeditious' reading due to the practical challenge of how to control the time readers allocate to their reading of a lengthy text (see Khalifa and (Weir 2009:45–47) for a full discussion of careful and expeditious reading). Expeditious reading skills are critical at the higher proficiency levels, especially in academic study and professional contexts, and computer-based text-removed summary completion tasks may offer an effective means of testing such skills.

Investigating the effect of context on processing

An important issue requiring further investigation is the effect of context on reader processing. It would be possible, for example, to give the same text to two groups of readers, each group having been presented with a significantly different context and purpose for reading. Comparison of reader recalls might show up the differences between the shared representations constructed by the two groups and a follow-up summary completion task could also be used to explore the effect of context and purpose.

It would also be interesting to study to what extent different text types lend themselves to varying representations depending upon context and purpose for reading. One might predict, for example, that there would be less potential for manipulating the reading context and purpose of a straightforward narrative text like Text A where the content is highly constrained. A more discursive text like Text B, however, may offer greater scope for manipulating

reading purpose and context and consequently generate variations in readers' mental representations, and one can see how this might be used to test various aspects of expeditious reading, e.g. reading a text quickly to find points against a proposal or solutions to a problem.

Current technology opens up various avenues for exploring the nature of text processing and the construction of meaning. For example, eye-tracking studies during the reading of the original source text could be correlated with similar studies conducted during completion of the summary completion task to further investigate the way readers read text to construct meaning and then map this onto an external meaning representation.

Properly developed, summary completion tasks may prove to be a valuable instrument for investigating various issues of comprehension. It may even offer a more appropriate tool for research purposes than those which are currently used since much research into comprehension makes use of traditional reading tests that were designed primarily for educational assessment rather than research purposes.

Investigating measurement issues

From a measurement perspective, the negative correlation between item sequence and item misfit observed in both summary completion tasks (but especially in Task A) is an issue which clearly requires further investigation. (See Pollitt and Taylor 1997 for some additional discussion of the measurement issues observed in this research.)

Test models generally assume ability to be fixed rather than to change during a test. For each of Tasks A and B, however, it would seem that the better the task is completed by a test taker, the easier it becomes. Equally, the worse the task is completed, the harder it becomes. It may that the mental representation continues to change as the task is being completed. Those whose representation is developing well are helped by the task, while those whose representation is failing are not helped. There is some previous discussion in the literature of how responses to early items in a test can affect comprehension of the rest of the text. Gordon and Hanauer (1995) explored in some detail the theory that the testing task itself functions as an information source to affect the ongoing construction of the test taker's mental model. Multiple-choice questions appear to be particularly prone to overstimulating comprehension and thus assisting the answering process. It is possible that these effects are frequently present in other types of comprehension test but that test development has in the past not been sufficiently detailed to allow them to be seen as clearly as in Tasks A and B. For this reason, it would be worth exploring more systematically to what extent these 'positive feedback' effects occur in other reading test formats, such as cloze tasks or sets of text-based multiple-choice questions.

A fundamental technical issue relating to summary completion concerns the extent to which the assumption of local independence of items is violated within summary completion (and probably other text-based test formats). Both classical approaches and latent trait models assume test items to be locally independent. The over-consistency of the items in Tasks A and B, however, suggests that the items in a summary completion task may not be functioning independently. In some cases, this effect may be cumulative so that a text of 40 items actually functions more as a single-item test with 40 marks. In other cases, the breakdown of local independence may become focused in subsets of items which receive greater weight than more discrete, independent items. Task A seems to be rather like the first of these cases, while Task B is similar to the second.

The result of such effects may be an increase in test reliability leading to the problem of the test appearing more reliable than it ought to. The paradox appears to be that in trying to make a reading test more valid (by reflecting more closely the process of comprehension), certain measurement characteristics risk being violated. The critical issue is to what extent a partial violation of measurement characteristics is a matter for concern. Once again, it could be that over-consistency of items is a feature of not only summary completion tasks, but of any integrated test of comprehension.

Investigating the use of summary completion in L2 reading assessment

Finally, although this study restricts itself to investigating text-removed summary completion to assess L1 reading comprehension ability, the potential application of this test format in assessing L2 reading comprehension ability is equally important. Text-removed summary completion is likely to prove a useful assessment tool among L2 learners in general, but it may have a special application in the growing field of young learner assessment where there is a need for tools that take account of the cognitive development of the test takers. A related issue for investigation in this context could be the relative merits of constructing the summary completion task for younger learners in the L1 rather than the L2.

Conclusion

As demonstrated through the literature review in Chapters 2 and 3 of this book, the past three decades have seen regular calls from within the reading research and language testing professional communities for more thorough investigation of the reading comprehension process together with systematic validation of both existing and new tests for assessing reading comprehension. A fundamental aim of the empirical research reported in subsequent

chapters of the volume was to bring together the two areas of applied linguistic theory – *reading* and *testing*.

If creating a text-level representation constitutes the highest level in a socio-cognitive processing model of reading, as recently proposed by researchers in reading assessment such as Enright et al (2000), Cohen and Upton (2006) and Khalifa and Weir (2009), then it seems plausible to suggest that a task that embodies a discourse-level structure of a text could serve as a useful instrument for measuring a skilled reader's ability to recognise the hierarchical structure of the whole text, how the different parts of it fit together and which parts are important to the writer or to reader purpose. Summarising tasks are generally considered to engage readers in precisely this sort of high-level processing because they require readers to identify and organise information that is key to overall meaning, sifting main ideas from supporting details and integrating these into a discourse structure that is consistent with writer/reader purpose.

A text-removed summary completion task avoids many of the disadvantages associated with traditional summarising tasks. The format first assumes and then evaluates the reader's ability to construct a text model representation of what is read and to form a relevant situation model, integrating and connecting the detailed information provided by the text into a coherent whole. In seeking to efficiently align text model and situation model, the design of text-removed summary completion tasks outlined in this volume represents a theoretical attempt to reconcile our current understanding of the nature of reading comprehension with the current demands of assessment and measurement theory. Furthermore, the findings from the empirical research reported above suggest that text-removed summary completion is capable of producing an assessment tool in which we can have confidence and whose results can be interpreted to draw valid and meaningful conclusions about reading comprehension ability.

Appendix 1

Text A: *Journey by Night* (short story)

Journey by Night

He stood alone, leaning against a post, and shifting his weight from one foot to the other. It was late, and the taxi-stand was empty. The street was silent. He looked up and down, hoping that some vehicle would come in sight, for he wanted to get home. But none came.

The silence began to pall. He started to whistle, but there was no mirth in it, and he soon stopped. Midnight, ten miles away from home! What was he to do? To begin to walk that distance was out of the question.

A dark cloud passed across the sky, hiding the few pale stars that had been there. The noise of a falling dust-bin reached his ear. Some dog must have been scattering its contents.

Instinctively his hand felt for his wallet. Yes, it was still there. If only he had a stick! But he had nothing with which he might protect himself. He began to walk up and down, up and down.

What was that in the distance? At last two headlights were drawing near. He stepped into the middle of the street and held up his hand, and the car stopped.

' Taxi?' he asked. 'Valencia?'

'Get in,' said the driver, opening the door.

He sat beside the driver, glad to be on his way home at last. He had felt so lonely while he had been waiting. If only someone would say something! In the semi-darkness of the car he turned to look at the other passengers, but no one else was there.

The driver said nothing to him as the car sped along.

Suppose . . .

No, he mustn't allow himself to think of that. He glanced at the driver, and again his hand went to his wallet. He had heard of passengers being attacked at night and robbed. But surely . . . No, that couldn't happen to him.

If only he could see the other man's face clearly! But he had no idea who the driver was. He kept his eye intently on him during the seemingly interminable journey.

Now they were approaching a spot where the road branched off in another direction. There were tall, dark bushes around. The car slowed down, and the driver was looking at him. Then the driver took something short and black from the side-pocket of the car. It looked like an iron tool. Would the driver attack him with that?

'Stop!' he heard himself screaming, and his heart beat so fast with fear that he could hardly breathe.

But the car did not stop. Faster and faster instead it went. Now they were nearing his destination. Did the driver intend to take him past and then . . .

'Put me down here,' he cried out.

Still with his eyes on the driver, he quickly stepped from the car as it came to a standstill. He fumbled in his wallet for his fare, but the taxi was no longer there.

'No night passengers for me again,' exclaimed the driver, as with a sigh of relief he hurriedly moved off. And his hand tenderly caressed the heavy spanner with which he had meant to defend himself had that queer passenger attacked him.

Source: Giuseppi, U (1973) Journey by Night, in Giuseppi, N and Giuseppi, U (Eds) (1973) *Backfire – A Collection of Caribbean Short Stories*, Macmillan Caribbean.

Appendix 2

Text B: *The rights and wrongs of treating anorexia* (newspaper editorial)

The rights and wrongs of treating anorexia

THE CASE of Samantha Kendall, the anorexia nervosa sufferer who discharged herself from hospital despite doctors' fears for her life, has highlighted the confusion in public thinking about this disturbing and perplexing disease. Ten years ago anorexia was still dismissed as nothing more than slimming gone too far. Today it is recognised as a treatable medical condition; but the degree to which treatment should be carried out without the patient's consent has become a topic of debate.

Researchers have suggested two psychiatric explanations behind the onset of anorexia. One is that the patient, faced with an unacceptably stressful or difficult adult life, is trying to retreat into childhood or avoid leaving it. Another is that choosing what to eat — and specifically choosing not to eat — is often an attempt to exert control by people who feel that their lives are too constrained in other ways. But the truth is that for all the resources that have been devoted to its study, the syndrome remains imperfectly understood.

It is beyond doubt, however, that anorexia is a severe psychiatric disorder. There is no other way to describe an illness that allows a patient to look in the mirror at her own emaciated, starved body, and see someone obese staring back. Severe sufferers often deny that they are trying to kill themselves, but the diet they are pursuing is all too likely to make death inevitable.

The 1983 Mental Health Act provides for sufferers from severe psychiatric disorders to be held in hospital for treatment against their will if there is a danger that they will do harm to themselves or others. Yet even though one in 10 anorexia sufferers dies, doctors are sometimes reluctant to use their powers under the law. This is often because of a fear that treatment by compulsion is self-defeating, since force-fed victims of anorexia often return to starvation diets when they get home.

There is clearly work to be done in making the treatment of extreme anorexia — which often involves leaving patients in isolation and without their clothes, and watching them as they eat and go to the lavatory — more humane. But the shortcomings of the available treatments should not obscure the fact that the alternative to treatment can sometimes be death. If doctors made more use of the powers available to them, lives could be saved.

Source: The rights and wrongs of treating anorexia, *The Independent*, 18 May 1994.

Appendix 3

Research protocol for the oral recall study of Texts A and B

Researcher: Thank you for agreeing to help me with this project.
I have been asked by the local examinations board to gather together some reading materials which could be used with Key Stage 3 students (that is 13 to14-year-olds) for the purpose of topic and discussion work in class.
Before deciding on the final choice of materials, I want to collect reactions to some of the texts and topics from slightly older students so we can be confident the texts will be appropriate for use with the intended age-group.
I'd like to ask you to read two short texts today.
What I'd like to know for each one is: firstly, how accessible you think the subject matter would be to students in the 13 to 14 age-group; and secondly, how you think students would react to the text after reading it. For example, how interesting do you think they would find it, and would it encourage them to think about and discuss the topic in question?
I'm going to record what we say for easy reference later. Is that OK with you? Do you have any questions?
Here is the first text for you to read through at your own pace.
(*hand informant text to read through at own pace*)

Researcher: Before we talk about the text, can I ask you first to fill in a few details for my own records on this sheet.
(*remove text – informant fills in brief details on age, A-levels, general interests*)

Researcher: Now let's just make sure you can remember what the text is about. Can you recall the details of the short story/newspaper article for me please?
(*free recall of text by informant – recorded*)

Researcher: I'd like to ask you a few more specific questions about the text.
(*probe recall questions as appropriate – response recorded*)

Researcher: So how accessible do you think the subject matter of this text would be to 13 to 14-year-olds?
(*informant responds – recorded*)

Researcher: And how do you think students would react to the text after reading it? Would it stimulate thinking and discussion?
(*informant responds – recorded*)

Researcher: Let's turn to the second text now – on a rather different topic. Here is the text for you to read.
(*hand informant text to read through at own pace*)

Researcher: Before we talk about the text, can I ask you to look at this list of topic areas for Key Stage 3. Could you please tick the topic area on the list which you think this text best fits into?
(*remove text – informant decides on best topic area*)

Researcher: Now let's just check you can remember what the text is about. Can you recall the details of the short story/newspaper article for me please?
(*free recall of text by informant – recorded*)

Researcher: I'd like to ask you a few more specific questions about the text.
(*probe recall questions as appropriate – response recorded*)

Researcher: So how accessible do you think the subject matter of this text would be to 13 to 14-year-olds?
(*informant responds – recorded*)
And how do you think students would react to the text after reading it? Would it stimulate thinking and discussion?
(*informant responds – recorded*)

Researcher: That's the end of the task. Thank you very much for your help.

Appendix 4

Probe questions for the oral recall study of Texts A and B

Text A: *Journey by Night*

1. Can you recall anything more of the scene at the start of the story?
2. Can you recall anything more about the car stopping to pick the man up?
3. Can you recall anything more about the passenger's thoughts and feelings after he got into the car?
4. Can you recall anything more of what happened during the journey?
5. Can you recall anything more about the car stopping?
6. Can you recall anything more of the driver's feelings at the end of the story?
7. Can you recall the title of the short story?

Text B: *The rights and wrongs of treating anorexia*

1. Can you recall anything in the editorial about a person called Samantha Kendall?
2. Can you recall anything in the editorial about how attitudes to anorexia have changed over the years?
3. Can you recall any explanations for anorexia given in the editorial?
4. Can you recall any effects of the illness on the patient which were described?
5. Can you recall anything mentioned about the current legal position on treating anorexia sufferers?
6. Can you recall any methods of treatment for extreme anorexia which were mentioned?
7. Can you recall the writer's point of view on treating extreme anorexia sufferers?
8. Can you recall the title of the editorial?

Appendix 5

Sample transcript from oral recall study

Transcript of a Text A recall (Subject HR015)

(**Note**: During analysis of the transcripts for the presence of text-based and/or summarising propositions, the use of synonymous expressions, paraphrasing, indirect speech, approximations, was considered acceptable, as well as verbatim rendition of the propositions concerned.)

In the following sample transcript, the use of underlining designates the presence of text-based propositions while the use of upper case designates the presence of summarising propositions. Inevitably, there was occasionally some degree of overlap between the two types of proposition.

Can you recall the details of the short story for me in your own words?

right + um + THERE WAS A MAN standing + obviously NEEDING TO GET HOME + late at night + um + probably quite WORRIED THAT + BECAUSE HE'S ON HIS OWN HE HASN'T GOT ANY BACKUP + QUITE WORRIED HE MAY GET ATTACKED + um + he doesn't want to walk all the way home + waiting for a taxi but none seems to appear + then finally he sees headlights of a car and thinks right + it's a taxi + thinking + not thinking straight + HE SUDDENLY THINKS RIGHT + TAXI + HOP IN + YOU KNOW + SAFE WAY TO GET HOME + um + then + THE PARANOIA IS STILL THERE and + HE'S STILL PARANOID OF ATTACK whatever + he's sat in the car + um + GETTING INCREASINGLY WORRIED THAT THIS TAXI DRIVER'S GOING TO TURN ROUND AND ROB HIM + um + he sees THE TAXI DRIVER PULLING + um + SOMETHING out of his pocket but he doesn't actually know what it is + but obviously he presumes the worst and presumes he's going to be attacked whatever + screams to the driver stop + um + for some reason IT DOESN'T + I'm not sure why + um + and then + then he asks the taxi driver to put him down there where wherever they are + um + GETS OUT OF THE CAR + um + presumably still to have to walk home and + er + the taxi driver mutters something about no more night passengers for me thinking that + you know + he'd actually pulled out a spanner to defend himself because HE WAS THINKING THAT THE PASSENGER WAS GOING TO ATTACK HIM + THAT SORT OF PARANOIA

Can you recall anything more of the scene at the start of the story?

um + IT WAS DARK + um + I have the impression that it was misty + I don't remember for certain but I have the impression it was + it gave across that impression + um + fairly quiet + NO-ONE ELSE AROUND + probably very late at night + early morning + um + um + no

Can you recall anything more about the car stopping to pick the man up?

um + yeah + he saw the headlights coming and hoped it was a taxi + it was ASKED TO GO SOMEWHERE + VALENCIA + AND THE TAXI DRIVER JUST SAID GET IN + so he did

Can you recall anything more about the passenger's thoughts and feelings after he got into the car?

he was thinking it's + you know + it's happened before as in being attacked + passengers being attacked as opposed to the drivers + um + and HE WAS JUST GETTING INCREASINGLY WORRIED THAT HE WAS GOING TO GET INTO SOME SORT OF TROUBLE

Can you recall anything more about what happened during the car journey?

um + drove along + they + he was going quite fast + they + oh they passed what would be a turning in the road and there was lots of high bushes + tall bushes + and IT WAS VERY KIND OF DESOLATE + OUT IN THE COUNTRY + no-one could see them or whatever + um + and the taxi driver pulled out this spanner but the passenger didn't know what it was

Can you recall anything more about the car coming to a stop?

um + I don't think so + um + no

Can you recall anything more of the driver's feelings at the end of the story?

um + he was I think quite relieved to have got rid of this + um + this passenger who HE THOUGHT WAS GOING TO ATTACK HIM + um + I don't remember any specific words that were used

Can you recall the title of the short story?

um + night journey + or something to do with that

Total number of text-based propositions = 32
Total number of summarising propositions = 17

Appendix 6

Task instructions for written recall study of Texts A and B

INSTRUCTIONS FOR TASK A1/B1

Please read the short story/newspaper editorial below.

While you are reading the short story/newspaper editorial, think about whether it would be all right to use as a reading passage for a Key Stage 3 English test.

For example, do you think the ideas in the passage or the language it is written in would cause any difficulty for 13/14-year-olds?

When you have finished reading, you will be asked a few questions about the passage.

INSTRUCTIONS FOR TASK A2/B2

1. Answer questions (i) to (iii) by putting a circle round A, B or C.

(i) Would 13/14-year-olds find the short story/editorial
- A very interesting?
- B quite interesting?
- C not at all interesting?

(ii) Were the ideas in the short story/editorial
- A generally difficult to understand?
- B sometimes difficult to understand?
- C generally easy to understand?

(iii) Was the language of the short story/editorial
- A generally difficult to understand?
- B sometimes difficult to understand?
- C generally easy to understand?

INSTRUCTIONS FOR TASK A2/B2 (cont.)

2. In the space below, please try to **write down as much of the short story/ newspaper editorial as you can remember**. You can write in short sentences or in notes, and you can use your own words. It is more important to write down all that you can remember from the short story/editorial than to use correct grammar and spelling. If you run out of space here, please continue on the other side of this sheet.

. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .

3. Now try to sum up in one or two sentences what the writer was trying to say in the short story/newspaper editorial.

. . . . . .
. . . . . .
. . . . . .

Appendix 7

Summaries of Texts A and B derived from readers' oral recalls

Summary of Text A (*Journey*)

The following summary of the short story was constructed by combining summarising and text-based propositions recalled by 50% of readers in the oral recall study.

A man was standing alone waiting late at night in a dark and lonely place. The man wanted to get home. The noise of a falling dustbin reached the man's ear and the man began to be concerned for his own safety. Instinctively the man's hand felt for his wallet.

A car came along and stopped. "Valencia?" asked the man. The man and the driver exchanged words briefly and the man got in and sat beside the driver. The man started to think about being attacked. The man had heard of passengers being attacked at night and he grew increasingly suspicious of the driver. If only the man could see the other man's face clearly. Then the driver reached down and took something short and black from the side-pocket of the car. It looked like an iron tool. Was the driver going to attack him with that? The man panicked and his heart beat so fast with fear that the man cried out: "Put me down here!" The car came to a standstill and the man got out quickly. The man fumbled in his wallet for the fare but didn't manage to pay.

The driver drove away hurriedly with a sigh of relief. "There'll be no more night passengers for me again," exclaimed the driver. The driver had been afraid of being attacked and had meant to defend himself if that strange passenger had attacked him.

Summary of Text B (*Anorexia*)

The following summary of the newspaper editorial was constructed by combining the high-frequency summarising and text-based propositions (top 33%) recalled by readers in the oral recall study.

Samantha Kendall is an anorexia nervosa sufferer who discharged herself from hospital. There have been changes in recent years in the way anorexia is regarded. Ten years ago anorexia was still dismissed as nothing more than slimming-gone-too-far. Today anorexia is recognised as a medical condition. The degree to which treatment should be carried out without a patient's consent has become a topic of debate.

Researchers have suggested two psychiatric explanations behind the onset of anorexia. One explanation is that the patient is trying to retreat into childhood or is trying to avoid leaving childhood. Another explanation is that choosing what to eat is often an attempt to exert control by people who feel their lives are constrained in other ways. The syndrome remains imperfectly understood.

Anorexia is certainly a severe psychiatric disorder. The illness allows a patient to look in the mirror at their own emaciated body and to see someone obese staring back.

The law makes it possible to force treatment on anorexia sufferers. The 1983 Mental Health Act provides for sufferers from severe psychiatric disorders. Sufferers can be held in hospital for treatment against their will. One in ten anorexia sufferers dies. Doctors use their powers under the law.

There is clearly work to be done in making the treatment of extreme anorexia more humane. The treatment often involves leaving patients without their clothes and watching patients eat and go to the lavatory.

Appendix 8

Task instructions for trialling summary completion Task A

INSTRUCTIONS FOR TASK A1

Please read the short story below.

While you are reading the short story, think about whether it would be all right to use as a reading passage for a Key Stage 3 English test.

For example, do you think the ideas in the story or the language it is written in would cause any difficulty for 13/14-year-olds?

When you have finished reading, you will be asked a few questions about the passage.

Journey by Night

He stood alone, leaning against a post, and shifting his weight from one foot to the other. It was late, and the taxi-stand was empty. The street was silent. He looked up and down, hoping that some vehicle would come in sight, for he wanted to get home. But none came.

The silence began to pall. He started to whistle, but there was no mirth in it, and he soon stopped. Midnight, ten miles away from home! What was he to do? To begin to walk that distance was out of the question.

A dark cloud passed across the sky, hiding the few pale stars that had been there. The noise of a falling dust-bin reached his ear. Some dog must have been scattering its contents.

Instinctively his hand felt for his wallet. Yes, it was still there. If only he had a stick! But he had nothing with which he might protect himself. He began to walk up and down, up and down.

What was that in the distance? At last two headlights were drawing near. He stepped into the middle of the street and held up his hand, and the car stopped.

' Taxi?' he asked. 'Valencia?'

'Get in,' said the driver, opening the door.

He sat beside the driver, glad to be on his way home at last. He had felt so lonely while he had been waiting. If only someone would say something! In the semi-darkness of the car he turned to look at the other passengers, but no one else was there.

The driver said nothing to him as the car sped along.

Suppose . . .

No, he mustn't allow himself to think of that. He glanced at the driver, and again his hand went to his wallet. He had heard of passengers being attacked at night and robbed. But surely . . . No, that couldn't happen to him.

If only he could see the other man's face clearly! But he had no idea who the driver was. He kept his eye intently on him during the seemingly interminable journey.

Now they were approaching a spot where the road branched off in another direction. There were tall, dark bushes around. The car slowed down, and the driver was looking at him. Then the driver took something short and black from the side-pocket of the car. It looked like an iron tool. Would the driver attack him with that?

'Stop!' he heard himself screaming, and his heart beat so fast with fear that he could hardly breathe.

But the car did not stop. Faster and faster instead it went. Now they were nearing his destination. Did the driver intend to take him past and then . . .

'Put me down here,' he cried out.

Still with his eyes on the driver, he quickly stepped from the car as it came to a standstill. He fumbled in his wallet for his fare, but the taxi was no longer there.

'No night passengers for me again,' exclaimed the driver, as with a sigh of relief he hurriedly moved off. And his hand tenderly caressed the heavy spanner with which he had meant to defend himself had that queer passenger attacked him.

INSTRUCTIONS FOR TASK A2

Please complete the following details about yourself:

FULL NAME: . . . . . .

FORM: . . . . . .

ENGLISH SET: . . . . . .

1. Answer questions (i) to (iii) by putting a circle round A, B or C.

   (i) Did you think the short story was
   - A very interesting?
   - B quite interesting?
   - C not at all interesting?

   (ii) Were the ideas in the short story
   - A generally difficult to understand?
   - B sometimes difficult to understand?
   - C generally easy to understand?

   (iii) Was the language of the short story
   - A generally difficult to understand?
   - B sometimes difficult to understand?
   - C generally easy to understand?

(Now turn over this sheet and do the remaining part of the task.)

2. Use one or two or three words to complete the summary of the short story in the box below. Do not make any changes to the punctuation.

Journey by Night

A man was standing late at night in a dark and lonely
place . Although he was ..... to reach ..... , (1) (2)
it was too ..... to think about ..... . He (3) (4)
began ..... in order to keep his spirits up , ... (5) (6)
he soon ..... . The ..... of the surrounding (7) (8)
area and the sound of a ..... nearby ..... (9) (10)
him and, anxious for his own ..... , he reached (11)
for ..... . (12)

Finally ..... stopped . Having been offered (13)
..... , the man ..... beside ..... . (14-16)
At first he felt ..... ; then he began to feel ..... (17)(18)
again ..... he couldn't see the ..... . He (19)(20)
remembered ..... of passengers ..... and he (21)(22)
grew increasingly ..... of the ..... . A (23)(24)
feeling of ..... overwhelmed the man ..... (25)(26)
suddenly ..... pulled out what looked like (27)
a ..... . He desperately wanted ..... to (28)(29)
stop but they seemed to ..... instead . He shouted (30)
to be ..... . Eventually they ..... and the (31)(32)
man ..... . He got out his wallet to pay the (33)
driver ..... the car drove away ..... he (34)(35)
had handed over the money.

..... hurried away with a sense of ..... , (36)(37)
vowing not to take any more ..... . The truth was (38)
that he had felt as ..... as ..... . (39)(40)

Appendix 9

Task instructions for trialling summary completion Task B

INSTRUCTIONS FOR TASK B1

Please read the newspaper editorial from *The Independent* newspaper below.

While you are reading the passage, think about whether it would be all right to use as a reading passage for a Key Stage 3 English test.

For example, do you think the ideas in the passage or the language it is written in would cause any difficulty for 13/14-year-olds?

When you have finished reading, you will be asked a few questions about the passage.

The rights and wrongs of treating anorexia

THE CASE of Samantha Kendall, the anorexia nervosa sufferer who discharged herself from hospital despite doctors' fears for her life, has highlighted the confusion in public thinking about this disturbing and perplexing disease. Ten years ago anorexia was still dismissed as nothing more than slimming gone too far. Today it is recognised as a treatable medical condition; but the degree to which treatment should be carried out without the patient's consent has become a topic of debate.

Researchers have suggested two psychiatric explanations behind the onset of anorexia. One is that the patient, faced with an unacceptably stressful or difficult adult life, is trying to retreat into childhood or avoid leaving it. Another is that choosing what to eat — and specifically choosing not to eat — is often an attempt to exert control by people who feel that their lives are too constrained in other ways. But the truth is that for all the resources that have been devoted to its study, the syndrome remains imperfectly understood.

It is beyond doubt, however, that anorexia is a severe psychiatric disorder. There is no other way to describe an illness that allows a patient to look in the mirror at her own emaciated, starved body, and see someone obese staring back. Severe sufferers often deny that they are trying to kill themselves, but the diet they are pursuing is all too likely to make death inevitable.

The 1983 Mental Health Act provides for sufferers from severe psychiatric disorders to be held in hospital for treatment against their will if there is a danger that they will do harm to themselves or others. Yet even though one in 10 anorexia sufferers dies, doctors are sometimes reluctant to use their powers under the law. This is often because of a fear that treatment by compulsion is self-defeating, since force-fed victims of anorexia often return to starvation diets when they get home.

There is clearly work to be done in making the treatment of extreme anorexia — which often involves leaving patients in isolation and without their clothes, and watching them as they eat and go to the lavatory — more humane. But the shortcomings of the available treatments should not obscure the fact that the alternative to treatment can sometimes be death. If doctors made more use of the powers available to them, lives could be saved.

INSTRUCTIONS FOR TASK B2

Please complete the following details about yourself:

FULL NAME: . . . . . .

FORM: . . . . . .

ENGLISH SET: . . . . . .

1. Answer questions (i) to (iii) by putting a circle round A, B or C.

   (i) Did you think the editorial was
   - A very interesting?
   - B quite interesting?
   - C not at all interesting?

   (ii) Were the ideas in the editorial
   - A generally difficult to understand?
   - B sometimes difficult to understand?
   - C generally easy to understand?

   (iii) Was the language of the editorial
   - A generally difficult to understand?
   - B sometimes difficult to understand?
   - C generally easy to understand?

(Now turn over this sheet and do the remaining part of the task)

2. Use one or two or three words to complete the summary of the newspaper editorial in the box below. Do not make any changes to the punctuation.

THE RIGHTS AND WRONGS OF TREATING ANOREXIA

A recent news case has raised an important issue concerning anorexia; a girl suffering from anorexia ..... hospital ..... (1) (2) her doctors' ..... . (3)

Changing attitudes

..... anorexia ..... regarded as a ..... , (4-6) it ..... generally accepted as a ..... . The current issue (7) (8) is ..... anorexia sufferers should be treated without ..... . (9) (10)

Understanding the disease

Different ..... have been offered for anorexia . Sufferers (11) may want to ..... childhood ..... they ..... (12-14) the pressures of adult life. Alternatively, people who feel their lives are ..... may choose what or what not to eat ..... it (15) (16) gives them a feeling of being ..... . Our understanding of the (17) disease , ... extensive research , is ..... . (18) (19)

The nature of the disease

It is clear that anorexia is a serious ..... disorder. Otherwise (20) extreme sufferers would not be able to ..... themselves (21) in ..... and ..... someone ..... . Severe (22-24) anorexia can lead to death although sufferers generally ..... (25) they are not trying to ..... . (26)

Treating anorexia

The ..... makes it possible to force treatment on sufferers . (27) ..... one in ..... sufferers ..... , doctors (28-30) are often ..... to use ..... because they think the (31) (32) treatment will be ..... in the long run. (33)

The way ahead

New ways of treating extreme anorexia are needed. Current methods - such as leaving patients ..... and watching them ..... - (34) (35) could be described as ..... . It may be the case , ... , (36) (37) that such treatment is ..... if the alternative is ..... . (38) (39)

Appendix 10

Final answer key for summary completion Task A (*Journey*)

For most items there was more than one acceptable response. Acceptable alternatives are separated by /. Letters and words in brackets indicate non-essential elements of an otherwise acceptable response.

Item	1	anxious/desperate/eager/wanted/wanting/in a hurry/longing
	2	(his) destination/house/home
	3	dangerous/dark (and eery/late)/(too) far (away)/late (at night)/ scary
	4	walking (home/it/there)
	5	whistling/to whistle
	6	although/but (then)/though
	7	ceased/gave up/stopped
	8	appearance/atmosphere/blackness/creepyness/darkness/eeriness/emptiness/(strange)feeling/loneliness/look/nature/noise(s)/ quietness/scariness/silence/sound(s)/stillness/weirdness
	9	bin/dustbin/rubbish bin (lid) (falling/crashing)
	10	disturbed/frightened/scared/startled/worried
	11	protection/safety/well-being
	12	(his) wallet
	13	(a/the) car/taxi/taxicab
	14	(a) lift/ride/seat/transport
	15	climbed in/got in/sat (down) in/stepped into/entered
	16	(the) driver/taximan
	17	safe/relieved/relaxed/pleased/happy/glad/better/at ease
	18	anxious/frightened/insecure/nervous/scared/tense/wary/worried/ uncomfortable/uneasy
	19	as/because/for
	20	(taxi)driver's face/driver's head/face of the driver
	21	(a/the) story/stories/tales/what he'd heard/he had heard/hearing
	22	attacked/mugged/robbed
	23	afraid/anxious/concerned/fearful/frightened/scared/paranoid/ suspicious/terrified/uneasy/wary/worried
	24	(taxi)driver/driver's intentions/driver's motives
	25	dread/fear/fright/horror/panic/paranoia/shock/terror
	26	as/when
	27	(the) driver

28 metal or iron tool/bar/object/pole/rod spanner/tool/weapon
29 (the) driver/(the) car/(the) taxi/them
30 accelerate/speed up/increase speed/quicken/(be going/getting/go) faster/gain speed
31 dropped off/let off/let out/put down/set down (here/there)
32 came to a stop/did stop/stopped
33 got off/got out/jumped out/left/leapt out/stepped out
34 but (instead)
35 before
36 the (taxi) driver
37 relief
38 night passengers/night customers/night travellers/passengers at night/nocturnal passengers
39 afraid/frightened/paranoid/scared/terrified/threatened
40 (his/the) passenger/(the) man (did)

Appendix 11

Final answer key for summary completion Task B (*Anorexia*)

For most items there was more than one acceptable response. Acceptable alternatives are separated by /. Letters and words in brackets indicate non-essential elements of an otherwise acceptable response.

Item	1	checked out (of)/discharged (herself from)/got out of/left/signed out (from) walked out of
	2	against/despite/even though/ignoring
	3	advice/advised her not to/concern(s)/didn't/approve/dismay/ fears/worries/wish(es)/warning(s)/opinion(s)
	4	although/though
	5	had been/has been/was (once)/used to be
	6	dieting gone too far/dieting gone wrong/dieting problem/heavy diet/slimming disorder/overdone slimming programme/severe diet/slimming disease/slimming disorder/slimming gone too far/ slimming problem
	7	has been/has got/ (today) is (however/now)/recently was
	8	(serious/treatable/curable) disease/illness/medical condition/mental disorder/psychiatric problem/psychological problem
	9	whether/how far/if/should
	10	a say/any say/being consulted/being asked/(his/her/their/patient's) consent/permission/approval/will
	11	explanations/reasons/theories
	12	extend/get back to/go back to/hide in/keep/lengthen/maintain/ never leave/not leave/prolong/re-enter/regain/rejoin/relive/restart/ retain/retreat to/return to/stay in
	13	and/as/because/if/or/when
	14	worry about/can't accept/can't cope with/can't face/can't handle/ can't take/don't want/don't like/fear/feel/hate/have difficulty with/want to escape/want to avoid/wish to leave
	15	(being) controlled/constrained/dependent on others/overrun/ restrained/ruled/constricted
	16	as/because/for/if/so
	17	control again/free/in charge/independent/powerful/their own
	18	despite/even from/even with
	19	imperfect/incomplete/limited/not finished/not sufficient
	20	brain/mental/mind/psychological/psychiatric

21 look at/see
22 (a/the) mirror(s)
23 observe/see
24 big/fat/obese
25 admit/are convinced/believe/claim/explain/say/state/think
26 commit suicide/die/kill themselves
27 mental health act/1983 act/government/law
28 although/(even) though
29 ten
30 can die/dies/end up dead
31 afraid/avoid/cautious/don't want to use/hesitant/loath/reluctant/scared/slow/unwilling/worried
32 force/forceable treatment/forced treatment/the (1983) law/their power(s)
33 ineffective/useless/no help/not worth it/pointless/unhelpful/unbeneficial/unsuccessful/wasted/worthless
34 alone/by themselves/in isolation/in a room/isolated/solitary/in separate rooms on their own/with no clothes/without (their) clothes/naked/bare/clothesless/unclothed
35 eat/at lavatory time/excrete/go to the loo/go to the toilet/use the lavatory
36 barbaric/cruel/inhumane/inhuman/mean/not humane/torture
37 however/though
38 a good idea/acceptable/accepted/called upon/humane/necessary/ok/preferable/required
39 death/dying/suicide/to die

References

Alba, J W and Hasher, L (1983) Is memory schematic?, *Psychological Bulletin* 93, 203–231.

Alderson, J C (1978) *A study of the cloze procedure with native and non-native speakers of English*, unpublished PhD thesis, University of London.

Alderson, J C (1979) The cloze procedure as a measure of proficiency in English as a foreign language, *TESOL Quarterly* 13, 219–227.

Alderson, J C (1980) Native and non-native speaker performance on cloze tests, *Language Learning* 30 (1), 59–76.

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*, London: Longman, 1–24.

Alderson, J C (1988) *Testing reading comprehension skills*, paper presented at the 6th Colloquium in Reading in a Second Language, Chicago, March 1988.

Alderson, J C (1990a) Testing reading comprehension skills (Part One), *Reading in a Foreign Language* 6, 425–438.

Alderson, J C (1990b) Testing reading comprehension skills (Part Two), *Reading in a Foreign Language* 7, 465–503.

Alderson, J C (2000) *Assessing Reading*, Cambridge: Cambridge University Press.

Alderson, J C (2005) *Diagnosing Foreign Language Proficiency: The Interface Between Learning and Assessment*, London: Continuum.

Alderson, J C and Lukmani, Y (1989) Cognition and reading: cognitive levels as embodied in test questions, *Reading in a Foreign Language* 5, 253–270.

Alderson, J C and Urquhart, A H (1984) *Reading in a Foreign Language*, London: Longman.

Alderson, J C, Clapham, C and Wall, D (1995) *Language Test Construction and Evaluation*, Cambridge: Cambridge University Press.

Alderson, J C, Figueras, N, Kuijper, H, Nold, G, Takala, S and Tardieu, C (2006) Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of The Dutch CEFR Construct Project, *Language Assessment Quarterly* 3 (1), 3–30.

Anderson, R C and Pearson, P D (1988) A schema-theoretic view of basic processes in reading comprehension, in Carrell, P L, Devine, J and Eskey, D E (Eds) *Interactive Approaches to Second Language Reading*, Cambridge: Cambridge University Press, 37–55.

Anderson, R C, Pichert, J W and Shirey, L (1983) Effects of the reader's schema at different points in time, *Journal of Educational Psychology* 75, 271–279.

Anderson, R C, Reynolds, R E, Schallert, D L and Goetz, E T (1977) Frameworks for comprehending discourse, *American Educational Research Journal* 14, 367–381.

Anderson, N J, Bachman, L F, Perkins, K and Cohen, A D (1991) An exploratory study into the construct validity of a reading comprehension test: triangulation of data sources, *Language Testing* 8, 41–66.

Andrich, D and Godfrey, J R (1978/1979) Hierarchies in the skills of Davis' Reading Comprehension Test, Form D: an empirical investigation using a latent trait model, *Reading Research Quarterly* 14, 182–200.

Armbruster, B B, Anderson, T H and Ostertag, J (1987) Does text structure/summarisation instruction facilitate learning from expository text?, *Reading Research Quarterly* 22, 331–346.

Bachman, L F (1982) The trait structure of cloze test scores, *TESOL Quarterly* 16 (1), 61–70.

Bachman, L F, Davidson, F, Ryan, K and Choi, I C (1995) *An Investigation into the Comparability of Two Tests of English as a Foreign Language*, Studies in Language Testing volume 1, Cambridge: UCLES/Cambridge University Press.

Badger, R and Yan, X (2012) The use of tactics and strategies by Chinese students in the Listening component of IELTS, in Taylor, L and Weir, C J (Eds) *IELTS Collected Papers 2: Research in Reading and Listening Assessment*, Studies in Language Testing volume 34, Cambridge: UCLES/Cambridge University Press, 454–486.

Barratt, N (2003) The change process at the paper level: Paper 3, Use of English, in Weir, C J and Milanovic, M (Eds) (2003) *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002*, Studies in Language Testing volume 15, Cambridge: UCLES/Cambridge University Press, 237–314.

Barrett, T C (1968) *Taxonomy of cognitive and affective dimensions of reading comprehension*, unpublished paper.

Bartlett, F C (1932) *Remembering*, Cambridge: Cambridge University Press.

Baxter, G P and Glaser, R (1998) Investigating the cognitive complexity of science assessments, *Educational Measurement: Issues and Practice* 17 (3), 37–45.

Bensoussan, M (1983) Multiple-choice modifications of the cloze procedure using word-length and sentence-length blanks, *Foreign Language Annals* 16, 189–198.

Bensoussan, M and Kreindler, I (1990) Improving advanced reading comprehension in a foreign language: summaries versus short-answer questions, *Journal of Research in Reading* 13, 55–68.

Bernhardt, E B (1991a) A psycholinguistic perspective on second language literacy, *Reading in Two Languages, AILA Review* 8, 31–44.

Bernhardt, E B (1991b) *Reading Development in a Second Language: Theoretical, Empirical and Classroom Perspectives*, Norwood: Ablex.

Biber, D, Conrad, S, Reppen, R, Byrd, P, Helt, M, Clark, V, Cortes, V, Csomay, E and Urzua, A (2004) *Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus*, *ETS TOEFL Monograph Series*, MS-25, Princeton: Educational Testing Service.

Birch, M (2007) *English L2 Reading: Getting to the Bottom*, Mahwah: Lawrence Erlbaum Associates.

Bloom, B S (1956) (Ed) *Taxonomy of Educational Objectives Handbook 1: Cognitive Domain*, London: Longman.

Bransford, J D and Johnson, M K (1973) Considerations of some problems of comprehension, in Chase, W G (Ed) *Visual Information Processing*, New York: Academic Press, 383–438.

Bransford, J D and McCarrell, N S (1977) A sketch of a cognitive approach to comprehension: some thoughts about understanding what it means to comprehend, in Johnson-Laird, P N and Wason, P C (Eds) *Thinking: Readings in Cognitive Science*, Cambridge: Cambridge University Press, 377–399.

Bransford, J D, Barclay, J R and Franks, J J (1972) Sentence memory: a constructive versus interpretive approach, *Cognitive Psychology* 3, 193–209.
Bransford, J D, Stein, B S and Shelton, T (1984) Learning from the perspective of the comprehender, in Alderson, J C and Urquhart, A H (Eds) *Reading in a Foreign Language*, London: Longman, 28–47.
Brereton, J L (1944) *The Case for Examinations*, Cambridge: Cambridge University Press.
Britton, J, Burgess, T, Martin, N, McLeod, A and Rosen, H (1975) *Development of Writing Abilities (11–18)*, London: MacMillan.
Brown, A L and Day, J D (1983) Macrorules for summarizing texts: the development of expertise, *Journal of Verbal Learning and Verbal Behaviour* 22, 1–14.
Brown, G (1994) Modes of understanding, in Brown, G, Malmkjaer, K, Pollitt, A and Williams, J (Eds) *Language and Understanding*, Oxford: Oxford University Press, 10–20.
Brown, G and Yule, G (1983) *Discourse Analysis*, Cambridge: Cambridge University Press.
Brown, G, Malmkjaer, K, Pollitt, A and Williams, J (1994) (Eds) *Language and Understanding*, Oxford: Oxford University Press.
Brown, J D and Rodgers, T (2002) *Doing Second Language Research*, Oxford: Oxford University Press.
Buck, G (1991) The testing of listening comprehension: an introspective study, *Language Testing* 8 (1), 67–81.
Caccamise, D, Franzke, M, Eckhoff, A, Kintsch, E and Kintsch, W (2007) Guided practice in technology-based summary writing, in McNamara, D S (Ed) *Reading Comprehension Strategies: Theories, Interventions and Technologies*, New York: Lawrence Erlbaum Associates, 375–396.
Carrell, P L (1983) Three components of background knowledge in reading comprehension, *Language Learning* 33, 183–207.
Carrell, P L (1984a) The effects of rhetorical organisation on ESL readers, *TESOL Quarterly* 18, 441–70.
Carrell, P L (1984b) Evidence of a formal schema in second language comprehension *Language Learning* 34, 87–112.
Carrell, P L (1987) Readability in ESL, *Reading in a Foreign Language* 4 (1), 21–40.
Carrell, P L (1991) Second language reading: reading ability or language proficiency?, *Applied Linguistics* 12, 159–179.
Carrell, P L, Devine, J and Eskey, D E (1988) (Eds) *Interactive Approaches to Second Language Reading*, Cambridge: Cambridge University Press.
Carroll, B J (1978) *Specifications for an English Language Testing Service*, London: The British Council.
Carroll, B J (1981) Specifications for an English language testing service, in Alderson, J C and Hughes, A (Eds), *Issues in Language Testing – ELT Documents 111*, London: The British Council, 66–110.
Carroll, J B (1972) Defining language comprehension: some speculations, in Carroll, J B and Freedle, R O (Eds) *Language Comprehension and the Acquisition of Knowledge*, Washington: Winston, 1–29.
Carver, R P (1977) Toward a theory of reading comprehension and rauding, *Reading Research Quarterly* 13, 8–64.
Carver, R P (1992) Effect of prediction activities, prior knowledge, and text type upon amount comprehended: using rauding theory to critique schema theory research, *Reading Research Quarterly* 27 (2), 165–174.

Carver, R P (1997) Reading for one second, one minute, or one year from the perspective of rauding theory, *Scientific Studies of Reading* 1 (1), 3–43.
Chafe, W L (1979) The flow of thought and the flow of language, in Givon, T (Ed) *Syntax and Semantics Volume 12: Discourse and Syntax*, New York: Academic Press, 159–181.
Chikalanga, I W (1992) A suggested taxonomy of inferences for the reading teacher, *Reading in a Foreign Language* 8, 697–710.
Chou Hare, V and Borchardt, K M (1984) Direct instruction of summarization skills, *Reading Research Quarterly* 20, 62–78.
Clapham, C (1996) *The Development of IELTS: A Study in the Effect of Background Knowledge on Reading Comprehension*, Studies in Language Testing volume 4, Cambridge: UCLES/Cambridge University Press.
Clymer, T (1968) What is reading? Some current concepts, in Robinson, H (Ed) *Innovation and Change in Reading Instruction*: *The Sixty-seventh Yearbook of the National Society for the Study of Education*, Chicago: The National Society for the Study of Education, 7–29.
Coady, J (1979) A psycholinguistic model of the ESL reader, in Mackay R, Barhman, B and Jordan, R R (Eds) *Reading in a Second Language*, Rowley: Newbury Press, 5–12.
Cobb, T (2006) *VocabProfile, The Compleat Lexical Tutor*, available online: www.lextutor.ca
Cohen, A D (1984) On taking language tests: what the students report, *Language Testing* 1, 70–81.
Cohen, A D (1986) Verbal reports as a source of information on reading strategies, *The ESP* 15, 1–12.
Cohen, A D (1993) The role of instructions in testing summarising ability, in Douglas, D and Chapelle, C (Eds) *A New Decade of Language Testing Research*, Alexandria: Teachers of English to Speakers of Other Languages, Inc., 132–160.
Cohen, A D (1994) English for academic purposes in Brazil: the use of summary tasks, in Hill, C and Parry, K (Eds) *From Testing to Assessment: English as an International Language*, London: Longman, 174–204.
Cohen, A D (1998) *Strategies in Learning and Using a Second Language*, Harlow: Longman.
Cohen, A D (2006) The coming of age of research on test-taking strategies, *Language Assessment Quarterly* 3 (4), 307–331.
Cohen, A, D (2013) Verbal Report, in Chapelle, C (Ed) *The Encyclopaedia of Applied Linguistics*, Oxford: Wiley-Blackwell, available online: sites.google.com/a/umn.edu/andrewdcohen/publications/research-methodology
Cohen, A, D and Upton, T (2006) Strategies in responding to the new TOEFL reading tasks, *TOEFL Monograph Series MS-33*, Princeton: Educational Testing Service.
Collins, A M and Loftus, E F (1978) A spreading activation theory of semantic processing, *Psychological Review* 82, 407–428.
Connor, U (1984) Recall of text: differences between first and second language readers, *TESOL Quarterly* 18, 239–256.
Connor, U and McCagg, P (1983) Cross-cultural differences and perceived quality in written paraphrases of English expository prose, *Applied Linguistics* 4, 259–268.
Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching and Assessment*, Cambridge: Cambridge University Press.

Courchene, R and Ready, D (1993) *Summary cloze: What is it? What does it measure?*, paper presented at the 15th Language Testing Research Colloquium, Cambridge/Arnhem, 2–4 August 1993.

Crain-Thoresen, C, Lippman, M Z and McClendon-Magnuson, D (1997) Windows on comprehension: reading comprehension processes as revealed by two think-aloud procedures, *Journal of Educational Psychology* 89 (4), 579–591.

Criper, C and Davies, A (1988) (Eds) *ELTS Validation Project Report*, English Language Testing Service Research Report Vol 1(i), The British Council and University of Cambridge Local Examinations Syndicate.

Crossley, S A and McNamara, D S (2008) Assessing L2 reading texts at the intermediate level: An approximate replication of Crossley, Louwerse, McCarthy and McNamara (2007), *Language Teaching* 41 (3), 409–429.

Crossley, S A, Louwerse, M M, McCarthy, P M and McNamara, D S (2007) A linguistic analysis of simplified and authentic texts, *Modern Language Journal* 91(2), 15–30.

Crothers, E J (1972) Memory structure and the recall of discourse, in Freedle, R and Carroll, J B (Eds) *Language Comprehension and the Acquisition of Knowledge*, Washington: Winston, 247–283.

Crothers, E J (1978) Inference and coherence, *Discourse Processes* 1, 51–71.

Davies, A (1981) Review of J Munby: Communicative Syllabus Design, *TESOL Quarterly* 15 (3), 332–344.

Davies, A (2008) *Assessing Academic English: Testing English Proficiency 1950–1989 – The IELTS Solution*, Studies in Language Testing volume 23, Cambridge: UCLES/Cambridge University Press.

Davies, E and Whitney, N (1984) Study skill 11: writing summaries, in Davies, E and Whitney, N (1984) *Study Skills for Reading: Students' Book*, London: Heinemann, 56–58.

Davis, F B (1944) Fundamental factors of comprehension in reading, *Psychometrika* 9, 185–197.

Davis, F B (1968) Research in comprehension in reading, *Reading Research Quarterly* 3, 499–545.

Davison, A and Green, G (1988) (Eds) *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered*, Hillsdale: Lawrence Erlbaum Associates.

de Beaugrande, R (1980) *Text, Discourse and Process*, London: Longman.

de Jong, J H A L and Verhoeven, L (1992) Modeling and assessing language proficiency, in Verhoeven, L and de Jong, J H A L (Eds) *The Construct of Language Proficiency*, Amsterdam/Philadelphia: John Benjamins Publishing Co, 3–19.

Devine, J (1988) A case study of two readers: models of reading and reading performance, in Carrell, P L, Devine, J and Eskey, D E (Eds) *Interactive Approaches to Second Language Reading*, Cambridge: Cambridge University Press, 127–139.

Drum, P A, Calfee, R C and Cook, L K (1981) The effects of surface structure variables on performance in reading comprehension tests, *Reading Research Quarterly* 4, 486–514.

Editorial (1980) Why comprehension?, *Reading Research Quarterly* 15 (2), 181–182.

Engineer, W D (1977) *An investigation of a reading model for English as a second language*, unpublished PhD thesis, University of Edinburgh.

Enright, M K, Grabe, W, Koda, K, Mosenthal, P, Mulcahy-Ernt, P and Schedl, M (2000) *TOEFL 2000 Reading Framework: A Working Paper, TOEFL Monograph Series MS-17*, Princeton: Educational Testing Service.

Ericsson, K A (1988) Concurrent verbal reports on reading and text comprehension, *Text* 8 (4), 295–325.

Ericsson, K A and Simon, H (1993) *Protocol Analysis*, Cambridge: MIT Press.

Ericsson, K A, Charness, N, Feltovich, P J and Hoffman, R R (Eds) (2006) *The Cambridge Handbook of Expertise and Expert Performance*, Cambridge: Cambridge University Press.

Faerch, C and Kasper, G (1987) (Eds) *Introspection in Second Language Acquisition Research*, Clevedon: Multilingual Matters.

Farr, R C and Carey, R F (1986) *Reading: What can be Measured?*, Newark: International Reading Association.

Farr, R C, Carey, R F and Tone, B (1986) Recent theory and research into the reading process: implications for reading assessment, in Orasanu, J (Ed) *Reading Comprehension: From Research to Practice*, Hillsdale: Lawrence Erlbaum Associates, 135–150.

Field, J (2004) *Psycholinguistics: The Key Concepts*, London: Routledge.

Field, J (2011) Cognitive validity, in Taylor, L (Ed) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press, 65–111.

Field, J (2012) The cognitive validity of the lecture-based question in IELTS Listening, in Taylor, L and Weir, C J (Eds) *IELTS Collected Papers 2: Research in Reading and Listening Assessment*, Studies in Language Testing 34, Cambridge: UCLES/Cambridge University Press, 391–453.

Field, J (2013) Cognitive validity, in Geranpayeh, A and Taylor, L (Eds) *Examining Listening: Research and Practice in Assessing Second Language Listening*, Studies in Language Testing volume 35, Cambridge: UCLES/ Cambridge University Press, 77–151.

Fortus, R, Coriat, R and Fund, S (1998) Prediction of item difficulty in the English subtest of Israel's inter-university psychometric entrance test, in Kunnan, A J (Ed) *Validation in language assessment: Selected Papers From the 17th Language Research Colloquium, Long Beach*, Mahwah: Lawrence Erlbaum, 61–87.

Frederiksen, C H (1972) Effects of task-induced cognitive operations on comprehension and memory processes, in Freedle, R and Carroll, J B (Eds) *Language Comprehension and the Acquisition of Knowledge*, Washington: Winston, 211–245.

Frederiksen, C H (1975) Representing logical and semantic structure of knowledge acquired from discourse, *Cognitive Psychology* 4, 371–458.

Freebody, P and Anderson, R C (1983) Effects of vocabulary difficulty, text cohesion, and schema availability on reading comprehension, *Reading Research Quarterly* 18, 272–293.

Freebody, P and Anderson, R C (1986) Serial position and rated importance in the recall of text, *Discourse Processes* 9, 31–36.

Freedle, R and Kostin, I (1993) *The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types: Main idea, inference, and supporting idea items*, TOEFL Research Reports No. RR-93-44, Princeton: Educational Testing Service.

Fries, C C (1945) *Teaching and Learning English as a Foreign Language*, Ann Arbor: University of Michigan Press.

Fries, C C (1963) *Linguistics and Reading*, New York: Holt, Rinehart and Winston.
Garnham, A (1987) *Mental Models as Representations of Discourse and Text*, Chichester: Ellis Horwood Ltd.
Garnham, A and Oakhill, J (1992) Discourse processing and text representation from a "mental models" perspective, *Language and Cognitive Processes* 7, 193–204.
Garnham, A and Oakhill, J (1994) *Thinking and Reasoning*, Oxford: Blackwell.
Garrod, S and Sanford, A J (1981) Bridging inferences and the extended domain of reference, in Long, J and Baddeley, A (Eds) *Attention and Performance IX*, 331–346.
Garrod, S and Sanford, A J (1994) Resolving sentences in a discourse context: How discourse representation affects language understanding, in Gernsbacher, M A (Ed) *Handbook of Psycholinguistics*, San Diego: Academic Press, 675–698.
Gass, S M and Mackey, A (2000) *Stimulated Recall Methodology in Second Language Research*, Mahwah: Lawrence Erlbaum Associates.
Gernsbacher, M A (1990) *Language Comprehension as Structure Building*, New Jersey: Erlbaum.
Gernsbacher, M A (1997) Two decades of structure building, *Discourse Processes* 25, 265–304.
Gernsbacher, M A, and Faust, M E (1992) The mechanism of suppression: a component of general comprehension skill, *Journal of Experimental Psychology: Learning, Memory and Cognition* 17, 245–262.
Gernsbacher, M A and Foerstch, J A (1999) Three models of discourse comprehension, in Garrod, S and Pickering, M (Eds) *Language Processing*, London: Psychology Press, 283–299.
Gernsbacher, M A, Varner, K R and Faust, M E (1989) Investigating differences in general comprehension skill, *Journal of Experimental Psychology: Learning, Memory and Cognition* 16, 430–445.
Gerrig, R J (1988) Text comprehension, in Sternberg, R and Smith, E (Eds) *The Psychology of Human Thought*, Cambridge: Cambridge University Press, 242–266.
Gerrig, R (1993) *Experiencing Narrative Worlds*, New Haven: Yale University Press.
Goldman, S, Golden, R and van den Broek, P (2007) Why are computational models of text comprehension useful?, in Schmalhofer, F and Perfetti, C (Eds) *Higher Level Language Processes in the Brain*, Mahwah: Lawrence Erlbaum Associates, 27–51.
Gomulicki, B R (1956) Recall as an abstractive process, *Acta Psychologica* 12, 77–94.
Goodman, K S (1967) Reading: a psycholinguistic guessing game, *Journal of the Reading Specialist* 6, 126–135.
Goodman, K S (1988) The reading process, in Carrell, P L, Devine, J and Eskey, D E (Eds) *Interactive Approaches to Second Language Reading*, Cambridge: Cambridge University Press, 11–21.
Gordon, C M (1987) *The effect of testing method on achievement in reading comprehension tests in English as a foreign language*, unpublished MA thesis, Tel-Aviv University.
Gordon, C M and Hanauer, D (1993) *Test answers as indicators of mental model construction*, paper presented at the 15th Language Testing Research Colloquium, Cambridge/Arnhem, 2–4 August 1993.

Gordon, C M and Hanauer, D (1995) The interaction between task and meaning construction in EFL reading comprehension tests, *TESOL Quarterly* 29, 299–324.
Gough, P B (1972) One second of reading, in Kavanagh, J F and Mattingly, I G (Eds) *Language by Ear and by Eye*, Cambridge: MIT Press, 331–358.
Gough, P B and Wren, S (1999) Constructing meaning: The role of decoding, in Oakhill, J and Beard, R (Eds) *Reading Development and the Teaching of Reading: A Psychological Perspective*, Malden: Blackwell, 59–78.
Gough, P B, Hoover, W A and Peterson, C L (1996) Some observations on a simple view of reading, in Cornoldi, C and Oakhill, J (Eds) *Reading Comprehension Difficulties: Processes and Intervention*, Mahwah: Lawrence Erlbaum Associates, 1–13.
Grabe, W (1988) Reassessing the term "interactive", in Carrell, P L, Devine, J and Eskey, D E (Eds) *Interactive Approaches to Second Language Reading*, Cambridge: Cambridge University Press, 56–70.
Grabe, W (1991) Current developments in second language reading research, *TESOL Quarterly* 25, 375–406.
Grabe, W (2000) Developments in reading research and their implications for computer-adaptive reading assessment, in Chalhoub-Deville, M (Ed) *Issues in Computer Adaptive Testing of Reading Proficiency*, Studies in Language Testing volume 10, Cambridge: UCLES/Cambridge University Press, 11–47.
Grabe, W (2001) Reading-writing relations: theoretical perspectives and instructional practice, in Belcher, D and Hirvela, A (Eds) *Reading and Writing Relations in L2 Contexts*, MI: University of Michigan Press, 15–47.
Grabe, W (2003) Using discourse representations for reading development, in Swanson, M and Hill, K (Eds) *Proceedings of the 2002 JALT Conference*, Tokyo: JALT Publications, 9–17.
Grabe, W (2009) *Reading in a Second Language: Moving from Theory to Practice*, Cambridge: Cambridge University Press.
Grabe, W and Kaplan, R B (1996) *Theory and Practice of Writing: An Applied Linguistic Perspective*, London: Longman.
Grabe, W and Stoller, F L (2002) *Teaching and Researching Reading*, London: Longman.
Graesser, A C and Kreuz, R J (1994) A theory of inference generation during text comprehension, *Discourse Processes* 16, 145–160.
Graesser, A C, Singer, M and Trabasso, T (1994) Constructing inferences during narrative text comprehension, *Psychological Review* 101, 371–395.
Graesser, A, McNamara, D S, Louwerse, M M, and Cai, Z (2004) Coh-Metrix: Analysis of text on cohesion and language, *Behavioral Research Methods, Instruments, and Computers* 36, 193–202.
Graesser, A C, McNamara, D S and Kulikowich, J M (2011) Coh-Metrix: providing multilevel analyses of text characteristics, *Educational Researcher* 40 (5), 223–234.
Gray, W S (1948) *On Their Own in Reading*, Glenview: Scott Foresman.
Gray, W S (1960) The major aspects of reading, in Robinson, H (Ed) *Sequential Development of Reading Abilities*, Supplementary Educational Monographs, No 90, Chicago University Press, 8–24.
Green, A (1998) *Verbal Protocol Analysis in Language Testing Research: A Handbook*, Studies in Language Testing volume 5, Cambridge: UCLES/ Cambridge University Press.

Green, A (2012) *Language Functions Revisited: Theoretical and Empirical Bases for Language Construct Definition Across the Ability Range*, English Profile Studies volume 2, Cambridge: Cambridge University Press.

Green, A, Ünaldi, A and Weir, C J (2010) Empiricism versus connoisseurship: establishing the appropriacy of texts for testing reading for academic purposes, *Language Testing* 27 (3), 1–21.

Grellet, F (1987) *Developing Reading Skills*, Cambridge: Cambridge University Press.

Haastrup, K (1987) Using thinking aloud and retrospection to uncover learners' lexical inferencing procedures, in Faerch, C and Kasper, G (1987) (Eds) *Introspection in Second Language Research*, Clevedon: Multilingual Matters, 197–212.

Hamp-Lyons, L and Kroll, B (1997) *TOEFL 2000 – Composition, Community and Assessment*, TOEFL Monograph 5, Princeton: Educational Testing Service.

Handschin, C H (1919) *Handschin Modern Language Test: Test A*, Yonkers-on-Hudson: World Book Company.

Harri-Augstein, S and Thomas, L (1984) Is comprehension the purpose of reading?, in Grundin, E H and Grundin, E U (Eds) *Reading: Implementing the Bullock Report*, London: Ward Lock International, 250–276.

Hawkey, R (2004) *A Modular Approach to Testing English Language Skills: The Development of the Certificates in English Language Skills (CELS) Examinations*, Studies in Language Testing volume 16, Cambridge: UCLES/ Cambridge University Press.

Hawkey, R (2009) *Examining FCE and CAE: Key Issues and Recurring Themes in Developing the First Certificate in English and Certificate in Advanced English Examinations*, Studies in Language Testing volume 28, Cambridge: UCLES/ Cambridge University Press.

Heaton, J B (1988) *Writing English Language Tests*, London and New York: Longman.

Hidi, S and Anderson, V (1986) Producing written summaries: task demands, cognitive operations, and implications for instruction, *Review of Educational Research* 56, 473–493.

Hidi, S and Baird, W (1986) Interestingness – a neglected variable in discourse processing, *Cognitive Science* 10, 179–194.

Hill, C and Parry, K (1992) The test at the gate: models of literacy in reading assessment, *TESOL Quarterly* 26 (3), 433–461.

Hinds, J (1977) Paragraph structure and pronominalization, *Papers in Linguistics* 10, 77–99.

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*, Clevedon: Multilingual Matters, 1–19.

Howatt, A P R (1984) *A History of English Language Teaching*, Oxford: Oxford University Press.

Hudson, T (2007) *Teaching Second Language Reading*, New York: Oxford University Press.

Huey, E B (1908) *The Psychology and Pedagogy of Reading*, New York: Macmillan.

Hughes, A (1989) *Testing for Language Teachers*, Cambridge: Cambridge University Press.

Hughes, A (1993) Testing the ability to infer when reading in a second or foreign language, *Journal of English and Foreign Languages* 10 (11), 13–20.

Hughes, A (2003) *Testing for Language Teachers*, Cambridge: Cambridge University Press, 2nd edn.
Hyland, K (2002) *Teaching and Researching Writing*, Applied Linguistics in Action Series, London: Longman.
Hymes, D H (1972) Models of the interaction of language and social life, in Gumperz, J J and Hymes, D H (Eds) *Directions in Sociolinguistics*, New York: Holt, Rinehart and Winston, 35–71.
Jarvella, R J (1971) Syntactic processing of connected speech, *Journal of Verbal Learning and Verbal Behavior* 10, 409–416.
Jenkinson, M D (1972) Sources of knowledge for theories of reading, in Melnik, A and Merritt, J (Eds) *Reading: Today and Tomorrow*, London: University of London Press Ltd, 102–117.
Johns, A M (1985) Summary protocols of "underprepared" and "adept" university students: replications and distortions of the original, *Language Learning* 35, 495–517.
Johns, A M and Mayes, P (1990) An analysis of summary protocols of university ESL students, *Applied Linguistics* 11, 253–271.
Johnson-Laird, P N (1983) *Mental Models*, Cambridge: Cambridge University Press.
Johnston, P (1981) *Prior knowledge and reading comprehension test bias*, unpublished doctoral dissertation, University of Illinois.
Johnston, P (1984) Prior knowledge and reading comprehension test bias, *Reading Research Quarterly* 19, 219–239.
Just, M A and Carpenter, P A (1987) *The Psychology of Reading and Language Comprehension*, Boston: Allyn and Bacon, Inc.
Khalifa, H (1997) *A study in the construct validation of the reading module of an EAP proficiency test battery: Validation from a variety of perspectives*, unpublished PhD thesis, University of Reading.
Khalifa, H and Weir, C J (2009) *Examining Reading: Research and Practice in Assessing Second Language Reading*, Studies in Language Testing volume 29, Cambridge: UCLES/Cambridge University Press.
Kintsch, W (1974) *The Representation of Meaning in Memory*, Hillsdale: Lawrence Erlbaum Associates.
Kintsch, W (1988) The use of knowledge in discourse processing: a construction-integration model, *Psychological Review* 95, 163–182.
Kintsch, W (1993) Information accretion and reduction in text processing: inferences, *Discourse Processes* 16, 193–202.
Kintsch, W (1998) *Comprehension: A Paradigm for Cognition*, Cambridge: Cambridge University Press.
Kintsch, W and van Dijk, T A (1978) Toward a model of text comprehension and production, *Psychological Review* 85, 363–394.
Kintsch, W and Yarborough, J C (1982) Role of rhetorical structure in comprehension, *Educational Psychology* 74 (6), 828–834.
Kintsch, E, Steinhart, D, Stahl, G, Matthews, C and Lamb, R (2000) Developing summarisation skills through the use of LSA-feedback, *Interactive Learning Environments* 8 (2), 87–109.
Kobayashi, M (1995) *Effects of text organization and test format on reading comprehension test performance*, unpublished PhD thesis, Thames Valley University, London.
Kobayashi, M (2002) Method effects on reading comprehension test performance: Text organization and response format, *Language Testing* 19, 191–218.

Koda, K (2005) *Insights into Second Language Reading: A Cross-linguistic Approach*, Cambridge: Cambridge University Press.
Laszlo, J, Meutsch, D and Viehoff, R (1988) Verbal reports as data in text comprehension research: An introduction, *Text* 8 (4), 283–294.
Lee, J F (1986) On the use of the recall task to measure L2 reading comprehension, *Studies in Second Language Acquisition* 8, 201–212.
Lehrer, A (1994) Understanding classroom lectures, *Discourse Processes* 17, 259–281.
Long, D L, Johns, C L and Morris, P E (2006) Comprehension ability in mature readers, in Traxler, M J and Gernsbacher, M A (Eds) *Handbook of Psycholinguistics*, New York: Academic Press, 2nd edition, 801–833.
Longacre, R E (1979) The paragraph as a grammatical unit, in Givon, T (Ed) *On Understanding Grammar*, New York: Academic Press, 115–134.
Lumley, T (1993) The notion of sub-skills in reading comprehension tests: An EAP example, *Language Testing* 10, 211–234.
Lund, R J (1991) A comparison of second language listening and reading comprehension, *Modern Language Journal* 75, 196–204.
Lunzer, E, Waite, M and Dolan T (1979) Comprehension and comprehension tests, in Lunzer, E and Gardner, K (Eds) *The Effective Use of Reading*, London: Heinemann Educational Books, 37–71.
Lyons, J (1977) *Semantics*, Cambridge: Cambridge University Press.
Mackey, A and Gass, S M (2005) *Second Language Research: Methodology and Design*, Abingdon: Routledge.
Mandler, J M and Johnson, N S (1977) Remembrance of things parsed: story structure and recall, *Cognitive Psychology* 9, 111–151.
Masi, S (2002) The literature on complexity, in Merlini Barbesi, L (Ed) *Complexity in Language and Text*, Pisa: PLUS-University of Pisa, 197–228.
McClelland, J L, Rumelhart, D E and Hinton, G E (1986) The appeal of parallel distributed processing, in Rumelhart, D E and McClelland, J L (Eds) *Parallel Distributed Processing: Explorations in the microstructure of cognition, Volume 1, Foundations*, Cambridge: MIT, 3–44.
McDonough, J and McDonough, S (1997) *Research Methods for English Language Teachers*, London: Arnold.
McKoon, G and Ratcliff, R (1992) Inference during reading, *Psychological Review* 99, 440–466.
McNamara, T (1996) *Measuring Second Language Performance*, London: Longman.
McNamara, T P, Miller, D L and Bransford, J D (1991) Mental models and reading comprehension, in Barr, R, Kamil, M L, Mosenthal, P and Pearson, P D (Eds) *Handbook of Reading Research Volume II*, New York: Longman, 490–511.
Mead, R (1982) Review of Munby: Communicative Syllabus Design, *Applied Linguistics* 3 (1), 70–78.
Melnik, A and Merritt, J (1972) (Eds) *Reading: Today and Tomorrow*, London: University of London Press Ltd.
Messick, S A (1989) Validity, in Linn, R L (Ed) *Educational Measurement*, New York: American Council of Education/Macmillan, 13–103, 3rd edn.
Meyer, B J F (1975a) Identification of the structure of prose and its implications for the study of reading and memory, *Journal of Reading Behavior* 7, 7–47.
Meyer, B J F (1975b) *The Organisation of Prose and its Effects on Memory*, Amsterdam: North-Holland Publishing Co.

Meyer, B J F (1985) Prose analysis: purposes, procedures, and problems, in Britten, B K and Black, J B (Eds), *Understanding Expository Text: A Theoretical and Practical Handbook for Analyzing Explanatory Text*, Hillsdale: Erlbaum, 11–64.

Meyer, B J F (1987) Following the author's top-level organisation: an important skill for reading comprehension, in Tierney, R J, Anders, P L and Mitchell, J N (Eds) *Understanding Readers' Understanding: Theory and Practice*, Hillsdale: Lawrence Erlbaum Associates, 59–76.

Meyer, B J F and Freedle R O (1984) Effects of discourse type on recall, *American Educational Research Journal* 21, 121–143.

Miles, M B and Huberman, A M (1994) *Qualitative Data Analysis: A Source Book of New Methods*, Thousand Oaks: Sage Publications.

Minsky, M (1975) A framework for representing knowledge, in Winston P H (Ed) *The Psychology of Computer Vision*, New York: McGraw-Hill, 211–277.

Mossenson, L, Hill, P and Masters, G (1987) *Tests of Reading Comprehension (TORCH)*, Hawthorn: Australian Council for Educational Research.

Mullis, I V S (1980) *Using the Primary Trait System for Evaluating Writing*, Report No. 10-W-51, Denver: Education Commission of the States.

Munby, J L (1978) *Communicative Syllabus Design*, Cambridge: Cambridge University Press.

Nassaji, H (forthcoming 2014) Lower level processes in L2 reading, *Language Teaching* 47 (1).

Neville, M H and Pugh, A K (1982) *Towards Independent Reading*, London: Heinemann Educational.

Oakhill, J and Garnham, A (1988) *Becoming a Skilled Reader*, Oxford: Blackwell.

Oakhill, J, Garnham, A, Gernsbacher, M A and Cain, K (1992) How natural are conceptual anaphors?, *Language and Cognitive Processes* 7, 257–280.

O'Brien, E J, Duffy, S A and Myers, J L (1986) Anaphoric inference during reading, *Journal of Experimental Psychology: Learning, Memory and Cognition* 12, 346–352.

O'Dell, F, Chandler, K, da Silva, L, Cotterill, S and Hogan M J (2013) *Pearson Test of English Academic Practice Tests Plus and CD ROM with Key Pack*, Harlow: Pearson Longman.

Ohno, M (2005) *Relationships between L2 reading proficiency and applications of macrorules in summary writing of Japanese high school students*, unpublished Master's thesis submitted to University of Tsukuba, Japan.

Ohno, M (2007) Relationships between L2 reading proficiency and applications of macrorules in summary writing of Japanese high school students, *JABAET Journal* 11, 103–124.

Oller, J W (1972) Assessing competence in ESL reading, *TESOL Quarterly* 4, 107–116.

O'Sullivan, B and Rugea, S (2011) *Review of the ILA Placement Test Usage*, internal report, London: The British Council.

Pearson, P D (1979) Methodological concerns in comprehension research, in Harste, J and Carey, R (Eds) *New Perspectives on Comprehension*, Bloomington: Indiana University School of Education, 153–170.

Pearson, P D (1985) Changing the face of reading comprehension instruction, *The Reading Teacher* 38, 724–738.

Pearson, P D (1986) Twenty years of research in reading comprehension, in Raphael, T E (Ed) *Contexts for School-based Literacy*, New York: Random House, 43–62.

Pearson, P D and Dunning, D B (1985) The impact of assessment on reading instruction, *Illinois Reading Council Journal* 3, 18–19.
Pearson, P D and Johnson, D D (1978) *Teaching Reading Comprehension*, New York: Holt, Rinehart and Winston.
Pearson Tests of English (2012) *The Official Guide to the Pearson Test of English Academic Pack*, Harlow: Pearson Longman.
Perfetti, C A (1985) *Reading Ability*, New York: Oxford University Press.
Perfetti, C A (1994) Psycholinguistics and reading ability, in Gernsbacher, M A (Ed) *Handbook of Psycholinguistics*, San Diego: Academic Press, 849–894.
Perfetti, C A (1997) Sentences, individual differences, and multiple texts: Three issues in text comprehension, *Discourse Processes* 23, 337–355.
Perrig, W and Kintsch, W (1985) Propositional and situational representations of text, *Journal of Memory and Language* 24, 503–518.
Pichert, J W and Anderson, R C (1977) Taking different perspectives on a story, *Journal of Educational Psychology* 69, 309–315.
Pocock, G N (1917) *Précis Writing for Beginners*, London: Blackie and Son.
Pollitt, A (1993) *Summary completion and assessing the ability to comprehend*, Internal EFL Research Report, Cambridge: University of Cambridge Local Examinations Syndicate.
Pollitt, A and Hutchinson, C J (1987) Calibrating graded assessments: Rasch partial credit analysis of performance in writing, *Language Testing* 4, 72–93.
Pollitt, A and Taylor, L (1993) Question level bias in cloze questions – an L1 transfer effect, in Huhta, A, Sajavaara, K and Takala, S (Eds) *Language Testing: New Openings*, University of Jyväskylä: Institute for Educational Research, 219–236.
Pollitt, A and Taylor, L (1997) Good reading tests aren't as good as they seem to be, in Huhta, A, Kohonen, V, Kurki-Suonio, L and Luoma, S (Eds) *Current Developments and Alternatives in Language Assessment – Proceedings of LTRC 96*, Jyväskylä: Institute for Educational Research, 203–223.
Pollitt, A and Taylor, L (2006) Cognitive psychology and reading assessment, in Sainsbury, M, Harrison, C and Watts, A (Eds) *Assessing Reading: From Theories to Classrooms*, Slough: National Foundation for Educational Research, 38–49.
Pollitt, A, Entwistle, N J, Hutchinson, C J and De Luca, C (1985) *What Makes Exam Questions Difficult?*, Edinburgh: Scottish Academic Press.
Pollitt, A, Hutchinson, C J, Napuk, A, Munro, L and Dickie, S (1990) *Assessment of Achievement Programme: Second Round English Language Monitoring Survey*, Report to Scottish Education Department.
Pressley, M and Afflerbach, P (1995) *Verbal Protocols of Reading: The Nature of Constructively Responsive Reading*, Hillsdale: Lawrence Erlbaum Associates.
Quinn T J (1993) The competency movement, Applied Linguistics and Language Testing: some reflections and suggestions for a possible research agenda, *Melbourne Papers in Language Testing*, 2 (2), 55–87.
Rastle, K (2007) Visual word recognition, in Gaskell, M G (Ed) *Oxford Handbook of Psycholinguistics*, Oxford: Oxford University Press, 71–87.
Ratteray, O M T (1985) Expanding roles for summarisation, *Written Communication* 2, 457–472.
Rayner, K, Pollatsek, A, Ashby, J and Clifton, C (2012) *The Psychology of Reading* Hove: Psychology Press, 2nd edn.
Richterich, R (1972) *A Model for the Definition of Language Needs of Adults Learning a Modern Language*, Strasbourg: Council of Europe.

Riesbeck, C K and Schank, R C (1978) Comprehension by computer: expectation-based analysis of sentences in context, in Levelt, W J M and Flores d'Arcais, G B (Eds) *Studies in the Perception of Language*, New York: Wiley, 247–294.

Riley, G L and Lee, J F (1996) A comparison of recall and summary protocols as measures of second-language reading comprehension, *Language Testing* 13 (2), 173–189.

Roach, J O (1945) *Some problems of oral examinations in modern languages: An experimental approach based on the Cambridge Examinations in English for foreign students*, UCLES report circulated to Oral Examiners and Local Examiners.

Robeson, F E (1913) *A Progressive Course of Précis Writing*, London, New York: Henry Frowde, Oxford University Press.

Robinson, H A (1966) The major aspects of reading, in Robinson, H A (Ed) *Sequential Development of Reading Abilities*, Supplementary Educational Monographs, No 90, Chicago: Chicago University Press, 22–32.

Rumelhart, D E (1975) Notes on a schema for stories, in Bobrow, D G and Collins, A M (Eds) *Representation and Understanding: Studies in Cognitive Science*, New York: Academic Press, 211–236.

Rumelhart, D E (1977) Toward an interactive model of reading, in Dornic S (Ed) *Attention and Performance VI*, Hillsdale: Lawrence Erlbaum Associates, 573–603.

Rumelhart, D E and McClelland, J L (1986) *Parallel Distributed Processing: Explorations in The Microstructure of Cognition, Volume 1, Foundations*, Cambridge, MA: MIT.

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.

Samuels, S J and Kamil, M L (1988) Models of the reading process, in Carrell, P L, Devine, J and Eskey, D E (Eds) *Interactive Approaches to Second Language Reading*, Cambridge: Cambridge University Press, 22–36.

Sanford, A J and Garrod, S C (1981) *Understanding Written Language: Explorations of Comprehension beyond the Sentence*, Chichester: John Wiley and Sons.

Sarig, G (1989) Testing meaning construction: can we do it fairly?, *Language Testing* 6, 77–94.

Schank, R C and Abelson, R (1977) *Scripts, Plans, Goals and Understanding*, Hillsdale: Lawrence Erlbaum Associates.

Schmalhofer, F and Glavanov, D (1986) Three components of understanding a programmer's manual: verbatim, propositional, and situational representations, *Journal of Memory and Language* 25, 279–294.

Schuetz, A (1953) Common-sense and scientific interpretation of human action, *Philosophy and Phenomenological Research* 14, 1–38.

Seddon, G M (1978) The properties of Bloom's taxonomy of educational objectives for the cognitive domain, *Review of Educational Research* 48, 303–323.

Seidlhofer, B (1990) Summary judgements: Perspectives on reading and writing, *Reading in a Foreign Language* 6 (2), 413–424.

Seliger, H W and Shohamy, E (1989) *Second Language Research Methods*, Oxford: Oxford University Press.

Sharkey, A J and Sharkey, N E (1992) Weak contextual constraints in text and word priming, *Journal of Memory and Language* 31, 507–524.

Shaw, S D and Weir, C J (2007) *Examining Writing: Research and Practice in Assessing Second Language Writing*, Studies in Language Testing volume 26, Cambridge: UCLES/Cambridge University Press.
Sherrard, C (1986) Summary writing – a topographical study, *Written Communication* 3, 324–343.
Shohamy, E (1984) Does the testing method make a difference? The case of reading comprehension, *Language Testing* 1, 147–170.
Singer, M (1990) *Psychology of Language*, Hillsdale: Lawrence Erlbaum Associates.
Singer, M (1994) Discourse inference processes, in Gernsbacher, M A (Ed) *Handbook of Psycholinguistics*, San Diego: Academic Press, 359–374.
Singer, M (2007) Inference processing in discourse comprehension, in Gaskell, M G (Ed) *Oxford Handbook of Psycholinguistics*, Oxford: Oxford University Press, 343–359.
Skehan, P (1984) Issues in the testing of English for specific purposes, *Language Testing* 1, 202–220.
Skehan, P (1988) Language Testing: Part I, *Language Teaching* 21, 211–221.
Smith, F (1971) *Understanding Reading*, New York: Holt, Rinehart and Winston.
Spache, G D (1964) *Reading in the Elementary School*, Boston: Allyn and Bacon.
Spilich, G J, Vesonder, G T, Chiesi, H L and Voss, J F (1979) Text processing of domain-related information for individuals with high and low domain knowledge, *Journal of Verbal Learning and Behavior* 18, 275–290.
Spivey, M, McRae, K and Joanisse, M (2012) (Eds) *The Cambridge Handbook of Psycholinguistics*, Cambridge: Cambridge University Press.
Spolsky, B (1994) Comprehension testing, or can understanding be measured?, in Brown, G, Malmkjaer, K, Pollitt A and Williams, J (Eds) *Language and Understanding*, Oxford: Oxford University Press, 141–152.
Stanovich, K E (1980) Toward an interactive-compensatory model of individual differences in the development of reading fluency, *Reading Research Quarterly* 16, 32–71.
Steffensen, M S (1988) Changes in cohesion in the recall of native and foreign texts, in Carrell, P L, Devine, J and Eskey, D E (Eds) *Interactive Approaches to Second Language Reading*, Cambridge: Cambridge University Press, 140–151.
Steffensen, M S and Joag-Dev, C (1984) Cultural knowledge and reading, in Alderson, J C and Urquhart, A H (Eds) *Reading in a Foreign Language*, London: Longman, 48–61.
Steffensen, M S, Joag-Dev, C and Anderson, R C (1979) A cross-cultural perspective on reading comprehension, *Reading Research Quarterly* 15, 10–29.
Stein, B S and Bransford, J D (1979) Constraints on effective elaboration, *Journal of Verbal Learning and Verbal Behavior* 18, 769–777.
Strang, R (1972) The nature of reading, in Melnik, A and Merritt, J (Eds) *Reading: Today and Tomorrow*, London: University of London Press Ltd, 67–101.
Swaffar, J K, Arens, K M and Byrnes, H (1991) *Reading for Meaning: an Integrated Approach to Language Learning*, Englewood Cliffs: Prentice Hall.
Tanenhaus, M K (1988) Psycholinguistics: an overview, in Newmeyer F J (Ed) *Linguistics: The Cambridge Survey, Volume III – Language: Psychological and Biological Aspects*, Cambridge: Cambridge University Press, 1–37.
Tannen, D (1979) What's in a frame? Surface evidence for underlying expectations, in Freedle, R O (Ed) *New Directions in Discourse Processing*, Norwood: Ablex, 137–181.

Taylor, L (1991) *Some aspects of the comparability problem for communicative proficiency tests*, unpublished M Phil dissertation, Research Centre for English and Applied Linguistics, University of Cambridge.
Taylor, L (1996) *An investigation of text-removed summary completion as a means of assessing reading comprehension ability*, unpublished PhD thesis, Research Centre for English and Applied Linguistics, University of Cambridge.
Thorndike, E L (1917/1972) Reading as reasoning: a study of mistakes in paragraph reading, *Journal of Educational Psychology* 8, 323–332. (Reprinted in Melnik, A and Merritt, J (1972) (Eds) *Reading: Today and Tomorrow*, London: University of London Press Ltd, 20–30)
Thorndike, E L (1973) Reading as reasoning, *Reading Research Quarterly* 2, 135–147.
Thorndyke, P W (1977) Cognitive structures in comprehension and memory of narrative discourse, *Cognitive Psychology* 9, 77–110.
Thurstone, L L (1946) Note on reanalysis of Davis' reading tests, *Psychometrika* 11, 185–188.
Tinker, M A and McCullough, C M (1962) *Teaching Elementary Reading*, New York: Appleton-Century-Crofts.
Touchstone Applied Science Associates (1991) *Degrees of Literacy Power Program*, New York: Touchstone Applied Science Associates.
Trabasso, T and Bouchard, E (2002) Teaching readers how to comprehend texts strategically, in Block, C and Pressley, M (Eds) *Comprehension and Instruction: Research-based Best Practices*, New York: Guilford Press, 176–200.
Urquhart, A H and Weir, C J (1998) *Reading in a Second Language: Process, Product and Practice*, New York: Longman.
Vacca, R T and Vacca, J L (1983) Two less than fortunate consequences of reading research in the 1970s, *Reading Research Quarterly* 18, 382–383.
Valencia, S and Pearson, P D (1987) Reading assessment: time for a change, *The Reading Teacher* 40 (8), 726–732.
van Dijk, T A (1977) *Text and Context*, London: Longman.
van Dijk, T A (1979) Relevance assignment in discourse comprehension, *Discourse Processes* 2, 113–126.
van Dijk, T A (1980) *Macrostructures*, Hillsdale: Lawrence Erlbaum Associates.
van Dijk, T A and Kintsch, W (1977) Cognitive psychology and discourse: recalling and summarizing stories, in Dressler, W U (Ed) *Current Trends in Text Linguistics*, New York: De Gruyter, 61–80.
van Dijk, T A and Kintsch, W (1983) *Strategies of Discourse Comprehension*, New York: Academic.
Varnhagen, C K (1991) Text relations and recall for expository prose, *Discourse Processes* 14, 399–422.
Venezky, R L (1984) The history of reading research, in Barr, R, Kamil, M L, Mosenthal, P and Pearson *Volume II* P D (Eds) *Handbook of Reading Research*, New York: Longman, 3–38.
Vincent, D (1985) *Reading Tests in the Classroom: An Introduction*, Slough: NFER-Nelson.
Weaver, C A and Kintsch, W (1991) Expository text, in Barr, R, Kamil, M L, Mosenthal, P and Pearson, P D (Eds) *Handbook of Reading Research Volume II*, New York: Longman, 230–245.
Weigle, S C (2002) *Assessing Writing*, Cambridge: Cambridge University Press.
Weir, C J (1983) *Identifying the language problems of overseas students in tertiary education in the United Kingdom*, unpublished PhD thesis, University of London.

Weir, C J (1990) *Communicative Language Testing*, Hemel Hempstead: Prentice Hall.

Weir, C J (1993) *Understanding and Developing Language Tests*, Hemel Hempstead: Prentice Hall.

Weir, C J (2005) *Language Testing and Validation: An Evidence-Based Approach*, Basingstoke: Palgrave Macmillan.

Weir, C J (2013a) Case study: A quantitative analysis of the context validity of the CPE reading passages used in translation tasks (1913–88), summary tasks (1930–2010) and comprehension question (MCQ/SAQ) tasks (1940–2010), in Weir, C J, Vidaković, I and Galaczi, E D (2013) *Measured Constructs: A History of Cambridge English Language Examinations 1913–2012*, Studies in Language Testing volume 37, Cambridge: UCLES/Cambridge University Press, 472–537.

Weir, C J (2013b) The measurement of reading ability 1913–2012, in Weir, C J, Vidaković, I and Galaczi, E D (2013) *Measured Constructs: A History of Cambridge English Language Examinations 1913–2012*, Studies in Language Testing volume 37, Cambridge: UCLES/Cambridge University Press, 180–256.

Weir, C J and Milanovic, M (Eds) (2003) *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002*, Studies in Language Testing volume 15, Cambridge: UCLES/Cambridge University Press.

Weir, C J and Porter, D (1994) The multi-divisible or unitary nature of reading: the language tester between Scylla and Charybdis, *Reading in a Foreign Language* 10, 1–19.

Weir, C J, Hughes, A and Porter, D (1990) Reading skills: hierarchies, implicational relationships and identifiability, *Reading in a Foreign Language* 7, 505–510.

Weir, C J, Yang, H and Jin, Y (2000) *An Empirical Investigation of the Componentiality of L2 Reading in English for Academic Purposes*, Studies in Language Testing volume 12, Cambridge: UCLES/Cambridge University Press.

Weir, C J, Hawkey, R, Green, A and Devi, S (2009) The cognitive processes underlying the academic reading construct as measured by IELTS, in Thompson, P (Ed) *IELTS Research Reports Volume 9*, London: British Council/IDP Australia, 157–189.

Weir C J, Bax, S, Chan, S, Field, J, Green, A and Taylor, L (2012) *The contextual parameters of CAE examinations and Coh-Metrix d*, unpublished project report for Cambridge ESOL.

West, R (1991) Development in the testing of reading, *English as a World Language* 1, 60–70.

Williams, J N (1993) *Products of Comprehension*, lecture series delivered during the M Phil in English and Applied Linguistics course, Research Centre for English and Applied Linguistics, University of Cambridge.

Williams, J N (1995) *Inferencing*, lecture delivered during the M Phil in English and Applied Linguistics course, Research Centre for English and Applied Linguistics, University of Cambridge.

Winograd, P N (1984) Strategic difficulties in summarizing texts, *Reading Research Quarterly* 19, 404–425.

Woodfield, H (2012) Pragmatic variation in learner perception: the role of retrospective verbal report in L2 speech act research, in César Félix-Brasdefer, J and Koike, D A (Eds), *Pragmatic Variation in First and Second Language Contexts: Methodological Issues*, Amsterdam: John Benjamins Publishing, 209–238.

Wu, R Y F (2011) *Establishing the validity of the General English Proficiency Test reading component through a critical evaluation on alignment with the Common European Framework of Reference*, unpublished PhD thesis, University of Bedfordshire.

Yalden J (1987) *Principles of Course Design for Language Teaching*, Cambridge: Cambridge University Press.

Zuck, L V and Zuck, J G (1984) The main idea: specialist and non-specialist judgements, in Pugh, A and Ulijn, J (Eds) *Reading for Professional Purposes: Studies and Practices in Native and Foreign Languages*, London: Heinemann, 132–135.

Zwaan, R A and Rapp, D N (2006) Discourse comprehension, in Traxler, M J and Gernsbacher, M A (Eds) *Handbook of Psycholinguistics*, New York: Academic Press, 2nd edition, 725–764.

Zwaan, R A, Langston, M C and Graesser, A C (1995) The construction of situation models in narrative comprehension: an event-indexing model, *Psychological Science* 6, 292–297.

Author index

Subject index