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 under-stood. 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 dis-connection 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.

Testing Reading Through Summary

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.

4

Introduction

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

5

Testing Reading Through Summary

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

7

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.

9

Testing Reading Through Summary

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

10