1 Introduction

Background: the Chinese operational context

Reading in English at undergraduate level in China

Before 1985, the required speed of reading in English for Chinese university students was 17 words per minute. A survey by the Ministry of Higher Education (later the State Education Commission) showed that only one third of university graduates acquired this 'reading ability'.

In 1985, the National College English Teaching Syllabus (NCETS) was introduced by the State Education Commission. In this syllabus, the English course, which is compulsory for all university students across the country, is divided into six bands. All the students must meet the requirements of Band 4, which include a reading speed of 50 wpm for careful reading and 80 wpm for quick reading with 70% comprehension. The requirements of Band 6 are aimed at students who have successfully completed Band 4 study, the target reading speed being 70 wpm for careful reading and 120 wpm for quick reading with 70% comprehension. Band 4 and 6 together constitute the basic grounding stage of the College English Course. This course focuses primarily on the development of students' linguistic competence with only limited attention being paid to the development of language skills and strategies.

Despite this focus on linguistic competence, the publication of this syllabus has had a positive impact on English language learning at tertiary level in China. This can be seen in the data available on the College English Test (CET) based on the NCET syllabus which was inaugurated in 1987.

A recent 4 year research study has shown the CET to be a valid and reliable measure of general linguistic competence (Yang and Weir forthcoming). It has had a powerful backwash effect on the numbers learning English at the foundation stage in the university system with the candidature growing to over two million by 1997. Since the inception of the test ten years ago there has been an improvement in the language competence of university students as attested by institutional and national performance levels.

However, an important stated aim of English language teaching at the tertiary level in China is to improve access to scientific and technical literature

through reading in English. Unfortunately linguistic competence is not the same as performance ability in language skills and strategies in reading. If providing students with the latter is the ultimate goal of English language teaching as specified in NCETS, then further steps need to be taken.

It was evident that many students and also their university authorities saw the foundation stage of English study as the end of English language learning and teaching. Furthermore, the end-users of the CET tended to misinterpret the value of CET and often expected too much from the certificate holders in terms of performance skills and strategies. Clearly the system was not providing university graduates with the requisite skills and abilities to access foreign academic and technical literature through the medium of English.

Rationale for the reading test project

It should be noted that to help achieve this criterial goal for English teaching in Chinese universities an EAP reading course is already stipulated in the syllabus for when the undergraduates finish foundation stage study after year 2. The problem is that the course is not given due attention and is neglected in many universities and institutions by university authorities, teachers and students.

The lack of an adequate and appropriate assessment tool is seen as a major reason for this neglect. A widespread tradition in China, as everywhere else, is 'what is tested is taught' and consequently 'what is not tested is not taught'. In this situation, an Advanced English Reading Test (AERT) could be the necessary catalyst for encouraging the achievement of the aim of the final stage of the college English course in China.

This project, therefore, set out to help achieve that end by developing an Advanced English Reading Test (AERT). The main benefits from this test would be:

- the availability of an appropriate tool for universities and teachers to monitor and evaluate students' performance in EAP reading
- the exertion of a much-needed positive backwash effect on English teaching in years 3 and 4 of college education in China in the sense of actually encouraging the teaching of reading skills and strategies where none may take place at present

Monitoring the impact of the reading project

Currently demographic statistics on the College English Test (CET – 4 and 6) are collected as a matter of routine and there is no reason why the same could not be done for the AERT. Scores by centre, region, institution, gender etc., are made available annually and can be compared to previous national and institutional averages.

Rigorous attempts are made to statistically equate the various forms of the CET test through IRT and an anchor test so comparisons can be made between

cohorts from year to year. In this way it is possible to monitor whether there can be said to be improvement at a national or institutional level over a period of time. A similar system could be set in place for the AERT.

Empirical data can also be collected from institutions (staff and students) and end-users by self report (questionnaire survey and interview) in order to triangulate the descriptive statistics emanating from test administrations.

Such data will enable the authorities to monitor the impact of the AERT on the gain in reading ability in years 3 and 4 of Chinese universities. The most effective design would collect data relating to the state the students are at when they begin year 3 and where they get to by the end of year 4. Implementation data which established the amount and nature of mediated instruction in reading that went on within institutions might enable some useful pedagogical lessons to be drawn.

Failing such comprehensive evaluation, a comparison from year to year of exit behaviour would provide useful though less comprehensive data. At the very least it might show whether the test was having an impact in terms of gains in population reading scores across the various types of reading being tested over a period of time. The value added for the Chinese economy might then be investigated. This would involve end users calculating what progress in terms of performances at different levels on the test would contribute financially to their organisations. The enhanced value of improvement in the different types of reading measured by the test might then be estimated.

Background: developing tests to measure the construct of reading in English for academic purposes

The specification and operationalisation of the construct

In the past ESL examination specifications were either absent or extremely limited. Typically one met a spuriously circular argument relating examinations to the textbooks used to prepare the students for the examinations. The textbooks were viewed as a benchmark for establishing what levels such as intermediate and proficient meant. These text book writers would conversely refer to the examinations as their point of reference for both the content and level for their coursebooks.

In recent years a number of the major examination boards have taken a more principled and systematic approach to the development of tests. The University of Cambridge Local Examinations Syndicate (UCLES) is a good example of this (see handbooks produced for each examination by UCLES and **Users Guide** prepared for Council of Europe.). They have attempted to provide clear specifications for each of their major examinations and establish systematic development procedures to faithfully implement these.

It is clear that a reading test/examination is only as good as the texts and tasks that are used to operationalise the construct it is intended to measure. Inadequacies or limitations in the texts and tasks employed will constrain the value of any comprehension test. Given the huge potential test population and national importance of the Advanced English Reading Test (AERT) in China it was imperative that we developed maximally valid operationalisations of what we believed to be the important elements of the construct of EAP reading in the form of texts and associated tasks. In order to develop a construct valid reading test we had to develop effective, efficient and replicable a *priori* and a *posteriori* procedures for test development. In this book we lay out a comprehensive set of procedures for the development of an advanced reading test in English as a Second Language.

Our initial studies indicated that reading research in this century has been constrained by a narrow view of reading (in the main focusing on careful reading at the local level) and by serious limitations in the tests used as research instruments (see Urquhart and Weir 1998). It was therefore imperative for us to develop a comprehensive specification with a sound theoretical and empirical base and implement this as faithfully as possible in the AERT. As well as helping us to develop a valid and reliable operational test, such research might also provide data which cast some light on the componentiality of the reading construct. Only through the development of a valid and reliable set of tests could we hope to resolve the issue of whether reading is a unitary activity or whether it is made up of separable components, for example: expeditious types of reading as in search reading, skimming, scanning for specifics, and careful reading at the global and local levels. Such a test might shed light on the relative contribution of the posited skills and strategies to the overall picture of a student's reading ability. It could also tell us about the relationship between the test components and inform us of the relative weaknesses and strengths of our students. Whether for formative or summative purposes such diagnostic evaluation might impact on whole educational systems as well as individual classrooms.

The data from the test development procedures described in this book are all grist to the construct validity mill. They can all shed light on what it is we are measuring and how well we are doing this. The more of these we can embrace in our research investigations into EAP reading the more valid and reliable the resulting data on academic reading ability.

Urquhart and Weir (1998: chapter 5.3) outline a principled set of procedures for investigating the componentiality of reading (see Figure 1.1 below). The a *priori* and a *posteriori* procedures outlined there were based on the development work behind the AERT in China and are discussed in detail in this volume. Such a methodology we feel is generic and should for the most part apply to all reading situations.

Figure 1.1

A methodology for investigating the EAP reading construct

A *priori* validation

Stage 1: Specification of the construct

The reading strategies and skills and the conditions under which these activities are performed might be established through:

- target situation analysis of target population's reading activities;
- theoretical literature review of reading processes;
- research literature review concerning the componentiality of reading;
- document analysis: EAP reading course-books/EAP reading tests.

Stage 2: Development of pilot tests to operationalise EAP reading specification

Systematic textmapping of appropriate texts:

- to establish the consensus information recoverable according to type of reading employed i.e. careful versus expeditious;

Produce pilot version of test(s):

- decide on most appropriate format in relation to operations;
- allow for attrition in texts and items by trialling extra;
- ensure intelligibility of rubrics;
- empirically establish timing;
- consider order of questions/process dimension;
- check layout;
- trial on small samples. Produce first draft of mark scheme;
- moderate tasks and mark scheme in committee.

A *posteriori* validation

Stage 3: Analysis of data on the test

Trial on reasonable sample
Item analysis
Establish item:

- facility values;
- discrimination indices;

Estimates of reliability

- marker reliability.

Estimates of internal validity

- internal consistency;
- correlations;
- principal component analyses;
- level of subtests: Means, t tests and cross-tabulation.

Estimates of external validity
Establish what items are testing through:

- qualitative expert judgement of items
- qualitative introspection/retrospection by test-takers
- feedback from test-takers (interview /questionnaire)

Revise

- administrator's instructions
- items
- timing
- rubrics
- mark schemes
- re-trial any new items

[Source: Reading in a Second Language, Urquhart A. H. and Weir C. J. (1998). Longman]

Each of these stages is discussed in detail in the following chapters in this book as we describe the steps we took to try to achieve construct validity in our measuring instruments in terms of operationalising the skills/strategies underlying the various types of reading we wished to establish the empirical existence of.

In the next chapter, we will examine in close detail the a *priori* procedures which led to the specification of the AERT (see 3.1 for details of the construct). These constituted the first stage in the development of a systematic approach to developing our test of English for Academic Purposes reading ability. On the basis of the specification we then developed the AERT using a principled set of procedures described in 4.1 – 4.4.

A brief outline of these a *priori* procedures employed to define the construct of reading both theoretically and operationally is provided in the next section.

A prior validation

Establishing the parameters of the reading construct

To establish a specification of operations and performance conditions to be tested we pursued a number of avenues in the development of the Advanced English Reading Test (AERT) for undergraduates in China. We carried out:

- a review of research and theories of reading;
- analysis of EAP reading tasks in Course-books;
- analysis of EAP reading tasks in Public Tests;
- needs analysis.

The investigation of the theoretical construct of EAP reading and the development of a test for this required systematic research. The research began with a review of the literature on existing theories of the reading process (see 2.1 below). The study of various models of reading shed light on the construct of reading from various aspects though it was noticeable that these theories are mostly premised on only one of the identified types of reading, that is, careful reading. Nevertheless such research drew our attention to the importance of reader driven processing at the text level and the importance of goal setting, as well as more traditional text driven processing at the word/sentence level.

We also examined the empirical (largely test based) research literature concerning the divisibility of the reading construct into components and the salient performance conditions under which these types of reading are performed (see 2.1). This product focused empirical research relating to the componentiality of reading ability points to at least a bidivisible view of reading with vocabulary loading on a separate factor in addition to general reading comprehension in nearly all cases. The data suggest that a partially divisible view of reading is preferable to a unitary view.

The empirical study also involved a survey of Chinese undergraduates' EAP reading needs as viewed by subject teachers of advanced reading in English in China (see 2.2 below).

For a view of what and how EAP reading is currently taught and tested, a survey was made of all the major EAP reading course books and EAP tests available (see 2.3 below) in terms of operations (skills/strategies) and performance conditions (length of texts, reading speed etc.).

The findings from these various strands of enquiry resulted in specifications of the types of reading to be included in the test as well as the conditions under which these reading activities should be performed (see 2.1 to 2.3 below).The types of EAP reading our research indicated we should include are listed below (see 3.1 for full details):

- careful reading for global comprehension of main ideas;
- careful local reading for understanding at the word level;
- expeditious reading (reading quickly, selectively and effectively) for global comprehension of main ideas;
- expeditious reading at the local level (scanning for specific details).

In addition our research suggested the following performance conditions had an important effect on performance on these tasks and must be carefully considered by the test developer:

- length of text must be appropriate for intended type of reading and time allocated be consonant with this;
- time allowed to complete to be empirically determined for each reading type tested;
- strict enforcement of such time controls at the passage/reading type level;
- nature of text must be accessible across three broad discipline areas;
- rhetorical organisation of texts appropriate for reading type;
- overtness of text organisation (markers of importance, textual sign-posting) for expeditious tasks;
- nature of vocabulary: appropriate degree of specialisation;
- topic familiarity: low to medium familiarity.

The test designed in line with the specified guidelines is expected to maximally operationalise the construct in question. Once appropriate operations and conditions are established these have to be implemented in a test.

Textmapping of main ideas in chosen texts

Having established the skills and strategies (see 2.1 – 2.3 and 3) we wanted to test located appropriate texts which lent themselves to the testing of these and which met the specified performance conditions (see 4.1), we moved on to constructing the test items.

The first step was to process the texts to establish the macro-propositions which might be extracted in line with the specified type of reading to be performed on that text (see 4.2). This utilisation-focused approach seeks to establish the main ideas of a passage through expeditious or careful 'textmapping' procedures.

Urquhart and Weir (1998) describe how in each textmapping an attempt should be made to replicate a single type of reading on a single text, e.g. *reading a text slowly and carefully to establish the main ideas*. The product of the particular reading of a text can be compiled in the form of a spidergram or as a linear summary. This is first done individually and then the extent of consensus with colleagues who have followed the same procedure is established. The objective of the procedure is to examine whether what we have decided is important, in line with the specified type of reading activity, matches what colleagues consider important.

Urquhart and Weir (1998) see this as a crucial first step in trying to ensure the validity of our tests. The answers to the questions developed equate with the important information in the text that could be extracted by the particular type of reading being assessed. An ability to answer the items should indicate that the candidate has understood the passage in terms of successful performance of the specified operation(s).

Procedures such as textmapping should enable us to determine in a principled fashion the content we might wish students to recover from a text according to the type of reading employed. However, the format, which acts as the vehicle for testing reading activities may constrain the operations and conditions we attempt to include. So as well as carefully specifying the latter we need to consider carefully the method we are going to use so as to minimise the influence of method on measurement of the trait (see 2.3 and 4.3). The cardinal rule remains however: we must first decide what types of reading we want to test and develop systematic procedures for deciding the micropropositions and/or macropropositions that we would expect candidates to extract from texts in performing these types of reading. Only then do we consider test formats and decide which will most faithfully mirror the procedures and allow the appropriate propositions to be extracted.

The mapping procedure will provide the content to be extracted for each of the types of reading in the test. It will also show whether each passage is suitable for its intended reading purpose. Where it is possible to produce a consensus textmap, this then needs to be converted into appropriate test items in the format selected. Where consensus is not achieved or the textmapping produces too few items these texts must be rejected!

A *posteriori* validation

Once the first version of the test was ready a *posteriori* validation procedures were applied at the trialling stage to determine statistically whether the test was working in the way it was intended to and to closely examine the construct underlying the test. The a *posteriori* statistical procedures employed in the two trials of the AERT in China are discussed in detail in 5.1 – 5.2 below. They indicate that the components in the test do seem to be measuring differing parts of the reading construct and students do perform differentially on different types of reading.

The stages of the quantitative and qualitative a *posteriori* validation of the AERT are briefly outlined next.

Trialling and quantitative analysis

It was important to trial the test on as broad a sample of the intended population in terms of ability as possible and then subject the results to statistical analysis to establish the test's value as a measuring instrument (see 5.1 – 5.2 below for detailed analysis).

The test data generated from the pilot and main trial were subjected to statistical analyses using the Statistical Package for the Social Science (SPSS). These included calculation of mean, discrimination and internal consistency as well as principal components analysis of the loading of test components, cross-tabulation of individual's performance on various components, ANOVA analyses to investigate differences in performance on components of the test in the whole sample and across disciplines.

It is important at the trialling stage to administer the research instrument to as normally distributed a sample as possible. This might mean purposefully sampling from top, medium and lower universities, institutions, schools and classes within these. Normally distributed data allows the researcher to apply the statistical analysis recommended below to establish how the items in the sub-tests are functioning. A skewed sample where the majority of students are too strong or too weak will not allow the researcher to do this.

To complement the result from the statistical analysis, qualitative data were collected through EAP reading experts' judgements on the skills and strategies tested in the AERT, students' introspection on the process of taking the AERT, and students' perceptions of the test conditions and the skills and strategies tested in the AERT (see 6.1 – 6.3 below).

Qualitative studies

The product of language tests tell us little of the processes that underlie reading and we need to employ different methodological procedures to investigate these. In particular introspective methods can help shed light on underlying thought processes. There are a number of problems associated with the method such as the time taken to administer and analyse, limited sampling and sensitivity to instructional variables. However, methods such as introspection and retrospection may offer insights into the perceived processes that take place during different types of reading and help us understand the nature of the differences in processing as well as the existence of such differences (Urquhart and Weir 1998). Such methods are considered in 6.1 – 6.3.

Internal statistical measures are necessary but not sufficient to establish the nature of the reading abilities under investigation. We needed to get a closer idea of what is actually happening during the test experience to accumulate evidence that the test is performing in an ecologically valid fashion i.e, in answering the items the students are processing text(s) as the test developer/

researcher intended them to. For example, if test takers were using test-taking strategies to avoid skimming or search reading through faulty item construction the test statistics would not necessarily tell us this. We needed to generate data on the process as well as the product. In 6.– 6.3 below we examine a number of ways in which this was done including survey, introspection and retrospection. A brief description of these is provided next by way of introduction.

Structured feedback from test takers

The intentions of the test developer are always mediated by the response of the test takers. Their attitude to facets of the test are as important as the evidence arising from the statistical data as it can often explain why things have happened in a certain way.

Data from structured questionnaires are important because they give us a broad based view of how the sample is responding to the test. A lot of the features examined through these sample questionnaires (e.g., familiarity with text topic) might impact adversely on the measurement of the construct if we have not done a proper job at the development stage. They act as a check on our ability to faithfully implement the test specification. As well as this broad-spectrum data we also need more in-depth information on our test items. This is provided by qualitative research procedures such as introspection, retrospection and expert judgement. As well as data relating to students perceptions of the instruments, texts and tasks, we are also interested in their views on what they thought the items in the test were actually getting them to perform in terms of the skills and strategies in our posited construct. Qualitative data obtained from introspection, retrospection and questionnaire survey provided us with process information on what the test-takers thought the test was testing. This can usefully complement the quantitative data obtained through test administration.

Introspection

An introspection study into the students' process of reading texts and answering the questions was carried out to find out what skills and strategies students were using in completing each section of the test (see 6.3 for details). The students were trained to think aloud onto tapes in a language laboratory while taking a test. Students were allowed to use L1 if they wanted to in their verbal reports. The data were then transcribed and content analysed in terms of the test operations.

Retrospection

A separate retrospection study enabled the researchers to obtain a larger data set (than is possible through the time consuming spoken protocols) to establish student perceptions of the skills and strategies used in the process of taking the test. This can be carried out in the large-scale trialling of the test. It can