INSTITUTE ASSESSMENT PROJECT

OCCASIONAL PAPER NO. 3

TEXTS AND TASKS: INVESTIGATING READING COMPREHENSION SKILLS IN ENGLISH AS A SECOND LANGUAGE

Doreen Wapshere

March 1995
Institute Assessment Project

Occasional Papers available:

Number 1: *Competency Based Assessment in Design Studies: Some Reflections*, Camille Ducker and Chris Klimek, August 1994

Number 2: *Retesting in Four Competency Based Courses in the Canberra Institute of Technology*, Joanna Karmel, November 1994
# TEXTS AND TASKS: INVESTIGATING READING COMPREHENSION SKILLS IN ENGLISH AS A SECOND LANGUAGE

## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRODUCTION</td>
<td>3</td>
</tr>
<tr>
<td>TESTING RECEPTIVE SKILLS</td>
<td>5</td>
</tr>
<tr>
<td>PROJECT 1: SAMPLING TEXTS</td>
<td>5</td>
</tr>
<tr>
<td>Method</td>
<td>7</td>
</tr>
<tr>
<td>Subjects</td>
<td>7</td>
</tr>
<tr>
<td>Text-types</td>
<td>7</td>
</tr>
<tr>
<td>Language tasks</td>
<td>7</td>
</tr>
<tr>
<td>Results</td>
<td>8</td>
</tr>
<tr>
<td>PROJECT 2: OPERATIONALIZING THE CONSTRUCT</td>
<td>10</td>
</tr>
<tr>
<td>Method</td>
<td>11</td>
</tr>
<tr>
<td>Results</td>
<td>12</td>
</tr>
<tr>
<td>CONCLUSION</td>
<td>14</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>15</td>
</tr>
</tbody>
</table>
INTRODUCTION

Two different types of assessment are used to measure the language ability of students studying English as a Second Language at Canberra Institute of Technology:

- proficiency tests, measured by the Australian Second Language Proficiency Ratings (ASLPR)
- competency-based tests, presented by the New South Wales Certificate in Spoken and Written English (CSWE) and Advanced Certificate in Spoken and Written English (ACSWE).

A scale which represents a continuum of language behaviour stretching from zero language through to native-like language is used to measure the ASLPR (Ingram, 1984). The scale is divided into nine levels of band descriptors of language behaviour (plus three intermediary levels which are not described), against which the language produced by students during oral interviews is compared. The language learners are given a proficiency rating for each of the four macro-skills: speaking, listening, reading and writing.

Competency in ESL is measured by the competencies taught and assessed in the three stages of the CSWE, which was registered and nationally accredited in 1992, and in the fourth stage, which constitutes the ACSWE.

The two types of assessment have certain similarities in that they are both criterion-referenced and are both based on suggested text types, but whereas the ASLPR measures the language that occurs during the performance of tasks, language in the CSWE is broken down into elements specified in performance criteria, each of which must be addressed by the tasks. To achieve a competency, the learners must satisfy all the performance criteria (Burrows, 1993). Thus, an essential difference between these two types of assessment is that the ASLPR measures total language ability, while the CSWE measures the ability to carry out specific text-based tasks.

Another very important difference between the ASLPR and the CSWE is that the former are based on language tests that are completely independent of any course or syllabus, while the latter is based on achievement assessment: the students have been formally taught the tasks they will be required to perform.
Completely different theories of what it means to know a language underpin the ASLPR and the CSWE. The ASLPR, a measure of general language proficiency, is based on the theory of language proficiency as a unitary ability (the Unitary Trait Hypothesis), which was rejected in 1983 (Oller, 1983). On the other hand, the CSWE is based on a social theory of language as text and is strongly influenced by the genre approach. Such a theory, where language is defined as essentially a variety of text types in many different contexts, also has its problems because this limited and narrow view of language does not take into account the need for learners to develop the strategic competence which will enable them to deal with the unexpected and the unknown in communicative situations. To state that certain communicative acts can only be performed by using specific linguistic criteria is a very strong claim. However, the principal funding bodies for adult ESL, the Department of Immigration and Ethnic Affairs (DIEA) and the Department of Employment, Education and Training (DEET) require both types of assessment for their records.

It was against this background of assessment in ESL that I carried out two research projects in 1993-1994.

The first project investigated the choice of texts made by ESL teachers to assess reading and listening skills at an intermediate level, ASLPR 2. Given the objective of the Adult Migrant Education Program (AMEP) to take its clients to the proficiency level of ASLPR 2 within an average of 510 hours of classroom instruction, reliable assessment is vital at this level. Students assessed as being at ASLPR 2 are considered to be functional in the English language by the funding bodies and are consequently no longer eligible for AMEP classes.

The second project was an investigation into the tasks used to test reading skills in the Australian Assessment of Communicative English Skills (access:), which is used off-shore to determine the English language proficiency of certain categories of non-English speaking migrants.

This occasional paper summarizes the principal findings of these two research projects.
TESTING RECEPTIVE SKILLS

There are two main problems associated with the assessment of reading and listening skills:

- the problem of sampling, that is the selection of written and spoken texts that will enable students to demonstrate their skills and
- the problem of operationalizing the construct, that is knowing whether the language skills being measured are, in fact, being operationalized.

PROJECT 1: SAMPLING TEXTS

For a reading test to have content validity, a representative selection of the texts that language learners would normally have to read should be selected, while for a reading test to have predictive validity, as large a number of texts as possible should be sampled in order to predict the reading ability of learners over a wide range of contexts. The normal procedure in the ESL departments is for teachers assessing reading skills to examine samples of texts for their syntactic and lexical complexity and familiarity of content so as to decide which level of proficiency or which stage of competency the texts will be appropriate for. In this they are also guided by the examples of text types deemed by test designers to be suitable for testing at certain levels. For example, texts suggested for use at ASLPR 2 are similar to those suggested for the measurement of reading competency at Stage 3 of Community Access in the CSWE. However, reading competency is measured by the ability to read four different text types: diagrammatic, procedural informational and formal letters, while the examples of text types for ASLPR 2 are more general (Figure 1). This is a deliberate strategy on the part of the test designers as the ASLPR was designed to measure general language proficiency and not the ability to carry out particular tasks. Indeed, the tasks must be varied to cater for the needs of individual learners, a policy which affects the reliability of the assessment.
In both methods of assessment, the texts are assumed to be equivalent; that is, any one text can be substituted for another and will sample the same underlying language ability. However, there is no evidence that this is indeed the case. Nor is there any evidence that texts considered to be simple or difficult by the teachers are found to be so by the students. Research has shown that choice of text can affect performance on reading tests and even reduce the effects of reading ability. This, in turn, affects the interpretation of the results and thus the reliability and validity of the tests. For example, the “everyday” reading texts suggested by the test designers as suitable texts to measure ASLPR are thought to be boring and irrelevant by ESL students with tertiary qualifications; as a result, they do not
respond well to this type of text, preferring the more expository texts, which are usually considered difficult by teachers.

Accordingly, two of the research questions explored in my first project were specifically concerned with the sampling of reading texts at an intermediate level of ESL.

1. Is it possible for teachers as test designers to select reading texts at the same level of complexity?
2. Are ESL learners with an academic/professional background better able to read expository texts than read texts met in “everyday” life?

Method

Subjects
A reading test was designed and trialled by 56 students from eighteen different language backgrounds, who were studying in three different Adult Migrant English Service (AMES) classes at the CIT. They had all completed secondary education in their countries and most had tertiary qualifications. The students were about to complete 510 hours of classroom instruction or were thought by their teachers to have a reading proficiency level of ASLPR 2.

Text-types
The reading test consisted of a large sample of twenty texts. Twelve of the texts satisfied the criteria for measuring reading proficiency at ASLPR 2, while the remainder were more specialized texts, representing those relevant to many professions and academic study. Included in the twenty texts were eight pairs of text types; for example, there were two advertisements, two sets of instructions for cleaning household appliances, two letters from a Real Estate agent regarding the same property, two handwritten postcards and so on.

Language tasks
Students were required to carry out various tasks when reading the texts: extract specific information, identify the purpose of the text, understand the gist, follow a sequence of events and transcode information from the diagrammatic texts.
Results

A split-half reliability coefficient was calculated, using the Spearman-Brown prophecy formula, to test reliability and internal consistency. With a coefficient of .86, the whole test was considered reliable.

The twenty texts were randomly placed to give three separate pairs of tests:

1. Two tests of ten texts each, containing an equal mix of “everyday” texts (those satisfying the criteria for measuring reading proficiency at ASLPR 2) and expository texts (those more representative of the texts studied by the students). Both tests were found to be reliable (Kuder-Richardson-21) and significantly highly correlated (Pearson’s $r$). These two reading tests were considered to be equivalent tests.

2. The whole test was divided a second time into “everyday” texts (ASLPR 2) and expository texts (academic/professional). These two tests were found to be significantly different (t-test). The students scored consistently higher on the expository texts.

3. The third test consisted of four pairs of texts considered by the teachers to be at the same proficiency level: the advertisements, the postcards, the instructions for cleaning two household appliances and the letters from the Real Estate agent. These were randomly selected to make two “similar” tests to assess reading proficiency according to the criteria of ASLPR 2. Both tests were found to be less reliable than the other tests (KR-21). Also, their means were highly significantly different (t-test), indicating that the two tests were not equivalent.

Furthermore, when a classical analysis of item facility was carried out on the eight pairs of texts in the whole test, it was found that only two pairs of texts - the postcards and the diagrammatic texts - had the same degree of difficulty. All the other pairs of texts, considered by the teachers to be equivalent, had quite different facility values (Figure 2).
These results have three important implications for testing reading skills at an intermediate level in ESL:

- a reading test that contains a number of different texts combining "everyday" language with more specialized language is both valid and reliable
- a reading test that contains only "everyday" texts will disadvantage students from an academic/professional background
- texts that appear to be of similar type and complexity are not equivalent; therefore, one text cannot be replaced by a text of similar type in order to equally sample the underlying trait.
PROJECT 2: OPERATIONALIZING THE CONSTRUCT

The second problem associated with the assessment of reading ability is whether the test items or questions really measure what they are intended to measure.

As designers of reading tests, teachers need to justify how they understand reading ability. If they see it as made up of a series of sub-skills or linguistic elements represented by performance criteria, then they need to be sure that the test items they design actually test these skills or criteria. One of the main problems in testing reading is that the students' answers are the only evidence available to us to show whether or not there has been comprehension. We have no idea what processes are used to arrive at the answers and usually assume that a correct answer indicates that the specified language skill has been used and that the performance criteria have been satisfied. Because reading is a mental activity, it is difficult to determine whether correct performance on specific texts indicates that the language criteria being measured are being used.

Thus, an important aspect of the construct validation of reading tests is to gather information on the processes used by candidates when they are taking the tests to try and discover whether they actually use the reading sub-skills or linguistic elements predicted by the test designers.

Most reading tests have tables of specifications detailed in terms of the components of ability to be tested. In English as a Second or Foreign Language these tables are often a selection of reading sub-skills which have been extracted from Munby's (1978) taxonomy of language skills. For example, the reading component of the English Language Test, now the International English Language Test (IELTS), is based on fourteen sub-skills selected from Munby's taxonomy. Similarly, the Australian Assessment of Communicative English Skills (access:) lists nine language operations as being tested by the items in the Reading Skills Test, and the performance criteria listed for each reading competency in the New South Wales CSWE represent the specifications of linguistic features which must be demonstrated for the students to achieve competency in reading at each stage.

In addition, there is an implication in most reading tests that there is a hierarchy of reading sub-skills, that is that some sub-skills are easier and are acquired before
others. For example, the performance criteria assessed to achieve the competency of *ability to read informational texts* in Stage 1 of the CSWE are as follows:

- understands overall meaning of text
- recognizes layout
- demonstrates understanding of key vocabulary
- locates required specific information.

These criteria are repeated at all stages of the certificate. However, one or two more are added on at each stage, indicating that the designers see the criteria as a cumulative hierarchy. For instance, the following performance criteria are added successively: "scans for gist" at Stage 2; "scans for main points" and "understands purpose" at Stage 3; and at Stage 4, "understands logical links in the text achieved through conjunctions and reference".

The second research project tested these assumptions by posing four questions.

1. How far do experienced ESL teachers and test designers agree on the relative difficulty of reading sub-skills?
2. How far do experienced ESL teachers and test designers agree on the reading sub-skills being tested by the test items?
3. Are experienced ESL teachers and test designers able to predict the difficulty of the test items?
4. Is there any correspondence between the reading sub-skills supposedly tapped by test items and test taker behaviour?

**Method**

A partial replication of the research carried out by Alderson and Lukmani (1989) and by Alderson (1990) was attempted, using the Reading Skills component of the access: test. Nine experienced teachers from the three ESL departments at CIT participated in the study together with the two test designers of the access: Reading Skills test.

141 students from the ESL Academic/Professional Department sat the test and a Rasch Item Response Theory (IRT) analysis was computed from their scores. The
IRT measures the relationship between candidates’ ability and item difficulty in logit values.

Finally, three students did the reading test introspectively. By using this “Think-Aloud” protocol, it was possible to determine the actual processes used by the three students to answer the questions.

Results
The following results were all statistically tested, using appropriate parametric and non-parametric methods.

1. The teachers’ rankings of the reading sub-skills according to difficulty were related. However, the teachers’ rankings did not correlate with the rankings of the test designers (Kendall’s W). Only two of the teachers agreed with the test designers on the relative difficulty of the reading sub-skills (Spearman’s rho). More importantly, there was no correlation between the sub-skills found to be difficult by the students and the rankings of either the teachers or the test designers (Kendall’s W). For example, the students found the sub-skills of “identifying the main ideas of the text” and “the purpose of the text” (Stage 3 in the CSWE) to be far more difficult than “understanding the logical relationships in the texts” (Stage 4 - the ACSWE).

2. Agreement between all the judges (teachers and test designers) on the reading sub-skills tested by the items in each of the four texts ranged from 23% to only 10% (Cohen’s KAPPA).

3. The teachers and the test designers were able to predict the difficulty of 50% of the test items (ANOVA). When the items were divided into twenty easy ones (negative logit values) and twenty difficult ones (positive logit values), it was found that teachers predicted 60% and test designers 67% of the easy items, but were less successful at predicting the difficult items (teachers, 3% and test designers, 19%).

4. The introspective study of the test taking processes used by the three students found that the predicted reading sub-skills were used for most
of the easy items although using those predicted sub-skills did not always mean that the students gave the expected answers. For example, two of the students used the predicted sub-skills to answer eight of the ten items on which more than half of the teachers agreed with the test designers on the sub-skills being tested. However, correct answers were given to only three of the eight items by these students. On the other hand, the three students often gave correct answers to items without using the predicted sub-skills. The test items where the students used different or additional sub-skills to those supposedly being tested tended to have high logit values. For example, all three students used the following four language operations to respond to one of the most difficult items in the test: *identifying explicitly stated information, identifying information stated in paraphrase, identifying participants and roles through use of grammatical cohesive devices and recognizing logical relationships between ideas or information within the text*, but according to the test designers, this particular item tested only the ability to *identify explicitly stated information*.

Finally, as a postscript to the first part of my paper on the problem of sampling texts in reading tests, there was an assumption on the part of the teachers who participated in the second study that the texts were arranged in a sequence of difficulty, that is from a simple “everyday” text to a semi-technical report. Although only 40% of the students at lower levels of language proficiency attempted Text 4, the semi-technical report, because of time constraints, all the students completed Texts 1, 2 and 3. From the logit values, it was found that Text 1, the “everyday” text in simple, familiar prose, was the most difficult for the students, while Text 3, a densely written newspaper report full of argument and opinion, was the easiest. Thus, it can be inferred again from these results that teachers cannot necessarily predict the relative difficulty of texts and that texts using “everyday” language do not provide appropriate evidence of the reading ability of academic/professional students.
CONCLUSION

The results of these studies have shown that language teachers designing reading tests:

- should not assume that texts which appear to be of similar type and complexity are of comparable difficulty.
- should not assume that it is possible to predict the difficulty of test items
- should not assume that certain reading sub-skills are more difficult than others
- should not assume that other teachers will agree on the reading sub-skills being tested by the test items or even interpret them in the same way
- should not assume that students will use the specified reading sub-skills to answer the items correctly or even answer correctly when they do use these sub-skills.

The findings of the two studies have serious implications for competency-based testing of reading ability as presented in the CSWE. Firstly, a demonstrated competence to read one text does not imply an ability to read other texts of a similar type. Secondly, the claim that certain reading sub-skills can be tested by specific items is not supported by the second study. Any claims that students have achieved competency in reading certain text types because all performance criteria have been observed are therefore suspect. The converse also applies. Finally, the studies indicate that any division of levels of competence in reading based on predictions of text and task difficulty and/or on a presumed hierarchy of performance criteria needs further investigation.
REFERENCES


