Email addresses: Low_ying_ping@seab.gov.sg
Keith_john_lenden-hitchcock@seab.gov.sg
Five key words: Assessing Language Use, Reading Comprehension

Assessing Language Use in Reading Comprehension?

Low Ying Ping, Keith John Lenden-Hitchcock
Singapore Examinations and Assessment Board

While it is a common practice for reading comprehension to be assessed using short-answer questions based on a passage, the criteria for marking could vary. It is generally recognised that in assessing students' responses to these questions, errors in language use, namely, grammar, spelling and punctuation, should not be penalised, so as to preserve the integrity or 'purity' of the reading comprehension construct. However, when reading comprehension is assessed as part of a large scale examination, there is concern among markers that failure to penalise for errors in language use would result in a washback effect whereby teachers place less emphasis on the teaching and learning of correct grammar, spelling and punctuation. This is especially pertinent when the examination is offered by relatively young learners at the primary level, where a strong foundation in the mechanics of using the language is seen to be very important. A research study was thus carried out on a sample of Grade 6 students to provide further insight into this issue. In particular, the study seeks to find out how the different ways of assessing language use in a Reading Comprehension paper affect the test scores across different ability groups. Also, do the test scores differ significantly if language use is not assessed, and in what way? This paper discusses the preliminary findings from the study.

Background

Reading comprehension can be assessed via various techniques. Common item types include selective deletion gap filling or cloze, short-answers answers based on a passage, information transfer tasks (e.g. from a passage into a diagram) - a variant of the short-answer question (Weir, 2005), multiple choice questions based on a passage, true-false items and so on. Most tests would incorporate a number of different techniques. Objective methods, for example, could complement more subjectively evaluated methods to give a more adequate representation of a candidate's proficiency (Alderson, 2000). The item types selected, as well as the marking criteria, depend on the purpose(s) that the test is intended to serve. Bachman identifies two major uses of language tests: firstly, to provide information for decision-making in educational contexts, and secondly, to provide data on linguistic abilities or attributes for research on language, language acquisition, and language teaching (1990).

The advantage of employing short-answer questions based on a passage in testing reading comprehension is that as constructed response items, the answers have to be sought and expressed by the student, rather than being provided. This facilitates the testing of higher order skills such as interpretation and evaluation, and allows the assessor to reasonably assume that should the student get the answer right, it is not for reasons other than that he has comprehended the text (Weir, 2005).

Such an item type is widely used in various English Language tests around the world. For example, the Progress in International Reading Literacy Study (PIRLS), the Programme for International Student Assessment (PISA) and the University of Cambridge International Examinations' IGCSE First Language English all include reading comprehension components structured around a passage with

accompanying questions. While their test purposes may differ,[1] the primary aim of all these reading comprehension tests is to measure reading literacy. Hence, for such tests, the "mechanical accuracy criteria (grammar, spelling, punctuation)" are not featured "in the scoring system as this affects the accuracy of the measurement of the reading construct" (Weir, 2005).

However, in assessing reading comprehension as part of a large scale, high stakes examination, failure to penalise for errors in language use could lead to the perception that correct grammar, spelling and punctuation is not important. This could result in a negative impact if teachers place less emphasis on the teaching and learning of correct grammar, spelling and punctuation. Such an undesirable effect of teachers teaching to the test is especially unwanted when the examination is offered by relatively young learners at the primary level, where a strong foundation in the mechanics of using the language is seen to be very important.

To address such pedagogical concerns, one might seek to assess language use together with reading comprehension. Yet, in assessing language use in a reading comprehension test, one might worry that the integrity of the reading comprehension construct would be compromised. But in practical terms, how much difference does keeping the assessment construct pure and conflating the two purposes of pedagogy and assessment make?

This paper offers a discussion of the preliminary results of a study that was carried out to find out if the test scores of a reading comprehension test differed significantly when language use was taken into account. And if language use was taken into account, would marking for language use holistically and marking it item by item make any difference to the test scores?

Methodology

For this study, a sample of 204 Grade 6 students from schools across a few countries studying English as a second language was given a reading comprehension test. These students were selected to represent the spectrum of language abilities, and had equal representation of boys and girls across the spectrum.

The test comprised a passage with ten reading comprehension items based on the passage. Each item had a maximum of 2 marks for Content. The total marks for the script was 20. Each script was marked in the following five ways:

Method 1:
Each item was first scored for content (0 mark, 1 mark or 2 marks, depending on the number of relevant points given by the student). Language use was not assessed at all.

Method 2:
For each item, marks were awarded for content in the same way as described in Method 1. Then, for each item, marks (0 mark, ½ mark or 1 mark, depending on the severity of the errors) were deducted for

[1] PISA "assesses how far students near the end of compulsory education have acquired some of the knowledge and skills that are essential for full participation in society." (PISA, 2010). PIRLS measures "trends in children's reading literacy achievement" where "reading literacy" is defined as understanding, using, and reflecting on written texts, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society." (PIRLS, 2010). The Assessment Objectives for reading in the IGCSE First Language English are to "understand and collate explicit meanings; understand, explain and collate implicit meanings and attitudes; select, analyse and evaluate what is relevant to specific purposes and understand how writers achieve effects". (CIE, 2010)

language errors. This could be interpreted to mean that language use comprises up to 50% of the assessment, since the marks deducted for language errors are capped at 10 marks (10 questions x 1 mark), although it should be noted that this assumption is simplistic as items scoring 0 in content would not be assessed for language at all.

Method 3A:
For each item, marks were awarded for content in the same way as described in Method 1. The entire script was then studied holistically to award an overall mark for language use (0-5 marks, based on a set of band descriptors). The content marks were added to the holistic language marks to obtain a maximum of 25, which was then scaled down to 20, so that the total marks for the script still amounted to 20, for ease of comparison across the methods. This meant that language use amounted to 20% of the assessment.

Method 3B:
Each item was marked as per Method 3A. That is, for each item, marks were awarded for content in the same way as described in Method 1. The entire script was then studied holistically to award an overall mark for language use (0-5 marks, based on a set of band descriptors). However, this time the content marks were scaled from a maximum of 20 to 15, so that when added to the holistic language mark, the total marks for the script still amounted to 20. This meant that language use amounted to 25% of the assessment.

Method 3C:
Each item was marked as per Method 3A. However, this time the content marks were scaled from a maximum of 20 to 10, and the holistic language marks were scaled from a maximum of 5 to 10. The two scaled scores were then added to obtain a maximum of 20. This meant that language use amounted to 50% of the assessment.

The five methods are summarized in Table 1 below:

Method	**Content**	**Language Use**	**Total score**	**Percentage of holistically scored marks for Language Use in Total Score**
1	Scored by item	Not assessed	Content (max of 20 marks)	0
2	Scored by item	Scored by item (negative marking)	Content (max of 20 marks) – Language errors (0-1m per item)	0% (but up to 50% in negative Language marking)
3A	Scored by item	Scored holistically	[Content (20 marks) + Holistic Language (5 marks)] scaled from max of 25 marks to 20 marks	20%
3B	Scored by item	Scored holistically	Content (scaled from max of 20 marks to 15 marks) + Holistic Language (5 marks)	25%
3C	Scored by item	Scored holistically	Content (scaled from max of 20 marks to 10 marks) + Holistic Language (scaled from max of 5 marks to 10 marks)	50%

Table 1: Summary of methods adopted for marking

Findings and Discussion

Since it is of concern that the reading comprehension construct would be muddied as a result of assessing language use, various ways of assessing language use together with content were studied by generating Pearson correlation coefficients between the various methods of marking.

With Method 1[2] as the control, the effect of marking using Method 2 was examined. In Method 2, each item was marked for content, then had marks deducted for language use errors. Surprisingly, the correlation coefficient between the content score and the Method 2 score was 0.97 (p=0.0001), showing that this way of integrating assessment of reading comprehension and language use did not muddy the reading comprehension construct significantly.

Next, Method 3A [$(Content + Holistic) \times \frac{20}{25}$] was examined. It was found that the correlation coefficient between the content score and the Method 3A score was 0.97 (p=0.0001). The correlations for Method 3A and Method 2 were comparable even though Method 3A accounted for a fixed proportion of 20% of the total score. This suggests that when language use is assessed holistically and hence separately from the content, the effective 20% weight on language in the total score for Method 3A and the variable weights of up to 50% weight on language have the same minimum effect. The extremely close correlation coefficients between content and final score for Method 2 and Method 3A suggest that these two methods are both fairly good approximations to marking the scripts as a purely reading comprehension construct.

For Method 3B ($\frac{15}{20} \times Content + Holistic$), the correlation coefficient with the content score was lower than that of Method 3A at 0.94 (p=0.0001), possibly due to the blurring of the reading comprehension construct, since language use now accounted for a higher proportion (25%) of the total score.

For Method 3C, with language weighted similarly to Method 2, but marked holistically, the correlation coefficient (0.77) with content was inevitably found to be lower than those of Method 3A and Method 3B, since language use now accounted for an even higher (50%) proportion of the total score. Predictably, it was also lower than that of Method 2, reinforcing the earlier findings that scoring language use separately as a holistic component and giving the component a greater weight in the total score results in greater interference of the reading comprehension construct.

Figure 1 below illustrates how a composite score based on content and holistic language correlates with the content score depending upon the weight placed on either of these components.

[2] Method 1 only assesses content. Hence, when 'content score' is mentioned, it also refers to the Method 1 score.

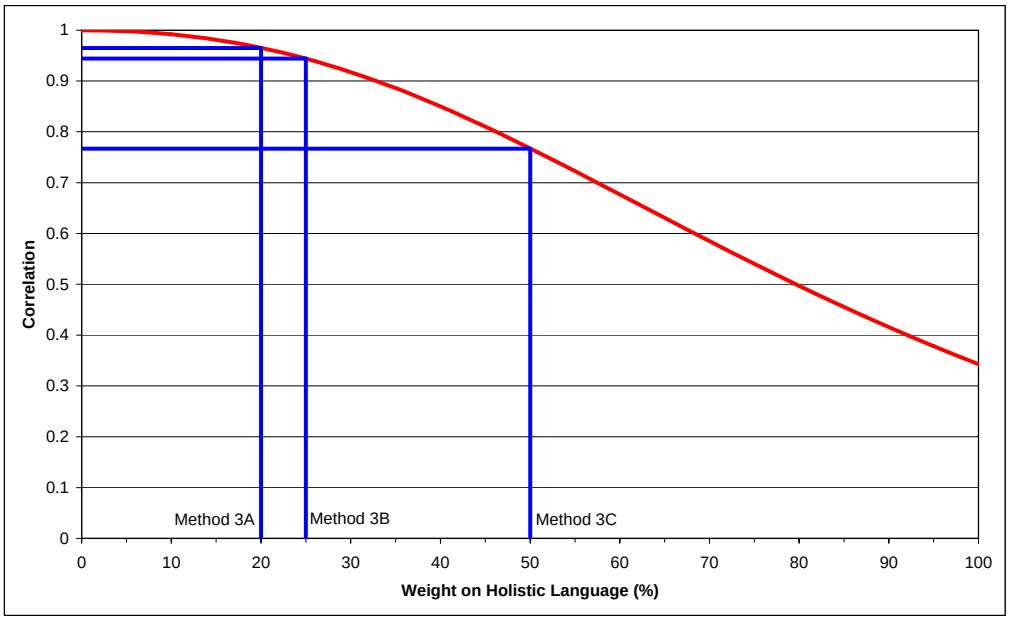

Figure 1: Correlation of composite score based on content and holistic language with content score

For example, a weight of 20% (Method 3A) on language corresponds to a correlation of 0.97, 25% (Method 3B) to 0.94 and 50% (Method 3C) to 0.77. The graph shows that the higher the language weight, the more muddied the reading comprehension construct. This strongly suggests that should language be assessed together with reading comprehension to address a possible negative impact on the teaching of grammar, spelling and punctuation, the language weight should not be too high. More studies however would need to be carried out to ascertain whether there is any difference in the ways the high ability and low ability groups are affected when the various methods of marking are used.

When the data was broken down by gender, the distribution of scores for boys and girls showed that whereas there was very little difference between boys and girls in their content scores (see Figure 2), a higher proportion of girls scored into the higher mark range for the holistic language score (see Figure 3). For example, about 45% of girls obtained a score that was higher than 3 but only about 30% of boys did the same. The better performance of the girls in language use is partly borne out in their higher holistic language mean score of 3.27 versus the boys' score of 3.15 though the difference of 0.12 is not statistically significant. This is not surprising since girls typically outperform boys in language subjects. The finding suggests that including language use in the assessment of reading comprehension would benefit the more able girls in comparison with boys of the same content ability. However, further study would be required to verify such a hypothesis.

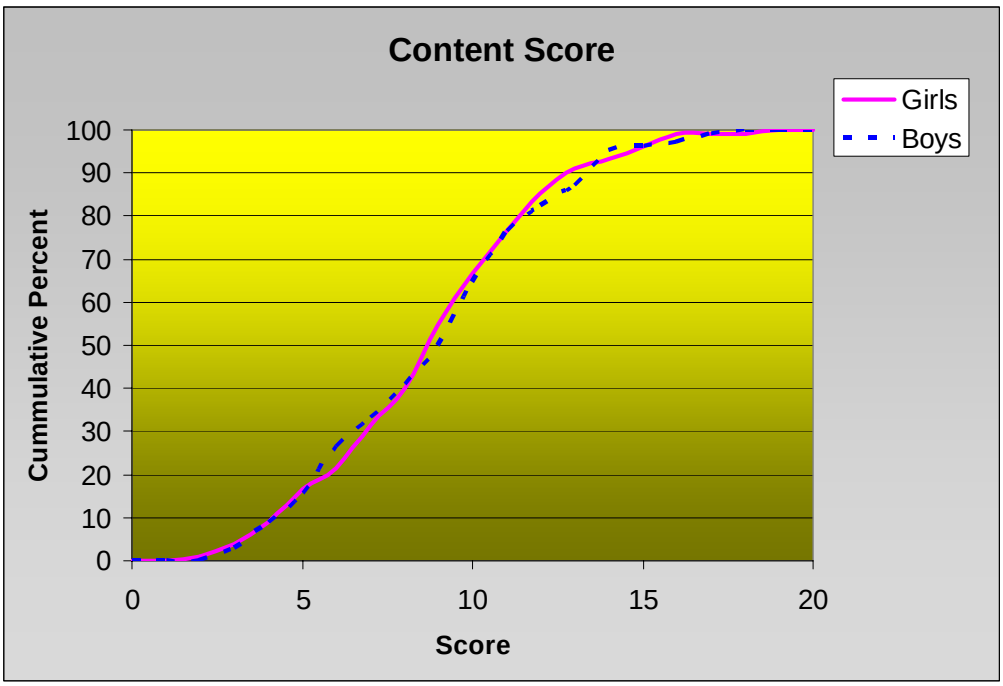

Figure 2: Cumulative frequency of content scores by gender

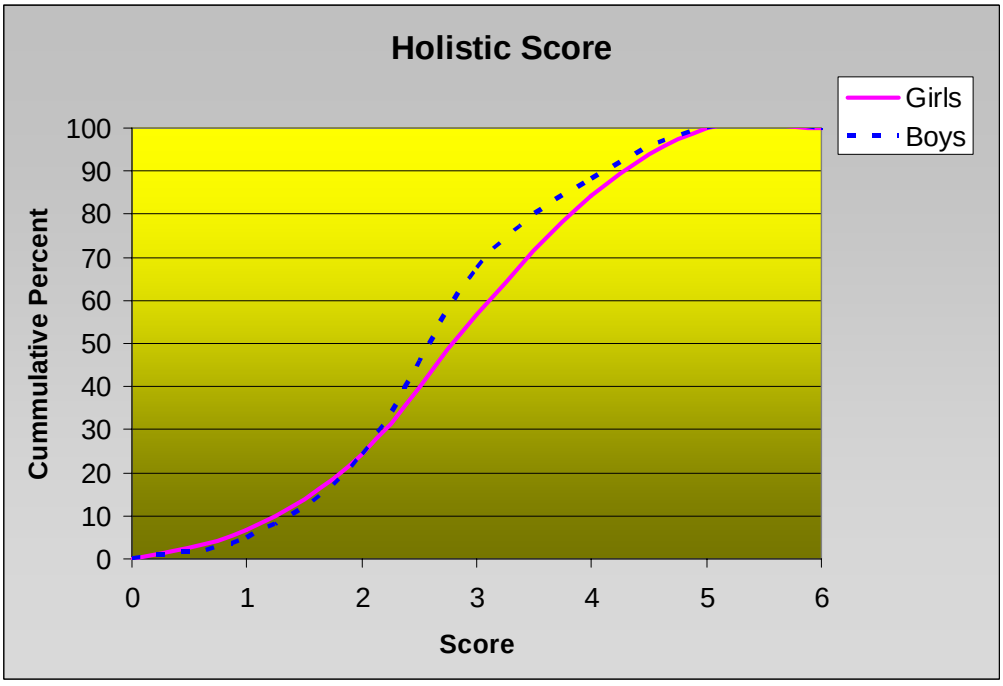

Figure 3: Cumulative frequency of holistic language scores by gender

However, despite the above findings, the correlation coefficients between the content score and the final score for each method of marking (see Table 2) showed that there were only some very slight

differences between the boys and the girls. This was true even of Method 2 and Method 3C, where language accounted for 50% of the assessment.

Method	Pearson Correlation Coefficient between content and final score Girls: N = 102	Pearson Correlation Coefficient between content and final score Boys: N = 102
2	0.98	0.97
3A	0.94	0.95
3B	0.96	0.97
3C	0.76	0.78

Table 2: Correlation coefficients by gender

The correlation between the content score and the holistic language score for boys and girls were also measured and found respectively to be 0.34 and 0.35. Importantly, the two figures do not differ significantly. Thus, while this shows that there is some correlation between language ability and being able to get the content correct, there is no evidence here that it affects boys and girls differently as long as the weight put on the language component is not too high.

Concluding Remarks

The findings of this study suggest that if the pedagogical reasons for wanting to include the assessment of language use within a reading comprehension paper are strong enough, it is possible to do so without compounding the measurement of the reading comprehension construct too much. However, it might be preferred to keep the weighting of the language use component lower to preserve the 'purity' of the construct as much as possible.

In this light, it could be suggested that Method 2 be adopted, since the correlation between content and final score is among the highest using this method. However, feedback from markers was that this way of assessing language use (by marking first for content, then deducting marks for language errors for each item) results in multiple penalties for the same mistake made in different items. Furthermore, as each item is marked individually for language, the range of marks is limited (0, ½ mark or 1 mark deducted for each item for language errors), resulting in difficulty in differentiating between students who committed language errors of varying degrees of severity (e.g. punctuation error, errors of agreement, and structural errors).

In contrast, the holistic method of marking language (Methods 3A, 3B and 3C) would not encounter the above problems cited for Method 2. Since the correlation between content and language for Method 3A is similar to that of Method 2, it might be preferable to adopt this method. In addition, the study suggests that the inclusion of the assessment of language use would have a slight negative bias against boys. Hence, it is preferable that the language use weighting be kept low, as is the case for Method 3A.

As this study was carried out on a small sample of second language users, further research is necessary to ascertain if the same results would be obtained with first language users. More research would also be conducted to study the gender differences in the effects of the various methods of marking, as well as how the different ability groups would be affected.

References

Alderson, J.C. (2000). Assessing Reading. Cambridge: Cambridge University Press.

Bachman, Lyle F. (1990). Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

Progress in International Reading Literacy Study (2010, May). Retrieved from http://timss.bc.edu/pirls2006/about.html

Programme for International Student Assessment (2010, May). Retrieved from http://www.oecd.org/pages/0,3417,en_32252351_32235918_1_1_1_1_1,00.html

University of Cambridge International Examinations (2010, May). Retrieved from http://www.cie.org.uk/qualifications/academic/middlesec/igcse/subject?assdef_id=852

Weir, Cyril J. (2005). Language Testing and Validation: An Evidence-Based Approach. Basingstoke, England: Palgrave MacMillan.