4-2012

A framework for predicting item difficulty in reading tests

Tom Lumley
ACER, tominoz2002@zoho.com

Alla Routitsky
ACER, alla.routitsky@acer.edu.au

Juliette Mendelovits
ACER, mendelovits@acer.edu.au

Dara Ramalingam
ACER, dara.ramalingam@acer.edu.au

Recommended Citation

Lumley, T., Routitsky, A., Mendelovits, J., & Ramalingam, D. (2012). A framework for predicting item difficulty in reading tests. https://research.acer.edu.au/pisa/5

A Framework for Predicting Item Difficulty in Reading Tests[1]

Tom Lumley, Alla Routitsky, Juliette Mendelovits and Dara Ramalingam
Australian Council for Educational Research

Results on reading tests are typically reported on scales composed of levels, each giving a statement of student achievement or proficiency. The PISA reading scales provide broad descriptions of skill levels associated with reading items, intended to communicate to policy makers and teachers about the reading proficiency of students at different levels. However, the described scales are not explicitly tied to features that predict difficulty. Difficulty is thus treated as an empirical issue, using a post hoc solution, while a priori estimates of item difficulty have tended to be unreliable. Understanding features influencing the difficulty of reading tasks has the potential to help test developers, teachers and researchers interested in understanding the construct of reading.

This paper presents work, conducted over a period of more than a decade, intended to provide a scheme for describing the difficulty of reading items used in PISA. Whereas the mathematics research in earlier papers in this symposium focused on mathematical competencies, the reading research concentrates on describing the reading tasks and the parts of texts that students are required to engage with.

The PISA 2000 Reading Literacy Framework (Kirsch et al., 2002) drew on work conducted by Kirsch and colleagues (e.g. Kirsch, 2001; Kirsch & Mosenthal, 1990), who developed a framework that was able to predict with a high degree of accuracy the difficulty of a range of items used in the context of adult literacy assessment. PISA differs from material used in those assessments described in because of what is included in the test: a wider range of text types (narrative texts, for example, form a significant proportion of the texts used); and a focus on wider range of functions and reading processes, especially the inclusion of the aspect (or reading process) of reflection and evaluation, which requires students to relate their prior knowledge to a features of the text’s content or form.

The rating scheme that was developed went through a series of stages, adding to and modifying the criteria proposed by Kirsch and Mosenthal, in order to take account of the nature of the tasks and texts used in PISA.

Initial attempts during the first PISA cycle (2000) to produce a scheme describing all variables focused on the framework variables of text format (continuous or non-continuous) and aspect (cognitive process), and on ways of applying variables differently to these variables In subsequent work, a scheme was sought that was applicable to all items and texts.

One important distinction derived from Kirsch and Mosenthal’s (1990) work concerns necessary (textual) information, the information that is required to do the task (*required information*), in contrast to target information, the information that the reader needs to provide to gain credit (*requested information*). The focus, then, is not on the difficulty of an entire text, but on the relevant parts, containing information needed by readers to respond to tasks. This distinction recognises that tasks drawing on the same text may vary very substantially.

[1] Paper presented at the American Educational Research Association Meeting, Vancouver, April 2012

In work conducted in 2008 and 2009, a scheme was developed composed of ten variables. This stage showed that agreement amongst raters using the scheme to rate a selection of 100 PISA reading items was modest and variable. In subsequent revisions to the scheme, one variable was replaced, and the four different levels, or steps of difficulty associated with each variable, were defined. A group of three experts, with extensive experience as item developers and coder training for PISA, produced a set of consensus ratings for a set of 84 items (100 score points) using the ten variables making up this revised scheme (Table 1).

Table 1. Revised PISA reading item difficulty scheme: proposed variables.

1. **Number of features and conditions**. This variable relates to the number of features that need to be located in the text in order to provide an adequate response to gain credit for the question.
2. **Proximity of pieces of required information**. This refers to the identification of pieces that need to be put together in order to answer the question. The rating is determined by the proximity of the relevant pieces of information to each other.
3. **Competing information**. This refers to information in the stimulus and/or in the distractors (if multiple choice) that the reader may mistakenly select, or that the reader may generate, because of its similarity in one or more respects to the target information.
4. **Prominence of necessary textual information**: the prominence of the necessary information - that is, information in the text that is needed to answer the question (even if it is not sufficient by itself to answer the question).
5. **Relationship between task and required information**: the relationship between the question (the whole task, including the multiple-choice options where relevant) and the required information - that is, the kind of answer required to gain credit.
6. **Semantic match between task and text**: the degree to which there is a semantic match between the wording of the task and the necessary information - that is, information in the text that is needed to answer the question (even if it is not sufficient by itself to answer the question).
7. **Concreteness of information**: the kind of information that readers must identify to complete a question (Mosenthal, 1998).
8. **Familiarity of information needed to answer the question**. This variable distinguishes tasks that focus on information inside or outside the text, or the text structure, that is close to the experience and concerns of the reader, from those focusing on what is likely to be remote and unfamiliar.
9. **Register of the text**. This refers to the stimulus or part of the stimulus that the reader needs to refer to in order to complete the task. It takes into account the implied relationship between the reader and the text, and the lexico-grammatical density.
10. **Extent to which information from outside the text is required to answer the question**. This variable deals with the extent to which the reader needs to draw on world knowledge, experience or personal beliefs and ideas and opinions in order to answer the question.

Simple correlations between ratings of each item according to its perceived difficulty on each of these criteria (variables), and the empirical difficulty of the item, were calculated. As a result, several of these variables were identified as having correlations with empirical difficulty above 0.5, namely

- 8. Familiarity of information needed to answer the question.
- 5. Relationship between task and required information.
- 3. Competing information (in task or text).

Likewise, a simple regression analysis, assuming that all variables would fit to the same regression line, identified these three variables as explaining the greatest amount of variance.

The ratings represented a consensus rating arrived at by a set of three experts, all of whom were involved in the test development process.

Prompted by this result, a new set of research questions was generated:

- How well do trained raters, with no previous knowledge of the reading item difficulty rating scheme, agree with each other and with the consensus of experts in the application of the scheme's variable to a set of PISA reading items?
- How well does the scheme developed in this project predict reading item difficulty?
- Which of the ten variables in the scheme are the best predictors? Is there evidence that any variables may be discarded?

Five raters were recruited. Four were Australians, with experience as coders of constructed response items for PISA or similar reading tests, but with no other familiarity with PISA or the process of test development. The fifth rater was Hungarian, and had worked with PISA for many years: in terms of test development experience, this rater had more in common with the expert raters. The Australian raters worked with the English source version of the PISA test materials, and the Hungarian with the translated PISA Hungarian test materials. An additional question, therefore, concerned the extent to which the scheme appeared to work with test materials in a language very different from English: was there evidence of disagreement, in the form of higher or lower ratings, or ratings that showed bias with individual variables.

When considering agreement amongst raters, two measures were considered. One measure looked at bias, or patterns of harshness by individual raters on individual variables in relation to a standard, defined as the consensus of the expert raters. The second measure considered absolute differences amongst the five raters.

The results of the bias analysis show that, using the consensus of the expert raters as a benchmark:

1) For four variables the Australian coders gave higher scores than the expert consensus (1. *Number of features and conditions*, 5. *Relationship between task and required information*, 6. *Semantic match between task and target information* and 10. *Use of information from outside the text*)
2) Two criteria are given a slightly lower code on average by the five coders than the expert consensus: *2. Proximity of pieces of required information* and 8. *Familiarity of information needed*.
3) The other four criteria were coded either very consistently by the raters in comparison to the expert consensus or did not show a clear pattern of bias.
4) The Hungarian rater was not consistently different from the Australian raters, and generally tended to give ratings that were closer to the ratings of the experts.

A measure of reliability coding, considering the absolute agreement for each rater compared to others, showed that there was reasonable agreement. The average discrepancy for each rater for each variable and each item, compared to all other raters, was in the vicinity of 0.8 of a step.

It may be concluded from this that training is necessary to use the rating scheme; and that those with a greater level of expertise in test development might be more likely to show greater agreement in its application than those with less experience. All raters commented that the task of producing so many ratings was taxing.

Results from the multiple regression analysis included a more surprising finding. The multiple regression with the highest explanatory power shows five variables with significant coefficients, as shown in Table 2.

Table 2.

Criterion	Coefficient
3. Competing information	.587
5. Relationship between task and required information	.566
7. Concreteness of information	-.386
8. Familiarity of information	.377
10. Reference to information from outside the text	.277

The adjusted R Square for this regression is 0.569: it explains about 57% of variability in difficulty of items.

Previous analyses had suggested that the framework feature, ‘Aspect’, might contribute significantly to difficulty, although this was not an intention of the framework. The multiple regression analysis showed that ‘Aspect’ has a coefficient which is not significantly different from zero, which means that the five variables listed in Table 2 account for all the predictive possibility offered by Aspect.

As with analyses based on earlier schemes, three variables emerged as especially important in explaining variance:

- 3. Competing information (in task or text).
- 5. Relationship between task and required information.
- 8. Familiarity of information needed to answer the question.

Two other variables were identified as contributing to explanation of variance:

- 10. The extent to which information from outside the text is required to answer the question.
- 7. Concreteness of information.

The multiple regression analysis showed that the remaining five variables did not add to the explanatory power of the rating scheme. In effect, this means that the same results can be obtained by reducing the number of variables by half, with a corresponding reduction in the amount of work required.

The first four variables referred to above correlated positively with item difficulty. The fifth variable, 7. Concreteness of information, which on its own correlated modestly but positively with item difficulty, was also found to be significant in the multiple regression analysis, but ratings had a *negative* relationship with item difficulty, once the four variables listed above were taken into account. Removing this variable lowered the explanatory power of the data, including the amount of variance explained by each of the other four variables listed here.

This finding requires comment. Essentially, what it is saying is that as an item becomes more difficult, and as the first four variables mentioned contribute to explanation of item difficulty, the degree of abstractness of the information readers need relates negatively, but significantly, to the item's difficulty.

One hypothesis for why this might be so is that PISA is designed for 15-year-olds. The assessment needs to cover the entire spread of ability of students of that age. The overall level of abstraction is not dramatically high in PISA. The consequence is that test developers do not generally aim to write items that are difficult on *all* features, nor easy on *all* features. A task that 'focuses on information that is very unfamiliar and remote from typical readers' experience' (8. Step 4); and that also requires a high level of interpretation to determine its scope and nature (5. Step 4); and that offers highly plausible competing, ambiguous or distracting information in the text and/or the question (3. Step 4); and that requires students to make links with outside knowledge with little support from the text or question (10. Step 4); is unlikely also to include a high level of abstractness; in effect, students are asked to bring complex conditions to bear on concrete situations. This is because such items are likely to be extremely difficult, and better suited to, say, university graduates. Experience with PISA suggests that extremely demanding items tend to produce poor data, as they are inaccessible to too large a proportion of the population. Conversely, a task that is relatively easy on those four variables (competing information, relationship between the task and the text, familiarity of information, reference to prior knowledge) is more likely to include a level of abstraction, in order to offer a realistic reading challenge.

Conclusion

This paper has identified that a rating scheme using ten variables to predict difficulty of items can be used with reasonable success by trained raters. Experts, those with experience of writing items, are more likely to agree than those without such expertise. There is a reasonable level of agreement amongst raters who undergo some training with the scheme. There is some evidence that an expert rater working in another language (Hungarian) is equally able to apply the scheme as those working in English; indeed, the data support the view that an expert may achieve higher levels of agreement with a consensus rating of experts than do regular raters with limited training.

The multiple regression analysis suggests that the rating scheme can be made more efficient by reducing the number of variables included from ten to five, without loss of predictive power of the framework. This analysis shows the surprising finding that the concreteness / abstractness of information is important in explaining item difficulty, but that once other factors are taken into account, it has a negative relationship with item difficulty. A possible explanation for this finding is that reading tasks for 15-year-olds cannot be hard on all variables, and that the level of abstraction tends to be reduced as the other variables increase in difficulty.

It is important also to note that the results should be treated with a degree of caution because of the size of the data set examined in this paper. That is, although it appears justified to recommend a reduction in the number of variables used for prediction, the regression coefficients for these five variables will probably change with a different set of items.

Further steps in this research project will involve the use of item developers applying the variables in the process of writing items, in order to manipulate their difficulty. If this can be achieved, the scheme can act as a tool to allow test developers to better target a test to the ability of the population and teachers to better target their teaching by understanding what features contribute to item difficulty.

References

Kirsch, I. 2001. *The International Adult Literacy Survey (IALS): Understanding What Was Measured. Research Report RR-01-25*. Princeton, NJ: Educational Testing Service.

Kirsch, I., & Mosenthal, P. B. (1990). Exploring document literacy: Variables underlying the performance of young adults. *Reading Research Quarterly*, *25*(1), 5-30.

Kirsch, I., deJong, J., Lafontaine, D., McQueen, J., Mendelovits, J., & Monseur, C. (2002). *Reading for change: Performance and engagement across countries: Results from PISA 2000.* Paris: Organisation for Economic Co-operation and Development.

Mosenthal, P. 1998. Defining prose task characteristics for use in computer-adaptive testing and instruction. *American Educational Research Journal, 35*, 2: 269-307.