The Roles of Language and Item Formats

This chapter focuses on factors other than the three Cs (mathematical content, competencies and context) that influence students' performances. Just as countries differ, students' experiences differ by their individual capabilities, the instructional practices they have experienced, and their everyday lives. The chapter examines some of these differences in the patterns of performance by focusing on three factors accessible through data from PISA 2003: language structure within PISA 2003 assessment items, item format, and student omission rates related to items.

INTRODUCTION

What role does the wording of the problems themselves play in the PISA findings for mathematics? There is a vast literature detailing the importance of language factors in mathematics learning (Ellerton and Clements, 1991). The literature on performance assessments suggests that the use of language in test questions can influence the difficulty of the question and therefore students' performance on assessments (O'Leary, 2001; Routitsky and Turner, 2003). In this chapter, different aspects of the use of language are investigated, including the length of the text (number of words) and therefore the amount of reading required to understand the question.

A question's language and format influences whether students answer correctly or ...

Question format also has a potential to influence students' performance through its structure and response demands (O'Leary, 2001; Routitsky and Turner, 2003). The types of questions asked and the types of responses required by students vary considerably with the PISA mathematics assessment. Some questions require students to provide one simple answer, such as just a number, or select an answer from a range of possible responses. Other questions require students to provide an answer and explain why or justify how they came to their particular conclusion. The reasoning demands and response constraints that each question type can have on student performance varies across countries due to their differences in curriculum, instructional practices and students' everyday experiences. The analysis in the second part of this chapter will investigate the relationship between the different types of questions used in PISA mathematics and their difficulty.

... whether they even attempt to answer it.

Another related issue in international assessments is the difference in omission rates, meaning the percentages of students who do not attempt to answer questions. Omission rates consider patterns of non-response that occur even after student data is conditioned for non-completion, due to time constraints of the testing situation. Beyond time, the responsiveness of students (patterns of missing values) can depend on item characteristics that can be intentionally varied or controlled such as item difficulty, item format, the mathematical content involved, the context of the item, the level of reading demand involved and the amount of information in the stimulus (Jakwerth, Stancavage and Reed, 1999). Of course, omission rates may also be influenced by factors other than specific item characteristics that are outside the control of a teacher or assessment designer, for example cultural factors; however such factors lie outside the scope of this report. The chapter concludes with an analysis of patterns of differences in student omission rates on PISA mathematics assessments.

THE USE OF LANGUAGE IN PISA MATHEMATICS QUESTIONS AND STUDENT PERFORMANCE

As PISA is the first large-scale international study to assess *reading literacy*, *mathematical literacy* and *scientific literacy*, it is particularly important to also consider the use of language in contextualising the questions. As in Chapter 4, the focus

in this chapter is on difficulty of the groups of questions within each country. Here, the questions are grouped by number of words.

The number of words in a question measures its reading load.

For the purpose of this analysis, all the PISA 2003 mathematics questions were closely analysed and classified according to the number of words used in the question. This was done using the English language version of the test questions by counting all of the words used in both the stimuli and the questions. However, the calculation of the number of words for each question was not straightforward in all cases. Questions where this process was straightforward include the so-called "single-question units" where there is no clear distinction between the stimulus and the question, the question was presented as a whole (*e.g.* CUBES Q1 and STAIRCASE Q1). Further, for many of the questions belonging to a unit with more than one question included, it was necessary for students to read the information in the stimulus in order to answer the question (*e.g.* both questions in the unit WALKING or two out of three questions in the unit GROWING UP Q2 and Q3). However, there were a few questions where information within the question itself was sufficient for students to answer the question without reading the stimulus, such as GROWING UP Q1.

It can be expected that different students would use different reading strategies. Hopefully, all students would read the stimulus before attempting to answer the questions. For GROWING UP Q1, reflective students may have laboured over the stimulus unnecessarily. However, careful reading of the stimulus would save them time when answering GROWING UP Q2 and Q3. Another strategy would be to quickly look through the stimulus, answer the first question, and then return to the stimulus again when answering the second and third questions. These strategies would influence the time required for each question, but not necessarily the difficulty of the questions. Whether the student would read the stimulus of GROWING UP carefully or just looked through it quickly, GROWING UP Q1 would still require only a simple subtraction of two numbers given in the question itself.

After careful consultation and consideration, it was decided that if information in the stimulus was required for students to answer the question, the number of words is counted as the sum of the number of words in the stimulus and the number of words in the question itself. If the information within the question is sufficient for students to be able to give the answer, then the number of words is counted as only the number of words in the question, including words in any graphic elements and words used to formulate answers for multiple-choice questions, if applicable.

WORD-COUNT AND QUESTION DIFFICULTY ACROSS COUNTRIES

Using the methodology detailed above, the correlation between the number of words and the question difficulty in OECD countries was 0.28. To find out more about the relationship between the number of words and the difficulty of the items, all items were divided into three categories: short (50 words or

less), medium (between 51 and 100 words), and long (more than 100 words). Of course, a text with more than 100 words in the English version will have more or fewer words according to the language into which it is translated. The English version is used as a basis of text length to approximate the measure of the real length of the texts presented to students. The three categories represent a hierarchy of the texts according to their lengths (word count).

Short and medium questions are equally difficult and long questions are the most difficult.

Figure 5.1 shows the average relative difficulty of questions in each of the three word-count categories for each country. The short and the medium-length questions are on average of similar difficulty, while the long questions are significantly more difficult across all countries. Variation between countries within the medium category is small, while variation is slightly higher for the long and short categories (see Table 5.1)

So, longer questions are on average more difficult in all countries, but does the different difficulty of longer and short/medium-length questions explain performance differences across countries? This was investigated, and no significant differences were found for the majority of the countries (See Annex A4).

The difficulty of questions for each country (in logits centered at 0) was defined as the dependent variable and word-count and country were defined as independent

Figure 5.1 ■ **Average relative difficulty of questions within each word-count category for each country**

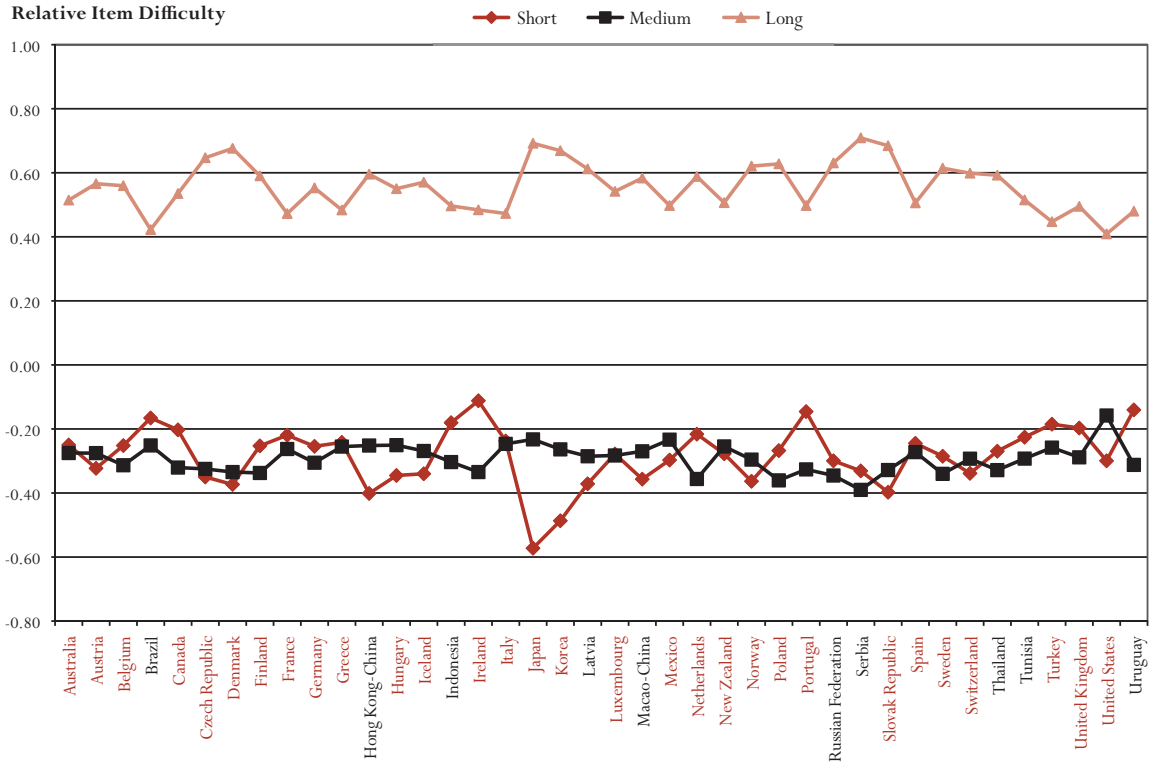


Table 5.1
Mean and standard deviation of difficulty of questions in each word-count category across countries

Word-count group	Number of questions included	Difficulty of questions included across countries (in logits)	
		Mean	(SD)
Short	21	-0.28	(0.09)
Medium	35	-0.29	(0.04)
Long	29	0.56	(0.08)

variables or factors (for a description of ANOVA see, for example, Rutherford, 2001). The results of this analysis (see Annex A4) show that while word-count categories across countries are significantly different, the interaction between countries and word-count is not significant. That is, between countries the variation within each category of word-count is indeed small. So there is really very little variation within each of the three word-count groups between countries (see Annex A4 for details). However, there were a few exceptions. The short questions are relatively easier for students in Korea and Japan (mean difficulty -0.49 and -0.57 logits respectively; both means are more than two standard deviations away from the overall average for the short questions). For the partner country Serbia the long questions are relatively more difficult (0.71 logits) while the medium-length questions are relatively easier (-0.39 logits). Finally, for the United States the long questions are relatively easier (0.41 logits) while the medium-length questions are relatively more difficult (-0.16 logits).

These few exceptions cannot be easily attributed to particular instructional and cultural differences. However, it is possible that this apparent effect of question length is confounded by other question characteristics that have already been analysed in Chapter 4, such as the mathematical content, the context in which the question is presented and the mathematical competencies required to answer the question. These factors are examined in the next section. It was also found that long questions are on average more difficult than medium-length (and short) questions, while there is no significant difference between the mean difficulty of medium-length and short questions (see Annex A4, Table A4.5).

WORD-COUNT AND THE CONTEXT IN WHICH A QUESTION IS PRESENTED

To what extent is the amount of reading involved in the question connected to the context in which that question is presented? This was investigated for all countries overall and results show that there is a small interaction between the context and the number of words used in the question, and that of these two factors, it is the number of words used that contributes more to the difficulty of the question. However, there are no differences among countries in this respect. The check for interaction between word-count and context in relation to question difficulty was also investigated through a full factorial analysis

of variance where the dependent variable was defined as question difficulty for each country (in logits centered at 0) and a factor for context was added (See Annex A4, Tables A4.3 and A4.4 for full results).

A question's context and length are related.

The results of this analysis show that there is a small but significant interaction between context and word-count which accounts for about 4% of the overall variance. At the same time, word-count as a main effect accounts for about 12 % of the overall variance and context as a main effect accounts for about another 4% of the overall variance. This can be interpreted to mean that word-count is a more important predictor of item difficulty than context. There were no significant interactions between context and country or between word-count and country, which means that country differences within each of the categories are insignificant.

Figure 5.2 illustrates the interaction between word-count and context. It shows quite different behaviour in the *educational and occupational* context area compared to all other areas in relation to the item difficulty for different categories of word-count.

The distribution of word-count for each context area was calculated to find out why the *educational and occupational* context area behaves differently compared to all other areas in relation to the item difficulty for different categories of

Figure 5.2 ■ **Context and length of question by average relative difficulty of questions**

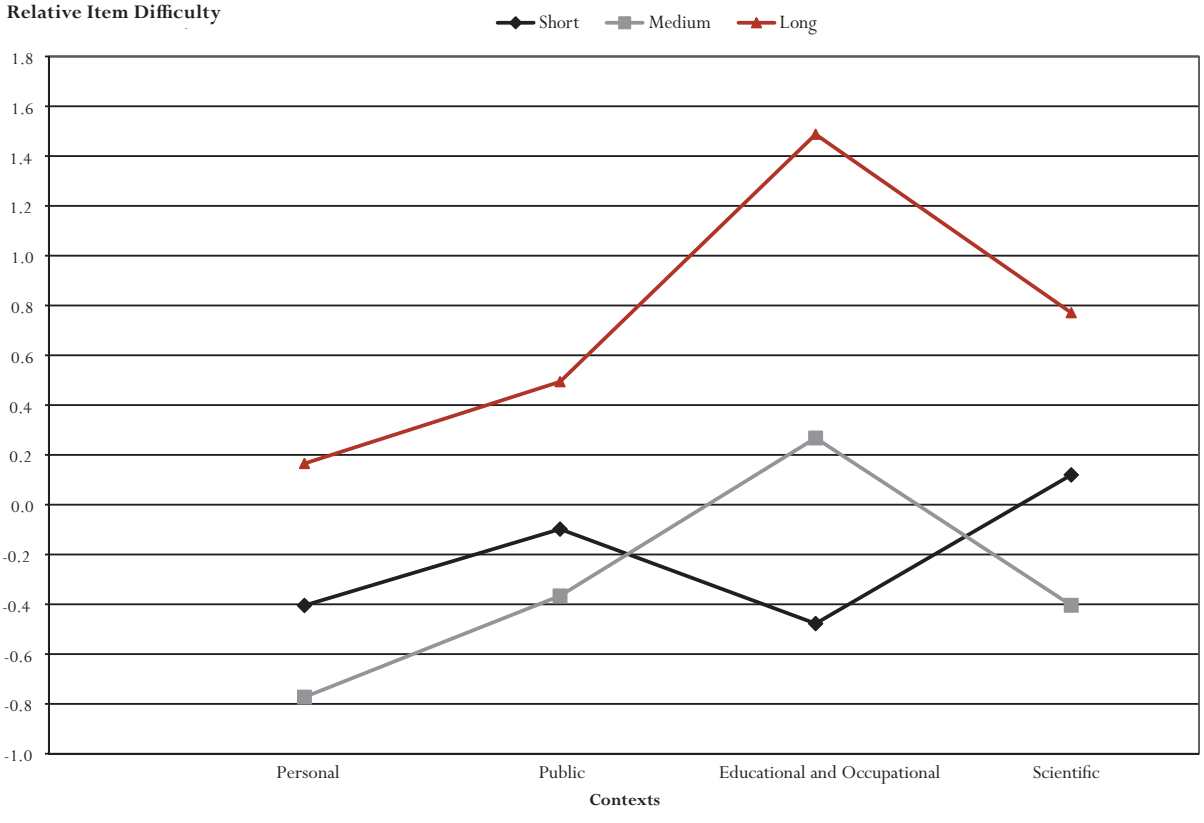


Table 5.2
Item distribution by context by word-count

Item group by word-count	Context			
	Educational and occupational	Personal	Public	Scientific
	Percentage of items in each category of word-count (number of items)			
Short	45% (9)	17% (3)	24% (7)	11% (2)
Medium	45% (9)	39% (7)	41% (12)	39% (7)
Long	10% (2)	44% (8)	34% (10)	50% (9)
Total	100% (20)	100% (18)	100% (29)	100% (18)

word-count. Table 5.2 shows that the distribution of items in the *educational and occupational* context area is quite different from the distribution of items in the other context areas. It has only two long items (10% of total items in the context area); while other content areas have eight to ten long items (34-50%).

It is worth noting that it is the number of items in each of the word-count categories, and not the average number of words, that explains the interaction between word-count and context. Table 5.3 shows that the average number of words in each of the word-count categories in the *educational and occupational* context area is quite similar to the average number of words in each of the word-count categories for the other context areas.

Therefore it is more likely that it is the distribution of items by word-count, and not the average number of words, that is responsible for the interaction between word-count and context.

Table 5.3
Average number of words by context by word-count

Item group by word-count	Context			
	Educational and occupational	Personal	Public	Scientific
	Average number of words in each category			
Short	32	34	30	30
Medium	74	74	65	86
Long	138	144	115	143

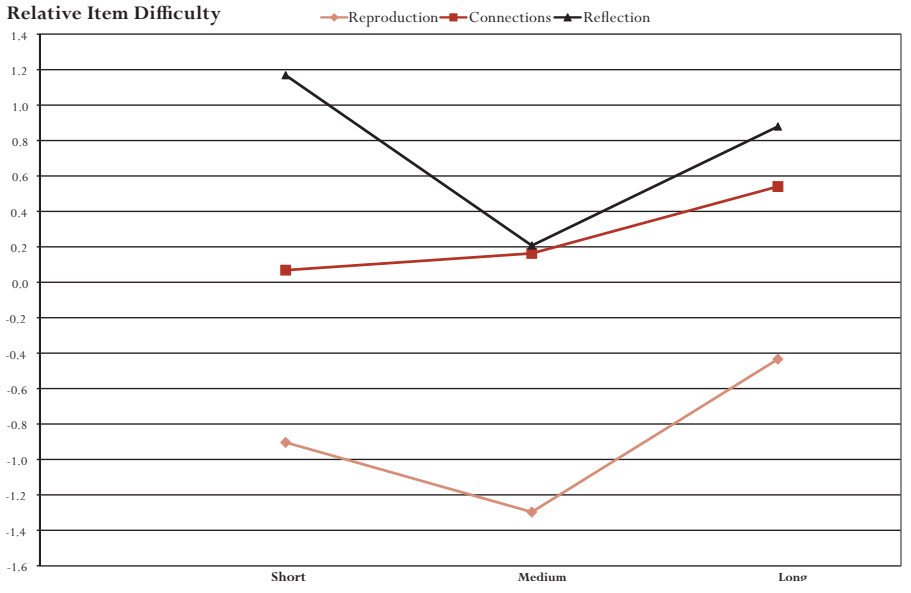
WORD-COUNT AND COMPETENCIES REQUIRED TO ANSWER THE QUESTION

The methodology used in this section is the same as in the previous section. First, full factorial analysis of variance was performed. As in the previous section, item difficulty for each country (in logits centered at 0) was defined as the dependent variable. Additional factors were country, word-count, and competencies. The results of this analysis (see Annex A4, Table A4.2) show that there is a small but significant interaction between competencies and word-count that accounts for about 2% of the overall variance. At the same time, competencies as a main effect account for about 23% of the overall variance, and word-count as a main effect accounts for another 5%. There were no significant interactions between competencies and country or word-count and country, which means that country differences within each of the categories are insignificant. Unlike in the previous section, it is not the word-count that is responsible for most of the variance, but competencies.

A question's competency cluster adds more to its difficulty than its word-count.

Figure 5.3 shows the relationship described above. In particular it shows that the differences between the *reproduction* cluster on the one hand and the *connection and reflections* clusters on the other hand are larger than the differences between the word-count categories within each of the competency clusters. This demonstrates the greater importance of the competencies compared to the word-count. Figure 5.3 also illustrates that the *connections* cluster behaves differently in relation to the item difficulties in each of the word-count categories. This illustrates the interaction between competencies and word-count.

Figure 5.3 ■ **Competency clusters and length of question by average relative difficulty of questions**



Figures 5.2 and 5.3 show that medium and short items appear to differ in average difficulty when subdivided between competency clusters or context areas. Unexpectedly, the short items look more difficult on average than the medium items. This happens in three out of four context areas (Figure 5.2) and in two out of three competency clusters (Figure 5.3). This gives us some indication that the shortest items are not always the easiest. It is possible that sometimes more explanations in the stimulus (provided that the explanations are not too long) make items easier.

Table 5.4 shows a somewhat expected item distribution within the word-count categories in each of the competency clusters. In the *reproduction* cluster only 16% of the items belong to the long category, while in the *connections* cluster long items make up 30% and in the *reflection* cluster they make up 68%.

Table 5.4
Item distribution by competencies by word-count

Item group by word-count	Competencies		
	Reproduction	Connections	Reflection
	Percentage of items in each category of word-count (number of items)		
Short	42% (11)	18% (7)	16% (3)
Medium	42% (11)	53% (21)	16% (3)
Long	16% (4)	30% (12)	68% (13)
Total	100% (26)	100% (40)	100% (19)

WORD-COUNT AND CONTENT

The interaction between content and word count is strong ...

Finally, the same methodology was applied to the content. The traditional topics described in detail in Chapter 4 were chosen as content categories rather than overarching ideas in order to make a more direct connection to traditional curriculum. The results of the full factorial analysis of variance (see Annex A4, Table A4.3) show that there is a larger interaction between content and word-count than between competencies and word-count or context and word-count. The interaction between content and word-count accounts for 9% of the overall variance. This means that in relation to word-count, the topics differ much more than competency clusters or context areas.

... but content is a better predictor of difficulty than word-count.

At the same time, word-count as a main effect accounts for only 3% of the overall variance while content as a main effect accounts for about 16% of the overall variance. This means that the traditional topics are more important predictors of item difficulty than the word-count, yet not as important as competencies.

As in previous sections, there were no significant interactions between context and country or between word-count and country.

Question length and difficulties vary considerably across content areas.

Figure 5.4 shows quite different behaviour of the five topics in relation to word-count. This corresponds to the high interaction between the word-count and the topics.

One observation from Figure 5.4 is that the variation of average item difficulty by word-count is higher for Data, Number and Measurement than for Algebra and Geometry. Although the number of items in each category (see Table 5.5) is not large enough to draw a conclusion, this evidence suggests that for the difficulty of the Algebra items the word-count is less important than formula manipulations and other algebraic cognitive demands. Similarly for Geometry items we can suppose that spatial cognitive demands influence item difficulty more strongly than reading demands, thus reducing the influence of the word-count variable. The results should be treated with caution given the small number of items in some combinations of word-count category and topics (see Table 5.5).

Short Measurement questions are the most difficult ones.

Table 5.5 shows a quite unequal distribution of items by word-count in each topic. On the one hand this distribution partially explains the high interaction between the content and the word-count. On the other hand it represents the relationship between the topics and the PISA framework. Nearly all Geometry items in PISA are related to two- and three- dimensional shapes, location and spatial relation, and symmetry and transformation. Thus it is not surprising that more than half of them are short items. Measurement, not surprisingly, also has only one long item.

Algebra questions are either long or medium length.

At the same time, Algebra items, when situated in a realistic context, require a somewhat wordy explanation of this situation and as a result the Algebra items do not have short items at all.

Figure 5.4 ■ **Traditional mathematics topics and length of question by average relative difficulty of questions**

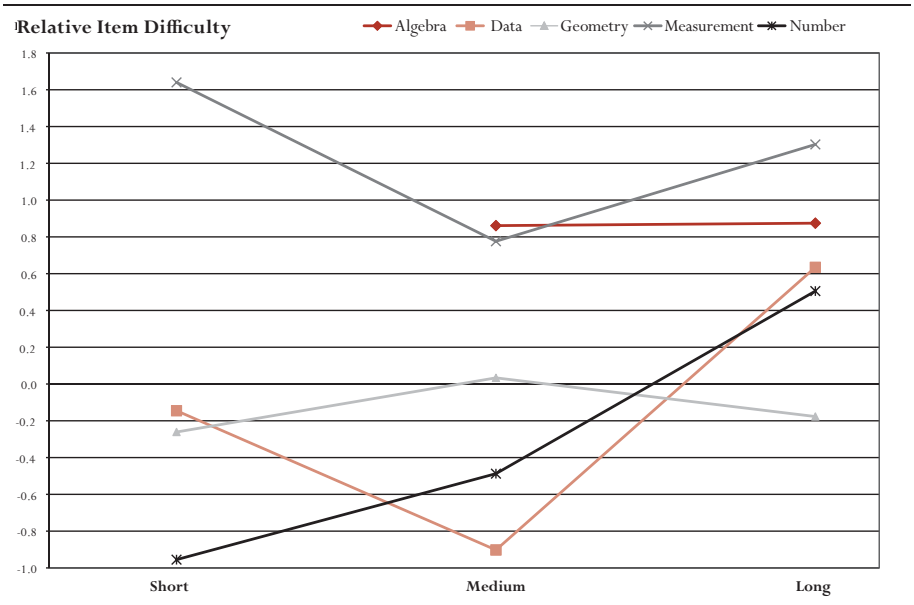


Table 5.5
Distribution of questions by traditional topic and length of question

Word-count category	Content (traditional topic most predominantly tested)				
	Algebra	Data	Geometry	Measurement	Number
	Percentage of questions in each word-count group (number of questions)				
Short	0% (0)	19% (5)	58% (7)	25% (2)	22% (7)
Medium	43% (3)	42% (11)	17% (2)	62% (5)	44% (14)
Long	57% (4)	39% (10)	25% (3)	13% (1)	34% (11)
Total	100% (7)	100% (26)	100% (12)	100% (8)	100% (32)

Data and Number have a quite even distribution of items across all three word-count groups, probably because it is easier to find authentic contexts for 15-year-olds for quantitative or statistical problems.

ITEM-FORMAT AND MATHEMATICS PERFORMANCE

How the question is asked can have an impact on its difficulty.

There is research evidence that item format can influence students' performance in different countries (O'Leary, 2001) and that this can vary for different levels of ability (Routitsky and Turner, 2003). In this section the analyses investigate how item-format associates with item difficulty and whether there is an interaction between item format and other features of the items discussed earlier in this report. The question of whether there is a format by country interaction is also studied. Differences between countries can pose important questions about instructional and assessment practices in these countries. For this purpose, a full factorial analysis of variance is used in the same way as it was used in previous sections (see Annex A5 for these results).

In the PISA 2003 initial report, mathematics items were represented by constructed response or selected response items.

Constructed response items can be subdivided into the following two categories (see Chapter 3 for the examples listed below):

- *Extended open constructed response*: response requires some explanation or justification of the answer referred to as the "extended response" type of "open constructed-response" items (see, for example, GROWING UP Q3).
- *Short answer*: response requires a number as an answer (see, for example, GROWING UP Q1 and EXPORTS Q1).
- *Multiple short answer*: response requires several numbers as an answer, and these answers were scored as one item (see, for example, SKATEBOARD Q3).

Selected response items can be subdivided into the following two categories:

- *Simple multiple choice items* (see, for example, COLORED CANDIES Q1).
- *Complex multiple choice items* (see, for example, CARPENTER Q1).

The analysis in this section is based on the categorical variable item-format with the five categories described above.

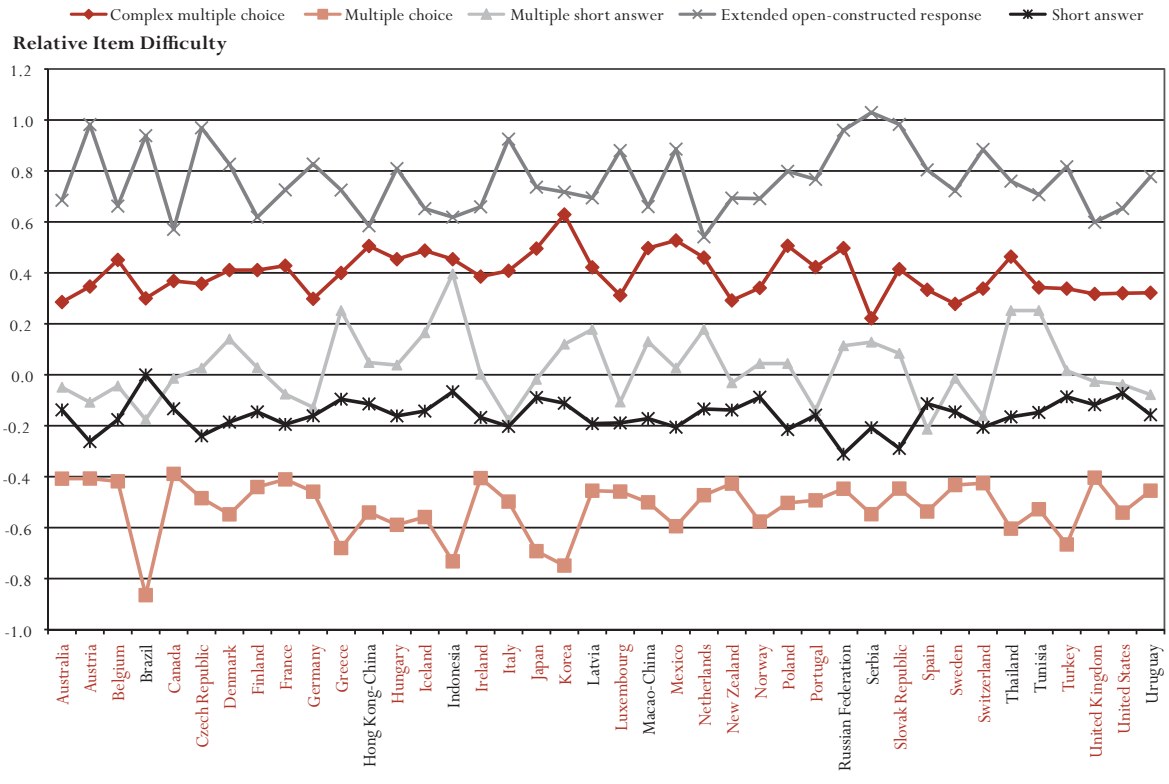
ITEM-FORMAT AND ITEM DIFFICULTY ACROSS COUNTRIES

The most difficult questions are extended response and complex multiple choice.

Figure 5.5 shows the average item difficulty for each item-format category in each country. It is not clear from this figure which item-format categories are significantly different from each other. Number of items, means and standard deviations are presented for each item-format category in Table 5.6.

Table 5.6 shows that on average the most difficult item type is *extended response* followed by *complex multiple choice*. The easiest item type is simple *multiple choice*. It also shows that the *extended response* type and the *multiple choice* type vary between countries more than the *complex multiple choice* type and the *short answer* type.

Figure 5.5 ■ **Average item difficulty (logits) by item-format by country**



Multiple comparisons for mean difficulties using Bonferroni adjustment (see Annex A5, Table A5.6) show that all item-format categories are significantly different from each other at the 0.01 level. However, analysis of variance with item-format and countries used as factors (see Annex A5, Table A5.1) shows that there is no interaction effect between countries and item-format.

There are a few countries for which the mean item difficulty in some of the item-format categories is more than two standard deviations away from the overall mean for these item types. For Brazilian students the multiple choice items appear to be relatively easier (-0.87 logits) while the short answer items appear to be relatively more difficult (0.00 logits). For the Russian Federation and the Slovak Republic the short answer items appear to be relatively easier (-0.31 and -0.29 logits respectively). For Serbia the complex multiple choice items appear to be relatively easier (0.22 logits) while the extended response items are relatively more difficult (1.03 logits). Finally, for Korea the complex multiple choice items are relatively more difficult (0.63 logits) and the multiple choice items are relatively easier (-0.75 logits). These differences might give some indication to the specialist in national assessment and curriculum where to look for strengths and weaknesses.

ITEM-FORMAT, THE THREE C'S AND WORD-COUNT

Content is still a stronger predictor of difficulty than question format.

As it was the case for the word-count, item-format shows a small but significant interaction with competencies and context (see Annex A5, Table A5.2): the competencies are a much stronger factor than item-format while context is weaker. Interactions between topics and item-format will not be discussed due to the very small number of items in each cell (see Annex A5, Table A5.3).

There is an interesting relationship between the item-format and the word-count. Although analysis of variance shows a strong interaction, this is mainly due to the fact that all complex multiple choice items have more than 50 words and therefore none of them belong to the category of "short" answer items.

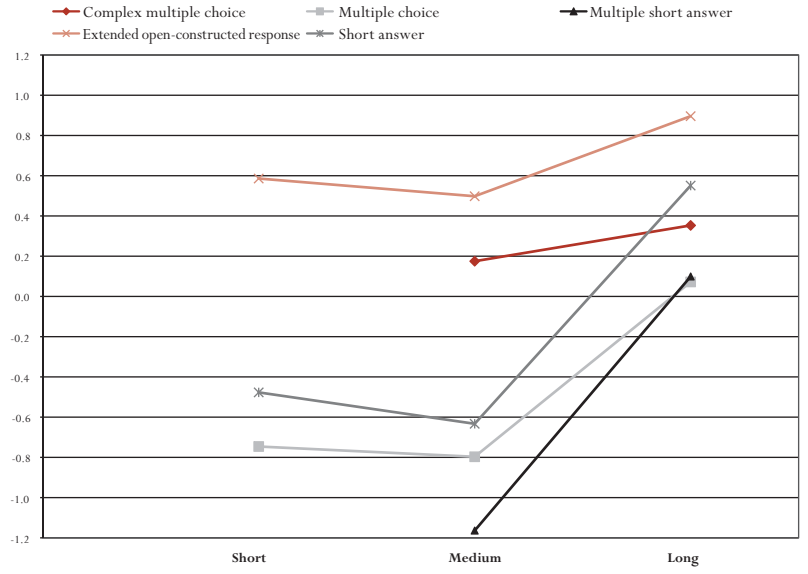
Table 5.6
Mean and standard deviation of item difficulty in item-format categories across countries

Item format type	Number of items	Item type difficulty across countries	
		Mean	(SD) in logits
Complex multiple choice	11	0.40	(0.08)
Multiple choice	18	-0.52	(0.11)
Short answer	37	-0.16	(0.06)
Multiple short answer	5	0.03	(0.13)
Extended response	14	0.76	(0.13)

Figure 5.6 shows that the multiple choice, short answer and extended response categories behave very similarly in relation to word-count.

Figure 5.6 also shows that the item-format is a stronger predictor of item difficulty than the word-count (see Annex A5, Table 5.4).

Figure 5.6 ■ **Average relative difficulty of questions by item-format and word-count**



DIFFERENCES IN ITEM-FORMAT AND OMISSION RATES

A question's content, context, format, word count and difficulty are all related to whether a student attempts to answer it or not.

Another issue related to item-format is differences in omission rates to items. The responsiveness of students in terms of patterns of missing values can depend on the different surface characteristics of the test item as well as on its difficulty. These item characteristics include: format, content, context, reading demand and amount and complexity of information in the stimulus. The results of this section will have important implications for assessment practices and instruction.

While examining Michigan's High School Proficiency Test, DeMars (2000) found an interaction between test consequences (high/low stakes) and item-format. She also argues that "motivation and performance may be influenced by item response format". There is also a general belief that non-response is somewhat higher for constructed response items than for multiple-choice items, although it could be an effect of item difficulty (Lord, 1975; Dossey, Mullis, and Jones, 1993).

To investigate the relationship between the amount of missing data and item format, data from the PISA 2003 field trial were examined. These data were used because they were coded to better reflect the nature of the missing data. The percent of missing data was calculated for each item in each of the following item types: multiple choice, extended response, and short answer. For this

analysis, the calculation of missing data identified and excluded non-reached items. In other words, missing data are here defined as "embedded missing" responses, and "trailing missing" responses are not included.

The correlation between the percentage of missing responses and item difficulty was calculated. This gives a measure of the degree to which missing data can be explained by item difficulty for each format type. For the multiple choice items the correlation was 0.331, for the extended response items the correlation was 0.499 and for the short answer items the correlation was 0.737.

The distribution of the omissioin rates by item-format shows that the amount of missing data in multiple choice items, which comprise the easiest set, varies from 1.66% to 17.62%. Here the relative item difficulty accounts for about 11% of the variation of missing data.

At the same time, the results show that the amount of missing data in extended response items, which comprise the most difficult set, varies from 9.78% to 57.54%. Here the relative item difficulty accounts for about 25% of the variation of missing data.

Finally, the results show that the amount of missing data in short answer items, which are slightly easier than extended response items, varied from 2% to 48%. Here the relative item difficulty accounts for about 54% of the missing data.

The scoring of student responses for PISA treats missing responses (excluding non-reached) as incorrect. This is based on the assumption that students omit an item because they do not know how to answer it. Such an assumption is supported more strongly when there is a strong relationship between item difficulty and omission rates. In this study, the short answer items fit this model best and the multiple choice items the least.

There also appear to be some other factors, other than item difficulty, particularly for the multiple choice and for the extended response items that contribute to the causes for missing data. There is a widespread belief that for multiple choice items, if students don't know the answer, they have the possibility of guessing, and therefore omission rates are low. This possibility does not exist for the extended response format types. What other factors might apply? Item difficulty is one possibility. Other particular factors that may contribute to missing data might include the reading load of the item. Further investigation is required.

To further investigate the relationship between missing data and item format type, the average amount of missing data for each item format type described above (multiple choice, extended response and short answer types) was calculated. The results are shown in Table 5.7. The results are reported separately for a sequence of item difficulty ranges to control for difficulty.

Table 5.7
Average percent of missing data by item difficulty for three item-format categories – PISA Field Trial 2003

	Average percent of missing data		
Difficulty range (logits)	**Multiple Choice**	**Short Answer**	**Extended Response**
Less than -2	2.90%	2.65%	N/A
Between -2 and -1	3.89%	8.98%	N/A
Between -1 and 0	4.96%	9.45%	17.80%
Between 0 and 1	6.30%	19.28%	21.44%
Between 1 and 2	9.12%	24.96%	29.00%
Between 2 and 3	6.82%	31.83%	33.62%
More than 3	N/A	28.35%	48.56%

This table shows that the percent of missing data in each difficulty range is the lowest for the multiple choice items and slightly lower for the short answer items than for the extended response items. In addition it shows that the general trend within each format type is that the more difficult the item, the more missing data is observed, but for the most difficult multiple choice items, the percent of missing data is lower than expected, which raises a question about guessing. Is there some difficulty threshold for the multiple choice items beyond which students will guess rather than omit the item?

Overall the PISA 2003 field trial data shows that while both item format type and item difficulty play significant roles in the amount of missing data, there are other features of items such as the length and complexity of the stimulus and the form in which choices are presented that also play a role.

CONCLUSION

This chapter examines the relationship between the difficulty of the PISA 2003 mathematics questions, and features such as the amount of reading required, the type of questions asked and the percentage of students who do not answer each question.

The correlation between question difficulty and the number of words in the question was weak because the question difficulty does not reflect small changes in the number of words used. The difficulty of the question is only influenced once the questions are of a certain length: on average there is no difference in difficulty between short questions (less than 50 words) and medium-length questions (between 50 and 100 words). Long questions (more than 100 words) were significantly more difficult across all countries. From this point of view, although words were counted in English and the meaning of "short", "medium" and "long" will be different in different languages, the three word-count categories were stable across all countries and, therefore, appropriate for analysis.

Content and competency are stronger predictors of difficulty than context or word count.

The analysis shows that there is a small but significant interaction between the number of words used in the questions and the context in which the question is presented, the mathematical content and the mathematical competencies required in answering the question. Further, the number of words used in the questions is a better predictor of question difficulty than the context in which the question is presented, but weaker than the mathematical content and the mathematical competencies required in answering the question. Another finding is that although on average short and medium-length questions are of the same difficulty, when subdivided by the context in which the question is presented, the mathematical content or the mathematical competencies, medium-length questions consistently show that they can be easier than short questions. One possible explanation for this finding is that as long as the stimulus and the question itself are not too wordy, some additional words, if appropriate, can help to solve the problem rather than make it more difficult.

In the PISA 2003 field trial data, a strong correlation was found between students' general performance in mathematics and their preferential performance on multiple choice items versus open-ended extended response items (see Routitsky and Turner, 2003). That is, lower ability students are performing better than expected on the multiple choice items and higher ability students are performing better than expected on the open-ended extended response items. This was partially explained by the combination of psychometric characteristics of these types of items (lower discrimination for multiple choice items) in the PISA 2003 field trial test and the wide range of students' abilities. This was avoided in the PISA 2003 Main Survey by keeping item discrimination in a narrower range. In this chapter the analysis is concentrated on the PISA 2003 Main Survey items. As with the word-count, the item-format shows a small but significant interaction with context and competencies. The three main item-format categories – multiple choice, extended open ended response and short answer – do not have an interaction with word-count. Generally, the item-format is a better predictor of item difficulty than the context and the word-count but weaker than the competencies. This means that format considerations should be treated with caution when tests are constructed, especially when students from a wide range of abilities are tested.

While format and difficulty are related to students not attempting to answer questions, word count and cultural bias also play a role.

In relation to omission rates, the findings of this chapter show that while both the item-format and item-difficulty play significant roles in the amount of missing data, there are other features of items such as the length and complexity of the stimulus and form in which choices are presented that also play a role. Cultural differences among countries may explain a portion of differential non-response due to cultural views about guessing when one does not know the answer.

In conclusion, the analyses carried out in this chapter show that PISA results can provide useful information about different features of questions, how these features relate to each other and how relevant they are to the difficulty of the questions. The most important factor influencing the difficulty of the PISA mathematics questions is the mathematical competencies or the cognitive

demands required in answering the question, followed by the mathematical content (represented as traditional topics), the type of question (item-format), and the amount of reading required to understand the question (word-count). The factor showing the weakest influence over the difficulty of the PISA mathematics questions was the context in which the question is presented. These findings can be helpful in developing both mathematics assessments and mathematics text books, as well as for classroom teachers when making a choice of questions for instruction and assessment.

Content is the only factor related to difficulty that showed important variation across countries.

The only factor that showed differentiation by countries was the mathematical content of the questions (as represented by the traditional topic). For all other factors there were no significant variations between countries within each factor: context in which the question is presented, word-count, mathematical competencies required in answering the question.