A Critical Comparison of the Contents of PISA and TIMSS Mathematics Assessments

Margaret Wu
University of Melbourne

June 2, 2009

Send all correspondence to:
Margaret Wu
University of Melbourne
Assessment Research Centre
234 Queensberry Street
Faculty of Education
Victoria, Australia, 3010
E-mail: m.wu@unimelb.edu.au

ABSTRACT

A study was carried out that compared Programme for International Student Assessment (PISA) 2003 Mathematics results with Trends in International Mathematics and Science Study (TIMSS) 2003 Grade 8 Mathematics results (Wu, 2008). It was found that Western countries generally performed better in PISA than in TIMSS, and Eastern European and Asian countries generally performed better in TIMSS than in PISA.

In this paper, TIMSS items are divided into two sets: one that fits the PISA framework and one that does not. In particular, many geometry and algebra items in TIMSS are “inner mathematics” (mathematics without a real-life context), and such items do not appear in the PISA test. Differential performances of six countries on each set of items are examined. In this way, the relative strengths and weaknesses of Western and Asian countries are identified. These strengths and weaknesses are then linked back to the contents of the PISA and TIMSS tests. There is strong evidence that differential performance between Western and Asian countries in PISA and TIMSS can be directly attributed to the types of items in the respective tests.

Keywords: International study, Mathematics, PISA, TIMSS

INTRODUCTION

A study was carried out that compared PISA 2003 Mathematics results with TIMSS 2003 Grade 8 Mathematics results, using country mean scores for 22 participants of both studies (Wu, in press). It was found that Western countries generally performed better in PISA than in TIMSS, and Eastern European and Asian countries generally performed better in TIMSS than in PISA. Furthermore, two factors, content balance and years of schooling, accounted for most of the variation in PISA country mean scores after controlling for TIMSS country mean score. Consequently, the rankings of countries in the two studies can be reconciled to a reasonable degree of accuracy.

The fact that content balance has a significant effect on country performance suggests that students in different countries have particular relative strengths and weaknesses. If these specific strengths and weaknesses are identified beyond the level of broad content categories, mathematics educators in each country can be informed of the specific skills students have or lack. This will also provide a further insight into what PISA and TIMSS are each assessing, beyond the usual rhetoric about curriculum-based and non-curriculum-based focus.

PISA AND TIMSS MATHEMATICS FRAMEWORKS AND TESTS

Comparisons of PISA and TIMSS mathematics frameworks can be found in a number of publications (American Institutes for Research, 2005; Hutchison & Schagen, 2007; National Center for Education Statistics, 2008; Neidorf et al., 2006). These comparisons tend to produce a descriptive list of similarities and differences between the two published frameworks, such as the classifications of the content domains and the cognitive domains. However, few published comparisons critically examined the differences. For example, would one framework lead to a test that is completely different from a test based on the other framework? Is one framework a subset of the other? If so, what is missing in that framework? Are the two frameworks essentially the same, other than nomenclature of the classifications? This paper looks at these issues and, in doing so, it is hoped that a better understanding is gained in relation to key differences between PISA and TIMSS at the level of the tests and items, not just at the level of broad aims and orientations, as reflected in the frameworks.

At a first glance, most would conclude that both PISA and TIMSS mathematics frameworks are comprehensive. It does not appear that one framework is necessarily a subset of the other, or that something is glaringly missing from either framework. However, the PISA mathematics framework suggests its more inclusive approach, with the following line at the beginning of the framework document:

> Rather than being limited to the curriculum content students have learned, the assessments focus on determining if students can use what they have learned in the situations they are likely to encounter in their daily lives (Organisation for Economic Co-operation and Development [OECD], 2003, p. 24).

The word “limited” suggests that PISA is attempting to be more inclusive in terms of coverage of the mathematics domain. It also suggests that the school curriculum, whether intended, implemented, or attained, does not focus on whether students can use what they have learned. Does this mean that TIMSS, being more curriculum-based, does not assess whether students can use what they have learned? The TIMSS mathematics framework states the following:

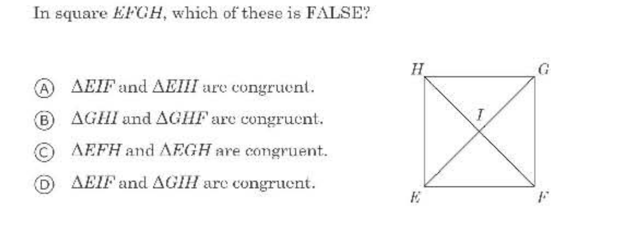
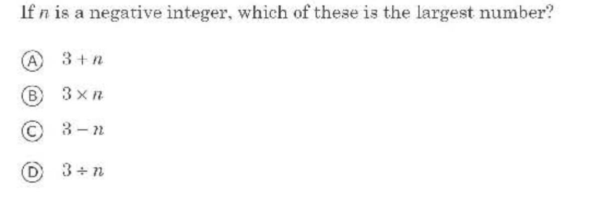
> While the assessment of abilities such as solving nonroutine problems and reasoning with mathematics will be of special interest, the factual, procedural, and conceptual knowledge that form the initial base for the development and

> implementation of these skills will also be assessed. Problem solving and communications are key outcomes of mathematics education that are associated with many of the topics in the content domains. They are regarded as valid behaviors to be elicited by test items in most topic areas (International Association for the Evaluation of Educational Achievement, 2003, p. 11).

It would appear that TIMSS does not preclude problem solving in the framework, but there may be less emphasis on this than in PISA, where the mathematics framework is almost entirely built on the *application* of mathematics competencies that may reasonably be expected of students. In short, one may expect a different balance in the two tests in terms of problem solving items and fact/procedural items, but one might not necessarily expect a complete absence of one type of item.

However, a close examination of the test items in the two surveys reveals that there is one class of items in TIMSS that does not appear in PISA. One might label this class of items as “content rich,” “formal mathematics,” or “naked mathematics.” Some of these items are illustrated in Figure 1.

Figure 1. Examples of TIMSS Items Not Appropriate for the PISA Test

Item A	Item B
 In square $EFGH$, which of these is FALSE? Ⓐ ΔEIF and ΔEIH are congruent. Ⓑ ΔGHI and ΔGHF are congruent. Ⓒ ΔEFH and ΔEGH are congruent. Ⓓ ΔEIF and ΔGIH are congruent. 	 If n is a negative integer, which of these is the largest number? Ⓐ $3 + n$ Ⓑ $3 \times n$ Ⓒ $3 - n$ Ⓓ $3 \div n$

If one checks the PISA framework, items shown in Figure 1 may come under the classification of competency list number 8: Using symbolic, formal, and technical language and operations (OECD, 2003, p. 41). While these items could fit the PISA framework at a theoretical level, they are not tasks that "(students) are likely to encounter in their daily lives (outside the classroom context)." These may be items that an adult working in a scientific field may encounter in his or her daily life, but not 15-year-old students studying at school. Consequently, items focusing on technical language and formal mathematics are typically not in the PISA test. This is not because the PISA framework does not cover such competencies, but because the operationalisation of the PISA framework as applied to 15-year-old students precludes these kinds of items. So, at one level one could say the PISA framework is comprehensive, but at another level, typically at the level of the construction of test items, a set of typical school mathematics items are excluded.

In contrast, when PISA items are examined against the TIMSS framework, no group of items appears to be misfitting the TIMSS framework to the extent that such items could not be included in the TIMSS test. However, there are more PISA items than TIMSS items where, within an item, the competencies tested cover multiple content domains.

Because formal mathematics items are not included in the PISA test, one may ask if, and how, country results may be affected, particularly in the context of differential performance of countries in PISA and TIMSS. This paper attempts to answer this question, by examining the differential performance of countries on the set of TIMSS items that do not fit with the PISA test.

METHODOLOGY

Because there is a large pool of released items from TIMSS 2003 Grade 8 assessment (99 items in total), we examined these items and divided them into two sets: one that fits the PISA test and one that does not. In particular, many geometry and algebra items in TIMSS are "inner mathematics" (mathematics without a real-life context), and such items do not appear in the

PISA test. Owing to the amount of the work involved in rescaling the item responses for all countries, we selected six countries in particular to examine their differential performances on the two sets of items. The six countries were Australia, England, and the United States (Western countries, abbreviated as West); and Hong Kong, Japan, and Korea (Asian countries, abbreviated as East). Item response modeling (Rasch model) was used to calibrate the item difficulties. The item difficulties for the 99 released items were estimated for each country, with the average difficulty for each country centred at zero. These item difficulties were compared across the six countries. If there was no differential item functioning, then an item that was relatively more difficult than another item for one country should be relatively more difficult in another country. On the other hand, if there was differential item functioning, then one item may be found to be relatively easier in one country, but relatively difficult in another country. We carried out the comparisons of item difficulties particularly with respect to items not fitting the PISA test, and with respect to Western countries and Asian countries.

RESULTS

Of the 99 released items, 42 were deemed not fitting the PISA test. That is, around half of the TIMSS items are not likely to appear in the PISA test. This proportion is surprisingly high. It could mean that a large part of mathematics taught in schools is not included in the PISA test. Table 1 shows our classifications so that other researchers can cross-check if desired. An “n” in the column headed “In PISA?” means that the TIMSS item is not likely to be in the PISA test, and an “N” means that the item is almost certainly not a PISA item. A “y” means that the TIMSS item could be a PISA item, and a “Y” means that the item is almost certainly a PISA item. This classification was made before any item calibration was carried out, to ensure that the classification process was independent of any knowledge of the relative item difficulties for each country.

Table 1. Classification of TIMSS 2003 Released Items into Categories of Appropriateness in the PISA Test

Item Seq	Unique ID	In PISA?	Item Seq	Unique ID	In PISA?	Item Seq	Unique ID	In PISA?
1	M012001	n	34	M022139	n	67	M032261	N
2	M012002	y	35	M022142	N	68	M032271	Y
3	M012003	n	36	M022144	n	69	M032403	N
4	M012004	y	37	M022146	y	70	M032447	y
5	M012005	N	38	M022148	y	71	M032489	y
6	M012006	Y	39	M022154	N	72	M032533	y
7	M012013	y	40	M022156	y	73	M032545	N
8	M012014	y	41	M022185	N	74	M032557	N
9	M012015	N	42	M022188	N	75	M032570	n
10	M012016	N	43	M022189	y	76	M032588	n
11	M012017	y	44	M022191	y	77	M032609	n
12	M012025	Y	45	M022194	y	78	M032612	N
13	M012026	N	46	M022196	N	79	M032643	N
14	M012027	n	47	M022198	n	80	M032647	y
15	M012028	N	48	M022199	N	81	M032649A	y
16	M012029	n	49	M022202	N	82	M032649B	y
17	M012030	y	50	M022227A	y	83	M032652	y
18	M012037	Y	51	M022227B	y	84	M032670	n
19	M012038	n	52	M022227C	y	85	M032671	y
20	M012039	N	53	M022251	N	86	M032678	N
21	M012040	Y	54	M022252	y	87	M032689	N
22	M012041	n	55	M022253	N	88	M032690	n
23	M012042	N	56	M022261A	y	89	M032693	N
24	M022002	N	57	M022261B	y	90	M032699	Y
25	M022004	n	58	M022261C	y	91	M032727	y
26	M022005	y	59	M032036	N	92	M032728	N
27	M022008	n	60	M032044	n	93	M032732	n
28	M022010	y	61	M032046	N	94	M032743	y
29	M022012	n	62	M032079	n	95	M032744	y
30	M022016	n	63	M032208	N	96	M032745	y
31	M022021	y	64	M032210	N	97	M032762	Y
32	M022127	Y	65	M032228	y	98	M032763	Y
33	M022135	y	66	M032233	y	99	M032764	Y

The second step in our analysis was to calibrate the item difficulties for each country, using item response modeling (Rasch model). The average item difficulty for each country was set at zero, so that the calibrated item difficulty for each item was a measure relative to the average item difficulty in that country. In this way, item difficulties across countries can be compared, even if the abilities of students vary across countries. That is, the calibrated item difficulties are relative item difficulties for each country; they are not absolute difficulties such as percentages correct. Table 2 shows the calibrated item difficulties for Australia and Hong Kong, arranged in order of the difference in item difficulties. The items at the top left part of Table 2 are those that Australian students found relatively easier (as compared to other items) than Hong Kong students did. The items at the bottom right part of Table 2 are those that Australian students found relatively more difficult than Hong Kong students did. It is worth noting that, as one scans down from the top left part to the right bottom part of Table 2, the number of “n” and “N” in the column “In PISA?” increases. In fact, of the first 20 items that Australian students found much easier (relative to other items) than Hong Kong students did, 16 could have been PISA items (16 “y” or “Y”). In contrast, of the 20 items that Australian students found much more difficult (relative to other items) than Hong Kong students did, 15 are not appropriate for the PISA test (15 “n” or “N”).

Figure 2 shows a plot of the average item difficulties for Western countries (AUS, ENG, USA) and average item difficulties for Asian countries (HKG, JPN, KOR), arranged in order by the magnitude of the difference between the average difficulties. That is, the items on the left side of the plot are those that Western countries found relatively easier than Asian countries did. The items on the right side of the plot are those that Western countries found relatively more difficult. Interestingly, the majority of the items on the right side of the plot are items that are

Table 2. Calibrated Item Difficulties (in IRT logits) for Australia and Hong Kong

Seq	Unique ID	AUS	HKG	Differ-ence	In PISA?
33	M022135	-0.415	2.36	-2.775	y
36	M022144	0.195	2.217	-2.022	n
97	M032762	0.058	1.906	-1.848	Y
18	M012037	-0.385	1.232	-1.617	Y
80	M032647	-0.441	1.13	-1.571	y
90	M032699	-2.164	-0.686	-1.478	Y
38	M022148	-1.217	0.191	-1.408	y
8	M012014	-2.16	-0.996	-1.164	y
16	M012029	-1.049	0.064	-1.113	n
30	M022016	0.165	1.276	-1.111	n
11	M012017	-0.782	0.307	-1.089	y
26	M022005	0.169	1.239	-1.07	y
27	M022008	0.619	1.685	-1.066	n
45	M022194	-0.875	0.15	-1.025	y
84	M032670	-2.484	-1.465	-1.019	n
94	M032743	-1.224	-0.24	-0.984	y
2	M012002	-1.505	-0.581	-0.924	y
95	M032744	-0.109	0.753	-0.862	y
10	M012016	-0.25	0.553	-0.803	N
58	M022261C	2.248	3.047	-0.799	y
4	M012004	-0.117	0.567	-0.684	y
44	M022191	-0.813	-0.161	-0.652	y
96	M032745	2.096	2.725	-0.629	y
63	M032208	-0.978	-0.402	-0.576	N
42	M022188	0.337	0.903	-0.566	N
17	M012030	-0.065	0.497	-0.562	y
98	M032763	1.901	2.435	-0.534	Y
7	M012013	-1.315	-0.819	-0.496	y
92	M032728	0.253	0.639	-0.386	N
85	M032671	-1.493	-1.122	-0.371	y
99	M032764	1.815	2.165	-0.35	Y
29	M022012	0.36	0.677	-0.317	n
43	M022189	-1.582	-1.277	-0.305	y
28	M022010	-0.509	-0.226	-0.283	y
54	M022252	-1.422	-1.159	-0.263	y
34	M022139	0.293	0.541	-0.248	n
87	M032689	0.546	0.768	-0.222	N
32	M022127	1.237	1.446	-0.209	Y
9	M012015	-0.909	-0.73	-0.179	N
21	M012040	-1.414	-1.254	-0.16	Y
62	M032079	0.638	0.797	-0.159	n
47	M022198	-0.26	-0.102	-0.158	n
56	M022261A	-0.185	-0.066	-0.119	y
66	M032233	1.776	1.863	-0.087	y
78	M032612	0.303	0.288	0.015	N
88	M032690	0.167	0.149	0.018	n
6	M012006	-1.068	-1.09	0.022	Y
71	M032489	-2.156	-2.186	0.03	y
76	M032588	-0.258	-0.297	0.039	n
53	M022251	0.991	0.949	0.042	N
70	M032447	-0.099	-0.183	0.084	y
15	M012028	-0.914	-1.091	0.177	N
12	M012025	-1.383	-1.567	0.184	Y
83	M032652	1.107	0.913	0.194	y
77	M032609	-1.112	-1.321	0.209	n
40	M022156	-0.101	-0.334	0.233	y
57	M022261B	1.151	0.836	0.315	y
1	M012001	-0.848	-1.184	0.336	n
25	M022004	-0.038	-0.424	0.386	n
82	M032649B	1.668	1.267	0.401	y
69	M032403	-0.172	-0.596	0.424	N
79	M032643	0.7	0.253	0.447	N
19	M012038	-1.542	-1.998	0.456	n
37	M022146	-0.444	-0.906	0.462	y
64	M032210	0.71	0.239	0.471	N
72	M032533	0.058	-0.425	0.483	y
68	M032271	-0.229	-0.75	0.521	Y
49	M022202	1.296	0.718	0.578	N
20	M012039	-0.169	-0.777	0.608	N
75	M032570	-0.888	-1.528	0.64	n
35	M022142	0.127	-0.525	0.652	N
59	M032036	0.488	-0.206	0.694	N
5	M012005	0.255	-0.443	0.698	N
89	M032693	1.103	0.398	0.705	N
74	M032557	1.664	0.878	0.786	N
67	M032261	0.537	-0.288	0.825	N
60	M032044	0.077	-0.77	0.847	Y
46	M022196	-0.059	-1.026	0.967	N
48	M022199	1.063	0.054	1.009	N
65	M032228	-0.207	-1.217	1.01	Y
13	M012026	0.362	-0.688	1.05	N
81	M032649A	0.733	-0.355	1.088	N
41	M022185	0.275	-0.845	1.12	N
24	M022002	1.827	0.701	1.126	N
31	M022021	0.413	-0.761	1.174	Y
86	M032678	0.434	-0.773	1.207	N
50	M022227A	0.006	-1.312	1.318	Y
52	M022227C	2.144	0.812	1.332	N

(continued)

Table 2. Calibrated Item Difficulties (in IRT logits) for Australia and Hong Kong (continued)

Seq	Unique ID	AUS	HKG	Difference	In PISA?
93	M032732	-0.122	-0.04	-0.082	n
91	M032727	-0.159	-0.084	-0.075	y
39	M022154	-0.229	-0.168	-0.061	N
14	M012027	-0.675	-0.622	-0.053	n
3	M012003	-1.153	-1.106	-0.047	n
22	M012041	-1.031	-1.018	-0.013	n
61	M032046	1.936	0.559	1.377	N
23	M012042	0.668	-0.907	1.575	N
55	M022253	0.094	-1.49	1.584	N
51	M022227B	1.713	-0.427	2.14	Y
73	M032545	2.404	-0.132	2.536	N

deemed unlikely to appear in the PISA test (mostly labeled “N”). The numerical values are shown in Appendix A.

In summary, both Table 2 and Figure 2 show that Asian countries have a tendency to perform relatively better on items that are deemed not appropriate for the PISA test. These items are mostly “content rich” items that involve formal mathematics.

REASONS FOR DIFFERENTIAL ITEM FUNCTIONING

It is interesting to examine each item in Table 2 and Figure 2 to identify the reasons for differential item functioning between Western and Asian countries. While the items Asian countries performed relatively well in tend to be those with formal mathematics, what types of items are those that Asian countries performed relatively poorly in, and what are the reasons for this poor performance? We take a look at four items as an illustration.

The item where there was the largest difference between item difficulties for Australia and Hong Kong was item 33 (M022135) in the TIMSS released item set, as shown in Figure 3.

Table 3 shows the item analysis for this item for Australia and Hong Kong. It can be seen that Australian students answered this question correctly almost twice as often as Hong Kong students did (57% compared with 26%). Further, more than twice as many Hong Kong students chose the incorrect answer D compared with Australian students.

Figure 2. A Plot of Relative Item Difficulties between West and East, and Whether the Item Could Be in the PISA Test

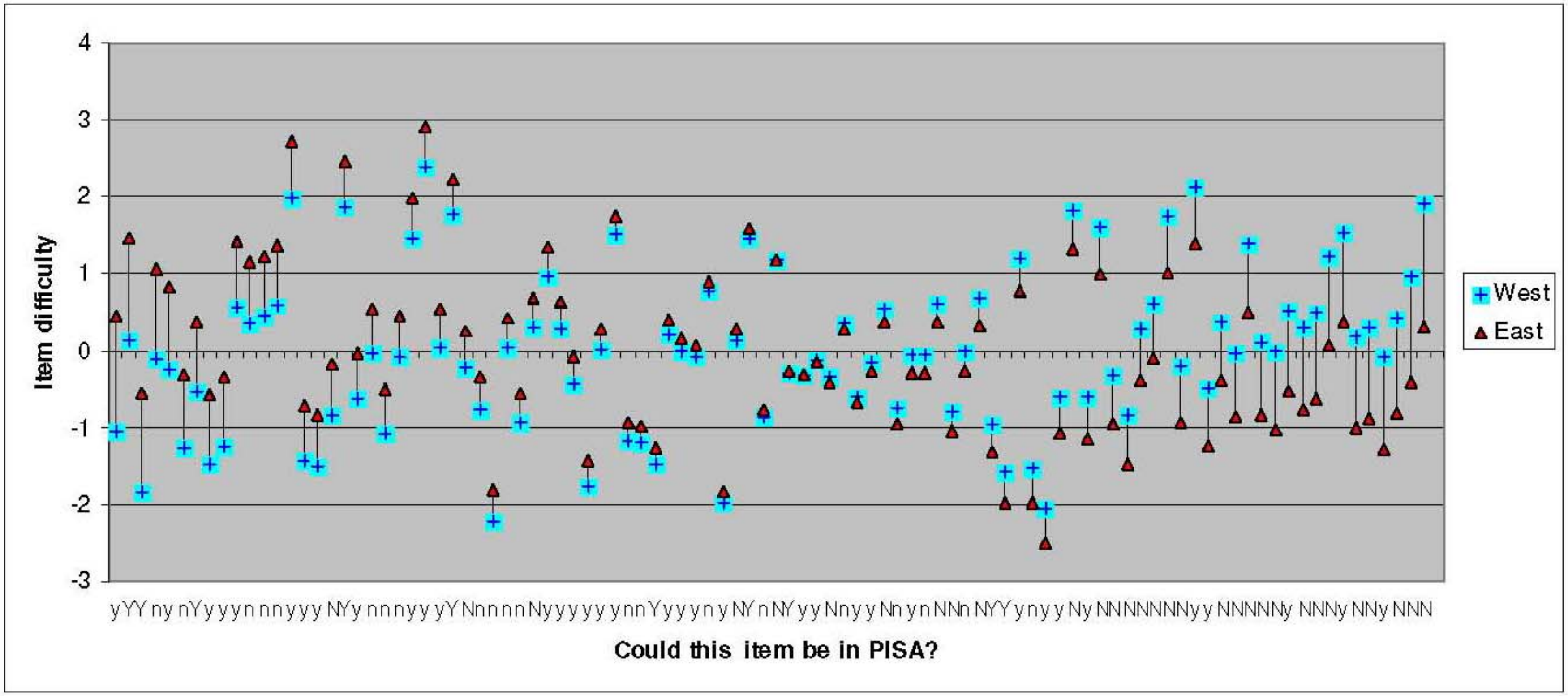
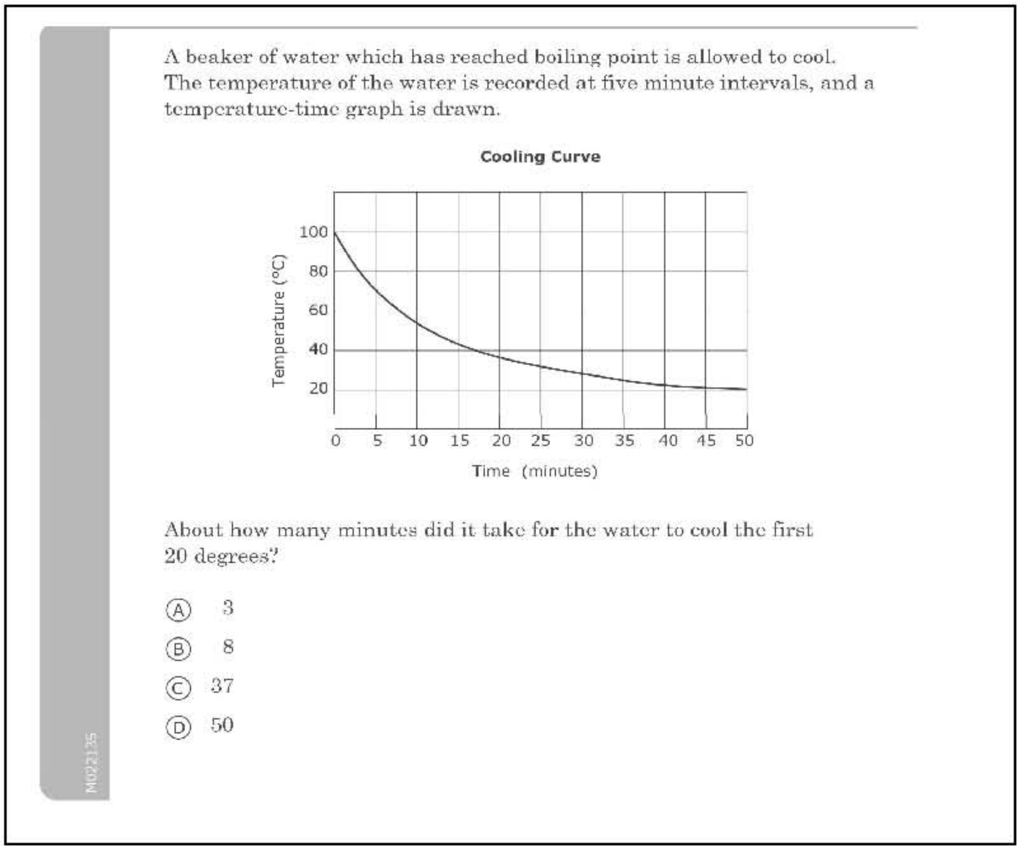


Figure 3. TIMSS 2003 Released Item M022135


Cooling Curve
100
80
60
40
20
Temperature (°C)
0
5
10
15
20
25
30
35
40
45
50
Time (minutes)
M022135

A beaker of water which has reached boiling point is allowed to cool. The temperature of the water is recorded at five minute intervals, and a temperature-time graph is drawn.

About how many minutes did it take for the water to cool the first 20 degrees?

Ⓐ 3

Ⓑ 8

Ⓒ 37

Ⓓ 50

Table 3. Item Statistics for Item M022135 for Australia and Hong Kong

Australia

```
item:33 (M022135)
Cases for this item    818
Discrimination  0.41
Item difficulty (in logits): -0.42
--------------------------------------
 Label     Score     Count   % of tot
--------------------------------------
   1        1.00      466      56.97
   2        0.00       91      11.12
   3        0.00       51       6.23
   4        0.00      198      24.21
   6        0.00        4       0.49
   9        0.00        8       0.98
======================================
```

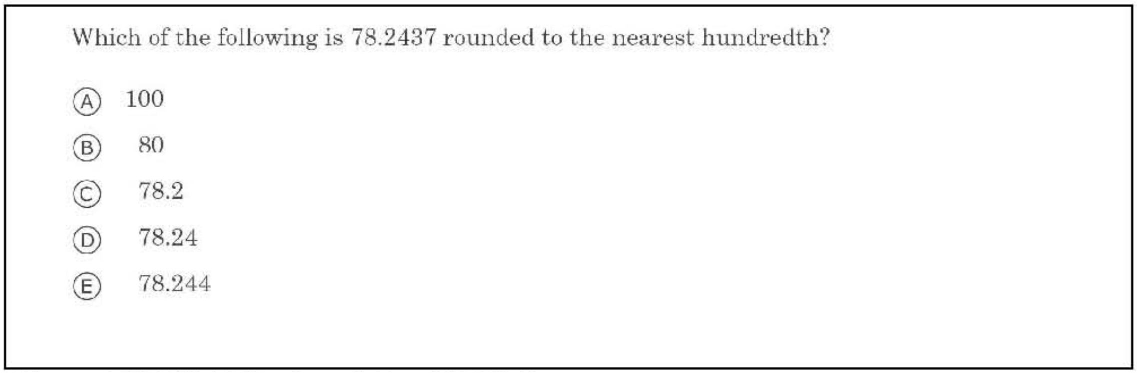
Hong Kong

```
item:33 (M022135)
Cases for this item    827
Discrimination  0.18
Item difficulty (in logits): 2.36
--------------------------------------
 Label     Score     Count   % of tot
--------------------------------------
   1        1.00      219      26.48
   2        0.00       46       5.56
   3        0.00       64       7.74
   4        0.00      493      59.61
   9        0.00        5       0.60
======================================
```

This suggests that many students in Hong Kong misread the question as “to cool *to* 20 degrees” instead of “to cool *the first* 20 degrees.” This is most likely because of language conventions, where the phrase “*the first* 20 degrees” is not a commonly used structure for Hong Kong students. In Chinese, it would typically be said as “to cool 20 degrees” (i.e., there is no explicit words “the first”). The question has been translated as “From the start till cooled the initial 20 degrees.” While this is a correct literal translation, it is not a sentence structure that people would normally use in speech. The word “initial” in this context sounds out of place in Chinese. This word can also mean “original,” or “the very beginning.” The word “till” may lead to an interpretation of “to cool to 20 degrees.” Simply put, in Chinese, there is no suitable literal translation of “to cool *the first* 20 degrees.” One would need to drop the word “the first” to make it sound like natural speech, but the translator clearly thought that this might disadvantage the students. By adding the word “initial,” while more “correct” in matching the English version, it made the sentence sound unnatural and foreign to Chinese speakers. The result was that students were still confused. In general, such differences in language usage are likely the cause for differential item functioning. This is not strictly a translation verification issue, because the translation may be literally correct, yet the sentence structure cannot be the same across languages, so the propensity to misunderstand the question will vary between language groups. The discrimination index of 0.18 for Hong Kong and 0.41 for Australia further supports the conjecture that the question was confusing to Hong Kong students.

The item that displayed the second largest difference between item difficulties for Australia and Hong Kong was item 36 (M022144). The item is shown in Figure 4.

Figure 4. TIMSS 2003 Released Item M022144



Which of the following is 78.2437 rounded to the nearest hundredth?

Ⓐ 100

Ⓑ 80

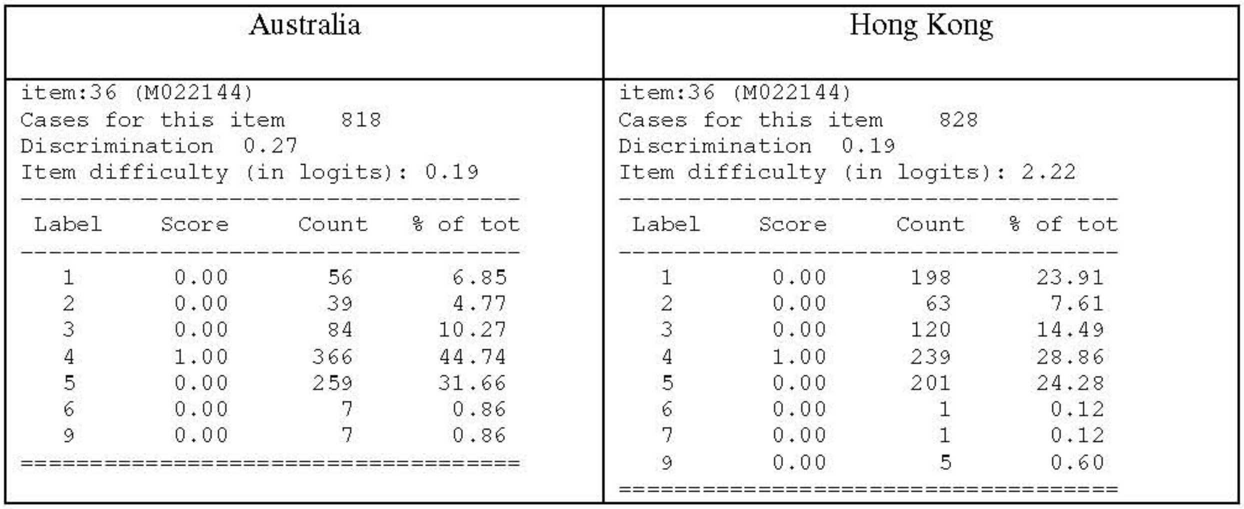
Ⓒ 78.2

Ⓓ 78.24

Ⓔ 78.244

The item statistics for Australia and Hong Kong for this item are presented in Table 4.

Table 4. Item Statistics for Item M022144 for Australia and Hong Kong



Australia

```
item:36 (M022144)
Cases for this item    818
Discrimination  0.27
Item difficulty (in logits): 0.19
------------------------------------
 Label    Score     Count   % of tot
------------------------------------
   1       0.00       56       6.85
   2       0.00       39       4.77
   3       0.00       84      10.27
   4       1.00      366      44.74
   5       0.00      259      31.66
   6       0.00        7       0.86
   9       0.00        7       0.86
====================================
```

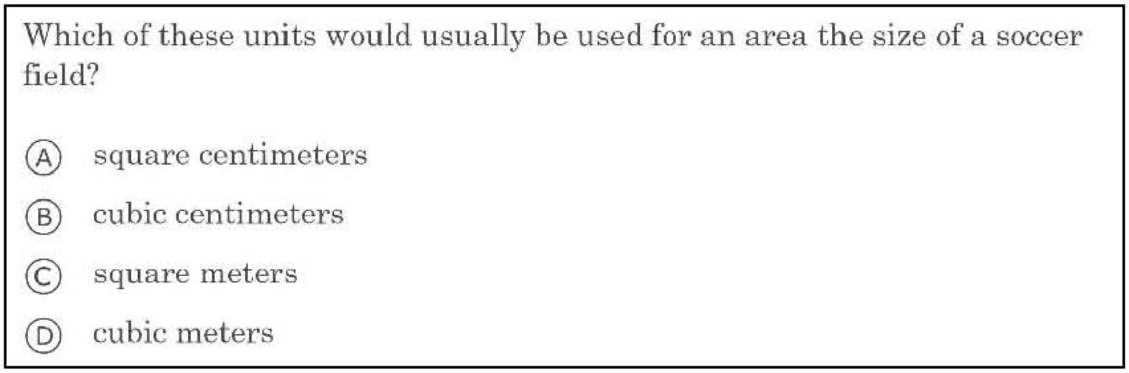
Hong Kong

```
item:36 (M022144)
Cases for this item    828
Discrimination  0.19
Item difficulty (in logits): 2.22
------------------------------------
 Label    Score     Count   % of tot
------------------------------------
   1       0.00      198      23.91
   2       0.00       63       7.61
   3       0.00      120      14.49
   4       1.00      239      28.86
   5       0.00      201      24.28
   6       0.00        1       0.12
   7       0.00        1       0.12
   9       0.00        5       0.60
====================================
```

For item M022144, Australia again outperformed Hong Kong by a large margin. The usage of the term “hundredth” for rounding numbers is not common in Chinese. Typically, it is the number of decimal places that is stated for rounding. The word “hundred” is typically only associated with whole numbers and not decimals. The discrimination index for Hong Kong (0.19) again shows that students were confused by the question. As for the previous item (M022135), the differential item functioning for item M022144 is caused by language.

A third item that displayed a large difference in item difficulties between Western and Asian countries was item 90 (M032699), as shown in Figure 5.

Figure 5. TIMSS 2003 Released Item M032699



Which of these units would usually be used for an area the size of a soccer field?

Ⓐ square centimeters

Ⓑ cubic centimeters

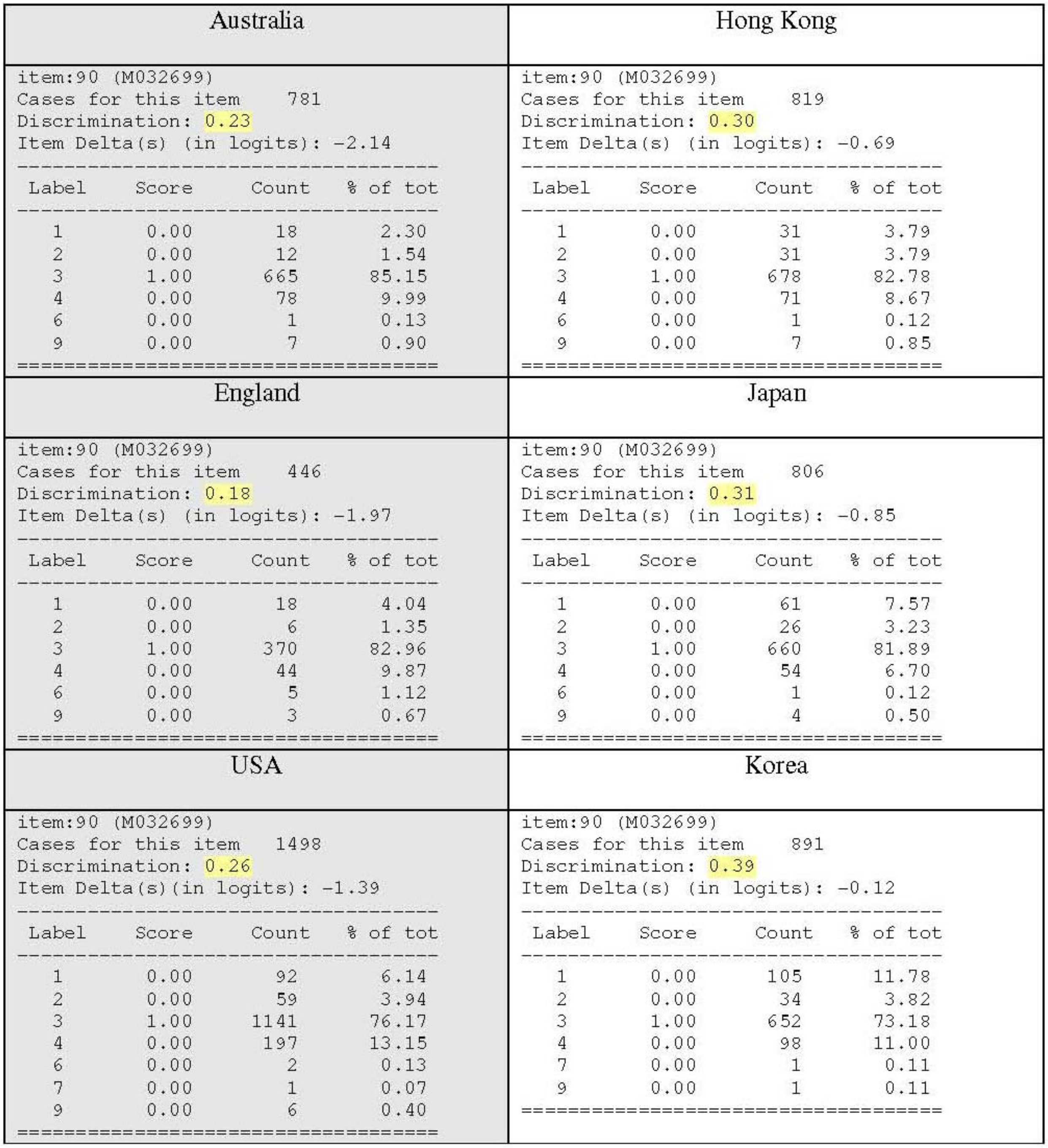
Ⓒ square meters

Ⓓ cubic meters

The item statistics for all six countries for this item are presented in Table 5.

It can be seen from Table 5 that all three Western countries found the item relatively easier than the Asian countries did (comparing item delta values). More interestingly, the discrimination indices for Western countries are all lower than the discrimination indices for Asian countries. This suggests that some students in Western countries know the correct answer because of their everyday life experience outside the mathematics classroom, while in Asian countries, students who know the correct answer to this item tend to have learned it from the mathematics classroom (most markedly in Korea). One question that one might ask is whether the item statistics will be similar had the context not been about a soccer field, but about the floor area of a concert hall. Nevertheless, the results for this item may lend some support to PISA’s suggestion that real-life mathematics is not necessarily the same as school mathematics, and promoting real-life mathematics is just as important. Note that we are not suggesting that TIMSS does not promote real-life mathematics. After all, this item appeared in the TIMSS test. The

Table 5. Item Statistics for Item M032699 for Western and Asian Countries



Australia

```
item:90 (M032699)
Cases for this item    781
Discrimination: 0.23
Item Delta(s) (in logits): -2.14
------------------------------------
 Label     Score      Count   % of tot
------------------------------------
   1        0.00       18        2.30
   2        0.00       12        1.54
   3        1.00      665       85.15
   4        0.00       78        9.99
   6        0.00        1        0.13
   9        0.00        7        0.90
====================================
```

Hong Kong

```
item:90 (M032699)
Cases for this item    819
Discrimination: 0.30
Item Delta(s) (in logits): -0.69
------------------------------------
 Label     Score      Count   % of tot
------------------------------------
   1        0.00       31        3.79
   2        0.00       31        3.79
   3        1.00      678       82.78
   4        0.00       71        8.67
   6        0.00        1        0.12
   9        0.00        7        0.85
====================================
```

England

```
item:90 (M032699)
Cases for this item    446
Discrimination: 0.18
Item Delta(s) (in logits): -1.97
------------------------------------
 Label     Score      Count   % of tot
------------------------------------
   1        0.00       18        4.04
   2        0.00        6        1.35
   3        1.00      370       82.96
   4        0.00       44        9.87
   6        0.00        5        1.12
   9        0.00        3        0.67
====================================
```

Japan

```
item:90 (M032699)
Cases for this item    806
Discrimination: 0.31
Item Delta(s) (in logits): -0.85
------------------------------------
 Label     Score      Count   % of tot
------------------------------------
   1        0.00       61        7.57
   2        0.00       26        3.23
   3        1.00      660       81.89
   4        0.00       54        6.70
   6        0.00        1        0.12
   9        0.00        4        0.50
====================================
```

USA

```
item:90 (M032699)
Cases for this item   1498
Discrimination: 0.26
Item Delta(s)(in logits): -1.39
------------------------------------
 Label     Score      Count   % of tot
------------------------------------
   1        0.00       92        6.14
   2        0.00       59        3.94
   3        1.00     1141       76.17
   4        0.00      197       13.15
   6        0.00        2        0.13
   7        0.00        1        0.07
   9        0.00        6        0.40
====================================
```

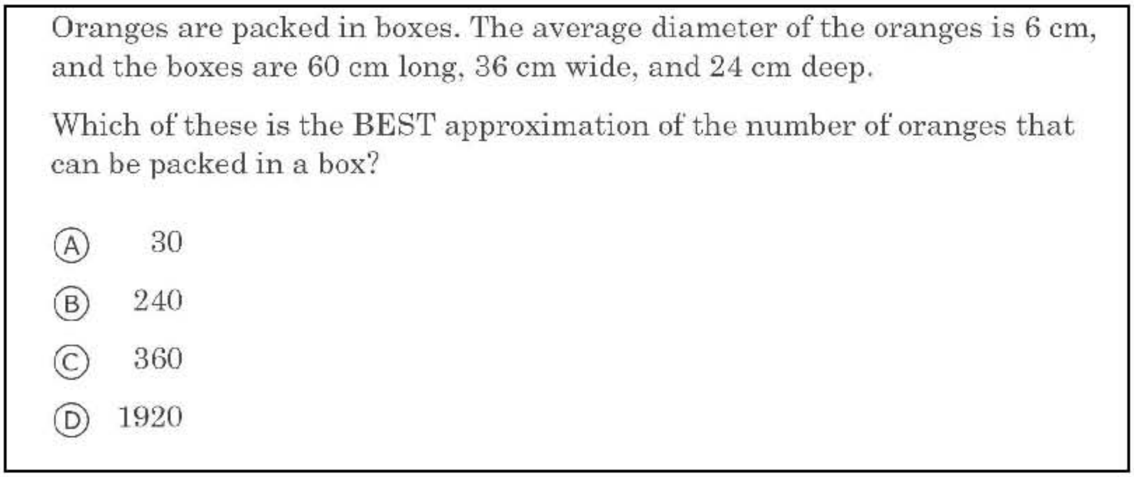
Korea

```
item:90 (M032699)
Cases for this item    891
Discrimination: 0.39
Item Delta(s) (in logits): -0.12
------------------------------------
 Label     Score      Count   % of tot
------------------------------------
   1        0.00      105       11.78
   2        0.00       34        3.82
   3        1.00      652       73.18
   4        0.00       98       11.00
   7        0.00        1        0.11
   9        0.00        1        0.11
====================================
```

finding from this item suggests that Western countries may have an advantage over Asian countries if the items are embedded in everyday life contexts.

A fourth item that demonstrated consistent differences between Western and Asian countries was item 80 (M032647). The item is shown in Figure 6.

Figure 6. TIMSS 2003 Released Item M032647



Oranges are packed in boxes. The average diameter of the oranges is 6 cm, and the boxes are 60 cm long, 36 cm wide, and 24 cm deep.

Which of these is the BEST approximation of the number of oranges that can be packed in a box?

Ⓐ 30

Ⓑ 240

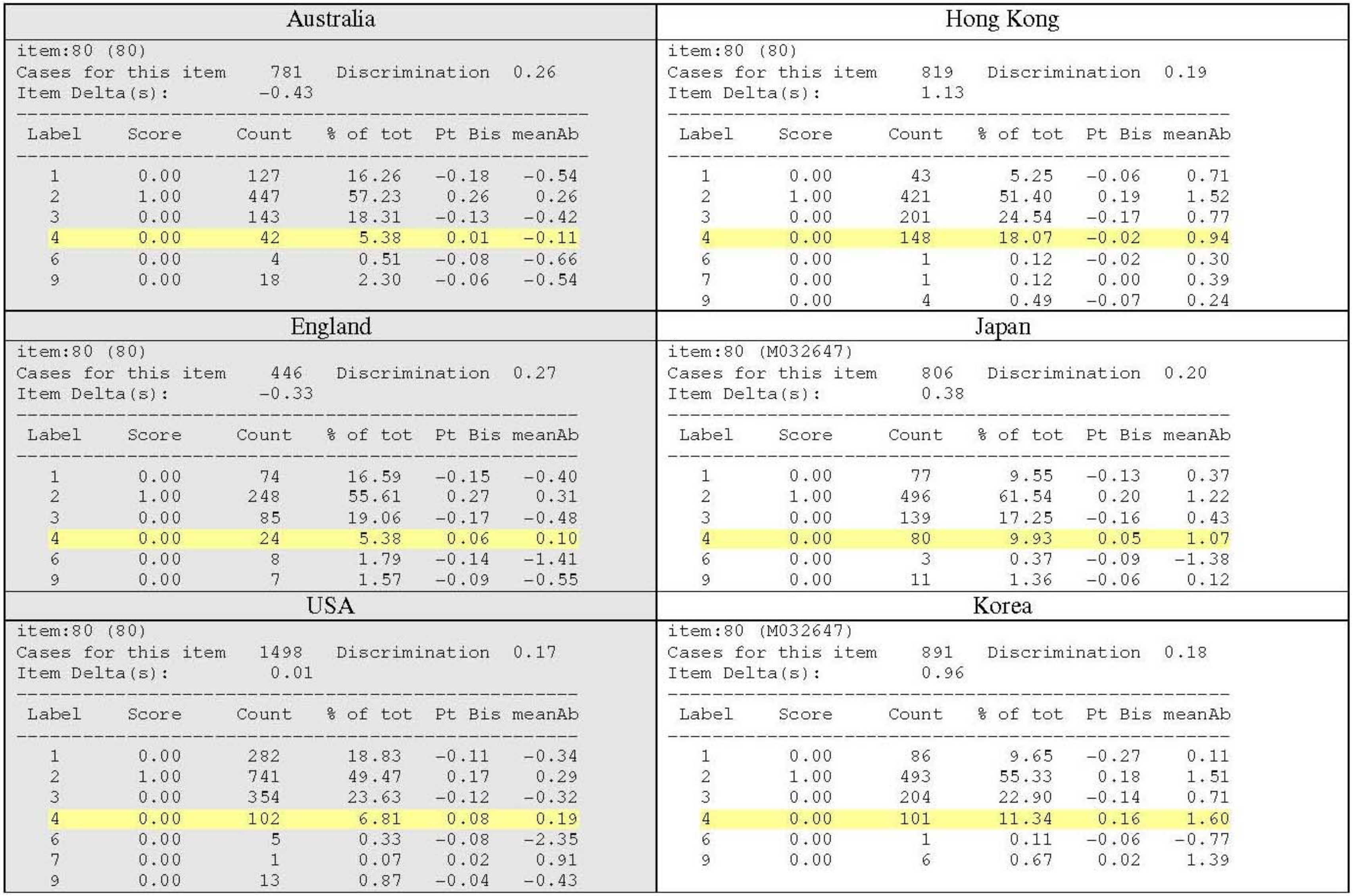
Ⓒ 360

Ⓓ 1920

This item is a real-life application of mathematics. The item statistics for the six countries are shown in Table 6. The three western countries performed relatively better on this item than the three Asian countries did (see item delta logits).

One interesting observation about this item is that there is evidence that some higher ability students chose the incorrect option D (labeled 4 in the item analysis). For all six countries, the average ability (column headed meanAb) for option D is somewhat closer to the average ability for the correction answer (option B [labeled 2]). In the case of Korea, the average ability for Option D is even higher than the average ability for Option B. To obtain the solution to this item, the mathematical skills required are not very high. It involves simple division and multiplication, where a quick estimation, rather than complex mathematical calculations of volume, can be used to obtain the answer. To get the incorrect answer of 1,920 (Option D), the likely error is that the radius of the orange is used instead of the diameter. A confusion between radius and diameter is likely if the computation of volume of shapes is carried out. It is baffling

Table 6. Item Statistics for Item M032647 for Western and Asian Countries



Australia

```
item:80 (80)
Cases for this item     781   Discrimination  0.26
Item Delta(s):        -0.43
--------------------------------------------------------------
 Label     Score      Count   % of tot  Pt Bis meanAb
--------------------------------------------------------------
   1        0.00       127      16.26   -0.18   -0.54
   2        1.00       447      57.23    0.26    0.26
   3        0.00       143      18.31   -0.13   -0.42
   4        0.00        42       5.38    0.01   -0.11
   6        0.00         4       0.51   -0.08   -0.66
   9        0.00        18       2.30   -0.06   -0.54
```

Hong Kong

```
item:80 (80)
Cases for this item     819   Discrimination  0.19
Item Delta(s):         1.13
-------------------------------------------------------------
 Label     Score      Count   % of tot  Pt Bis meanAb
-------------------------------------------------------------
   1        0.00        43       5.25   -0.06     0.71
   2        1.00       421      51.40    0.19     1.52
   3        0.00       201      24.54   -0.17     0.77
   4        0.00       148      18.07   -0.02     0.94
   6        0.00         1       0.12   -0.02     0.30
   7        0.00         1       0.12    0.00     0.39
   9        0.00         4       0.49   -0.07     0.24
```

England

```
item:80 (80)
Cases for this item     446   Discrimination  0.27
Item Delta(s):        -0.33
-------------------------------------------------------------
 Label     Score      Count   % of tot  Pt Bis meanAb
-------------------------------------------------------------
   1        0.00        74      16.59   -0.15   -0.40
   2        1.00       248      55.61    0.27    0.31
   3        0.00        85      19.06   -0.17   -0.48
   4        0.00        24       5.38    0.06    0.10
   6        0.00         8       1.79   -0.14   -1.41
   9        0.00         7       1.57   -0.09   -0.55
```

Japan

```
item:80 (M032647)
Cases for this item     806   Discrimination  0.20
Item Delta(s):         0.38
-------------------------------------------------------------
 Label     Score      Count   % of tot  Pt Bis meanAb
-------------------------------------------------------------
   1        0.00        77       9.55   -0.13     0.37
   2        1.00       496      61.54    0.20     1.22
   3        0.00       139      17.25   -0.16     0.43
   4        0.00        80       9.93    0.05     1.07
   6        0.00         3       0.37   -0.09    -1.38
   9        0.00        11       1.36   -0.06     0.12
```

USA

```
item:80 (80)
Cases for this item    1498   Discrimination  0.17
Item Delta(s):         0.01
-------------------------------------------------------------
 Label     Score      Count   % of tot  Pt Bis meanAb
-------------------------------------------------------------
   1        0.00       282      18.83   -0.11   -0.34
   2        1.00       741      49.47    0.17    0.29
   3        0.00       354      23.63   -0.12   -0.32
   4        0.00       102       6.81    0.08    0.19
   6        0.00         5       0.33   -0.08   -2.35
   7        0.00         1       0.07    0.02    0.91
   9        0.00        13       0.87   -0.04   -0.43
```

Korea

```
item:80 (M032647)
Cases for this item     891   Discrimination  0.18
Item Delta(s):         0.96
-------------------------------------------------------------
 Label     Score      Count   % of tot  Pt Bis meanAb
-------------------------------------------------------------
   1        0.00        86       9.65   -0.27     0.11
   2        1.00       493      55.33    0.18     1.51
   3        0.00       204      22.90   -0.14     0.71
   4        0.00       101      11.34    0.16     1.60
   6        0.00         1       0.11   -0.06    -0.77
   9        0.00         6       0.67    0.02     1.39
```

why some high-ability students made this mistake. Nevertheless, more students from Asian countries made this mistake. Further, the answer of 1,920 is almost nonsensical, because that is an enormous number of oranges to be placed in one box. From the item statistics one may conjecture that more higher-ability students in Asian countries than in Western countries tried to work through a mathematical solution to this question and paid no attention to sense making of the answer they obtained. This item again illustrates the contrast between a practical approach and a theoretical approach to solving mathematical problems by Asian and Western students. Because PISA contains more applied questions, Western students may have a relative advantage doing the PISA test, provided the questions can be solved using simple mathematics and a common sense approach.

Many other items that show differential item functioning in relation to Western and Asian countries also reveal interesting observations about students in these countries. Unfortunately, owing to the limited space for this paper, we are not able to provide more examples here. Appendix A shows the list of items with respect to differential item functioning between East and West. Further explorations of item performance can be carried out following the order of items in this list.

DISCUSSIONS AND CONCLUSIONS

A review of TIMSS 2003 released items showed that almost half the items were not deemed likely to appear in the PISA test, owing to the lack of application contexts for the items. These items are typically content-rich items that involve formal mathematics. An analysis of relative item difficulties showed that three Asian countries performed relatively better in these formal mathematics items than three Western countries did. This may be a contributing factor toward the observation that Western countries generally performed relatively better in PISA than in TIMSS, because TIMSS contains more context-free items involving formal mathematics.

Further, two interesting observations are made after a closer examination of differential item functioning on items where Western countries performed relatively better. First, some cultural and linguistic issues are identified. Because the source text is in English, some sentence structure and vocabulary used in English are not readily translatable into other languages. The consequence is that students in Asian countries get confused with unusual phrases and terms, and misunderstand some of the questions.

Second, students from Western countries appear to perform relatively better on everyday real-life context mathematics items where students bring their knowledge and experience from outside the classroom. In contrast, students from Asian countries tend to rely on knowledge gained in the mathematics class. Consequently, because most items in PISA are word problems in an everyday life context, it is not surprising that Western countries performed relatively better in PISA than they did in TIMSS. The observation that many students, particularly in Asian countries (and more markedly in Korea), disconnect mathematics problems from everyday life sends an important message to mathematics educators about the importance of linking mathematics to the real world. This message has been actively promoted by PISA. However, a caution is needed. An almost exclusive emphasis on real-life mathematics, particularly at the 15-year-old level, will likely restrict mathematics assessment to a set of items with lower mathematical content, and thus lead to an assessment that does not reflect all the mathematics topics taught in schools (that may be for *future* use by the students).

More generally, the findings from this paper not only help us identify the reasons for differential performance of countries in PISA and TIMSS, they also throw some light on important issues in test construction, particularly in the context of international studies. The presence of differential item functioning gives an interesting insight into the mathematical thinking of students and cultural/linguistic characteristics in different countries, yet it threatens the validity of the studies at the same time. Steps must be taken to ensure a fair assessment, both for the countries involved and for mathematics education.

ACKNOWLEDGEMENT

I am grateful to Professor Frederick Leung at the University of Hong Kong for providing the Chinese version of the TIMSS released items.

APPENDIX A: AVERAGE ITEM DIFFICULTIES (IN IRT LOGITS) FOR WESTERN AND ASIAN COUNTRIES

Seq	Unique ID	West	East	Difference	In PISA?
33	M022135	-0.28	2.61	-2.89	y
38	M022148	-1.05	0.44	-1.49	y
97	M032762	0.12	1.47	-1.35	Y
90	M032699	-1.84	-0.55	-1.29	Y
30	M022016	-0.11	1.07	-1.18	n
80	M032647	-0.25	0.82	-1.08	y
77	M032609	-1.27	-0.31	-0.96	n
18	M012037	-0.54	0.38	-0.92	Y
85	M032671	-1.49	-0.58	-0.91	y
94	M032743	-1.25	-0.35	-0.90	y
26	M022005	0.55	1.43	-0.88	y
88	M032690	0.35	1.15	-0.80	n
34	M022139	0.43	1.23	-0.80	n
27	M022008	0.58	1.37	-0.79	n
96	M032745	1.99	2.73	-0.74	y
2	M012002	-1.43	-0.73	-0.70	y
54	M022252	-1.51	-0.84	-0.67	y
63	M032208	-0.84	-0.18	-0.66	N
98	M032763	1.85	2.45	-0.60	Y
45	M022194	-0.63	-0.05	-0.58	y
36	M022144	-0.05	0.53	-0.58	n
75	M032570	-1.09	-0.51	-0.57	n
60	M032044	-0.09	0.44	-0.53	n
82	M032649B	1.46	1.98	-0.52	y
58	M022261C	2.40	2.91	-0.51	y
4	M012004	0.03	0.53	-0.51	y
99	M032764	1.76	2.23	-0.47	Y
10	M012016	-0.22	0.24	-0.46	N
14	M012027	-0.77	-0.34	-0.43	n
84	M032670	-2.22	-1.81	-0.41	n
47	M022198	0.03	0.42	-0.39	n
16	M012029	-0.95	-0.56	-0.39	n
42	M022188	0.29	0.67	-0.38	N
83	M032652	0.95	1.33	-0.38	y
81	M032649A	0.28	0.63	-0.36	y
11	M012017	-0.44	-0.09	-0.35	y
43	M022189	-1.76	-1.44	-0.32	y
95	M032744	0.01	0.29	-0.28	y
66	M032233	1.50	1.75	-0.25	y
3	M012003	-1.18	-0.94	-0.24	n
22	M012041	-1.20	-0.98	-0.22	n
53	M022251	1.17	1.18	-0.02	N
68	M032271	-0.29	-0.27	-0.01	Y
65	M032228	-0.31	-0.33	0.02	y
72	M032533	-0.12	-0.15	0.03	y
39	M022154	-0.34	-0.41	0.07	N
29	M022012	0.34	0.27	0.07	n
28	M022010	-0.61	-0.68	0.07	y
56	M022261A	-0.14	-0.26	0.12	y
87	M032689	0.53	0.36	0.17	N
1	M012001	-0.74	-0.95	0.21	n
70	M032447	-0.07	-0.29	0.23	y
25	M022004	-0.06	-0.28	0.23	n
64	M032210	0.61	0.36	0.25	N
9	M012015	-0.79	-1.05	0.26	N
93	M032732	0.00	-0.28	0.28	n
78	M032612	0.69	0.33	0.36	N
6	M012006	-0.95	-1.32	0.38	Y
12	M012025	-1.57	-1.98	0.41	Y
57	M022261B	1.20	0.76	0.44	y
19	M012038	-1.53	-1.98	0.46	n
71	M032489	-2.04	-2.50	0.46	y
44	M022191	-0.61	-1.07	0.47	y
24	M022002	1.82	1.32	0.50	N
7	M012013	-0.61	-1.15	0.54	y
61	M032046	1.61	0.98	0.63	N
46	M022196	-0.33	-0.97	0.64	N
15	M012028	-0.83	-1.49	0.66	N
67	M032261	0.28	-0.39	0.67	N
79	M032643	0.61	-0.11	0.72	N
74	M032557	1.74	1.01	0.73	N
69	M032403	-0.19	-0.93	0.74	N
52	M022227C	2.13	1.38	0.75	y
37	M022146	-0.49	-1.25	0.76	y
86	M032678	0.38	-0.39	0.77	N
20	M012039	-0.02	-0.85	0.83	N
49	M022202	1.40	0.49	0.90	N
23	M012042	0.11	-0.83	0.94	N
55	M022253	-0.02	-1.02	1.00	N
31	M022021	0.51	-0.53	1.04	y
35	M022142	0.30	-0.76	1.06	N
59	M032036	0.48	-0.62	1.10	N

Seq	Unique ID	West	East	Difference	In PISA?
21	M012040	-1.47	-1.26	-0.21	Y
17	M012030	0.20	0.40	-0.20	y
91	M032727	0.00	0.17	-0.17	y
40	M022156	-0.07	0.07	-0.14	y
62	M032079	0.76	0.90	-0.14	n
8	M012014	-1.97	-1.84	-0.13	y
92	M032728	0.13	0.26	-0.13	N
32	M022127	1.47	1.57	-0.11	Y
76	M032588	-0.87	-0.78	-0.09	n

Seq	Unique ID	West	East	Difference	In PISA?
89	M032693	1.22	0.07	1.15	N
51	M022227B	1.54	0.38	1.15	y
5	M012005	0.18	-1.00	1.18	N
41	M022185	0.30	-0.90	1.19	N
50	M022227A	-0.08	-1.29	1.21	y
13	M012026	0.41	-0.83	1.24	N
48	M022199	0.96	-0.41	1.37	N
73	M032545	1.90	0.30	1.60	N

REFERENCES

American Institutes for Research. (2005). *Reassessing U.S. International Mathematics Performance: New Findings from the 2003 TIMSS and PISA*. Washington, DC: American Institutes for Research.

Hutchison, G., & Schagen, I. (2007). Comparisons between PISA and TIMSS – Are We the Man with Two Watches? In Loveless, T. (Ed.), *Lessons Learned – What International Assessments Tell Us about Math Achievement*. Washington, DC: The Brookings Institution.

International Association for the Evaluation of Educational Achievement. (2003). *TIMSS Assessment Frameworks and Specifications, 2003*. Chestnut Hill, MA: TIMSS International Study Centre.

National Center for Education Statistics. (2008). *Comparing NAEP, TIMSS, and PISA in Mathematics and Science*. Retrieved May 2008, from http://nces.ed.gov/timss/pdf/naep_timss_pisa_comp.pdf.

Neidorf, T.S., Binkley, M., Gattis, K., & Nohara, D. (2006). *Comparing Mathematics Content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 Assessments (NCES 2006-029)*. U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Organisation for Economic Co-operation and Development. (2003). *The PISA 2003 Assessment Framework – Mathematics, Reading, Science and Problem Solving Knowledge and Skills*. Paris: Organisation for Economic Co-operation and Development.

Wu, M.L. (in press). *A Comparison of PISA and TIMSS 2003 Achievement Results in Mathematics and Science*. Prospects, UNESCO.