

THE IMPERIAL VS. METRIC STUDY

(IMS)

Mark Wilson & Yiyu Xie

University of California, Berkeley

October 2004

BACKGROUND

The United States is one of the few countries that use a non-metric system of measurement (which we will call the “imperial” system due to its origins in the British empire). When international tests on mathematics or science are distributed, the language of the tests is translated to suit each participating country, but the measurement units are usually not converted from metric units. Thus, the preponderance of the imperial system in general usage in the US and in the school curriculum in particular may be unfavorable to American students taking the assessments using the metric system. Search for literature on this topic shows that no research has been undertaken in the past 20 years to study how this unfamiliarity with the metric system might put American students at a disadvantage on international assessment (Calsyn, 2002). One article from *Education Week* outlined the findings from the Second International Mathematics and Science Study where it was hypothesized that the relative poor performance of U.S. students on the arithmetic and measurement portions of the assessment might be caused by their inexperience in using the metric system (Bridgman, 1984).

The purpose of this study is to investigate possible effects on test performance of using metric measurement units (e.g., meters, liters, etc.) compared to the more familiar imperial units (e.g., feet, gallons, etc.).

PILOT STUDY

A pilot study, using the PISA 2000 data, was set up to see if some existing data might shed some light on the issue. Australia was chosen as a comparison with USA, being an English-speaking country that uses metric units but that in other ways could be seen as being similar to the United States.

The study, which utilized a differential item functioning (DIF) design, was conducted on a subset of the PISA 2000 data, composed of students' responses from Australia and the U.S. to 31 math items. The goal of this study was to explore whether these items function the same for both countries. If DIF items are detected, then we can check to see if there is a relationship between an item's DIF status and the presence or absence of metric units.

The data were organized in a way so that only students with a complete set of valid responses from both countries were included in the analyses. The original PISA 2000 database contains 5176 and 3846 students from Australia and the U.S., respectively. Some students did not have responses to any of the 31 items, and therefore were eliminated from our study. As a result, the DIF analyses were carried out with a sample of 4936 students in total (2838 from Australia and 2098 from the U.S.).

The following is a list of models we tried out using the data set. Fit results are shown in Table 1.

1. A regular partial credit model (some items have polytomous responses)
2. An additional facet, country effect, was added to the partial credit model.
3. An exploratory DIF analysis (in addition to the country effect, each item was examined as to whether it behaves the same for Australia and the U.S.)

Insert Table 1 about here

By comparing the model deviance in rows 1 and 2 in Table 1 using the likelihood ratio Chi-squared test of parameter variation, we can see statistical significance for a country effect between the two countries ($\chi^2=351$, $df=1$, $p<.001$). Overall, Australian students' performances are estimated to be about 0.71 logits ($SE=.009$) higher than the U.S. students on these math items.

Results from the comparison of model 2 with model 3 indicate that there are statistically significant DIF (country by item) effects ($\chi^2=176$, $df=31$, $p<.001$). Thirteen out of the 31 items behave differently for the two countries. However, there is no systematic pattern showing one country always performing better than the other. There are 7 items on which American students perform relatively better and 6 items on which Australian students perform relatively better (see Table 2 for details).

Insert Table 2 about here

Since the main purpose of investigating the DIF effect of this data was based on the speculation that using metric units in the test might cause the difference in students' performances, a careful examination of the items was performed to check if there was any use of metric units in items displaying DIF effects and to what extent they were used in non-DIF items. Table 2 also shows in which items metric units were used. Clearly,

there is only a little overlap between the items involving metric units and those with DIF effects. Among the four DIF items that involved metric units, two showed Australian students performing relatively better, and two showed U.S. students performing relatively better. Moreover, there were nine items with DIF effects that did not involve metric units. Therefore, this analysis does not give clear support either way on the question of whether using the metric units might hamper the performance of American students.

The limitations of the pilot study are that (a) the DIF investigation necessarily confounds country effects with potential unit effects, and (b) some items involve metric units only “superficially” – that is, there is no need to understand anything about such units to get the item right, they are used only as labels.

A better data collection design will help disentangling these effects:

1. items using imperial units can be designed to match current PISA items
2. items that use metric and/or imperial units can be designed to do so to different extents, from “superficial” involvement to actually be involved in the calculation (and there should be an adequate number of items in each category).

With such a design, we can carry out analyses to see the extent of effects from these two causes. In the next section we give details of the design of this study, called the Imperial vs. Metric Study (IMS).

DESIGN

The IMS implemented a new design within the PISA 2003 study for U.S. students. The PISA 2003 main study contains 13 test booklets. Each booklet is made up of 4 test clusters. There are 7 mathematics clusters ($M_1 - M_7$), 2 science clusters ($S_1 -$

S₂), 2 problem solving clusters (P₁ – P₂) and 2 reading clusters (R₁ – R₂). The clusters are allocated in a rotated design to the thirteen booklets (see Table 2.1 for the allocation of clusters to booklets). Each cluster contains approximately 12 test items, equivalent to about 30 minutes of test material.

Item Clusters

Similarly as in the main study, the IMS contains 4 extra test booklets with 4 clusters in each booklet. The IMS is focused on mathematics items (as this is where the metric units occur). Among the 85 mathematics items in the main study, there are 20 items that have metric units in both the item itself and the source material¹ (“full metric”) and a further 7 that have metric units in the source, but not in the item (“part metric”). Matching imperial versions of these 27 items (see the section of Item Adaptation for details) have been constructed. The 27 adapted items were allocated to 2 clusters (I₁ – I₂). The other 2 clusters used in the IMS are clusters M₂ and M₆ from the main study, chosen because they do not include any metric items: they will serve as link items, used for technical purposes, to allow us to link the booklets together. Table 3 also shows how the imperial clusters are assigned to booklets.

Insert Table 3 about here

Spiraling the Booklets

The method of spiraling of the booklets for the PISA 2003 main study and the IMS is constructed by splitting the sequence into two parts:

- (a) the PISA sample for the main study consists of the first 5 of every 6 booklets

¹ I.e., the items are motivated by source material, sometimes shared across items (see sample two items in Appendix 1).

(b) the Metric/Imperial sample for the IMS is the sixth of every 6 booklets.

The booklets from the two sets were placed in order into six slots: booklets 1-13 spiraled through the first five slots out of every 6; booklets 14-17 spiraled through the sixth slot.

Thus, the first segment of the spiral looks like this (where regular numbers refer to the main study booklets and italic numbers refer to the IMS booklets):

1, 2, 3, 4, 5, *14*, 6, 7, 8, 9, 10, *15*, 11, 12, 13, 1, 2, *16*, 3, 4, 5, 6, 7, *17*, 8, 9, 10

This method of spiraling of the booklets is to make sure that we have an adequate number of students responding to each booklet.

Item Adaptation

In the process of unit conversion, efforts were made to keep the item difficulties unchanged by keeping the item content as constant as possible. For some of the items, simply changing the metric units to the corresponding imperial ones is enough, e.g., meters to yards. In this case, the unit is used “superficially” as described before. For some other items, it is necessary to change the numbers in addition to the units, e.g., 170 centimeters to 70 inches. In Appendix 1, some examples of item modification are shown.

DATA ANALYSIS

The IMS used the random coefficients multinomial logit (RCML) model (Adams, Wilson & Wang, 1997) to analyze the data. The scaling was done with the *ConQuest* software (Wu, Adams & Wilson, 1997) – note that this is the same software used for the international analyses of the PISA data. The results can be summarized in two parts, the comparison of items and the comparison of students.

Comparing Items

There are a total of 112 mathematics items in the calibration – 85 from the seven mathematics clusters (including 27 items that are the unadjusted metric items) and 27 adapted imperial items.

Appendix 2² lists the item parameter estimates, standard errors, and fit statistics for the 112 items. Most of the items fit the Rasch model reasonably well. Large positive T values are common, but this is to be expected due to the relatively large sample size. The most important fit indicator is the mean square (MNSQ), where values either considerably larger than 1.0 or considerably smaller indicate that there are a few items that are misfitting. All items have weighted mean square statistics that fall within regular guidelines.³ Some items have fit indices that are close to the standard limits. For example, the estimated parameters for items M266Q01 and I266Q01 (both based on the unit named “Carpenter”), at both item level and step level, have indices quite close to the limits. This is a link item from PISA 2000. It would be worthwhile to check how this item performed in the earlier administration to see if this is a consistent finding.

Figure 1 graphically places all focus items, the 27 items that have the metric unit of measurement in them and their matching imperial version of the items, on a common scale from approximately –4 to 4 logits. The items are re-labeled from 1 to 27 in the Figure so that the long item names do not obscure one another. The threshold estimates for them are shown in the right hand part of the Figure under the headings “Metric” and

² Note that in this Appendix, items starting with M are original mathematics items (i.e., the unadjusted metric items), while items starting with I are the imperial items. Having the same item label indicates the pairs of metric and imperial items. Thus the 86th item in the table is I302Q01, and its metric pair is the 74th item, M302Q01. The suffix Q01 indicates that this is the first question pertaining to the source material for Unit 302.

³ Standard control limits are .75 and 1.33 (Adams & Khoo, 1993)

“Imperial”. Items M302Q01 and I302Q01 (both based on the unit named “Car Drive”) are the easiest items – in fact, the estimates of their difficulty parameters are almost 2 logits below others. This is an item that requires the examinee to read a value from a map. This result may suggest that this item is too easy for 15-year-olds though it may be useful as a starter item for the test.

Insert Figure 1 about here

Table 4 lists the estimates of item difficulty parameters for these 27 pairs of items, and the step parameters for polytomous items. Both Figure 1 and Table 4 suggest that except for a few items, most of the imperial items behave similarly to their counterparts (differences are within .3 logits). From the Table, we can see that the items that differ to a substantial extent are 150Q03 (unit name “Growing Up”), 810Q03 (unit name “Bicycles”) and 124Q03 (unit name “Walking”).

Insert Table 4 about here

Ideally, we would find that we had made the conversion to imperial units without affecting the underlying difficulty of the items. However, this is difficult to accomplish. Hence, we analyze the results for the discrepant pairs below. Furthermore, we will carry out later analyses using both the full set of items, and the set without these items.

Discussion of selected items⁴

Growing Up Q03. The modification for this unit of items was to change the numbers and units on the Y-axis of a graph representing the relationship between height and age. Item I150Q03 appears to be more difficult than M150Q03. The difference is approximately .7

⁴ Note that we are not showing the text of the items in order to maintain test security.

logits. The other two questions related to the same graph, 150Q01 and 150Q02, do not show much difference between the item difficulty parameters from the metric and imperial versions (.12 and .08 logits respectively). To understand why 150Q03 behaves quite differently from the other two when the modification was applied to the common stem material, we broke down the counts of response categories and show the frequencies below in Table 5. This is a free-response type of question. Each response was given a code by readers as follows: Codes 01 and 02 receive zero credit whereas codes 11, 12 and 13 get full credit. Code 99 denotes a missing response.

Insert Table 5 about here

The major differences are in codes 02 and code 11. Code 11 grants full credit when students give correct answers using everyday language, not mathematical language. Code 02 is assigned to incorrect responses that do not refer to the characteristics of the graph. Thus, it seems that the unit of measurement is not the key to the discrepancy here. The item format allows the students to write freely about why the growth rate for girls slows down after 12 years of age. Absent a systematic explanation, it may be that the sample of students who received the imperial booklets randomly includes a larger proportion of students who give unrelated answers like “girls mature early”.

Bicycles Q03. This is the item with the largest discrepancy between its imperial parameter estimates and their counterparts among the focus items. The difference lies in the step parameters. The step parameter is an indication of the easiness or difficulty for students to make a certain step in a polytomous item (i.e., it describes the log-odds of step k compared to step k). The larger the estimated value is, the more difficult it is to make

the step. The imperial version of the item has a larger value for the first step, and a smaller value for the second step. This means that for the imperial version, it is more difficult for students to make the first step, which is to the partial credit score of 1. In contrast, once they have achieved the first step, the second step, to the full credit score of 2, is relatively easier to achieve. Figure 2 shows the category probability curves for this item for both the metric and imperial cases. Notice that the probability curves for categories 0 and 2 are approximately the same under the two conditions, but that the probability of a 1 is lower for the imperial item.

Insert Figure 2 about here

Indeed, it seems that the numbers we changed in this item (in addition to the change in the unit of measurement) made the step difficulty harder. The essential feature of this item is that the student must recognize that the relationship of 84 centimeters to 840 meters is the same as that of .84 meters to 840 meters.⁵ Trying to change this to an imperial version is quite difficult, as the imperial units of measurement that are most closely analogous to centimeters and meters are inches and yards, which have a ratio of 1:36 rather than 1:100. For this reason it was very difficult to maintain both the nature of the item and the relationship among the numbers to be used in the item.

Thus, the modification we made to this item has added a confounding factor (decimal vs. duodecimal) to the item parameter values.

Walking Q03. The difference between I124Q03 and M124Q03 lies in the third step parameter estimates. It is easier for students to make the third step, which is to full credit,

⁵ The numbers used here are changed from the item itself.

in the imperial version of the item (1.82 logits) than in the metric version of the item (2.49 logits). This should not be too surprising, when we consider the modification we made to this item. No number was changed, only the unit of measurement was changed. The metric version of the question asks for answers in two ways, meters per minute and kilometers per hour. It is quite straightforward to adapt the first into the imperial system; it became yards per minute. However, for the second, there is no imperial equivalent to the prefix “kilo-“ used in the metric system. We first used “thousand of yards per hour” based on the most direct conversion. Then we changed it to “hundred of yards per hour” because “hundred of ...” is more common than “thousand of ...”. We believe that this explains the relative easiness for students to achieve the correct answers in both ways in the imperial version.

Comparing Students

A total of 9610 U. S. students participated in PISA 2003. Among them, 8027 received the metric booklets, and 1583 received the imperial booklets. For each student, five plausible values were drawn from the latent distribution and the mean of the five means from the 5 sets of plausible values was used to calculate the sample mean. The plausible-value method is designed to provide reliable indices of student proficiency.

By virtue of the randomized spiraled sampling design, we can assume that the sample of students who received the two forms of the test were randomly equivalent. However, as we also have a core of items that does not contain any measurement units (we call them “non-focus items”), we can check up on how closely this assumption is reflected in their actual results. Since the non-focus items also include two clusters of

items that are common to both samples, we can compare the two distributions on the same scale. Thus, the results for the two samples on the non-focus items are shown in the top panel of Table 6. Because of the large sample sizes, z test was performed to compare the sample means. The difference between the two sample means is not statistically significant ($z=.616$, $p=.54$) from 0, and neither is the difference between the spreads of the samples ($F=1.02$, $p=.69$). This is further evidence that the two samples are equivalent.

Insert Table 6 about here

The second panel of Table 6 shows the means and standard deviations of the two samples using the complete set of items for each (i.e., the non-focus items plus the metric items for the metric group, and the non-focus items plus the imperial items for the imperial group). Again the difference between the two sample means is not statistically significant ($z=.077$, $p=.94$), and neither is the difference between the spreads of the samples ($F=1.043$, $p=.85$). Because one might wonder whether these results were due to the non-focus items overwhelming the effects of the focus items, we ran this comparison again, this time for only the focus items (i.e., the metric items for the metric group, and the imperial items for the imperial group). The difference between the two sample means is, again, not statistically significant ($z=-.093$, $p=.93$), and neither is the difference between the spreads ($F=1.032$, $p=.78$). Note that the analyses reported so far used all of the metric and imperial items, including the ones identified above as showing some DIF between the two samples. We also want to check the comparison if the DIF items were removed from the analyses. We re-ran the analyses deleting those items from the focus items set. The fourth panel of Table 6 shows the results. Results are similar again, no

significant differences between the means ($z=-.406$, $p=.68$) and spreads ($F=1.059$, $p=.92$).

Finally, we re-ran the analyses deleting the DIF items from the complete set of items.

The results were found to be identical in interpretation to that of all items – no statistically significant differences have been found in the sample means ($z=-.003$, $p=.99$) and sample spreads ($F=1.052$, $p=.90$).

Finally, the above comparisons of samples of students were repeated using sampling weights to ensure that appropriate inferences of population characteristics can be drawn. The numbers of students used in these analyses are 7582 and 1500 for metric and imperial groups, respectively. The results are shown in Table 7. Thus, students measured with the metric items are found to be not disadvantaged by the use of the metric items compared to others who took the imperial items. Nor are they advantaged by the differences in the item sets.

Insert Table 7 about here

CONCLUSION

This study shows no evidence that choice of measurement unit (imperial vs. metric) has a deleterious effect on American students' performance on mathematics for the 15-year-olds in the PISA study. Students who received the items with metric units performed similarly to their peers who received the items with imperial units. With just a few exceptions, items in two forms yielded similar difficulty estimates. There is no systematic pattern of the relative difficulty of the imperial version of the items compared with the metric version. Some item discrepancies were observed due to: (a) differences in the nature of the two systems (e.g., decimal vs. duodecimal, no equivalent wording of

the units), and (b) difficulties in the modification process (e.g., no comparable scoring guides for some incorrect approaches to an item).

REFERENCES

- Adams, R. J. & Khoo, S. (1993). *Quest: The interactive test analysis system* [computer program manual]. Hawthorn: Australian Council for Educational Research.
- Adams, R. J., Wilson, M. & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21(1), 1-23.
- Bridgman, A. (1984). International math assessment finds U.S. students 'average'. *Education Week*, May 16, 1984.
- Calsyn, C (2002). *The results of the research for literature on U.S. student performance with the metric system*. Unpublished document. Westat.
- Wu, M. L., Adams, R. J. & Wilson, M. (1997). *ConQuest: Generalized item response modeling software* [computer program]. Melbourne: Australian Council for Educational Research.

Table 1
Model fits for the three estimated models

Model	# of Parameters	Deviance
1. Partial Credit (PC)	41	80882
2. PC + country	42	80531
3. PC + country + country by item	73	80355

Table 2
 Presence of DIF and usage of metric units in PISA 2000 items

Item	DIF presence	Australia better	U.S. better	Metric unit used
1				
2	Y		Y	
3	Y		Y	Y
4				Y
5				Y
6	Y	Y		Y
7	Y	Y		
8				
9	Y	Y		
10	Y		Y	
11	Y		Y	
12	Y		Y	
13				
14				
15				Y
16				Y
17				Y
18				Y
19				
20				
21	Y	Y		
22	Y	Y		
23	Y		Y	Y
24				Y
25				Y
26				
27				
28				
29	Y		Y	
30				Y
31	Y	Y		Y

Table 3
Test booklet design for the PISA 2003 main study and IMS

Study	Booklet	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Main	1	M1	M2	M4	R1
	2	M2	M3	M5	R2
	3	M3	M4	M6	PS1
	4	M4	M5	M7	PS2
	5	M5	M6	S1	M1
	6	M6	M7	S2	M2
	7	M7	S1	R1	M3
	8	S1	S2	R2	M4
	9	S2	R1	PS1	M5
	10	R1	R2	PS2	M6
	11	R2	PS1	M1	M7
	12	PS1	PS2	M2	S1
	13	PS2	M1	M3	S2
IMS	14	I1	M2	I2	M6
	15	M2	I2	M6	I1
	16	I2	M6	I1	M2
	17	M6	I1	M2	I2

M1 indicates mathematics cluster 1, and so on; S is for science, R for reading, PS for problem solving, and I for imperial.

Table 4
Item parameter estimates for the focus items

#	Label	Metric				Imperial			
		δ_i	δ_{i1}	δ_{i2}	δ_{i3}	δ_i	δ_{i1}	δ_{i2}	δ_{i3}
1	302Q01	-3.92				-4.13			
2	302Q02	-1.80				-1.79			
3	302Q03	1.05				1.21			
4	266Q01	0.97	0.50	1.44		0.81	0.36	1.27	
5	464Q01	2.03				2.17			
6	810Q01	-1.10				-1.12			
7	810Q02	-0.78				-0.99			
8	810Q03	2.03	1.55	2.51		1.88	3.07	0.69	
9	421Q01	-1.01				-1.03			
10	421Q03	1.13				0.96			
11	421Q02	2.10				2.19			
12	547Q01	-1.62				-1.62			
13	124Q01	-0.28	-1.48	0.92		-0.32	-1.61	0.98	
14	124Q03	1.63	0.03	2.35	2.49	1.38	0.19	2.12	1.82
15	305Q01	-0.09				-0.13			
16	462Q01	0.78	-1.31	2.87		0.98	-1.19	3.15	
17	406Q01	2.00				1.95			
18	406Q02	2.15				2.35			
19	406Q03	2.04				2.34			
20	150Q01	-0.34				-0.21			
21	150Q03	-0.31				0.39			
22	150Q02	-0.77	-1.83	0.29		-0.85	-1.79	0.09	
23	474Q01	-0.78				-0.80			
24	273Q01	0.12				0.17			
25	828Q01	0.49				0.76			
26	828Q02	-0.26				-0.13			
27	828Q03	0.99				0.88			

δ_i is the item difficulty parameter and δ_{i1} to δ_{i3} are the step difficulty parameters for polytomous items.

Table 5
Comparison of response frequencies for one pair of items

Response Category	Frequency for M150Q03	Percent	Frequency for I150Q03	Percent
01	41	1.66	46	2.91
02	652	26.33	517	32.66
11	715	28.88	340	21.48
12	120	4.85	71	4.49
13	69	2.79	13	.82
99	879	35.5	596	37.65
Total	2476	100	1583	100

Table 6
Means and standard deviations for the samples of students
who received the different test forms

Sample	Mean (S.E.)	Std. Dev. (S.E.)
Non-focus items		
Metric	.0007 (.0127)	1.0886 (.0198)
Imperial	-.0228 (.0360)	1.0779 (.0472)
All items		
Metric	-.0064 (.0136)	1.0752 (.0185)
Imperial	-.0089 (.0297)	1.0526 (.0399)
Focus items		
Metric	-.0040 (.0182)	1.0704 (.0198)
Imperial	-.0003 (.0354)	1.0539 (.0399)
Focus items w/o 3 DIF items		
Metric	-.0067 (.0185)	1.0640 (.0191)
Imperial	.0095 (.0354)	1.0340 (.0369)
All items w/o 3 DIF items		
Metric	-.0041 (.0131)	1.0750 (.0185)
Imperial	-.0040 (.0286)	1.0479 (.0391)

Table 7
Means and standard deviations for the samples of students
who received the different test forms with sampling weights

Sample	Mean (S.E.)	Std. Dev. (S.E.)
Non-focus items ¹		
Metric	-.0009 (.0156)	1.1049 (.0192)
Imperial	-.0200 (.0310)	1.0971 (.0412)
All items ²		
Metric	-.0108 (.0135)	1.0826 (.0187)
Imperial	-.0231 (.0281)	1.0626 (.0416)
Focus items ³		
Metric	-.0131 (.0136)	1.0654 (.0183)
Imperial	.0022 (.0354)	1.0489 (.0399)
Focus items w/o 3 DIF items ⁴		
Metric	-.0135 (.0141)	1.0496 (.0174)
Imperial	-.0015 (.0374)	1.0325 (.0380)
All items w/o 3 DIF items ⁵		
Metric	-.0078 (.0128)	1.0793 (.0188)
Imperial	-.0194 (.0284)	1.0617 (.0411)

¹ z=.550, p=.58; F=1.014, P=.63

² z=.395, p=.69; F=1.038, P=.82

³ z=-.404, p=.69; F=1.032, P=.78

⁴ z=-.300, p=.76; F=1.033, P=.79

⁵ z=.372, p=.71; F=1.033, P=.79

Logit	Person	Metric Items	Imperial Items
			16.2
3			
		14.3 16.2	
		8.2	
	X		14.3 18 19
	XX 18		5 11
2	XX 5 11 14.2 17 19		8.2
	X		17
	XX		8.1
	XXX 4.2		14.2
	XXX		4.2
	XXXX 8.1		
	XXX		3
1	XXXXXX 3 10 13.2		13.2
	XXXXX 27		10 27
	XXXXXX		25
	XXXXXXXX		
	XXXXXXXXXX 25		
	XXXXXXXXXX 22.2		21
	XXXXXXXXXX 4.1		22.2 24
0	XXXXXXXXXX 24		4.1 14.1
	XXXXXXXXXX 14.1 15		15 26
	XXXXXXXXXX 26		20
	XXXXXXXXXX 20 21		
	XXXXXXXXXX		
	XXXXXXXXXX		
	XXXXXXXXXX 7 23		23
-1	XXXXXXXXXX		7
	XXXX 6 9		6 9
	XXXX		16.1
	XXXX 16.1		
	XXX 13.1		
	XX 12		12 13.1
	XX 2		2
-2	X 22.1		22.1
	X		
	X		
	X		
	X		
-3			
-4	1		1

Each 'X' represents 59.7 cases.

Figure 1. Threshold map for the focus items

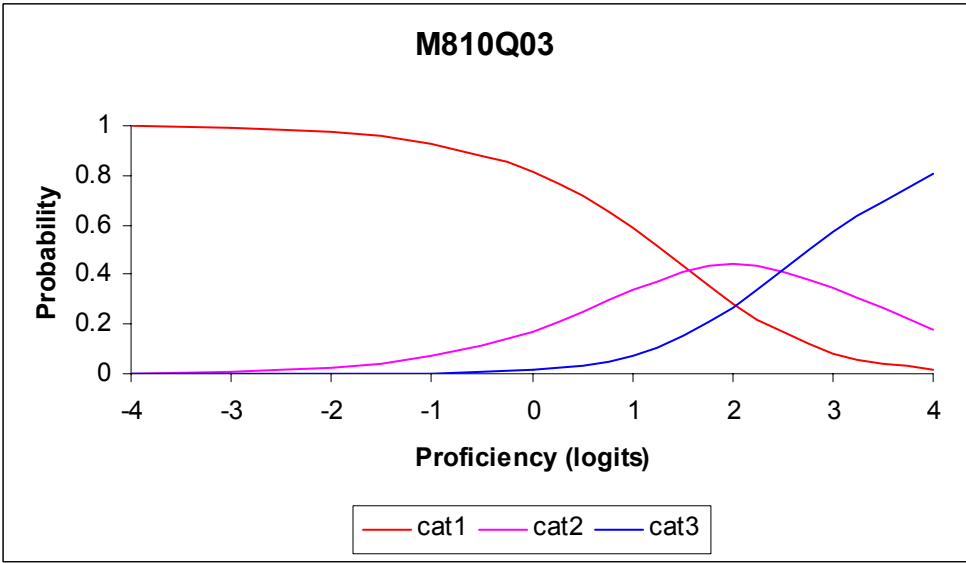
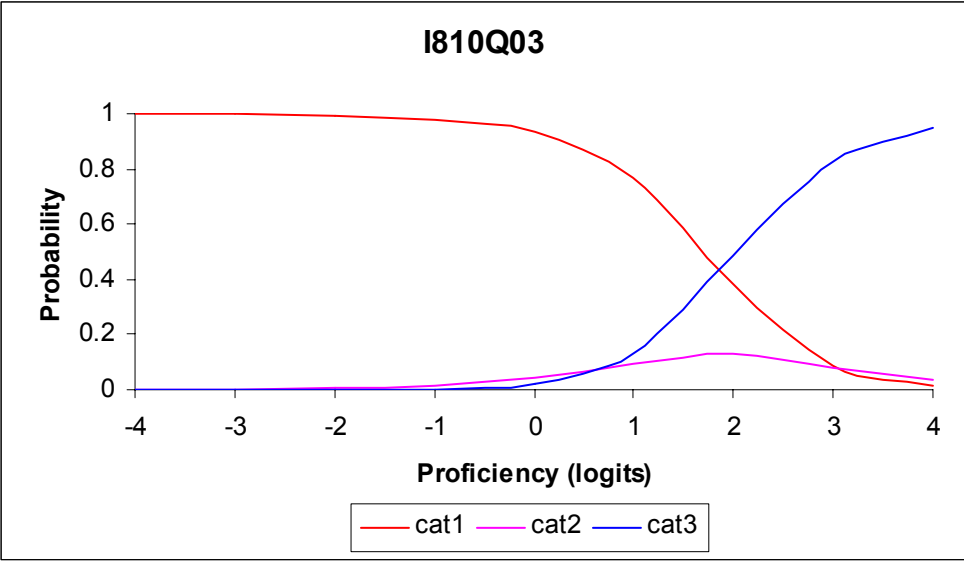
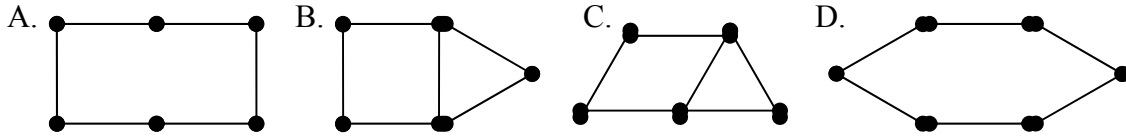


Figure 2 Category probability curves for an item for the imperial and metric conditions

Appendix 1
Item modification

1. Peter used 6 matches to form the following shapes. Each match is 5cm (**inches**) long.



Circle either “Correct” or “Incorrect” for each statement in the table.

A has the largest area	Correct / Incorrect
B and C has the same area	Correct / Incorrect
A and D has the same area	Correct / Incorrect
C has the smallest area	Correct / Incorrect

2. Jasmine’s home has a distance of 3 kilometers (**2 miles**) from the nearest bus station.

It usually takes Jasmine 10 minutes to ride a bicycle to the bus station from her home.

Question 1: what is Jasmine’s riding speed? Given your answer in kilometers per hour, **kph (miles per hour, mph)**.

Question 2: One day, Jasmine’s bicycle broke down and her father drove her to the bus station. The average speed of the car is 60 kph (**40 mph**). How many minutes can she save compared to her routine?

Appendix 2
Item parameter estimates, standard errors, and fit statistics

Item level Parameters		Weighted Fit				
Item ID	Estimate	Error	MNSQ	CI	T	
1	M033Q01	-1.170	0.060	1.05	(0.94, 1.06)	1.5
2	M467Q01	0.044	0.053	0.92	(0.96, 1.04)	-4.2
3	M810Q01	-1.095	0.063	1.02	(0.94, 1.06)	0.6
4	M810Q02	-0.784	0.059	1.00	(0.95, 1.05)	0.0
5	M810Q03	2.034	0.052	0.91	(0.91, 1.09)	-2.1
6	M833Q01	1.427	0.061	1.05	(0.94, 1.06)	1.8
7	M402Q01	0.352	0.054	1.07	(0.96, 1.04)	3.3
8	M402Q02	1.211	0.060	0.99	(0.95, 1.05)	-0.2
9	M179Q01	0.960	0.040	0.99	(0.94, 1.06)	-0.4
10	M464Q01	2.028	0.072	0.83	(0.92, 1.08)	-4.4
11	M564Q01	0.401	0.054	1.07	(0.96, 1.04)	3.3
12	M564Q02	0.612	0.055	1.07	(0.96, 1.04)	2.9
13	M145Q01	-0.815	0.045	0.86	(0.96, 1.04)	-7.9
14	M408Q01	0.732	0.044	0.96	(0.96, 1.04)	-2.2
15	M520Q01	-0.583	0.027	1.08	(0.95, 1.05)	3.1
16	M520Q02	-0.029	0.042	0.88	(0.97, 1.03)	-8.2
17	M520Q03	0.072	0.042	1.02	(0.97, 1.03)	1.0
18	M446Q01	-0.404	0.043	1.00	(0.97, 1.03)	-0.2
19	M446Q02	3.295	0.089	0.92	(0.86, 1.14)	-1.1
20	M192Q01	0.725	0.026	1.12	(0.95, 1.05)	4.7
21	M702Q01	0.382	0.028	1.17	(0.95, 1.05)	6.2
22	M034Q01	1.017	0.046	0.95	(0.96, 1.04)	-2.3
23	M423Q01	-1.400	0.048	1.03	(0.95, 1.05)	1.0
24	M555Q02	0.079	0.042	0.95	(0.97, 1.03)	-3.2
25	M305Q01	-0.086	0.053	1.18	(0.96, 1.04)	8.8
26	M510Q01	0.244	0.055	1.05	(0.96, 1.04)	2.6
27	M474Q01	-0.783	0.056	1.07	(0.95, 1.05)	3.0
28	M124Q01	-0.280	0.043	0.93	(0.94, 1.06)	-2.2
29	M124Q03	1.626	0.043	0.90	(0.92, 1.08)	-2.6
30	M434Q01	1.450	0.063	1.11	(0.94, 1.06)	3.2
31	M505Q01	1.649	0.064	1.04	(0.93, 1.07)	1.2
32	M462Q01	0.783	0.054	1.19	(0.93, 1.07)	4.7
33	M438Q01	0.441	0.054	1.23	(0.96, 1.04)	10.6
34	M438Q02	0.379	0.054	0.97	(0.96, 1.04)	-1.4
35	M547Q01	-1.616	0.072	0.96	(0.92, 1.08)	-1.0
36	M806Q01	-0.523	0.055	0.96	(0.96, 1.04)	-1.7
37	M800Q01	-2.115	0.073	1.09	(0.91, 1.09)	2.0
38	M421Q01	-1.008	0.062	0.85	(0.94, 1.06)	-5.5
39	M421Q03	1.130	0.060	1.03	(0.95, 1.05)	1.1
40	M421Q02	2.100	0.073	0.95	(0.91, 1.09)	-1.1
41	M704Q01	-1.759	0.071	0.96	(0.92, 1.08)	-0.9
42	M704Q02	1.636	0.068	0.87	(0.93, 1.07)	-3.7
43	M571Q01	0.536	0.056	0.99	(0.96, 1.04)	-0.6
44	M559Q01	-0.185	0.054	1.01	(0.96, 1.04)	0.7
45	M144Q01	-0.018	0.055	1.02	(0.96, 1.04)	1.0
46	M144Q02	1.660	0.066	1.03	(0.93, 1.07)	0.9
47	M144Q03	-1.391	0.062	0.93	(0.94, 1.06)	-2.3
48	M144Q04	0.857	0.060	0.96	(0.95, 1.05)	-1.7

49	M413Q01	-0.294	0.054	1.03	(0.96, 1.04)	1.3
50	M413Q02	-1.125	0.060	0.84	(0.94, 1.06)	-5.7
51	M413Q03	0.427	0.058	0.90	(0.96, 1.04)	-4.5
52	M406Q01	2.004	0.074	0.87	(0.91, 1.09)	-3.0
53	M406Q02	2.152	0.085	0.81	(0.90, 1.10)	-3.8
54	M406Q03	2.044	0.080	0.91	(0.91, 1.09)	-2.0
55	M150Q01	-0.335	0.055	0.91	(0.96, 1.04)	-4.3
56	M150Q03	-0.306	0.056	0.95	(0.96, 1.04)	-2.2
57	M150Q02	-0.770	0.042	1.14	(0.94, 1.06)	4.4
58	M598Q01	-0.832	0.058	1.15	(0.95, 1.05)	5.6
59	M710Q01	0.966	0.058	1.04	(0.95, 1.05)	1.4
60	M411Q01	0.013	0.054	0.91	(0.96, 1.04)	-4.8
61	M411Q02	0.053	0.054	1.01	(0.96, 1.04)	0.5
62	M496Q01	0.069	0.042	0.92	(0.97, 1.03)	-5.2
63	M496Q02	-0.577	0.044	1.03	(0.96, 1.04)	1.5
64	M484Q01	-0.482	0.044	0.84	(0.97, 1.03)	-9.5
65	M155Q02	-0.847	0.030	1.06	(0.95, 1.05)	2.2
66	M155Q01	-1.019	0.047	0.95	(0.96, 1.04)	-2.1
67	M155Q03	1.658	0.039	1.06	(0.93, 1.07)	1.5
68	M155Q04	-1.081	0.032	1.12	(0.95, 1.05)	4.3
69	M442Q02	0.523	0.045	0.86	(0.97, 1.03)	-8.2
70	M509Q01	-0.238	0.043	0.99	(0.97, 1.03)	-0.6
71	M420Q01	-0.486	0.043	0.93	(0.97, 1.03)	-4.2
72	M468Q01	-0.481	0.046	1.05	(0.96, 1.04)	2.7
73	M447Q01	-0.775	0.044	0.99	(0.96, 1.04)	-0.7
74	M302Q01	-3.921	0.147	0.99	(0.74, 1.26)	-0.0
75	M302Q02	-1.802	0.068	0.98	(0.92, 1.08)	-0.5
76	M302Q03	1.052	0.060	0.90	(0.95, 1.05)	-3.8
77	M603Q01	0.402	0.054	1.05	(0.96, 1.04)	2.3
78	M603Q02	0.497	0.081	0.92	(0.94, 1.06)	-2.9
79	M266Q01	0.967	0.040	1.27	(0.94, 1.06)	7.9
80	M513Q01	0.268	0.058	1.02	(0.96, 1.04)	1.1
81	M828Q01	0.488	0.061	0.96	(0.96, 1.04)	-1.7
82	M828Q02	-0.255	0.056	1.07	(0.96, 1.04)	3.3
83	M828Q03	0.991	0.063	0.97	(0.95, 1.05)	-1.2
84	M803Q01	1.237	0.065	0.88	(0.94, 1.06)	-4.1
85	M273Q01	0.117	0.054	1.09	(0.96, 1.04)	4.6
86	I302Q01	-4.127	0.199	0.99	(0.65, 1.35)	-0.0
87	I302Q02	-1.790	0.085	1.06	(0.90, 1.10)	1.3
88	I302Q03	1.206	0.077	0.86	(0.93, 1.07)	-3.9
89	I266Q01	0.812	0.049	1.26	(0.92, 1.08)	6.2
90	I464Q01	2.167	0.098	0.83	(0.88, 1.12)	-2.8
91	I810Q01	-1.118	0.078	0.99	(0.93, 1.07)	-0.3
92	I810Q02	-0.989	0.076	0.99	(0.93, 1.07)	-0.2
93	I810Q03	1.882	0.068	1.07	(0.85, 1.15)	0.9
94	I421Q01	-1.033	0.079	0.84	(0.93, 1.07)	-4.6
95	I421Q03	0.959	0.075	1.06	(0.93, 1.07)	1.9
96	I421Q02	2.189	0.095	0.95	(0.88, 1.12)	-0.8
97	I547Q01	-1.624	0.090	1.07	(0.90, 1.10)	1.3
98	I124Q01	-0.315	0.056	0.91	(0.92, 1.08)	-2.3
99	I124Q03	1.376	0.051	0.96	(0.90, 1.10)	-0.7
100	I305Q01	-0.132	0.068	1.14	(0.95, 1.05)	5.4
101	I462Q01	0.979	0.069	1.16	(0.91, 1.09)	3.2
102	I406Q01	1.945	0.093	0.87	(0.89, 1.11)	-2.5
103	I406Q02	2.352	0.113	0.84	(0.85, 1.15)	-2.3
104	I406Q03	2.341	0.109	0.93	(0.86, 1.14)	-1.0
105	I150Q01	-0.214	0.069	0.88	(0.95, 1.05)	-4.7

106	I150Q03	0.394	0.071	0.90	(0.95, 1.05)	-3.9
107	I150Q02	-0.848	0.053	1.07	(0.92, 1.08)	1.7
108	I474Q01	-0.798	0.071	1.14	(0.94, 1.06)	4.6
109	I273Q01	0.174	0.068	1.00	(0.95, 1.05)	0.1
110	I828Q01	0.755	0.077	0.96	(0.94, 1.06)	-1.2
111	I828Q02	-0.130	0.070	1.08	(0.95, 1.05)	3.0
112	I828Q03	0.879	0.077	1.01	(0.94, 1.06)	0.3

Step level			Weighted Fit				
Parameters							
Item	Step	Estimate	Error	MNSQ	CI	T	
5	M810Q03	0		0.88	(0.94, 1.06)	-3.7	
5	M810Q03	1	-0.479	0.067	0.94	(0.93, 1.07)	-1.9
5	M810Q03	2	0.479*	1.03	(0.84, 1.16)	0.4	
9	M179Q01	0		0.92	(0.95, 1.05)	-3.7	
9	M179Q01	1	-0.597	0.054	0.95	(0.97, 1.03)	-2.8
9	M179Q01	2	0.597*	1.07	(0.92, 1.08)	1.7	
15	M520Q01	0		1.06	(0.95, 1.05)	2.2	
15	M520Q01	1	1.108	0.065	1.01	(0.90, 1.10)	0.2
15	M520Q01	2	-1.108*	1.08	(0.95, 1.05)	3.5	
20	M192Q01	0		1.07	(0.96, 1.04)	3.0	
20	M192Q01	1	-1.314	0.042	1.04	(0.98, 1.02)	3.2
20	M192Q01	2	0.024	0.048	1.03	(0.96, 1.04)	1.5
20	M192Q01	3	1.290*	1.08	(0.90, 1.10)	1.6	
21	M702Q01	0		1.18	(0.95, 1.05)	7.0	
21	M702Q01	1	1.425	0.077	1.01	(0.88, 1.12)	0.1
21	M702Q01	2	-1.425*	1.12	(0.95, 1.05)	4.5	
28	M124Q01	0		0.99	(0.92, 1.08)	-0.3	
28	M124Q01	1	-1.201	0.052	0.97	(0.98, 1.02)	-2.9
28	M124Q01	2	1.201*	0.92	(0.95, 1.05)	-3.0	
29	M124Q03	0		0.92	(0.95, 1.05)	-3.4	
29	M124Q03	1	-1.592	0.059	0.96	(0.97, 1.03)	-2.7
29	M124Q03	2	0.728	0.101	0.97	(0.86, 1.14)	-0.3
29	M124Q03	3	0.864*	0.96	(0.74, 1.26)	-0.3	
32	M462Q01	0		1.22	(0.93, 1.07)	6.0	
32	M462Q01	1	-2.088	0.058	1.12	(0.95, 1.05)	4.6
32	M462Q01	2	2.088*	1.00	(0.85, 1.15)	-0.0	
57	M150Q02	0		1.09	(0.91, 1.09)	1.9	
57	M150Q02	1	-1.055	0.051	0.99	(0.98, 1.02)	-0.5
57	M150Q02	2	1.055*	1.08	(0.96, 1.04)	3.3	
65	M155Q02	0		1.07	(0.94, 1.06)	2.4	
65	M155Q02	1	0.373	0.052	1.00	(0.94, 1.06)	-0.1
65	M155Q02	2	-0.373*	1.02	(0.96, 1.04)	1.0	
67	M155Q03	0		0.94	(0.94, 1.06)	-2.0	
67	M155Q03	1	0.065	0.059	0.93	(0.93, 1.07)	-1.9
67	M155Q03	2	-0.065*	1.18	(0.90, 1.10)	3.4	
68	M155Q04	0		1.03	(0.93, 1.07)	0.9	
68	M155Q04	1	-0.378	0.044	1.06	(0.96, 1.04)	3.0
68	M155Q04	2	0.378*	1.13	(0.96, 1.04)	6.4	
79	M266Q01	0		1.24	(0.96, 1.04)	9.9	
79	M266Q01	1	-0.472	0.054	1.06	(0.96, 1.04)	2.9
79	M266Q01	2	0.472*	1.11	(0.91, 1.09)	2.4	
89	I266Q01	0		1.23	(0.94, 1.06)	7.6	
89	I266Q01	1	-0.454	0.067	1.06	(0.95, 1.05)	2.3
89	I266Q01	2	0.454*	1.13	(0.90, 1.10)	2.3	

93	I810Q03	0			1.07	(0.87, 1.13)	1.1
93	I810Q03	1	1.190	0.149	1.02	(0.75, 1.25)	0.2
93	I810Q03	2	-1.190*		1.03	(0.82, 1.18)	0.4
98	I124Q01	0			0.98	(0.89, 1.11)	-0.4
98	I124Q01	1	-1.294	0.067	0.95	(0.97, 1.03)	-3.4
98	I124Q01	2	1.294*		0.89	(0.93, 1.07)	-3.2
99	I124Q03	0			0.90	(0.94, 1.06)	-3.3
99	I124Q03	1	-1.190	0.074	0.96	(0.96, 1.04)	-2.0
99	I124Q03	2	0.743	0.124	0.99	(0.83, 1.17)	-0.0
99	I124Q03	3	0.447*		1.09	(0.77, 1.23)	0.8
101	I462Q01	0			1.20	(0.92, 1.08)	4.7
101	I462Q01	1	-2.170	0.074	1.13	(0.94, 1.06)	3.7
101	I462Q01	2	2.170*		0.95	(0.77, 1.23)	-0.4
107	I150Q02	0			1.05	(0.87, 1.13)	0.7
107	I150Q02	1	-0.939	0.066	0.98	(0.97, 1.03)	-1.0
107	I150Q02	2	0.939*		1.03	(0.95, 1.05)	1.1

An asterisk next to a parameter estimate indicates that the parameter is constrained.