P23

DIFFERENCES BETWEEN MULTIPLE-CHOICE AND CONSTRUCTED RESPONSE ITEMS IN PIRLS 2001

Dirk Hastedt IEA Data Processing Center, Hamburg, Germany

Keywords:
PIRLS, Item types, scaling

Abstract
In international large-scale surveys, constructed response (CR) items are increasingly being used and multiple-choice (MC) items are being used less frequently. In the 1991 IEA Reading Literacy Study, only 6 out of 108 items were CR items. In PIRLS 2001, there were 46 MC and 52 CR items. This change makes it worthwhile evaluating the differences between multiple-choice and constructed response items in terms of the scaling

Introduction

In this chapter the differences between multiple-choice and constructed response items used in PIRLS 2001 have been evaluated. Earlier research work lead to "…the conclusion that item format exerts no significant influence on the outcome in large-scale surveys of reading, where a major purpose is to establish levels of comprehension in comparable groups of students." (Elley and Mangubhai 1992). Since there are major differences in terms of costs when administering constructed response items compared to multiple-choice items, it is important to ask if multiple-choice items and constructed response items a) measure the same thing and, b) if constructed response items provide additional information. First the two groups of items have been compared by looking at item characteristics based on different item response models: 1) percent correct, 2) RASCH item parameters and 3) IRT item difficulties. In a second step the influence on student scores based on the different item types has been evaluated. Student scores were calculated based on multiple-choice items and constructed response items separately. Then differences between countries and between boys and girls on the scores on the different item types were compared.

The items

The easiest way of comparing item difficulties is to examine the percentage correct for each item respectively for different groups of items. When looking at the percentage correct for the PIRLS items, the different missing codes used in PIRLS must be considered. In that study a distinction was made between items that were

- not administered to students (because of the rotated test design),
- invalid missing codes (because students checked 2 options of a multiple-choice item, or similar problems),
- omit codes (because students did not answer an administered item)
- not reached codes (because student were not able to finish their test booklet in the time given)

In PIRLS 'invalid' and 'omit' codes were considered as wrong responses whereas 'not administered' and 'not reached' codes were removed when calculating item parameters. Removing 'not reached' codes takes away the position effect of the items. This was necessary because items appeared in different positions in the different test booklets. When calculating student abilities, the 'not reached' codes were treated as wrong because the time given to finish the test booklets was very generous and students who did not finish the test within the given time should be penalized. For details about the treatment of missing codes in PIRLS, the reader is referred to Chapters 8 and 11 in the PIRLS Technical Report (Martin and Mullis 2003). Furthermore, for some constructed response items, up to 2 score points could be given. This means that students awarded 2 points should be regarded as having achieved a full credit, whereas 1 score point means partially correct. This partial credit approach needs to be taken into account when calculating the percent of score points gained by a student. When calculating the percentage correct for all items and averaging these up for all multiple-choice items and constructed response items respectively, the results were an average correct for constructed response items of 38.3 percent and for multiple-choice of 49.1 percent.

Consequently, when defining the difficulty by the percentage of score points gained by the student, constructed response items were harder than the multiple-choice items.

An easier way of looking at item difficulties is to calculate RASCH item difficulties (Baker 1992). Several programs are available to calculate item parameters based on the RASCH model. For this article, the ConQuest software (Wu 1998) was used. For the calculation of the item parameters, all 151,761 students from all countries participating in PIRLS were read into the ConQuest software. Since there were items that were scored polytomously, the partial credit model (Wright and Masters 1982) was used for calculating the item parameters. The item difficulties were specified to add up to zero. When averaging the RASCH item difficulties for all multiple-choice items the mean was -0.41. The mean of the RASCH item difficulties of the constructed response items was 0.36. Thus, when applying the RASCH model, the constructed response items were also harder than the multiple-choice items.

For scaling the PIRLS 2001 data, a 3-parameter IRT (Baker 1992) model was applied. This model – unlike the Rasch model – also takes different item discrimination and students' guessing into account. For obvious reasons, guessing only had to be considered for the multiple-choice items.
In the PIRLS 2001 technical report (Martin, Mullis and Kennedy 2003), the item parameters calculated with PARSCALE^{TM} (Muraki and Bock 1997) have been published. The sum of the item difficulties was also fixed to be zero. When calculating the average item difficulty for the multiple-choice items, a value of –0.18 was obtained and for the constructed response items it was 0.01. Thus, even after taking guessing into account the constructed response items were still harder than the multiple-choice items, but the difference was less than with the Rasch or classical percentage correct approaches.

In Table 1 the percentage correct, the RASCH item difficulties and the IRT item difficulties for each of the items have been presented.

Table 1: Percentage correct, Rasch item difficulties and three parameter difficulties for each item

CR items	percent correct	ConQuest results	betas from international technical report	MC items	percent correct	ConQuest results	betas from international technical report
R011A01C	60.0	-1.26	-1.23	R011A02M	44.5	-0.02	0.26
R011A03C	53.6	-0.70	-0.86	R011A05M	63.2	-1.57	-0.96
R011A04C	38.3	0.42	0.00	R011A06M	64.4	-1.70	-1.11
R011A07C	21.7	-0.16	-0.37	R011A10M	42.1	0.19	0.14
R011A08C	51.0	-0.40	-0.68	R011L01M	66.3	-1.86	-2.50
R011A09C	37.4	0.51	0.09	R011L02M	35.4	0.61	0.81
R011A11C	35.5	0.65	0.15	R011L05M	33.3	0.76	0.75
R011L03C	43.5	0.06	-0.29	R011L07M	34.4	0.70	0.66
R011L04C	30.6	1.01	0.52	R011L09M	59.2	-1.14	-0.74
R011L06C	33.6	0.70	0.25	R011L11M	48.4	-0.26	-0.01
R011L08C	25.7	1.35	0.72	R011N01M	53.5	-0.72	-0.28
R011L10C	23.8	1.36	0.77	R011N02M	43.7	0.01	0.31
R011L12C	25.6	1.40	0.76	R011N03M	58.6	-1.15	-0.64
R011N07C	26.8	1.14	0.64	R011N04M	34.8	0.62	0.45
R011N08C	34.1	0.69	0.24	R011N05M	42.3	0.12	0.24
R011N10C	18.4	1.86	1.05	R011N06M	26.9	1.18	0.88
R011N12C	28.1	1.02	0.49	R011N09M	53.6	-0.71	-0.42
R011N13C	37.4	0.51	0.10	R011N11M	38.1	0.41	0.45
R011R04C	56.8	-0.94	-1.00	R011R01M	43.5	0.06	0.10
R011R05C	52.6	-0.59	-0.59	R011R02M	35.4	0.62	0.62
R011R06C	43.1	0.19	-0.14	R011R03M	62.8	-1.48	-1.10
R011R07C	35.2	0.66	0.20	R011C04M	36.2	0.58	0.46
R011R08C	32.5	0.85	0.35	R011C05M	51.7	-0.54	0.01
R011R09C	39.9	0.39	-0.03	R011C07M	54.1	-0.73	-0.28
R011R10C	36.6	0.77	0.37	R011C09M	31.7	0.90	0.69
R011R11C	29.5	0.43	0.01	R011C12M	41.8	0.21	0.43
R011C01C	42.6	0.12	-0.08	R011C13M	40.0	0.34	0.48
R011C02C	31.0	0.93	0.40	R011F01M	54.7	-0.77	-0.34
R011C03C	49.7	-0.39	-0.39	R011F02M	53.1	-0.64	-0.57
R011C06C	40.8	0.26	-0.03	R011F03M	53.9	-0.70	-0.46

R011C08C	32.3	0.81	0.35	R011F04M	60.7	-1.30	-0.71
R011C10C	20.9	0.90	0.39	R011F05M	51.0	-0.47	-0.18
R011C11C	34.0	0.77	0.27	R011F11M	39.5	0.38	0.25
R011F06C	43.9	0.05	-0.24	R011F13M	44.0	0.09	0.21
R011F07C	28.7	1.01	0.53	R011H01M	65.0	-1.80	-1.45
R011F08C	42.2	0.18	-0.10	R011H02M	64.7	-1.76	-1.43
R011F09C	51.0	-0.33	-0.46	R011H05M	61.2	-1.39	-1.01
R011F10C	59.6	-1.17	-1.08	R011H06M	50.7	-0.50	-0.60
R011F12C	23.4	1.39	0.76	R011H11M	54.6	-0.78	-0.44
R011H03C	28.5	1.06	0.87	R011M01M	59.1	-1.21	-0.51
R011H04C	57.9	-1.08	-1.11	R011M02M	65.4	-1.84	-1.10
R011H07C	47.9	-0.23	-0.55	R011M03M	40.4	0.26	0.33
R011H08C	39.4	0.32	0.01	R011M05M	55.9	-0.91	-0.41
R011H09C	51.9	-0.45	-0.68	R011M08M	35.7	0.61	0.85
R011H10C	31.5	0.96	0.48	R011M09M	55.1	-0.84	-0.53
R011M04C	24.0	1.45	0.79	R011M13M	53.1	-0.64	-0.05
R011M06C	47.4	-0.18	-0.34				
R011M07C	51.8	-0.57	-0.55				
R011M10C	64.3	-1.70	-1.35				
R011M11C	27.7	1.20	0.58				
R011M12C	24.5	1.36	0.73				
R011M14C	42.1	0.19	-0.14				
Mean	**38.3**	**0.36**	**0.01**		**49.1**	**-0.41**	**-0.18**

From Table 1 it can be seen that the multiple-choice items appeared to be easier when guessing was not taken into account, which is the case if percentage correct or RASCH item difficulties were examined. When guessing was taken into account, the difference between the item difficulties became marginal.

The student scores

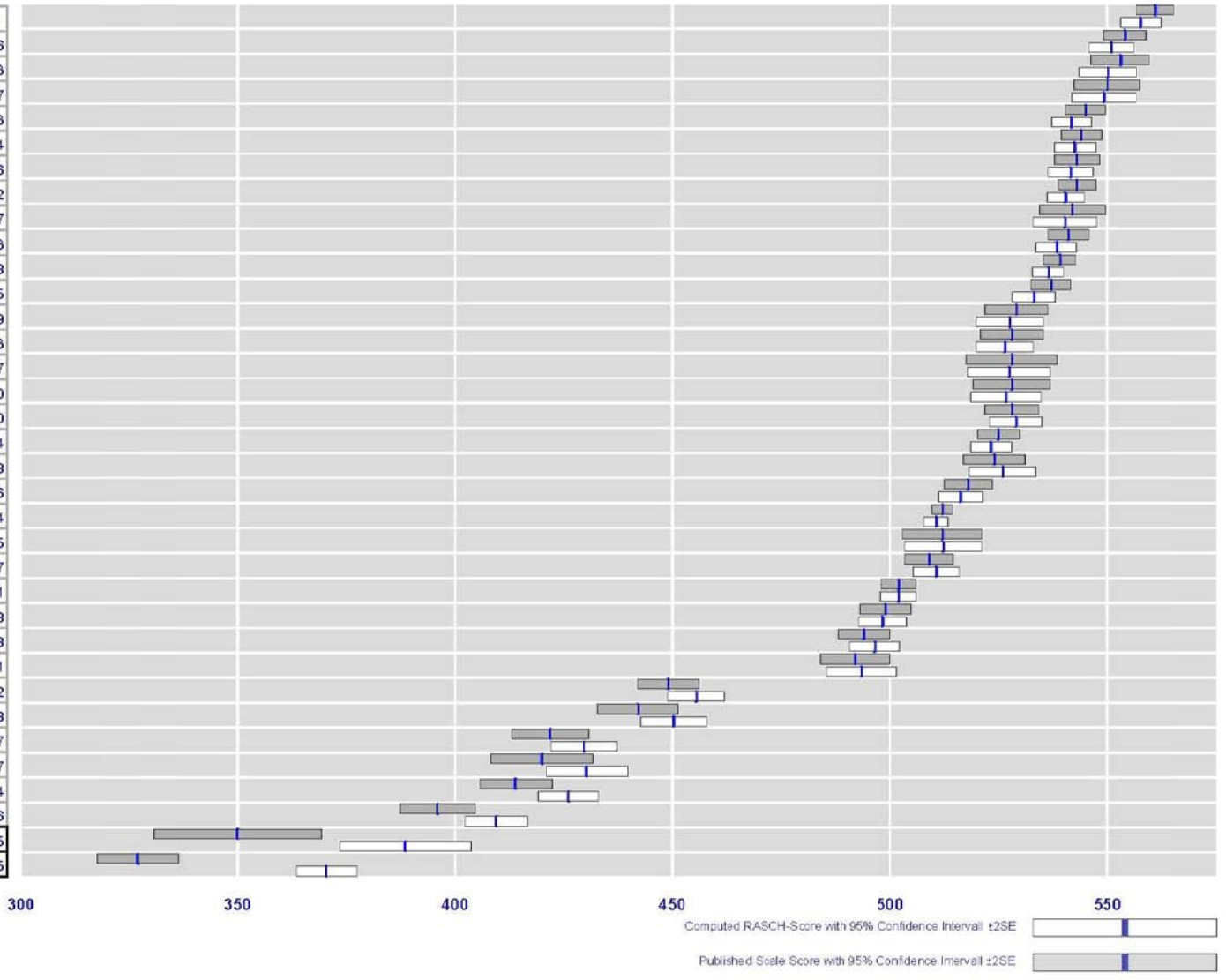
In this section the influence of the item type on the student scores for different countries and for different groups of students has been examined.

In PIRLS plausible value scores (Rubin 1987) have been calculated as it was described in chapter 11 of the PIRLS 2001 technical report (Martin, Mullis and Kennedy 2003). This multiple imputation methodology had to be used to derive reliable estimates of student performance in a study like PIRLS where a matrix sample design has been applied and students have been administered only a subset of the survey items.
Since no program was available to calculate plausible value student scores based on a three-parameter model, scores based on the RASCH model were calculated using the CONQUEST software. To ensure that the use of the RASCH scores had no significant influence on the subsequent results, plausible values were calculated based on all items together without any distinction between the item types. For this purpose, the item parameters calculated with the CONQUEST software before were used and plausible values were calculated by anchoring the item parameters onto the previously calculated item parameters. In contrast to the international PIRLS scaling process, no background variables were used for the conditioning model which caused that the reliability of the scores is reduced.. The calculated plausible values were standardized to an international mean of 500 with standard deviation of 100, similar to the international PIRLS scores.

These plausible values based on the RASCH model were then compared to the international scores reported in the international PIRLS report (see Exhibit 1.1 in the PIRLS 2001 international report). For this purpose, weighted means for all countries and their standard errors were calculated using the Jackknifing procedure (for details about the Jackknifing procedure, please refer to Chapter 12 in the PIRLS 2001 Technical Report).
The results have been presented in Figure 1.

Compare International IRT-Scores to International RASCH-Scores

Country	Published Scale-Score	Computed RASCH-Score
SWEDEN	561 / 2,2	558 / 2,3
NETHERLANDS	554 / 2,5	551 / 2,6
ENGLAND	553 / 3,4	550 / 3,3
BULGARIA	550 / 3,8	549 / 3,7
LATVIA	545 / 2,3	542 / 2,3
CANADA	544 / 2,4	543 / 2,4
LITHUANIA	543 / 2,6	542 / 2,6
HUNGARY	543 / 2,2	540 / 2,2
UNITED STATES	542 / 3,8	540 / 3,7
ITALY	541 / 2,4	538 / 2,3
GERMANY	539 / 1,9	536 / 1,8
CZECH REP.	537 / 2,3	533 / 2,5
NEW ZEALAND	529 / 3,6	528 / 3,9
SCOTLAND	528 / 3,6	526 / 3,3
SINGAPORE	528 / 5,2	527 / 4,7
RUSSIAN FED.	528 / 4,4	527 / 4,0
HONG KONG	528 / 3,1	529 / 3,0
FRANCE	525 / 2,4	523 / 2,4
GREECE	524 / 3,5	526 / 3,8
SLOVAK REP.	518 / 2,8	516 / 2,6
ICELAND	512 / 1,2	511 / 1,4
ROMANIA	512 / 4,6	512 / 4,5
ISRAEL	509 / 2,8	511 / 2,7
SLOVENIA	502 / 2,0	502 / 2,1
NORWAY	499 / 2,9	498 / 2,8
CYPRUS	494 / 3,0	497 / 2,8
MOLDOVA, REP.OF	492 / 4,0	493 / 4,1
TURKEY	449 / 3,5	455 / 3,2
MACEDONIA, REP.OF	442 / 4,6	450 / 3,8
COLOMBIA	422 / 4,4	430 / 3,7
ARGENTINA	420 / 5,9	430 / 4,7
IRAN, ISLAMIC REP.	414 / 4,2	426 / 3,4
KUWAIT	396 / 4,3	409 / 3,6
MOROCCO	350 / 9,6	389 / 7,5
BELIZE	327 / 4,7	371 / 3,5



The important result is that the means of the calculated RASCH reading scores did not differ significantly from the means of the international scores for all countries except for the two lowest scoring countries, Morocco and Belize.

There might be different reasons why there were differences in the scores for these 2 countries, but the most obvious one might be that these two countries were not only the lowest scoring countries, but also the countries with most students below the international 10th percentile. In Belize 50.1 percent of the students had an international RASCH score below the international 10th percentile of 365. In Morocco there were 42.0 percent of students scoring below the international 10th percentile. Since there were very few items that were easy enough to measure the students' ability in that range accurately, it was no surprise that the scores differ depending on the scaling model used. (The percentages of students scoring below the 10th international percentile have been given in Table 2.

Table 2: Percentages of students below the international 10th percentile

Country	Percentage of students below the international 10th percentile
BELIZE	50.1
MOROCCO	42.0
KUWAIT	31.0
ARGENTINA	24.0
IRAN, ISLAMIC REP.	23.3
COLOMBIA	21.8
MACEDONIA, REP. OF	20.5
TURKEY	14.5
ISRAEL	7.7
NORWAY	7.5
CYPRUS	7.5
MOLDOVA, REP. OF	6.9
ROMANIA	6.9
SINGAPORE	6.7
NEW ZEALAND	6.6
SLOVENIA	5.4
ICELAND	5.3
SCOTLAND	5.2
SWEDEN GR. 3	4.5
ENGLAND	3.7
SLOVAK REP.	3.6
UNITED STATES	3.6
BULGARIA	3.5
FRANCE	3.3
GREECE	3.0
RUSSIAN FEDERATION	2.6
HONG KONG	2.6
GERMANY	2.4
CANADA	2.4
ITALY	1.9
LITHUANIA	1.8
CZECH REP.	1.8
HUNGARY	1.7
SWEDEN	1.4
LATVIA	1.2
NETHERLANDS	0.7

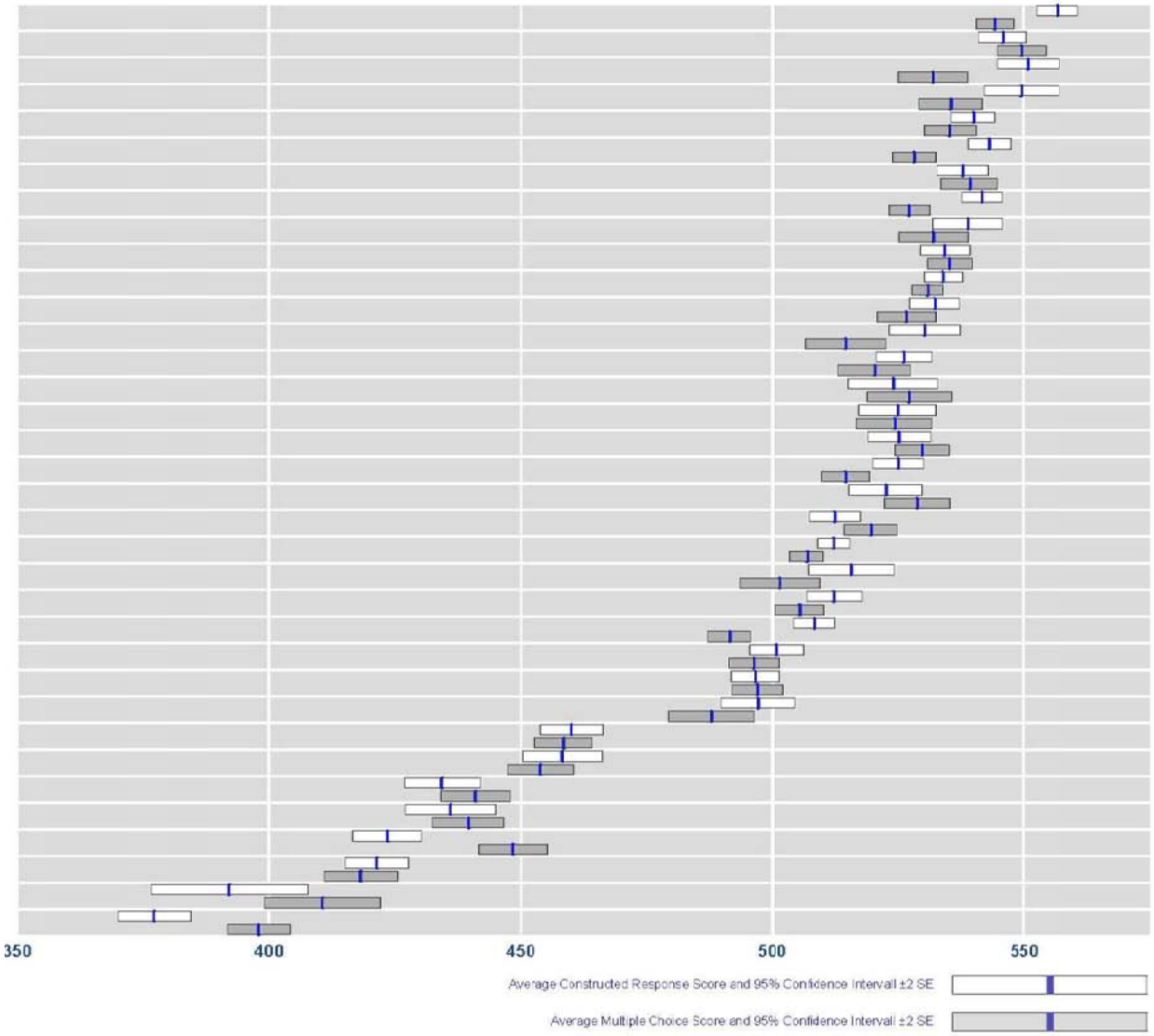
In a second step, two sets of plausible values were calculated with the CONQUEST software – one based on the multiple-choice items only, and the second based on the constructed response items only. For both sets of

plausible values, the same procedure as described before was used, but for each set only the multiple-choice items or the constructed response items were used for the calculation of the item parameters, as well as for the calculation of the students' plausible values. It must noted that the reliability – especially of the multiple choice scores – was lower than for the international scores.

In Figure 2 the national mean scores based on the constructed response items only (CR scores) and their standard errors, and national mean scores based on the multiple-choice items only (MC scores) and their standard error have been presented.

Compare Constructed Response Scores to Multiple Choice Scores

Country	CR Scores	MC Scores
SWEDEN	557 / 2,0	544 / 1,9
NETHERLANDS	546 / 2,4	550 / 2,4
ENGLAND	551 / 3,0	532 / 3,4
BULGARIA	550 / 3,7	535 / 3,2
LATVIA	540 / 2,2	535 / 2,5
CANADA	543 / 2,2	528 / 2,2
LITHUANIA	538 / 2,6	539 / 2,8
HUNGARY	542 / 2,0	527 / 2,1
UNITED STATES	539 / 3,5	532 / 3,4
ITALY	534 / 2,4	535 / 2,2
GERMANY	534 / 1,9	531 / 1,6
CZECH REP.	532 / 2,5	527 / 2,9
NEW ZEALAND	530 / 3,6	514 / 4,0
SCOTLAND	526 / 2,7	520 / 3,6
SINGAPORE	524 / 4,5	527 / 4,2
RUSSIAN FED.	525 / 3,8	524 / 3,8
HONG KONG	525 / 3,1	530 / 2,8
FRANCE	525 / 2,6	515 / 2,3
GREECE	522 / 3,6	529 / 3,3
SLOVAK REP.	512 / 2,6	520 / 2,6
ICELAND	512 / 1,6	507 / 1,7
ROMANIA	516 / 4,3	502 / 4,0
ISRAEL	512 / 2,7	505 / 2,4
SLOVENIA	508 / 2,0	491 / 2,1
NORWAY	501 / 2,7	496 / 2,5
CYPRUS	497 / 2,4	497 / 2,6
MOLDOVA, REP.OF	497 / 3,7	488 / 4,2
TURKEY	460 / 3,1	458 / 2,9
MACEDONIA, REP.OF	458 / 4,0	454 / 3,3
COLOMBIA	434 / 3,8	441 / 3,4
ARGENTINA	436 / 4,5	440 / 3,5
IRAN, ISLAMIC REP.	424 / 3,4	449 / 3,4
KUWAIT	421 / 3,1	418 / 3,7
MOROCCO	392 / 7,8	411 / 5,8
BELIZE	377 / 3,7	398 / 3,1



As can be seen from Figure 2, there were at least 10 countries where the two scores differed significantly: Sweden, England, Bulgaria, Canada, Hungary, New Zealand, France, Slovenia, Iran and Belize. The results of these countries have been framed.
For 8 of these 10 countries the constructed response item scores were above the international average of 500. For these 8 countries the multiple-choice score was lower than the constructed response score. The remaining 2 countries with significant different scores had a score below the international average and, furthermore, the multiple-choice score was higher than the constructed response score.

No cultural or other factors could be found between the countries with significantly higher constructed response scores. The 8 countries did not differ compared to the other countries in terms of more students reaching the top of the multiple-choice item scale which could result in a ceiling effect and consequently lower scores on that scale. No clusters of countries with similar cultures could be found. For example, Canada and the US tended to behave quite similarly in the analysis but the difference for Canada was twice as high as for the US.
A more detailed analysis of these differences would be desirable. For example, it would seem to be desirable to calculate 3-parameter IRT models for each country individually and then examining the item parameters might cast some light on different behaviors of student guessing. Nevertheless, from these first results it can be seen that there were some countries where it mattered if a test consisted of multiple-choice items or of constructed response items. The differences found with the PIRLS scores were up to 19 score points.

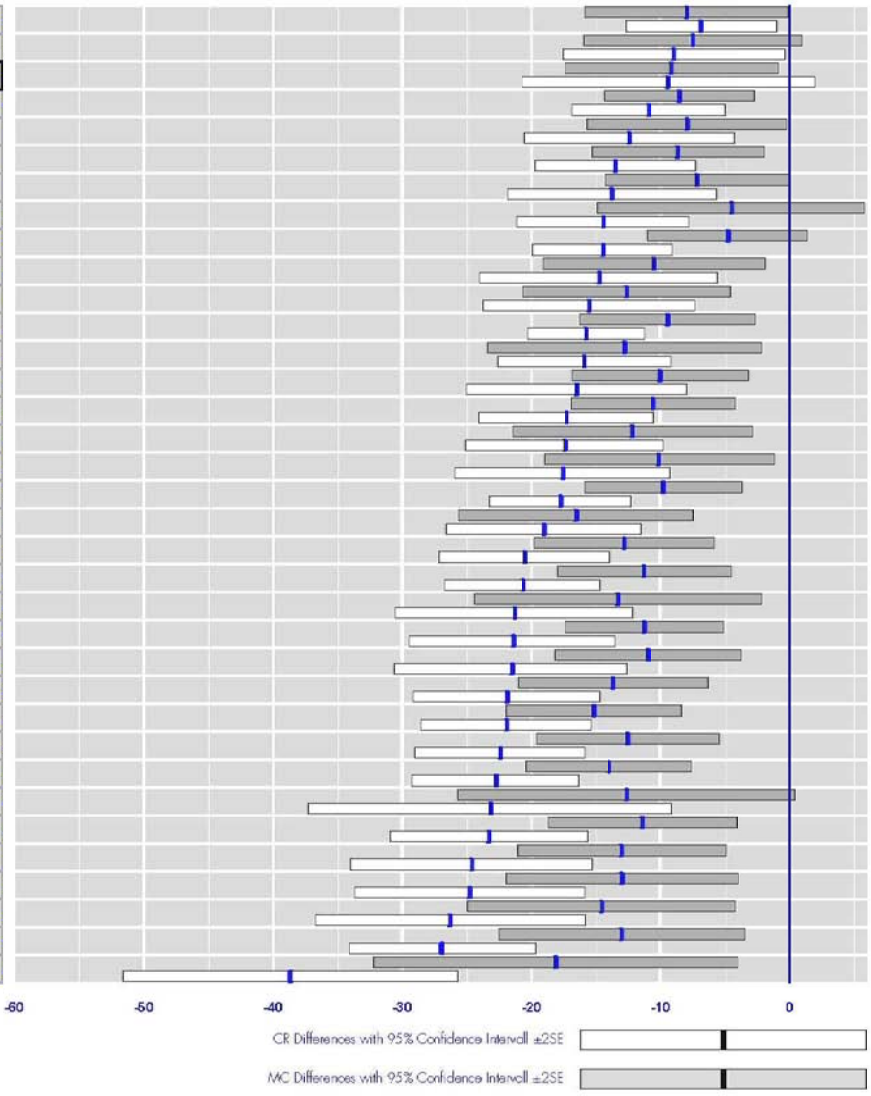
Gender differences

In this section an examination has been made of the differences between boys and girls in the countries with respect to their scores based on the multiple-choice items and the constructed response items respectively.

First, the mean ability of girls and boys were calculated for each country based on the scores calculated with the multiple-choice items only and also for the scores calculated with the constructed response items only. Then a regression analysis was performed to calculate the differences of the boys' means compared to the girls' means, with the appropriate standard errors for the differences. The result of comparing these boy-girl differences based on the different sets of items has been presented in Figure 3. The boys' reading achievement in relation to the girls' reading achievement based on the constructed response items have been depicted by a white bar. If the girls in the country performed better than the boys, this has been indicated by a negative number. The same differences but based on the scores of the multiple-choice items only have been depicted by grey bars.

Compare the Gender Differences based on Multiple Choice Scores and Constructed Response Scores

	MC			CR		
Country	Girls Scores	Boys Scores	Differences	Girls Scores	Boys Scores	Differences
ITALY	539 / 2,8	531 / 3,1	-8 / 4,0	538 / 2,8	531 / 2,9	-7 / 2,9
COLOMBIA	445 / 4,0	437 / 4,0	-8 / 4,2	439 / 4,5	430 / 4,1	-9 / 4,3
MOROCCO	416 / 5,7	407 / 6,4	-9 / 4,1	397 / 8,4	388 / 8,2	-9 / 5,7
RUSSIAN FEDERATION	529 / 4,2	520 / 3,9	-9 / 2,9	530 / 4,0	519 / 4,1	-11 / 3,0
FRANCE	519 / 3,1	511 / 3,0	-8 / 3,9	531 / 3,1	519 / 3,4	-12 / 4,1
NETHERLANDS	554 / 3,0	545 / 2,8	-9 / 3,3	553 / 2,8	539 / 3,0	-14 / 3,1
ROMANIA	505 / 3,7	498 / 5,0	-7 / 3,6	522 / 4,1	509 / 5,3	-14 / 4,1
CZECH REP.	529 / 4,1	524 / 3,7	-5 / 5,2	540 / 3,0	525 / 3,0	-14 / 3,3
GERMANY	533 / 2,0	528 / 2,4	-5 / 3,1	541 / 2,2	527 / 2,4	-15 / 2,7
ARGENTINA	445 / 4,4	434 / 3,9	-11 / 4,3	443 / 5,2	429 / 5,0	-15 / 4,6
SLOVAK REP.	526 / 3,2	513 / 3,3	-13 / 4,0	520 / 3,0	505 / 3,5	-16 / 4,1
HUNGARY	532 / 2,4	522 / 2,9	-9 / 3,4	549 / 2,2	534 / 2,4	-16 / 2,3
LITHUANIA	545 / 3,8	533 / 3,8	-13 / 5,3	546 / 3,2	530 / 3,0	-16 / 3,4
BELIZE	403 / 3,5	393 / 3,5	-10 / 3,4	386 / 4,3	369 / 4,2	-17 / 4,3
TURKEY	464 / 3,3	453 / 3,3	-11 / 3,1	469 / 3,7	452 / 3,4	-17 / 3,4
UNITED STATES	538 / 3,7	526 / 4,6	-12 / 4,7	547 / 3,7	530 / 4,3	-17 / 3,8
SCOTLAND	525 / 4,6	515 / 3,8	-10 / 4,4	535 / 2,9	517 / 3,9	-18 / 4,2
CANADA	533 / 2,6	523 / 2,7	-10 / 3,0	552 / 2,6	534 / 2,5	-18 / 2,7
LATVIA	544 / 3,7	527 / 3,1	-17 / 4,5	550 / 3,1	531 / 2,7	-19 / 3,8
MACEDONIA, REP. OF	461 / 3,9	448 / 3,6	-13 / 3,5	469 / 4,6	448 / 4,0	-21 / 3,3
ICELAND	513 / 2,2	501 / 2,5	-11 / 3,4	523 / 2,2	502 / 2,1	-21 / 3,0
GREECE	535 / 4,7	522 / 3,9	-13 / 5,6	533 / 3,6	512 / 4,8	-21 / 4,6
HONG KONG	535 / 3,0	524 / 3,2	-11 / 3,1	536 / 3,2	514 / 4,1	-21 / 4,0
NORWAY	502 / 3,0	491 / 3,2	-11 / 3,6	512 / 3,6	491 / 3,4	-22 / 4,5
ISRAEL	512 / 3,0	499 / 3,1	-14 / 3,7	523 / 3,2	501 / 3,3	-22 / 3,6
SLOVENIA	499 / 2,8	484 / 2,7	-15 / 3,4	519 / 2,5	497 / 2,7	-22 / 3,3
SWEDEN	551 / 2,4	538 / 2,7	-13 / 3,5	568 / 2,6	546 / 2,6	-22 / 3,3
SINGAPORE	534 / 4,3	520 / 4,7	-14 / 3,2	536 / 4,5	513 / 4,9	-23 / 3,2
IRAN, ISLAMIC REP.	454 / 4,9	442 / 4,4	-13 / 6,5	434 / 4,4	411 / 5,2	-23 / 7,0
ENGLAND	537 / 3,8	526 / 4,0	-11 / 3,7	562 / 3,3	539 / 3,8	-23 / 3,8
MOLDOVA, REP. OF	494 / 4,9	481 / 4,4	-13 / 4,0	509 / 4,5	485 / 4,1	-25 / 4,7
BULGARIA	542 / 3,0	529 / 4,6	-13 / 4,5	562 / 4,1	537 / 4,4	-25 / 4,4
NEW ZEALAND	522 / 5,0	507 / 4,5	-15 / 5,2	544 / 4,4	517 / 4,4	-26 / 5,2
CYPRUS	504 / 2,8	491 / 4,0	-13 / 4,8	511 / 2,9	484 / 3,1	-27 / 3,6
KUWAIT	428 / 5,7	410 / 4,5	-18 / 7,0	442 / 4,5	403 / 4,4	-39 / 6,5



From Figure 3 it can be seen that in all countries girls performed better than boys - which was also reported in the PIRLS 2001 International Report. For all countries except Italy, the difference based on the constructed response items only was higher than the difference based on the multiple-choice items only. For the countries Colombia, Czech Republic, Germany and Iran only the gender differences based on the constructed response items were significant. For Morocco, only the gender difference based on the multiple-choice scores was significant. It must be noted here that since gender was not included in the population model for calculating the plausible values, the gender differences were not absolutely correct and might be underestimated.
With the data presented here, however, no significant differences could be found for the different item types with respect to gender although the general statement that the differences between boys and girls tended to be higher when looking at the constructed response items holds.

Conclusion

The fact that percentage correct of the multiple-choice items in PIRLS 2001 tended to be lower than the percentage correct for constructed response items seems to be mainly effected by the guessing that took place for multiple-choice items. In general, scores based on multiple-choice items or on constructed response items seemed to be quite comparable. But there were differences for some countries with respect to the item type. Consequently, it appears to be a good approach, as done in PIRLS, to administer multiple-choice items as well as constructed response items.

REFRENCES

BAKER, Frank B., Item response theory: parameter estimation techniques, Dekker 1992

ELLEY, W.R. and MANGUBHAI, F., Multiple-choice and Open-ended Items in Reading Tests: Same or Different?, Studies in Evaluation 18, 191-199, 1992

MARTIN, Mick, MULLIS, Ina V.S., et al., "PIRLS 2001 International Report", Boston College, 2003

MARTIN, Mick, MULLIS, Ina V.S., KENNEDY, Ann M., "PIRLS 2001 Technical Report", Boston College, 2003

MURAKI, Eiji, BOCK, R.D., PARSCALE IRT Item Analysis and Test Scoring for Rating-scale Data, Scientific Software International, Inc. Chicago 1997

RUBIN, D.B., Multiple Imputation for Nonresponse in Surveys, John Wiley and Sons, New York 1987

WRIGHT, Benjamin D., MASTERS, Geofferey N., Rating Scale Analysis, Mesa Press Chicago, 1982

WU, Margaret L., ACER ConQuest: generalized item response modeling software manual, ACER press 1998