Model fit and robustness?

-

A critical look at the foundation of the PISA project

Svend Kreiner, Dept. of Biostatistics, Univ. of Copenhagen

TOC

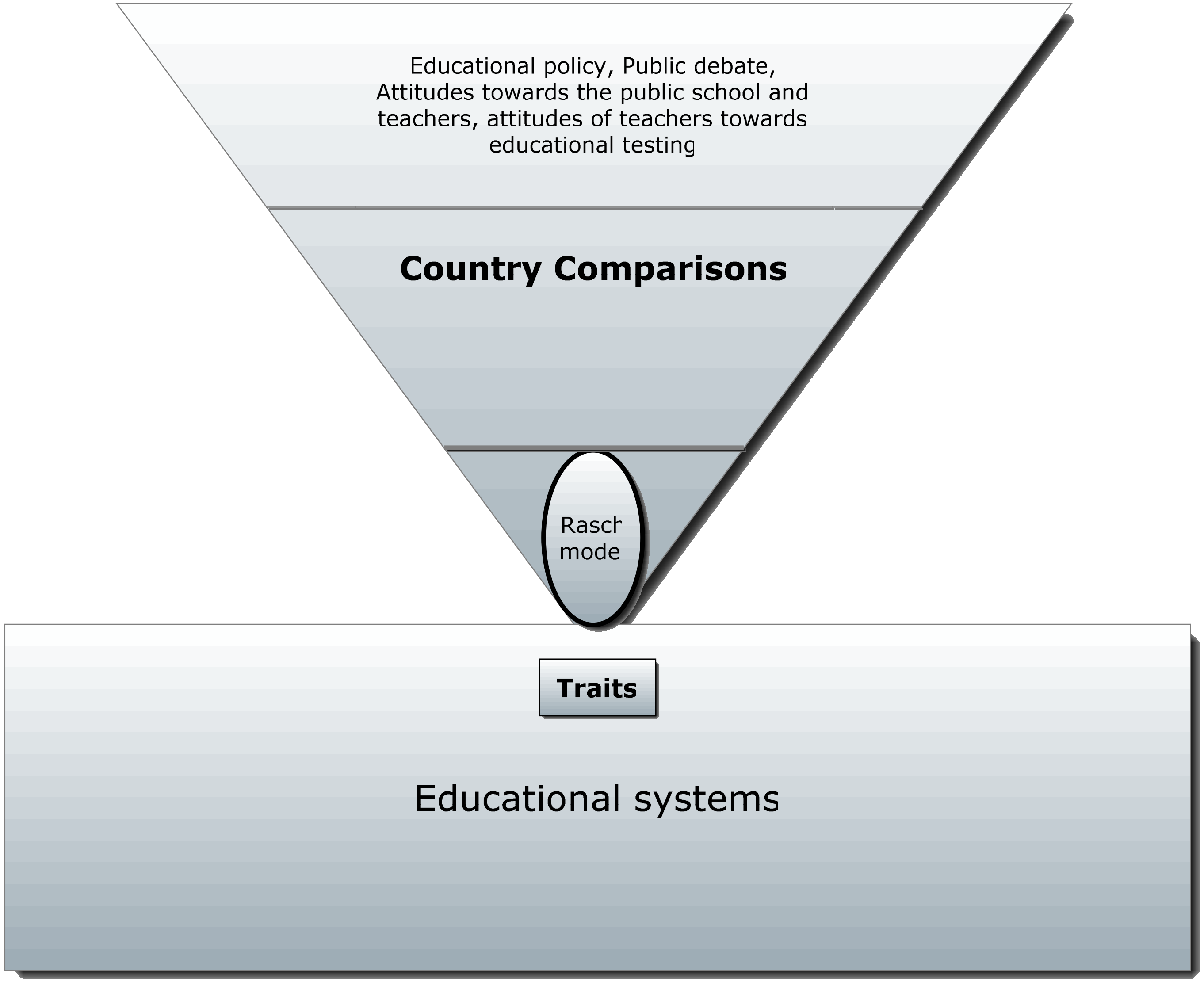
The PISA project and PISA data

PISA methodology

Rasch item analysis of data from 56 countries

Analysis of robustness towards model item and/or model errors

The PISA project



Educational policy, Public debate, Attitudes towards the public school and teachers, attitudes of teachers towards educational testing

Country Comparisons

Rasch mode

Traits

Educational systems

PISA data on reading

Test results from 398,750 students from 56 countries

Eight reading units (texts) with 28 items (questions)

69 % of responses to reading items are missing

Average number of item responses pr. Student = 8.3

The final data set.

Item set 1	Item set 2	Summary test results
183,569 students with imputed responses to all items		Plausible values
92,635 students with observed responses	92,635 students with imputed responses	Plausible Values
91,941 students with imputed responses	91,941 students with observed responses	Plausible Values
30,605 student with responses to all items		Plausible values

The Rasch analysis

	PISA	Kreiner & Christensen
Study population	15,000 OECD student	30,605 (Booklet 6)
Overall fit		Andersen's Conditional likelihood ratio (CLR) test
Item fits	Infit	Infit
Item discrimination	Point-Biserial correlation	Item-restscore correlation
DIF	Informal comparison of item parameters in different countries	CLR test Mantel-Haenszel Kelderman's CLR test for separate items

Overall test results

The CLR test rejects the hypothesis that item parameters are the same among students with scores from 1-13 and students with scores from 14-24

(CLR = 5371.0; df = 24; p < 0.00005)

The CLR test rejects the Rasch models for all countries except Lichtenstein (n = 34)

The CLR test rejects the hypothesis that item parameters are the same in all countries

(CLR = 27389.0; df = 1320, p < 0.00005)

Estimates of item parameters and item fit statistics based on Booklet 6 data on 20 PISA items.

```
                                           Item-restscore gamma
  item      thresholds     infit    p    observed  expected    p        chi      df    p         clr     df  p
 R055Q02   0.461 ......    0.968  0.000    0.554     0.522   0.000     896.1   684 0.000      680.2    55 0.000
 R055Q03  -0.185 ......    0.894  0.000    0.633     0.537   0.000    1099.3   684 0.000     1213.2    55 0.000
 R055Q05  -0.971 ......    0.822  0.000    0.719     0.559   0.000    1230.9   684 0.000      630.5    55 0.000
 R067Q01  -2.009 ......    0.978  0.055    0.601     0.590   0.154    1208.6   684 0.000      891.3    55 0.000
 R067Q04  -0.451   0.528   1.242  0.000    0.457     0.578   0.000    2644.4  1316 0.000     3230.5   110 0.000
 R067Q05   0.296  -0.777   1.233  0.000    0.544     0.643   0.000    2913.7  1316 0.000     3515.3   110 0.000
 R104Q01  -1.419 ......    0.899  0.000    0.681     0.572   0.000    1373.6   684 0.000      885.8    55 0.000
 R104Q02   1.088 ......    1.166  0.000    0.325     0.512   0.000    1335.4   684 0.000     1284.7    55 0.000
 R104Q05   0.937   2.963   0.998  0.782    0.563     0.531   0.000    2997.6  1103 0.000     2222.3   110 0.000
 R111Q01  -0.497 ......    0.971  0.000    0.586     0.545   0.000    1049.4   684 0.000      594.9    55 0.000
R111Q06B   1.228   0.019   1.123  0.000    0.565     0.620   0.000    1883.8  1262 0.000     2231.8   110 0.000
 R219Q02  -1.263 ......    0.891  0.000    0.663     0.567   0.000    1488.2   684 0.000     1104.4    55 0.000
 R220Q01   1.046 ......    0.871  0.000    0.653     0.513   0.000    1108.5   632 0.002      868.7    55 0.000
 R220Q04  -0.004 ......    1.003  0.545    0.539     0.532   0.253     932.5   684 0.000      717.3    55 0.000
 R220Q05  -1.113 ......    0.923  0.000    0.670     0.563   0.000     994.8   684 0.000      368.7    55 0.000
 R220Q06  -0.193 ......    1.055  0.000    0.498     0.537   0.000    1044.4   684 0.000     1075.3    55 0.000
 R227Q01   0.392 ......    1.132  0.000    0.398     0.524   0.000    1183.3   684 0.000     1494.5    55 0.000
R227Q02T  -0.803   1.248   1.099  0.000    0.501     0.559   0.000    2856.8  1316 0.000     3359.9   110 0.000
 R227Q03   0.099 ......    0.913  0.000    0.615     0.530   0.000    1216.9   684 0.000      656.5    55 0.000
 R227Q06  -0.618 ......    0.862  0.000    0.679     0.548   0.000    1354.7   684 0.000     1500.0    55 0.000
```

Conclusions:

1) The fit of PISA items to the Rasch model is rejected

2) Very strong evidence of DIF for all items

3) Country ranks may be confounded

The technical report on model fit and DIF

"The interpretation of a scale can be severely biased by unstable item characteristics from one country to the next" (page 86)

“particular attention was paid to the fit of the items to the scaling model, item discrimination and item-by-country interaction” and “consistency of item parameter estimates across countries was of particular interest” (page 147)

PISA now (2011) says that

“PISA anticipates a certain degree of differential functioning”

“… the Technical report shows the relative standing of countries is largely invariant to the choice of items. Even if one … ranks countries based on their performance on the preferred items of any country, the result remains largely invariant.

Are country ranks robust towards the lack of fit between item responses and the Rasch model?

What do we mean, when we say that the ranking is robust towards the inadequacy of the Rasch model?

How do we assess robustness?

Can we be sure, that there is a “true” ranking if item responses do not fit a unidimensional IRT model?

Are “plausible values” really plausible.

Analysis of robustness (assuming a true ranking exist)

Analysis of invariance

Comparison of ranking by the Rasch model and ranking by models with fewer errors

Invariance

The fundamental property of Rasch models

Except for random error, measurement results should be the invariant across all subsets of items

Country rankings should – except for random errors be the same for the complete set of items and for items from reading units

R055+R067, R104+R111, and R219+R220+R227

Item set 1: R055+R104+R111+R227 **Item Set 2: R067+R219+R220**

Information items, Interpretation items and Reflection items

Invariance summary for five countries

		Item no.			Item set		Item type		
Rank	Country	0	1	2	1	2	Inform	interpr	reflect
7	Canada	3	12	21	8	5	25	18	2
16	France	17	28	14	23	18	16	15	23
17	Denmark	36	20	5	7	32	6	8	37
22	Japan	40	11	9	13	36	8	30	28
23	UK	14	23	30	22	30	23	26	18

Invariance?

A Monte Carlo study of invariance

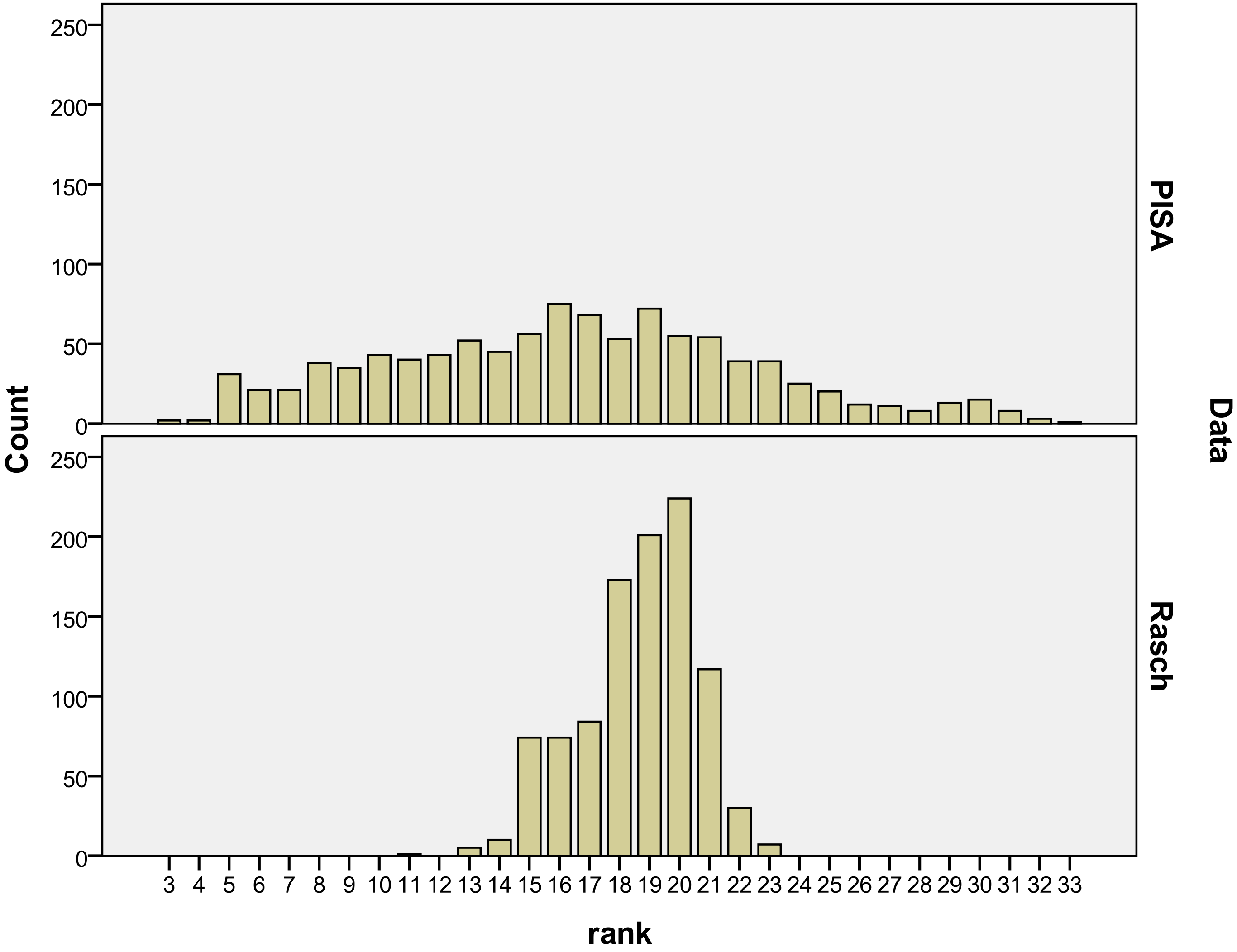
Analysis of rankings by 1000 random sets of 14 items.

<u>Two data sets</u>

PISA data

Simulated data from a Rasch model defined by the estimated parameters in the PISA data.

The variation of the Danish rank



Cases weighted by n

Variation of ranks in 1000 random subsets of 14 items

PISA

country	rank	mean	s.d.	range
Canada	7	8.16	3.44	3 - 21
France	16	16.35	3.77	5 - 27
Denmark	17	16.40	6.12	3 - 33
Japan	22	21.53	5.04	6 - 36
UK	23	22.06	2.78	15 - 30

Rasch

country	rank	mean	s.d.	range
Canada	9	8.39	1.27	6 - 13
France	15	16.39	1.94	10 - 23
Denmark	19	18.63	1.95	11 - 23
Japan	22	22.19	1.29	18 - 25
UK	25	25.25	0.55	23 - 27

Invariance?

Extreme rankings

Analyses of DIF identify items that favour and disfavour specific countries. These subsets define extreme ranks in PISA's items.

country	rank	mean	s.d.	range	extreme range
Canada	7	8.16	3.44	3 - 21	3 - 29
France	16	16.35	3.77	5 - 27	2 - 40
Denmark	17	16.40	6.12	3 - 33	3 - 42
Japan	22	21.53	5.04	6 - 36	4 - 39
UK	23	22.06	2.78	15 - 30	8 - 36

Ranking by other models

The ranks should be defined by estimates of person parameters or plausible values generated by the true model.

PISA's Rasch based ranks are robust if they are close to the ranks defined by the true model.

Since the true model has not and can not be identified it is not possible to assess the robustness relative to ranking be the true model.

Instead we assess the robustness by comparison with ranking by better models.

All models are wrong, but some models are more wrong than other models.

The worst model

The Rasch model

|

A better fitting model

The Rasch model + uniform DIF

|

An even better fitting model

Rasch model + uniform DIF and uniform local response dependence among items from the same reading units

|

The hypothetical true model

Modelling PISA responses

Initial model: The Rasch model. (BIC = 540293)

Find the better model by stepwise addition of uniform DIF terms until BIC starts to increase. (BIC = 526188)

Find the even better model by stepwise addition of uniform LD terms until BIC starts to increase. (BIC = 518950)

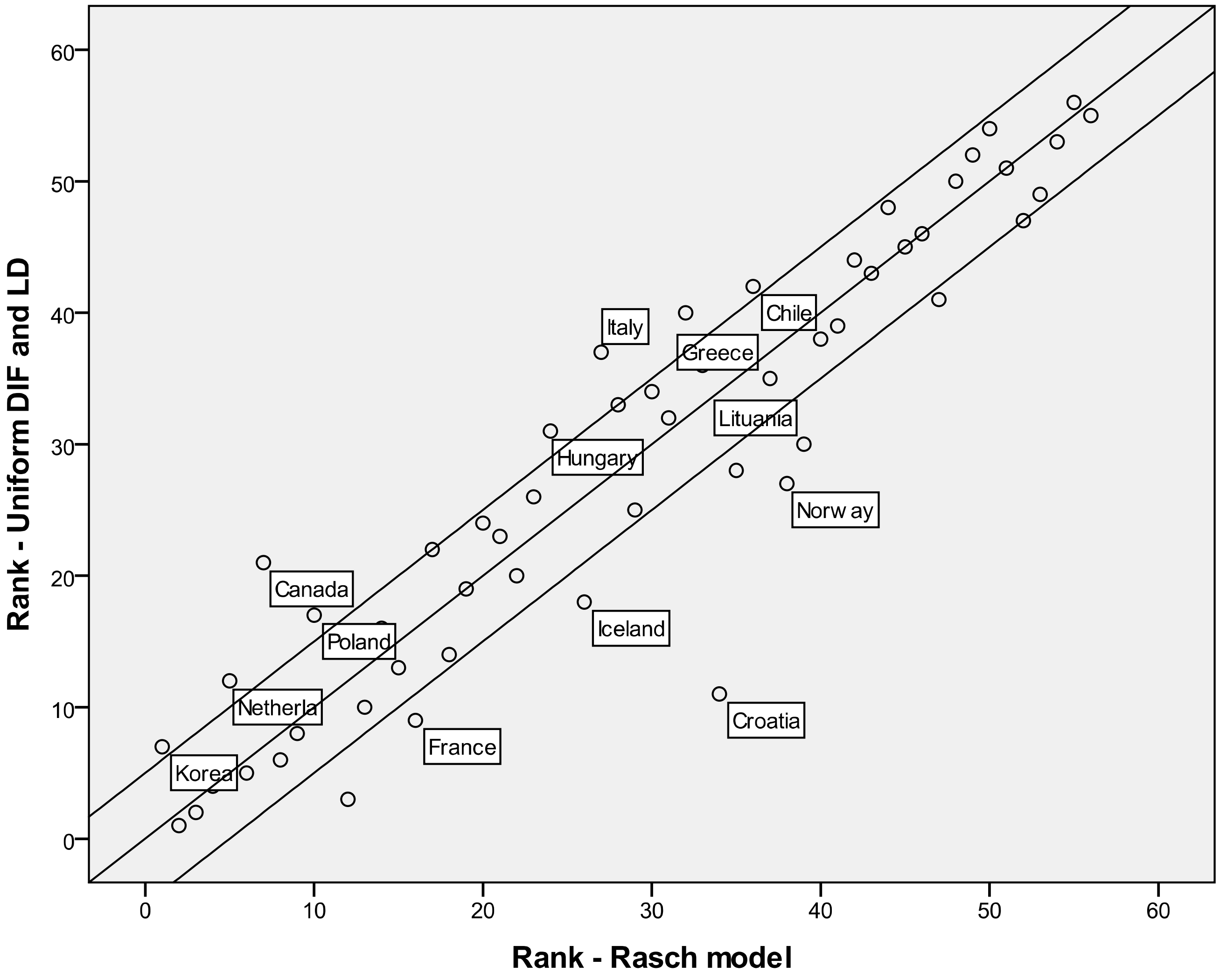
Estimate the mean and variance of the latent variable in the different countries under these models

How do the ranks defined by these statistics compare to the ranks defined by the Rasch model?

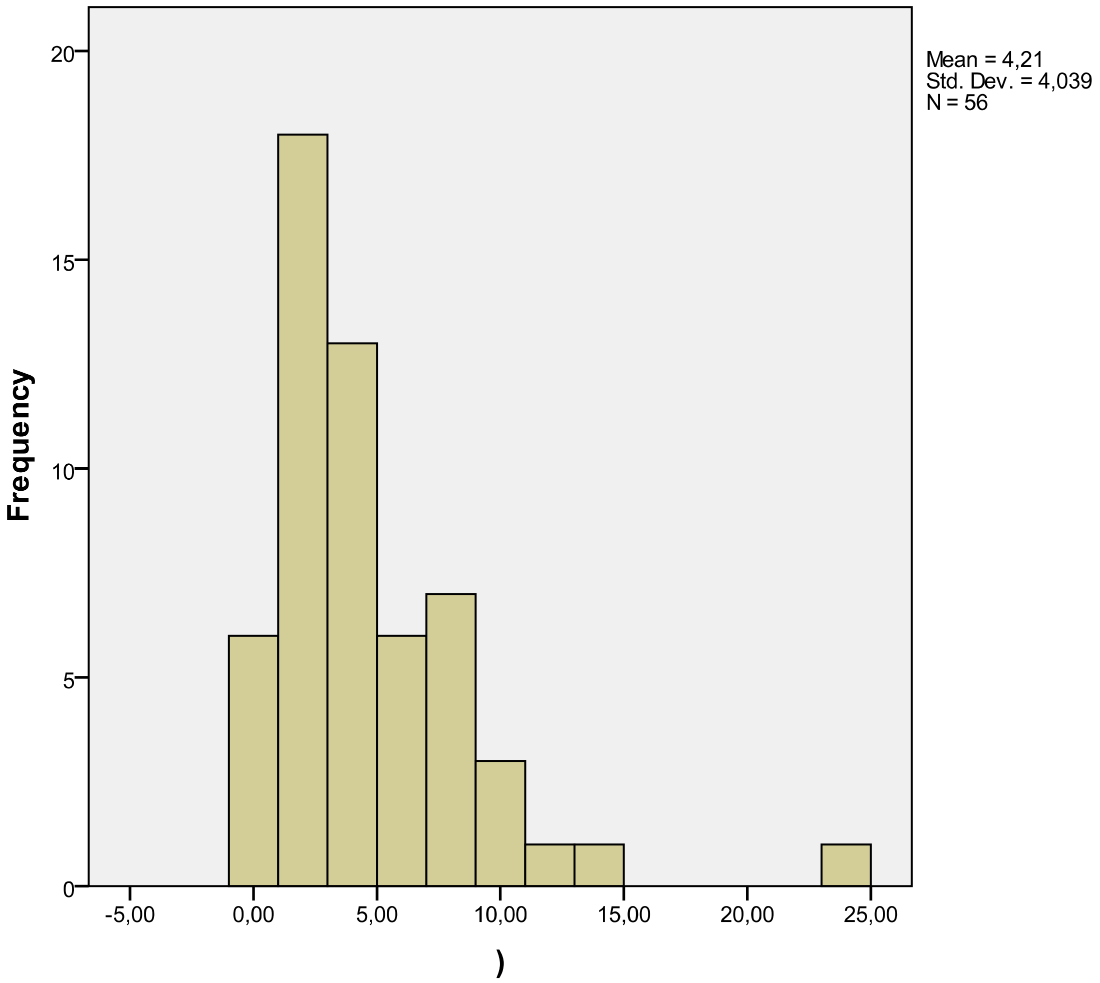
Overview of DIF and local dependence among items according to the loglinear Rasch models

Item	Country DIF	Local dependence
R055Q02	yes	R055Q03 and R055Q05
R055Q03	yes	R055Q02 and R055Q05
R055Q05	no	R055Q02 and R055Q03
R067Q01	yes	R067Q04 and R067Q05
R067Q04	yes	R067Q01 and R067Q05
R067Q05	yes	R067Q01 and R067Q04
R104Q01	no	R104Q02 and R104Q05
R104Q02	yes	R104Q01
R104Q05	yes	R104Q01
R111Q01	yes	R111Q06B
R111Q06B	yes	R111Q01
R219Q02	yes	
R220Q01	no	R220Q04, R220Q05 and R220Q06
R220Q04	yes	R220Q01 and R220Q06
R220Q05	no	R220Q01, R220Q04 and R220Q06
R220Q06	yes	R220Q01 and R220Q05
R227Q01	yes	R227Q02T
R227Q02T	yes	R227Q01, R227Q03 and R227Q06
R227Q03	yes	R227Q02T and R227Q06
R227Q06	yes	R227Q02T and R227Q03

Comparisons of ranks by the Rasch model and a model with uniform DIF and LD



Distribution of differences of ranks by the Rasch model and a model with uniform DIF and LD



Conclusions

Item responses do not fit a Rasch model

Ranking by PISA's items is not invariant and there are considerable differences for some countries between ranking by the Rasch model and ranking by models that take DIF and local dependence into account.

So what do you think?

Do *you* think that PISA's ranking of countries is robust towards the misfit of items and the Rasch model

PISA on the DIF issue

”A statistical index called *differential item functioning* was used to detect items that worked differently in some countries. … As a result, some items were excluded from scaling as if they had not been administered in that country. Table A1.4 lists the items that are excluded from the national scaling for each country”

Kirsch et.al. (2002) *Reading for change. Performance and Engagement across countries. Results from PISA 2000.* OECD

Adams et.al. (2007) say that

“an item with sound characteristics in each country but that shows substantial item-by-country interactions may be regarded as a different item (for scaling purposes) in each country (or in same subsets of countries)”.

The technical report says that

"The interpretation of a scale can be severely biased by unstable item characteristics from one country to the next" (page 86)

“particular attention was paid to the fit of the items to the scaling model, item discrimination and item-by-country interaction” (page 147)

“consistency of item parameter estimates across countries was of particular interest” (page 147)

due to DIF, "out of 179 mathematics, reading and science items, 28 items were omitted in a total of 38 occurrences for the computation of national scores" (page 104).

PISA now (2011) says that

“PISA anticipates a certain degree of differential functioning”

“results show that the impact of deviations of the reality from the PISA model is negligible and that DIF does not put at risk the comparability of the results”

“that country means are largely invariant against the choice of items has been tested extensively”

“… the Technical report shows the relative standing of countries is largely invariant to the choice of items. Even if one … ranks countries based on their performance on the preferred items of any country, the result remains largely invariant.

Schleicher (2011) to the Danish Radio