МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2014
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2014
Proceedings of the Forty Third Spring Conference
of the Union of Bulgarian Mathematicians
Borovetz, April 2–6, 2014

WHAT EVERY MATHEMATICIAN AND MATH TEACHER SHOULD KNOW ABOUT PISA

Erich Neuwirth

Political discussion about education PISA is quite often used as an argument, but almost nobody knows what PISA points really measure. We will discuss the statistical model behind PISA and also see that there were some problems in the past.

1. Introduction. Why (not only) mathematicians and math teachers should be knowledgeable about PISA?
A short list of interesting topics

- PISA is used in discussions about education
  - one should be able to judge the validity of assertions
  - not everything which could be studied is published
  - a large data set is available, so one should be able to do one's own research
- PISA has influence on math education in schools
  - Which test questions make sense

Some interesting questions and facts about PISA:

- What does a difference of 5 points between two countries indicate?
- 5 points in science have a differerent meaning from 5 points in reading
- Every student gets 5 different point values in each domain
- Students with identical answers still might get different point values
- Not all students are tested in all domains
- Children of people with higher education will receive better point values for domains in which they are not tested

2. What is PISA. In the words of OECD

> PISA is an international study which aims to evaluate education systems worldwide by testing the skills and knowledge of 15-year-old students
> – OECD Web site

How is PISA working?
Who is taking part and how are data collected?

- 65 countries (2012), 34 OECD and 31 partner countries
- 4500+ students per country (except Liechtenstein, . . .)
- Student age 15 years
- in 2012 485.000+ students
- in 2012
  - 109 math items
  - 44 reading items
  - 53 science items
- Each student works with a test booklet with 4 blocks of test items for 2 hours
- context questionnaire: economic situation, education of parents . . . (1/2 hour)
- country specific questions
  - (in context questionnaire, e.g. marks in school)

Not all test items can be published because some need to be reused to be able to compare results from different periods.

PISA web site has exemplary test items.

3. How PISA scores are computed. PISA is using a model from psychology, to be precise from psychometrics, the Rasch model.

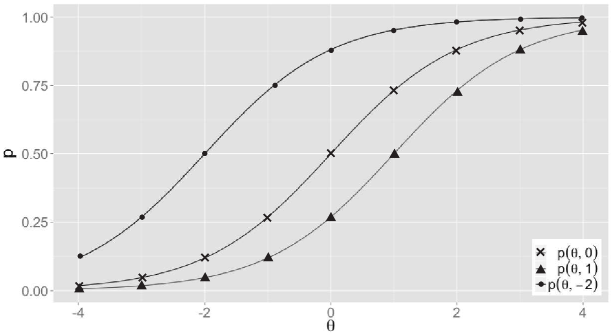
A simplified description:
Each test person has an ability θ and each test item has a difficulty ξ. Ability and difficulty are expressed as real numbers. According to the Rasch model there is a function $p(\theta, \xi)$ giving the probability that a person with ability θ solves an item with difficulty ξ.

A central assumption of the Rasch model is that there is no preference for items depending on the abilities. If we have 2 items A and B and give these items to two groups of people with different abilities and then study only the subgroups of persons solving exactly one of the two items, then the probabilty for solving item A has to be identical for both groups.

As consquence of some more (reasonable and justifiable) assumptions we have

$$p(\theta, \xi) = \frac{e^{\theta-\xi}}{1+e^{\theta-\xi}}$$

The dependency of the probabiity of solving an item on ability for items of different difficulty is given by the following graph:



To work with the Rasch model we need 2 auxiliary functions logit und prob defined this way:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right),$$

$$\text{prob}(l) = \frac{e^l}{1+e^l}.$$

These functions are inverse.

The naming of the parameter is meant to make reading easier. logit computes logit values from probabilites (therfore argument p), and prob computes probabilites from logits (therefore argument l).

Simplifying we can say that logit spreads out probabilites close to 0 or 1. If an item is very easy, such that average students solve it with a probability of 0.95, then it cannot be solved with probability higher than 1 even by the smartest students. For an item with probabiluty 0.5 for average students, smarter students can solve it with much higher probabilities.

But there is an even stronger reason to use these functions:

Let us assume we have 2 test persons with abilities θ_1 and θ_2 and a set of items with difficulties ξ_i, $i = 1, \ldots, n$. Then the probabilites for solving the items are

$$p_{1i} = \text{prob}(\theta_1 - \xi_i),$$

$$p_{2i} = \text{prob}(\theta_2 - \xi_i)$$

and the corresponding logits are

$$l_{1i} = \text{logit}(p_{1i}) = \theta 1 - \xi_i,$$

$$l_{2i} = \text{logit}(p_{2i}) = \theta 1 - \xi_i.$$

For the differences of the logits we therefore have

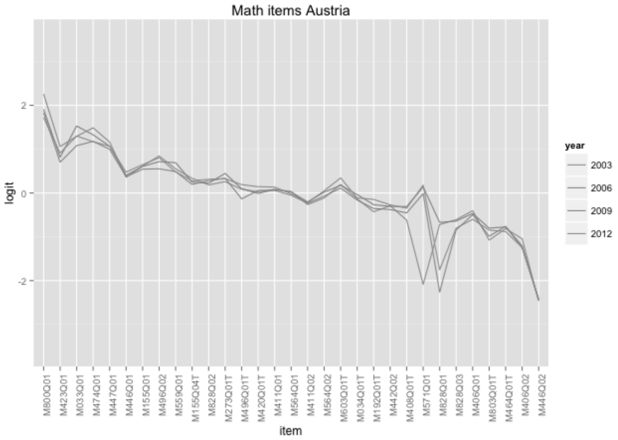
$$l_{1i} - l_{2i} = \theta_1 - \theta_2.$$

Comparing abilities of the the test persons therefore is independent of the set of items used when using logits. This is called Rasch homogeneity of the items.

This condition is not fulfilled all by itself, it is a property of the set of test items. In PISA, there are extensive pretests to ensure this property of the test items.

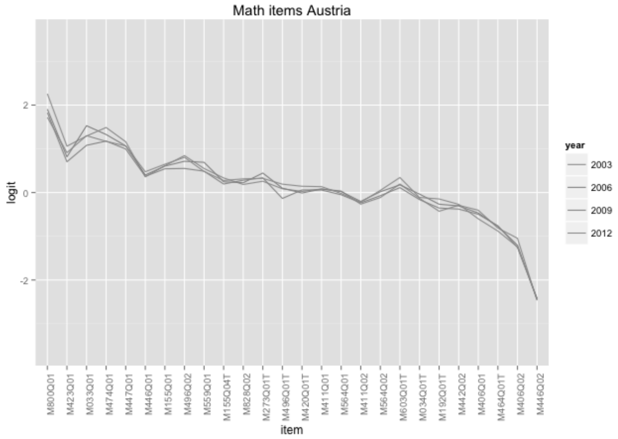
If this condition is fulfilled and we order the items according to difficulty in different countries, the order of the items has to be essentially the same in all countries.

In the following chart, the items have been ordered according to the average difficulty in all countries participating in PISA in all test periods so far. The value is the logit of the probability of solving by Austrian students. The values should be decreasing when going to the right because items harder on the average also need to be harder in Austria.



We see that some items do not fit the pattern in 2000 and 2003. PISA in such cases allows items to be removed for “irregular” countries, but this was not done in this special case.

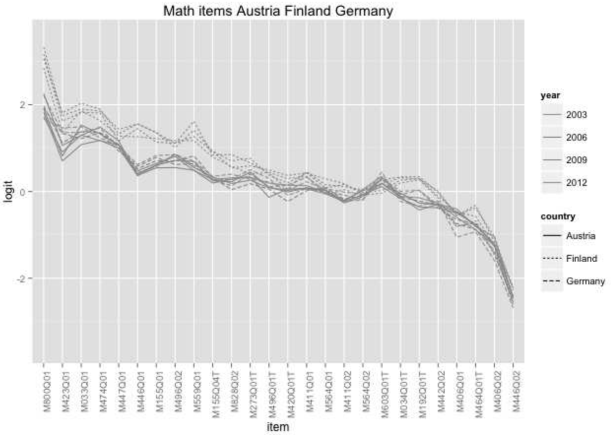
Removing these items we get:



In this case, the graph is essentially decreasing.

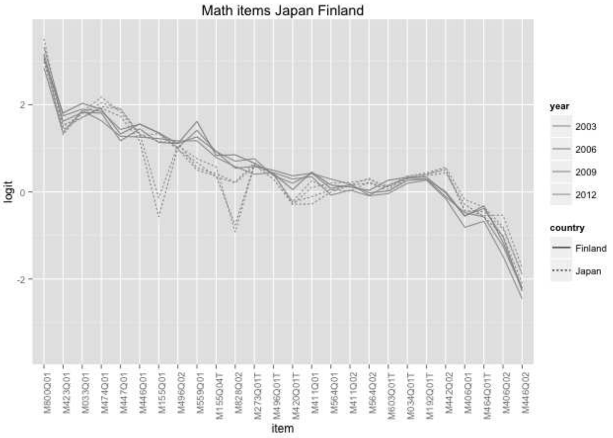
According to the Rasch model, these curves should be parallel, only be shifted up and down, for different countries

Comparing a few different countries yields the following graph.



The curves are approximately parallel.

Comparing Japan and Finland, however, does not produce parallel curves.



Rasch homogeneity seems not to be fulfilled when comparing countries with rather different cultures.

4. Computing country scores. Accepting the Rasch model assumptions we can compute statistical estimates of the abilities of the participating students. But PISA does things in a slightly different way.

PISA assumes that the abilities for students with the same economic and cultural background (education of parents, economic situation of family ...) are normally distributed. Under this assumption, the dependency of the ability on the background variables is estimated. Using these variables and the test answers, 5 plausible values for the ability are computed for each student (and each domain, namely mathematics, reading, and science). All published PISA results are based on these values. All analyses then are performed 5 times with each of these sets of plausible values to cross validate the results.

Why do we need complex statistical methods for PISA?

- in each country, only a subpopulation of all students is tested
- different students have to solve different test items
- Some students don't even get reading items but nevertheless are assigned a reading score

Since the students tested are a selection from the whole student population, statistical sampling methods (including weighted sampling) has to be used. There were some problems with weighting in Austria in 2000.

A later study ([1]) could identify the reasons for the incorrect results, and OECD officially corrected the results.

As a consequence of the Rasch model, it is possible to compute comparable scores even when different students have to solve different items.

Background variables allow to estimate plausible values for students and domains for which the students did not have to solve items.

The Rasch model computes scores on the logit-scale, roughly speaking in the range -3 to 3. To allow more easily understood presentation, these values are rescaled to achieve a mean value of 500 and a standard deviation of 100 for all OECD countries (for PISA 2000).

The conversion factors are:

domain	gender	factor
reading	m	79.4
	f	80.2
math		77.9
science		93.2

Simplifying we might say that the logits of the students are multiplied with these factors and then shifted to achieve a mean value of 500 for PISA 2000.

The PISA documents do not explain in detail why the conversion factors for reading are different for females and males.

Using these factors, we can convert PISA points of countries into differences of probabilties of solving items.

domain	item	+PISA	+p
reading	50%	3.2	1%
math	50%	3.1	1%
science	50%	3.8	1%
reading	50%	10	3.14%
math	50%	10	3.20%
science	50%	10	2.60%
reading	66%	10	2.78%
math	66%	10	2.84%
science	66%	10	2.38%
reading	75%	10	2.34%
math	75%	10	2.40%
science	75%	10	2.00%

When in one country a math item is solved by 50% of the students, the same item will be solved with potibability 53.2% in a country with a PISA score 10 points higher.

10 more PISA points in science, however, would raise the probability only by 2.6%.

5. Some results. Here are the Bulgarian results (Bulgaria did not participate in PISA 2003)

year	read	math	science
2000	430.4	429.6	448.4
2003	NA	NA	NA
2006	401.9	413.4	434.1
2009	429.1	428.1	439.3
2012	436.1	438.7	446.4

For comparison purposes, here are the Austrian results,

year	read	math	science
2000	492.1	502.5	504.7
2003	490.7	505.6	491.0
2006	490.2	505.5	510.8
2009	470.3	495.9	494.3
2012	489.6	505.5	505.8

The Bulgarian results seem to indicate improvement since 2006, but on a moderately low level.

The Austrian table shows that the 2009 result was an outlier. The results were practically constant for all the other years.

So we might say that the Austrian results were essentially constant for 5 PISA periods.

REFERENCES

[1] E. Neuwirth, W. Grossmann, I. Ponocny. PISA 2000 und PISA 2003: Vertiefende Analysen und Beitrage zur Methodik. Leykam. Graz, 2006.

[2] OECD. PISA-Website mit Daten und Ergebnissen. `ww.pisa.oecd.org` OECD, Paris, 2013.

Erich Neuwirth
Faculty of Computer Science
University of Vienna
e-mail: erich.neuwirt@univie.ac.at

КАКВО ТРЯБВА ДА ЗНАЯТ ЗА PISA МАТЕМАТИЦИТЕ И УЧИТЕЛИТЕ ПО МАТЕМАТИКА?

Ерих Нойвирт

PISA (Програма за международно оценяване на ученици) се използва твърде често като аргумент в политически дискусии за образованието. Но почти никой не знае какво се измерва (оценява) с точките, получени по PISA. В статията се обсъжда статистическият модел, който се използва в PISA. Посочват се и проблемите, които са съществували в PISA в миналото.