Was Mathematik(lehr)er über PISA wissen sollten

ERICH NEUWIRTH (UNIV. WIEN)

PISA wird in der aktuellen bildungspolitischen Diskussion oft als Argument verwendet. Allerdings herrscht große Unklarheit darüber, was denn die PISA-Punkte bedeuten. Wir diskutieren das statistisch-mathematische Modell hinter PISA, zeigen, welche Annahmen PISA zugrunde liegen und wo es in der Vergangenheit auch zu methodischen Problemen gekommen ist.

1. Einleitung

Warum sollen (nicht nur) Mathematik(lehr)er über PISA Bescheid wissen?
Eine kurze Aufzählung, warum man sich mit PISA beschäftigen sollte:

- Schüler fragen in der Schule, was PISA eigentlich misst und bedeutet
  - Wer soll die Schülerfragen beantworten?
- Mit PISA wird in der Bildungspolitik argumentiert
  - man sollte die Argumente bewerten können
  - nicht alles, was untersucht werden kann, wird auch veröffentlicht
  - viele Daten sind zugänglich, also sind auch eigene Auswertungen möglich
- PISA hat Einfluss auf den Unterrricht
  - Welche Aufgaben sind sachlich sinnvoll
- *Persönliche Anmerkung:*
  Universitär mathematisch Ausgebildete sollten diese erweiterte Kompetenz jedenfalls haben

Einige vielleicht überraschende Fragen und Feststellungen zu PISA:

- Was bedeutet ein Unterschied von 5 Punkten zwischen zwei Ländern?
- 5 Punkte Unterschied bei Naturwissenschaft bedeuten etwas anderes als beim Lesen.
- Jeder Student bekommt in jeder Domäne 5 verschiedene Werte zugewiesen.
- Schüler mit identischen Testergbnissen können verschiedene PISA-Scores haben.
- Nicht alle Schüler werden in allen Domänen getestet.
- Akademikerkinder ohne Lesetest bekommen im Lesen bessere Scores als Arbeiterkinder ohne Lesetest.
- Es ist sachlich unmöglich und daher Unfug, wenn Politiker verlangen, dass das PISA-Testergebnis auch in die Schulnoten einfließen soll.

2. Was ist PISA

PISA ist ein international durchgeführter Test, mit dem die Leistungen und kognitiven Fertigkeiten der Schüler im Alter von etwa 15 Jahren für die teilnehmenden Länder systematisch erhoben werden soll.

In den Worten der 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

Wie wird PISA durchgeführt?
Wer nimmt teil und wie werden die Daten erhoben?

- 65 Länder (2012), 34 OECD und 31 Partnerländer
- 4500+ Schüler pro Land (außer Liechtenstein, ...)
- Schüleralter 15 Jahre (9. Schulstufe)
- im Jahr 2012 485.000+ Schüler
- in Jahr 2012
  - 109 Mathematik-Items
  - 44 Lese-Items
  - 53 Naturwissenschafts-Items
- jeder Schüler arbeitet mit einem Testheft mit 4 Abschnitten 2 Stunden lang
- Kontextfragebogen zur sozioökonomischen Situation (Ausbildung der Eltern, Einkommen, Migrantenstatus, ...) (1/2 Stunde)
- Zusätzliche Fragen länderspezifisch
  - (im Kontextfragebogen, in Österreich z.B. Schulnoten)

Nicht alle PISA-Aufgaben können publiziert werden, da es notwendig ist, bei späteren Tests Aufgaben aus früheren Tests zu verwenden, damit die Vergleichbarkeit der Testergebnisse gesichert ist.

Freigegebene Musterbeispiele finden man am Internet, beispielsweise unter `https://www.bifie.at/node/264`

3. Wir werden PISA-Scores ermittelt

PISA verwendet ein Modell aus der Psychologie, genauer gesagt aus der Psychometrie, nämlich das Rasch-Modell.

Vereinfacht gesagt sieht das Modell so aus:
Man schreibt jeder Person eine Fähigkeit θ und jeder Aufgabe (in der Sprache der Psychometrie heißen Aufgaben Items) eine Schwierigkeit ξ zu. Schwierigkeit und Fähigkeit werden als Zahl ausgedrückt. Das Rasch-Modell setzt dann voraus, dass es eine mathematische Funktion $p(\theta,\xi)$ gibt, mit der sich die Lösungswahrscheinlichkeit für gegebene Fähigkeit und Schwierigkeit errechnen lässt.

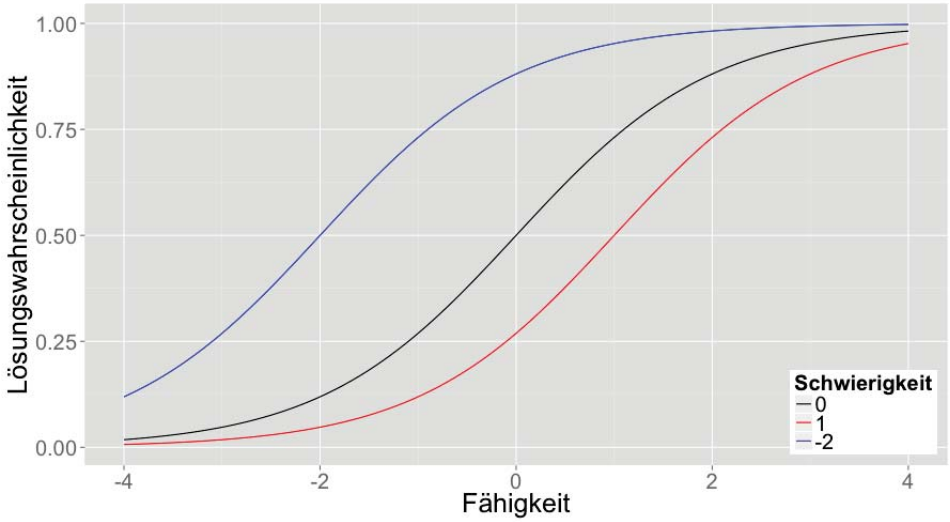
Wenn wir das in anderem Kontext veranschaulichen wollen, dann können wir uns eine Gruppe von Bogenschützen vorstellen. θ drückt dann aus, ob ein Bogenschütze besser oder schlechter ist, und ξ hängt von der Scheibengröße ab. Größere Scheiben sind leichter zu treffen und haben daher einen kleineren Wert für ξ.

Eine zentrale Annahme des Rasch-Modells ist, dass es keine „Bevorzugung“ von Items in Abhängigkeit von den Fähigkeiten gibt. Wenn man beispielsweise 2 Items A und B hat und diese Items 2 Gruppen von Personen mit verschiedenen Fähigkeiten vorlegt und dann aus beiden Gruppen nur jene Personen untersucht, die genau eine der beiden Aufgaben gelöst haben, dann verlangt das Rasch-Modell, dass die Wahrscheinlichkeit, Aufgabe A gelöst zu haben, für beide Gruppen gleich ist.

Aus weiteren vernünftigen und statistisch rechtfertigbaren Annahmen folgt dann, dass die Formel für die Berechnung der Lösungswahrscheinlichkeiten so aussieht:

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

Stellt man die Abhängigkeit der Lösungswahrscheinlichkeit von der Fähigkeit für verschieden schwierige Aufgaben dar, dann erhält man folgendes Bild:



Zur Arbeit mit dem PISA-Modell benötigen wir 2 Hilfsfunktionen, logit und prob, die so defininert sind:

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

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

Diese beiden Funktionen sind invers.

Die Bezeichung der Funktionsargumente ist als Lesehilfe gemeint. logit rechnet Wahrscheinlichkeiten (probabilities) in die logit-Skala um, daher p als Funktionsargument. prob rechnet Werte der logit-Skala in Wahrscheinlichkeiten um, daher l als Funktionsargument.

Vereinfacht gesagt dient die Funktion logit dazu, Wahrscheinlichkeiten nahe an 0 oder 1 zu „spreizen“. Wenn ein Beispiel sehr leicht ist, also beispielsweise von 95% der Schüler mit durchschnittlichen Fähigkeiten gelöst wird, dann kann es auch von ganz besonders befähigten Schülern nicht mit mehr als 100% Wahrscheinlichkeit gelöst werden. Bei einer 50%-Aufgabe ist aber eine Steigerung um mehr al 5% noch möglich.

Der tieferliegende Grund, diese Funktionen zu verwenden liegt aber woanders.

Nehmen wir folgende Situtation an: Wir haben 2 Testpersonen mit Fähigkeiten θ_1 und θ_2 und mehrere Aufgaben mit Schwierigkeiten ξ_i, $i = 1,\ldots,n$. Dann sind die Lösungswahrscheinlichkeiten

$$p_{1i} = \text{prob}(\theta_1 - \xi_i)$$
$$p_{2i} = \text{prob}(\theta_2 - \xi_i)$$

und die zugehörigen logits sind

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

Für die Differenzen der logits gilt daher

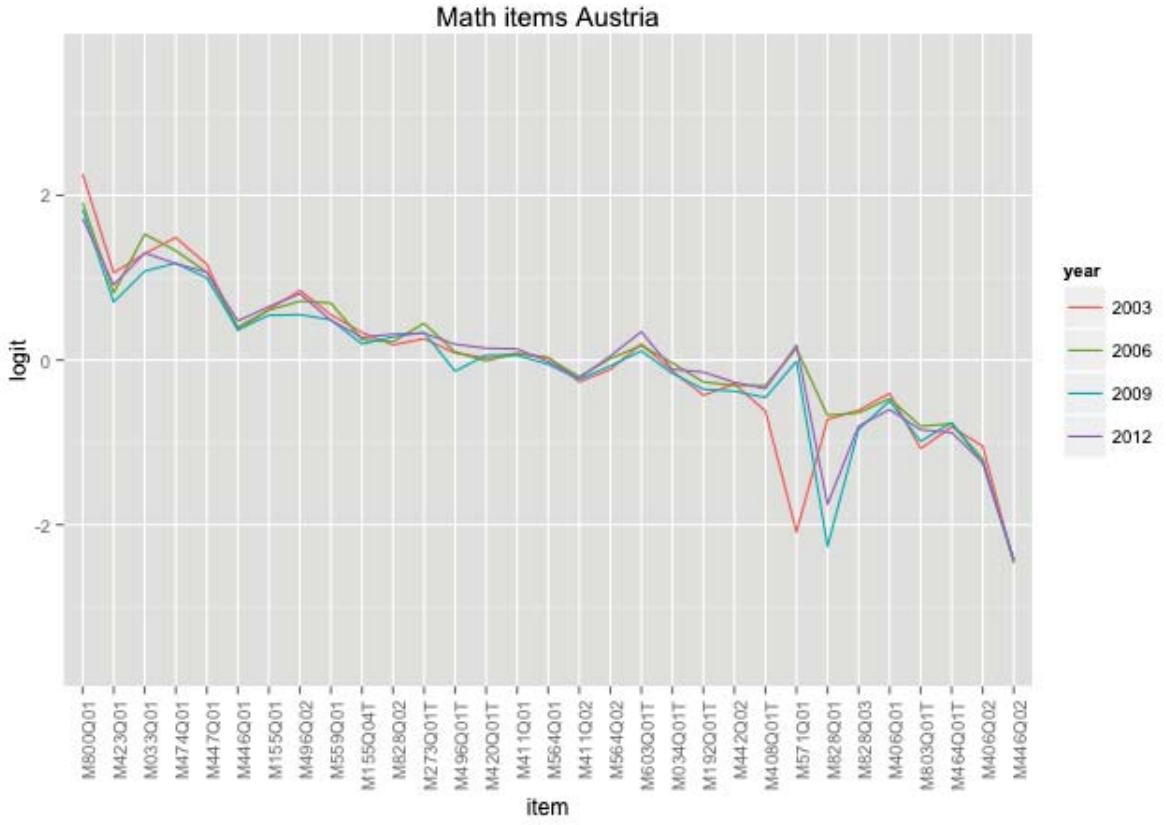
$$l_{1i} - l_{2i} = \theta_1 - \theta_2$$

Der Vergleich der Fähigkeiten der Testpersonen ist – wenn man die logits verwendet – also von den konkret verwendeten Aufgaben unabhängig möglich. Man spricht in diesem Zusammenhang von der Rasch-Homogenität der Items.

Diese Bedingung ist natürlich nicht „von selbst“ erfüllt, sie ist eine Eigenschaft der Zusammenstellung der Testaufgaben. Bei PISA wird in der Vorbereitung der Tests vorgesorgt, dass die Aufgaben diese Bedingung im wesentlichen erfüllen.

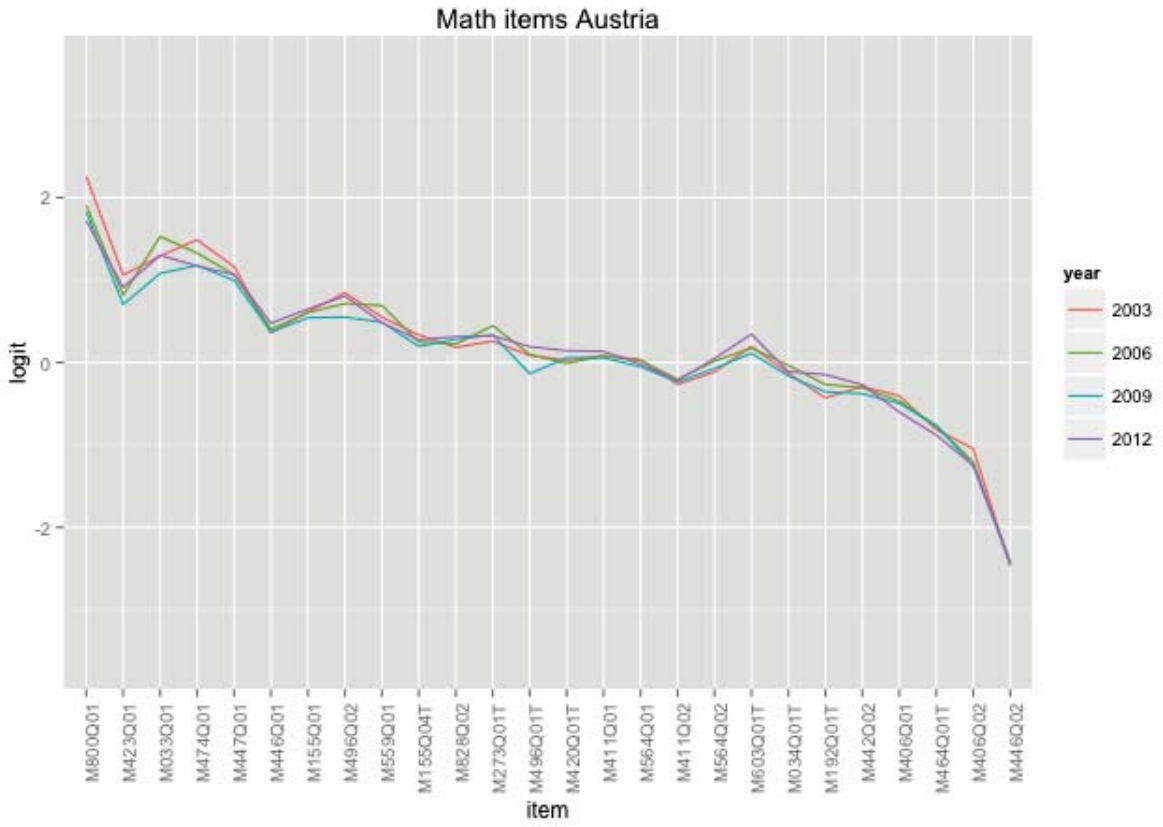
Wenn diese Bedingung erfüllt ist und man die dieselben Aufgaben in verschiedenen Ländern nach Schwierigkeit reiht, dann muss man dieselbe Reihenfolge erhalten.

In der folgenden Grafik wurden die Aufgaben nach der durchschnittlichen Schwierigkeit über alle Länder, die immer an PISA teilgenommen haben, geordnet und dann die logits der Lösungswahrscheinlichkeiten in Österreich eingetragen. Die Werte müßten nach rechts hin immer kleiner werden, weil Aufgaben, die im Länderschnitt schwieriger sind (also kleinere Lösungswahrscheinlichkeiten haben) auch in Österreich kleinere Lösungswahrscheinlichkeiten haben sollten.



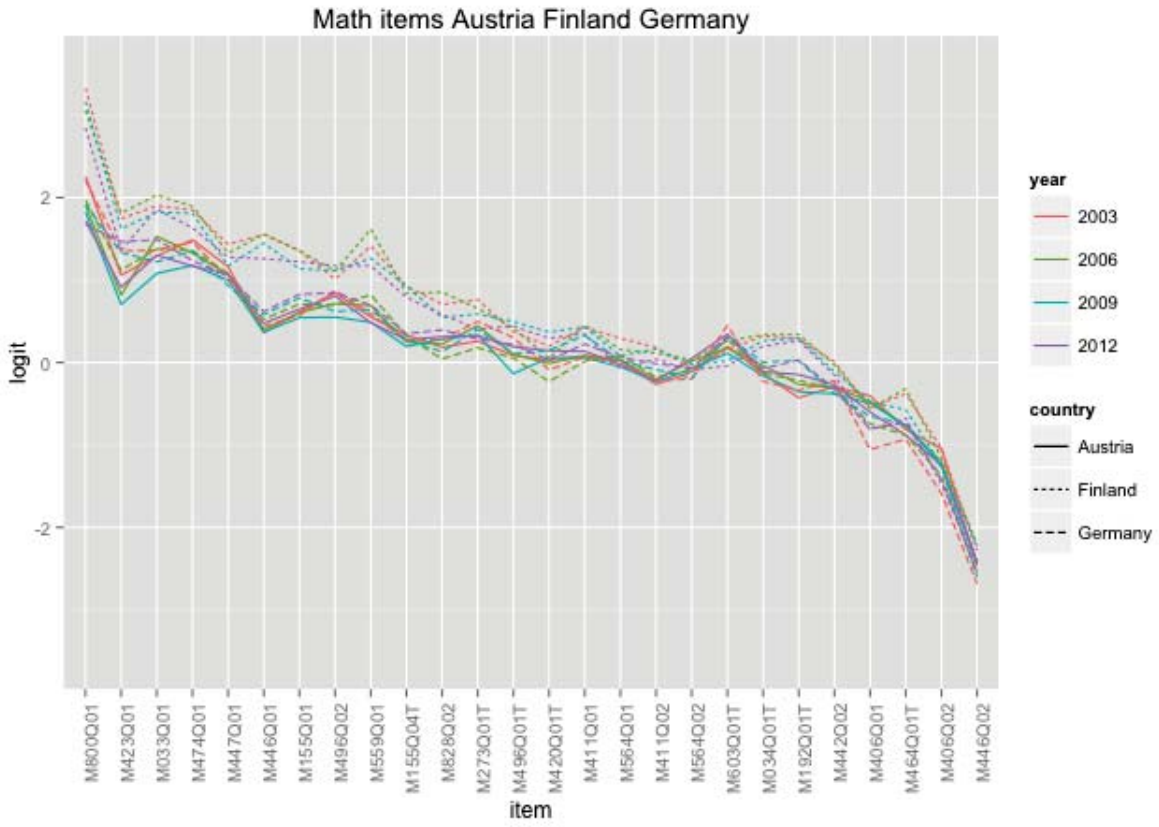
Man sieht, dass bei den Ergebnissen 2000 und 2003 Items „aus der Reihe tanzen“. Eigentlich sieht PISA in solchen Fällen vor, dass solche Items für einzelne Länder in der Berechnung nicht berücksichtig werden (laut Rasch-Modell ist das möglich). Aus ungeklärten Gründen ist das mit diesen Items für Österreich aber nicht geschehen.

Entfernt man diese Items, dann erhält man folgendes Bild:



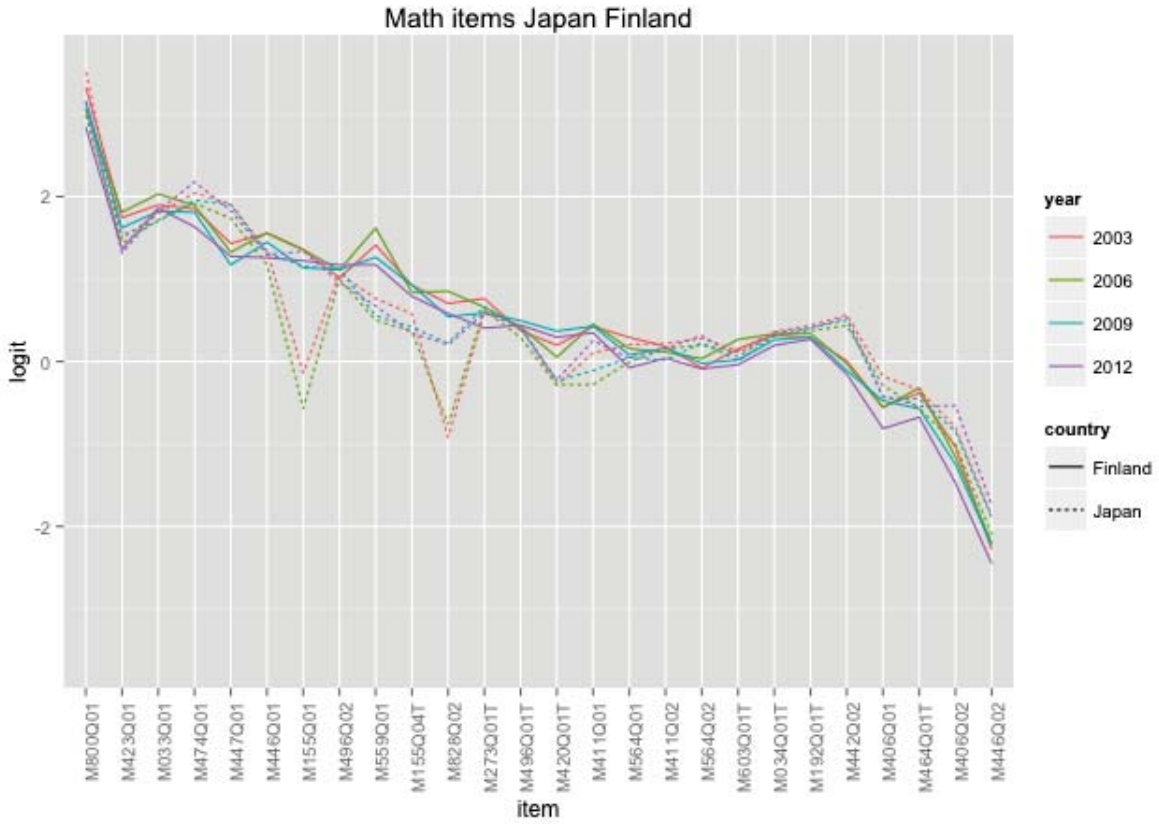
Die Monotoniebedingung ist dann einigermaßen erfüllt.

Laut Rasch-Modell müssten diese Kurven für veschiedene Länder und verschiedene Testzeitpunkte annähernd parallel sein. Ein Vergleich mehrerer Länder ergibt Folgendes:



Die Kurven sind annähernd parallel.

Sieht man sich jedoch die Grafiken von Finnland und Japan im Vergleich an,



dann sieht man, dass einige Items „nicht ins Muster passen“. Die Rasch-Homogenität der Items scheint beim Vergleich von Ländern mit extrem anderen Kulturen (vor allem Korea und Japan) doch nicht mehr gegeben zu sein.

4. Berechnung der Länderscores

Unter den Modellannahmen des Rasch-Modells kann man mit statistischen Methoden die Fähigkeiten der einzelnen Schüler schätzen. Allerdings geht PISA da noch etwas anders vor.

PISA nimmt an, dass die Fähigkeiten der Schüler von mit den Daten erhobenen Hintergrundvariablen wie Bildung der Eltern, wirtschaftliche Situation der Eltern usw. abhängen und bei „konstanten Nebenbedingungen“ einer Normalverteilung folgen. Unter diesen Annahmen wird die mathematisch formulierte Abhängigkeit der Fähigkeitsparamter von den Hintergrundvariablen geschätzt. Danach werden dann für die Schüler unter Verwendung des Testergebnisses und der Werte der Hintergrundvariablen „plausible Werte“ errechnet. Für jeden Schüler werden in jeder Domäne (Lesen, Mathematik, Naturwissenschaften) 5 plausible Werte errechnet und alle tiefergehenden Analysen in 5 Varianten durchgeführt und so konsolidierte Ergebnisse errechnet.

Warum ist bei PISA ziemlich komplexe statistische Methodik notwendig?

- In den einzelnen Ländern wird nur ein Teil der Schüler getestet
- Nicht alle Schüler erhalten dieselben Testaufgaben
- Manche Schüler lösen zwar keine Leseaufgaben, erhalten aber trotzdem einen Lesescore

Da die getesteten Schüler nur eine Auswahl aus allen Schülern sind, muss man statistische Stichprobenverfahren verwenden. Dabei ist es in Österreich 2000 zu einer Fehlgewichtung gekommen. Das errechnete Gewicht für die männlichen Berufsschüler war viel zu gering. Da die Berufsschüler die schlechtesten Testleistungen erbracht haben und zu gering gewichtet wurden, wurde das Österreich-Ergebnis zunächst überschätzt. Erst eine später durchgeführte Nachanalyse (Neuwirth et al (2006)) konnte diesen Methodenfehler richtigstellen. Die OECD hat daraufhin die Ergebnisse für Österreich 2000 offiziell korrigiert.

Das Rasch-Modell sichert, dass trotz verschiedener Aufgabenkombinationen die PISA-Punkte für die einzelnen Schüler vergleichbar sind.

Die Verwendung der Hintergrundvariablen macht es möglich, plausible Werte beispielsweise für Lesen auch für jene Schüler zu errechnen, die gar keine Leseaufgaben lösen mussten.

Das Rasch-Modell liefert zunächst Scores auf der logit-Skala, also im Bereich von -3 bis 3. Um die Werte besser darstellen zu können, werden sie reskaliert und zwar so, dass der Mittelwert aller OECD-Länder beim Test 2000 den Wert 500 und die Standardabweichung den Wert 100 ergaben.

Die entsprechenden Umrechungsfaktoren sind:

Domäne	Geschlecht	Faktor
Lesen	m	79.4
	w	80.2
Mathe		77.9
Science		93.2

Vereinfacht gesagt werden die logit-Scores der Schüler mit diesen Faktoren multipliziert und dann noch durch eine Addition eine Konstanten verschoben, um den Mittelwert 500 für die Ergebnisse von 2000 zu erreichen.

In den PISA-Dokumenten wird übrigens nicht sehr deutlich erklärt, warum die Umrechnungsfaktoren für Lesen für die Geschlechter verschieden sind.

Mit Hilfe dieser Tabelle kann man jedenfalls PISA-Punkte von Ländern in Unterschiede von Lösungswahrscheinlichkeiten umrechnen.

Domäne	Beispiel	+PISA	+p
Lesen	50%	3.2	1%
Mathe	50%	3.1	1%
Science	50%	3.8	1%
Lesen	50%	10	3.14%
Mathe	50%	10	3.20%
Science	50%	10	2.60%
Lesen	66%	10	2.78%
Mathe	66%	10	2.84%
Science	66%	10	2.38%
Lesen	75%	10	2.34%
Mathe	75%	10	2.40%
Science	75%	10	2.00%

Wenn in einem Land eine bestimmte Mathematik-Aufgabe von 50% der Schüler gelöst wird und ein anderes Land einen um 10 höheren PISA-Score hat, dann wird in diesem Land die Aufgabe von 53.2% der Schüler gelöst. 10 PISA-Punkte mehr in den Naturwissenschaften bedeuten aber nur eine um 2.6% höhere Lösungswahrscheinlichkeit (bei einer 50%-Aufgabe).

5. Einige Ergebnisse

Leider werden nicht alle in den PISA-Analysen verwendeten Rohdaten auch vollständig publiziert. Deswegen können nicht alle Analysen, die von Interesse wären, auch durchgeführt werden.

Und gelegentlich werden über PISA-Ergebnisse auch Behauptungen aufgestellt, die einer genaueren Analyse nicht standhalten.

Man konnte z.B. in Zeitungen lesen, dass die österreichischen AHS-Schüler bei PISA schlechter abschneiden als Finnland insgesamt.

Das lässt sich einfach mit der folgenden Tabelle widerlegen:

Jahr	Gruppe	Lesen	Mathe	Naturw
2000	Finn alle	546.5	536.2	537.7
	Ö AHS	564.2	564.2	571.8
	Ö BHS	544.3	552.6	556.0
	Ö alle	492.1	502.5	504.7
2003	Finn alle	543.5	544.3	548.2
	Ö AHS	571.6	570.7	566.1
	Ö BHS	544.7	553.5	540.2
	Ö alle	490.7	505.6	491.0
2006	Finn alle	546.9	548.4	563.3
	Ö AHS	568.6	568.5	578.1
	Ö BHS	541.5	553.1	558.1
	Ö alle	490.2	505.5	510.8
2009	Finn alle	535.9	540.5	554.1
	Ö AHS	550.4	568.0	568.3
	Ö BHS	515.7	541.6	540.3
	Ö alle	470.3	495.9	494.3

Zum Zeitpunkt dieser Analyse waren die Österreich-Daten noch nicht in der notwendigen Feingliederung verfügbar.

Zum Abschluss sehen wir uns noch alle bisher vorliegenden PISA-Ergebnisse an. In der Diskussion der Ergebnisse für 2012 in der Öffentlichkeit wurde mehrfach gesagt und geschrieben, dass wir uns im Vergleich zu 2009 merkbar verbessert hätten. Das stimmt zwar im Prinzip. Allerdings gab es 2009

einen Boykott-Aufruf und die OECD stellte fest, dass die Ergebnisse von damals nur sehr bedingt für Langfristvergleiche taugen.

Hier nun die österreichischen Ergebnisse:

Jahr	Lesen	Mathe	Naturw
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

Getrennt nach Geschlechtern sehen die Ergebnisse so aus:

Jahr	L w	L m	M w	M m	N w	N m
2000	509.2	475.8	492.5	512.0	502.2	507.1
2003	514.4	467.1	501.8	509.4	492.3	489.7
2006	512.9	468.3	494.0	516.6	507.0	514.5
2009	490.5	449.3	486.5	505.7	490.5	498.3
2012	508.0	471.1	494.5	516.7	501.5	510.1

Die beiden Tabellem zeigen, dass in Lesen und Mathematik das Jahr 2009 ein Ausreißer war. Bei allen anderen Testzeitpunkten ist das Ergebnis praktisch konstant.

Bei dieser Datengrundlage von einer Verbesserung zu sprechen erscheint sachlich kaum gerechtfertigt.

Literatur

Erich Neuwirth, Wilfried Grossmann und Ivo Ponocny. *PISA 2000 und PISA 2003: Vertiefende Analysen und Beiträge zur Methodik* Leykam. Graz (2006).

OECD. *PISA-Website mit Daten und Ergebnissen* `www.pisa.oecd.org` OECD. Paris (2013).

bifie. *bifie-Website mit österreichischen Ergebnissen* `www.bifie.at` bifie. Salzburg (2013).