ISSN 0940-5151

Economic Quality Control
Vol 16 (2001), No. 2, 227 – 253

OECD PISA - An Example of Stochastic Illiteracy?

Elart von Collani

Abstract: PISA stands for "Programme for International Student Assessment" performed by the OECD and consisting of a cyclic evaluation of the new generations' basic competences. The first evaluation took place in 2000 and the results were published on 4 December 2001. According to the official statements an overall of 32 countries participated in this first study. The publication of the results in Germany caused some great excitement, as Germany the "country of arts and science" was ranked at the lower edge not so far from Brazil. Politicians, teachers, scientists and any other people expressed their opinions, looked for those responsible for the bad state and demanded immediate actions for improving the education in Germany.

Bavaria, the German state which pretends to have the best educational system, at least in Germany, claimed that the Turkish and Jugoslavian students in Germany had caused the bad grade. Other being more cautious demanded better teachers, another educational system and better pupils. The nationwide discussion arouse my curiosity and thus, I tried to get some details on PISA and found them e.g. on the homepage of the OECD, the renowned Max-Planck-Institut Berlin (http://www.mpib-berlin.mpg.de/pisa) and other national institutions in the participating countries.

The study has an ambitious aim and is essentially a statistical one. Therefore, a thorough evaluation of the quality of the study and its results should stand at the beginning of any discussion. This paper looks at PISA exclusively from a statistical point of view and makes an attempt to evaluate its quality. It is an attempt, because only very limited information on the applied statistical procedures were available despite of the fact that hundred of pages describing the study are made available through internet.

1 Aim and Methods of PISA

The following information are taken from papers published on the internet and constitute the official statements of the national bodies being responsible for the study.

The study aims at making available early indicators for the national governments of the participating countries, which can be used for improving the national educational systems. The program refers to "reading literacy", "mathematical literacy" and finally to "scientific literacy". The study is performed with 15 year old students, who, in most of the OECD countries, are still obtaining compulsory education.

According to the official document [1] there were 32 participating countries. In each of them between 4.500 and 10.000 students, forming a representative sample, took part in the evaluation. The competence tests have been developed by the participating countries and are a mixture of multiple choice questions and questions the students have to answer on their own. The test takes about two hours, moreover each student has to fill in a questionnaire about her/himself and the same holds for the headmasters of the participating schools.

According to [1] about 180.000 students underwent the test in spring 2000. In Germany a supplementary study comprising about 50.000 students has been performed for comparing the educational systems in the different German federal states.

For guaranteeing comparable conditions, there were quality control activities. In Germany independent control persons visited 35 schools acknowledging correctness.

Reading the official German document [1], one experiences a first surprise when counting the countries in the tables and figures displaying the test results. There are only 31 countries and not as announced 32, and even a thorough search in [1] for the 32nd country was not successful. The 31 in the order of the average achievements in "reading literacy" are the following:

Table 1: PISA-results in reading literacy as displayed in [1].

1	Finland	2	Canada	3	New Zealand
4	Australia	5	Ireland	6	Korea
7	United Kingdom	8	Japan	9	Sweden
10	Austria	11	Belgium	12	Iceland
13	Norway	14	France	15	United States
16	Denmark	17	Switzerland	18	Spain
19	Czech Republic	20	Italy	21	Germany
22	Liechtenstein	23	Hungary	24	Poland
25	Greece	26	Portugal	27	Russian Federation
28	Latvia	29	Luxembourg	30	Mexiko
31	Brazil				

The second surprise appears when trying to figure out the sample sizes used in the various countries. In the official report for Germany [1] neither the exact German sample size nor those of the other countries can be found.

Turning to the competence tests themselves and looking for the evaluation scheme in order to get some feeling for the possible probability distribution one cannot find the slightest hint of it. There is a percent distribution of different types of tasks, but nothing specifying how the student's achievements were evaluated.

2 Participating Countries

The discrepancy about the participating countries may be not essential for validity of the study, but it characterizes the care it has been performed. The OECD homepage *http://www.pisa.oecd.org/* should solve the mystery about the missing 32nd country. And in fact the relevant page "Participating countries" contains 32 countries where the non-members of OECD are marked by an asterisk.

Table 2: Participating countries in alphabetical order as given by the official OECD-document.

1	Australia	2	Austria	3	Belgium
4	Brazil*	5	Canada	6	China*
7	Czech Republic	8	Denmark	9	Finland
10	France	11	Germany	12	Greece
13	Hungary	14	Iceland	15	Ireland
16	Italy	17	Japan	18	Korea
19	Latvia*	20	Luxembourg	21	Mexico
22	The Netherlands	23	New Zealand	24	Norway
25	Poland	26	Portugal	27	Russian Federation*
28	Spain	29	Sweden	30	Switzerland
31	United Kingdom	32	United States		

However, a closer look at the two tables increases the mystery. The announcement of the OECD contains The Netherlands and China, which are not included in the German document. On the other hand, the German document contains Liechtenstein, which is not listed by the OECD.

The search in the OECD and the German internet documents for the reasons of this discrepancy remained in vain. A partial solution was supplied by the official document of Ireland [4] with Table 1.1. "Countries Participating in Pisa 2000".

Table 3: Participating countries as displayed in [4].

OECD Countries						Non-OECD Countries	
1	Australia	11	Hungary	21	Norway	29	Brazil
2	Austria	12	Iceland	22	Poland	30	Latvia
3	Belgium	13	Ireland	23	Portugal	31	Liechtenstein
4	Canada	14	Italy	24	Spain	32	Russian Federation
5	Czech Republic	15	Japan	25	Sweden		
6	Denmark	16	Korea	26	Switzerland		
7	Finland	17	Luxembourg	27	United States		
8	France	18	Mexico	28	United Kingdom		
9	Germany	19	New Zealand				
10	Greece	20	Netherlands*				

* The school response rate for the Netherlands was too low to permit the computation of reliable student achievement estimates.

From the remark given on the bottom of the table, we conclude that there are requirements formulated with respect to the sampling plan and that the Netherlands did not meet these requirements. The questions at which stage China was excluded from the program, and at which stage Liechtenstein was included remain unanswered. The other question, why the (statistical) requirements set by the OECD for the study are not mentioned very clearly in each of the national studies remains unanswered, too.

3 Sample Sizes

The official report of the United States "Outcome of Learning" [6] proved to contain much more information about the sample design than the other reports. Appendix 1 of the U.S.-report contains the *Table A1.1.-Coverage of target population, student and school sample, and participation rates, by country: 2000* with the statement of sample sizes used in the different countries.

Table 4: Sample sizes by country as stated in the United States' Report.

Country	Number	Country	Number
Australia	5.154	Mexico	4.600
Austria	4.745	Netherlands	2.503
Belgium	6.648	New Zealand	3.667
Belgium (Flemish)	3.874	Norway	4.147
Belgium (French)	2.774	Poland	3.639
Canada	29.461	Portugal	4.517
Czech Republic	5.343	Spain	6.214
Denmark	4.212	Sweden	4.416
Finland	4.864	Switzerland	6.084
France	4.647	United Kingdom	9.250
Germany	4.983	England	4.099
Greece	4.672	Northern Ireland	2.825
Hungary	4.883	Scotland	2.326
Iceland	3.372	United States	3.700
Ireland	3.786	Brazil	4.885
Italy	4.984	Latvia	3.915
Japan	5.256	Liechtenstein	314
Korea	4.982	Russian Federation	6.701
Luxembourg	3.434		

The sample sizes in the different countries range from 314 (Liechtenstein) to 29.461 (Canada). Only in 19 countries the announced target of a sample size between 4.500 and 10.000 is met. In one country (Canada) the sample size is considerably larger and in the others smaller than anticipated.

Another list of the sample sizes at least of the OECD-countries is provided in the reports of England and Northern Ireland [8] and [9].

Table 5: Sample sizes by country as given by England.

Australia	5176	Japan	5256
Austria	4745	Korea	4982
Belgium	6670	Luxembourg	3528
Canada	29687	Mexico	4600
Czech Republic	5365	New Zealand	3667
Denmark	4235	Norway	4147
England	4120	Poland	3654
Finland	4864	Portugal	4585
France	4673	Ireland	3854
Germany	5073	Spain	6214
Greece	3644	Sweden	4416
Hungary	4887	Switzerland	6100
Iceland	3372	United Kingdom	9340
Italy	4984	United States	3846

There are 28 sample sizes given in the British reports. From these only 11 numbers coincide with those given in the US report, which cites an OECD report as source. Some of the differences are small, but others are large. In the case of Greece the British report states $n = 3.644$, while the US-source gives the number of $n = 4.672$.

One possible explanation would be that in various stages of data analysis students were excluded or included into the data set. If this is correct then one has to state that any subsequent data manipulation is extremely dangerous and should be avoided by any means.

There is another noteworthy information contained in Table B2 of the British reports:

Table 6: Sample sizes in England for each literacy domain

Literacy domain	Girls	Boys	Total*
Reading literacy	2034	2033	4120
Mathematical literacy	1131	1130	2292
Science literacy	1140	1117	2284

* The total includes 1% of students who did not give information on their gender.

Table 7: Sample sizes in Northern Ireland for each literacy domain

Literacy domain	Girls	Boys	Total*
Reading literacy	1361	1468	2849
Mathematical literacy	757	816	1586
Science literacy	765	793	1566

* The total includes 1% of students who did not give information on their gender.

The considerable difference in the sample size between the different literacy domains cannot be seen from the results as displayed in the reports due to data transformations which are kept secret. There are some vague hints from which one can conclude that the different tests were performed with varying sample size, but no clear information.

The sample size is one of the very relevant information for any statistical study. If it is not stated in a study utmost caution is advised. The facts

- that the exact sample sizes in most of the national documents are missing, and
- that there are different sample sizes stated in different documents

must arouse doubts in the reliability of the OECD-study.

4 The Sample Design

Besides the sample size the sample design is of eminent importance for a sampling procedure. Especially for the extremely heterogeneous characteristic as students' competence an inappropriate sampling design will necessarily yield totally wrong results. According to the national reports, e.g. [1], a "representative sample" was used in the different countries. However, the meaning of "representative" remains in the different documents vague and

4.1 General Comments

For instance in [2], there are some comments and explanations concerning the notion of a "representative sampling design". It is said that particularly for small sample sizes a random sample and a representative sample contradict each other to a certain extent. To clear the situation both notions are explained below.

A random sample of size n from a population of size N is called "random sample", if any of the possible n-tupels of different population elements has the same probability to be drawn.

Assume the population elements numbered from 1 to N, thus it may be represented by the set $\{1, 2, \ldots, N\}$ and let $\vec{X} = (X_1, \ldots; X_n)^T$ denote the sample, then

$$P_{\vec{X}}\Big(\{(i_1, \ldots, i_n)\}\Big) = \frac{1}{\binom{N}{n}} \tag{1}$$

with $1 \leq i_1 < \ldots < i_n \leq N$ characterizes a random sample.

Next, divide the population in m disjoint sub-populations of size $N_1, N_2, \ldots, N_m$:

$$\{1, \ldots, N_1\}, \{N_1 + 1, \ldots, N_1 + N_2\}, \ldots, \{\sum_{i=1}^{m-1} N_i + 1, \ldots, \sum_{i=1}^{m} N_i\} \tag{2}$$

The above defined sub-populations are often called "strata". Consider for stratum no. i a random sample $(X_{i,1}, \ldots, X_{i,n_i})$ with stratum sample size

$$n_i \quad \text{for } i = 1, \ldots, m \tag{3}$$

where n_i denotes the closest integer to $\frac{N_i}{N} n$ with $\sum_{i=1}^{m} n_i = n$. Let each population element be denoted by a capital X, then the overall sample is given by

$$(X_{1,1}, \ldots, X_{1,n_1}, X_{2,1}, \ldots, X_{2,n_2}, \ldots, X_{m,1}, \ldots, X_{m,n_m}) \tag{4}$$

and consists of m sub-samples where the following relation holds between the sub-populations and the respective sub-samples:

$$\frac{N_i}{N} \approx \frac{n_i}{n} \quad \text{with } n = \sum_{i=1}^{m} n_i \tag{5}$$

In this case the sample (4) is called representative for the partition (2). Often a sample of type (4) is called proportional stratified sample.

There might be two different reasons for defining strata and using a representative sample.

- Some dimensions are not only of interest for the total population, but also the strata. In this case the strata need not to be determined, but are given. For example, the extended PISA-study in Germany aimed at comparing the educational systems of the different German federal states. Thus, each state constitutes a given stratum.
- If the variability of the aspect of interest in the total population is large, one can divide the population into strata by means of some other aspects where each stratum is more homogeneous with respect to the dimension of interest than the whole population. In this case the problem is to define strata characteristics appropriately, which means:
  - The actual values of the characteristics defining the strata have to be known for each element of the population.
  - There should be a strong relation between strata characteristics and dimension of interest, for instance, expressed by a high value of the correlation coefficient.
  - The strata should be relatively homogenous with respect to the strata characteristics.

  The last condition in conjunction with the second one implies that within each stratum the variability of the dimension of interest is reduced compared to the total population, thus enabling more accurate stochastic procedures. If the last two conditions are not fulfilled, stratification does neither increase nor decrease accuracy.

As one can see from the above, there is no contradiction, neither for large nor for small sample sizes, between a random sample and a representative sample. The only difference is the fact that in the latter case the overall random sample is divided into random sub-samples. Without randomness of the sample no statistical procedure will result in meaningful results, therefore, assuring randomness is one of the most decisive problems when applying statistical procedures.

4.2 The U.S. Sample Design

More than half of the participating countries and among them the four non-OECD members did not publish a national report. There are national reports from the following countries:

Table 8: Countries which published in December 2001 national reports.

Austria	Canada	Czech Republic
Denmark	Finland	Germany
Ireland	New Zealand	Norway
Spain	Sweden	Switzerland
United Kingdom	United States	

Because of language problems only the reports of English and German speaking countries could be considered for this investigation. As already mentioned, these national reports contain only very few and vague information about the sample design actually used. Among them the best documented report comes from the United States and, therefore, it is taken here to illustrate the proceedings, the problems and the deficiencies of the studies.

The sampling design used in the U.S.-study is outlined in Appendix 1 of [6]. The sampling design is based on three stages.

- The first stage was a sample of "primary sampling units" (geographic areas referred to as PSUs).
- The second stage was a sample of schools within PSUs.
- The third stage was a sample of students from the set of all students enrolled in the school who were born in the calendar year 1984.

4.2.1 The first stage of sampling

There are no information given about the principles for dividing the United States into PSUs. Neither the overall number of PSUs is stated nor the reason for the sample size of $n = 52$. From a detail about the second stage of sampling one can conclude that at least 33 states were represented in some form in the sample of PSUs.

The significance of the PSUs remains unclear. In contrast to other countries which used a stratification, the PSUs do not constitute strata, as only a sample of them were considered.

The question about the PSUs which are not considered for the study remains unanswered. If the number of 33 states represented in the sample is correct, then 17 states, i.e. one third, are not represented in the U.S. study.

Thus, because of lack of information, a quality evaluation of this stage is not possible, although it could be decisive for the whole study.

4.2.2 The second stage of sampling

During the second stage, a total sample of 220 schools was selected from within the sampled PSUs. It is stated that the selected schools were located in 33 different U.S. states.

However, there are no information given about the sample sizes for each selected PSU and the mode of sampling is not mentioned explicitly. It is said:

> "In the United States, the public and private schools selected for PISA constituted a nationally representative sample of all schools in the country enrolling 15-year-olds."

Note that the meaning of "representative" is totally unclear here, because no stratification has been mentioned so far.

In the second stage, besides the 220 schools as a supplement, replacement schools were selected. Each school in the original sample was assigned to up to two replacement schools selected from the set of "neighboring schools" on the sampling frame. Again there are no information, e.g. about the meaning of "neighboring school's".

The replacement schools were necessary in case original sample schools would fail to get into the sample. The following illustrates that many of the schools selected were in fact not be included in the sample.

- Ten of the 220 schools in the original sample were ineligible because they did not have any students born in 1984. At this point the question about school types arises and, moreover, the claim of a representative sample becomes doubtful. It seems to be rather strange that about 5% of the sampled schools have no 15-year old students although they should constitute a "representative sample" "of all schools in the country enrolling 15-year-olds".
- A further 82 schools refused to participate. In other words almost 5% of the schools selected for the sample were not eligible and 39% of the eligible schools selected for the sample refused to cooperate.

In Appendix 1 of the national U.S. report it is written:

> "A minimum response rate target of 85 percent was required for initially selected educational institutions. In instances in which the initial response rate of educational institutions was between 65 and 85 percent, an acceptable school response rate could still be achieved through the use of replacement schools."

In the case of the United States the lower bound of 65% was not reached. Consequently, the United States should have been excluded from the study. However, this was not the case!

"Thirty-two replacement schools agreed to participate with the result that 160 schools in total agreed to participate in the study." This formulation in Appendix 1 of [6] implies that the 32 replacement schools added to the sample were not randomly selected, but asked for participation. The exact proceeding is unclear and it is unknown to which extent the resulting sample of 160 schools is "representative" for the nationwide schools. Moreover, the resulting sample of schools is by no means a "random sample", which is a necessary condition for obtaining meaningful results by applying statistical procedures.

4.2.3 The third stage of sampling

The third stage consisted of selecting a random sample of "up to 35" of the students born in 1984 for each of the 160 schools. There are no further information about the sampling procedure, e.g. in how many school no random sample could be drawn because of a too small number of 15-year-olds.

However, the following is said about the student response rate in the third stage of sampling:

> "Following data collection, decisions by the international Technical Advisory Group (made up of technical advisors from the PISA Consortium) reduced the number of "participating" schools based on the student response rates within schools.
>
> - Schools with more than 50 percent student participation were classified as "responding schools".
> - Schools in which 25 to 50 percent of sampled students participated were classified as "partially responding".
> - Schools with less than 25 percent student participation were treated as "nonresponding", and data from these schools were deleted from the database.
>
> In the United States the number of (original/replacement) schools falling into these categories was as follows:
>
> - responding (116/29);
> - partially responding (7/1);
> - nonresponding (5/2).

> For the purpose of calculating school response rates only the 145 responding schools (116 originals plus 29 replacements) were counted. On this basis the school response rate before replacement was 56 percent, and after replacement became 70 percent."

Remark:
Maintaining after all these manipulation the claim of a random sample or a "representative" sample in whatever sense seems to be rather inappropriate. Moreover, the question arises how the stated numbers were calculated:

- From the originally sampled 220 schools only 116 schools were classified as responding schools. Thus the response rate before replacement was 52.7%.
- From the originally sampled 220 schools, only 210 were eligible implying that the response rate of the eligible schools is 55.2%.
- Finally, the number of 70% is also not reproducible, as it is not said how the 32 replacement schools were selected for the sample.

The fact that the statements made in the report are at best only approximately reproducible characterizes the quality of the report. •

Evidently, the definition of a population is of utmost importance when making statements about it. The report states: "Eligible students were defined as those born in 1984 ...". For determining the average competences a sample of the defined population should be used.

Taking as basis 160 schools and 35 students per school yields an overall sample size of about 5600 students with some 4752 students from the 145 responding schools. Evidently, these 4752 students constitute not the originally planned sample, but at least a sample in whatever sense from the population of 15-year-olds in the United States. However, even this reduced sample was subsequently step by step further reduced as follows:

- Some 221 of these students were defined as ineligible and/or were withdrawn.
- Exclusion decisions by the schools resulted in a further 211 students being excluded from the assessment.
- Moreover, 620 students failed to take the assessment due to absence and /or parent/student refusals.

Thus, at the end only 3700 students from the 145 responding schools were assessed. The question, which part of the "eligible students" are represented by these students remains open.

The result of the assessment is commented in the report in the following way:

> "The weighted number of students assessed, expressed as a percentage of the weighted number of eligible students, gave the student response rate of 85 percent, a rate which exceeds the PISA international standard of 80 percent. In addition, 146 students in the partially responding schools took the assessment giving a total of 3846 students taking PISA assessment in the United States. All 3846 students are included in the international database."

Remark:
The value of 85% for the student response rate in the case of the United States is another doubtful issue. Although not stated in the document, it probably starts with the 4752 students of the responding schools and could be calculated in the following way:

$$\frac{4752 - 221 - 211 - 620}{4752 - 221 - 211} \cdot 100 = 85.6\% \tag{6}$$

If one would take the original sample of students from the 160 schools which had agreed in participating in the study, then the response rate would be much less. Moreover, the proceeding to define the eligible students wrongly as those born in 1984 and subsequently screening the sample for those students which cannot "meaningfully participate in the assessment" opens way for any manipulation and constitutes a serious methodological mistake. •

Additionally, the 146 students of partially responding schools, who took the assessment and were included in the international database, cause a major problem for a statistical analysis. They represent within their schools a minority of the "eligible students" and it is to be expected that their agreement to participate is not independent of their intellectual capabilities. Therefore, it is to be expected that including them means to risk an additional bias.

There is another problem concerning disabled students or the definition of the "population" to be assessed. In Germany there are special schools for these students called *Sonderschule*. As reported in the United States and in many other countries these students were declared ineligible. According to the German report [1] a different procedure was applied at Sonderschulen:

> "In Sonderschulen a shortened version of the international test was used, and also the questionnaire was reduced to a minimum so that the test-time took only twenty minutes. There was no second test-day in Sonderschulen."

The question arises whether there were different national definitions of eligible students on the one hand and whether there were different national tests used in the various countries on the other hand.

The notion of a "weighted number of eligible students", the definition of "responding", "partially responding", "nonresponding" schools and the proceeding of stepwise reducing the sample increase the doubts that it is possible to specify the part of the 15-year-olds in the United States which is represented by the sample.

A final comment in the U.S. report reads as follows:

> "While the students response rate exceeds both NCES and PISA standards, the school response rate of 56 percent before replacement fails to meet these standards. In the case of PISA a rate of 65 percent was required. The United Kingdom and the Netherlands also fell below the PISA standard for response rates."

5 Technical Standards

In [10] some technical standards for the national studies are given. Two of these technical standards are the following:

- A minimum sample size of 4.500 assessed students must be selected from a minimum of 150 schools.
- Further, if a minimum sample size of students is obtained (4.500), the sample must not depart significantly from a self-weighted design.

Still, according to Table A1.1 of the U.S. report, the total population of 15-year-olds in Iceland amounts to 4.062 and in Liechtenstein to 415. Moreover, Greece, Ireland, Japan, Korea, Luxembourg, Poland, Portugal and the United States did not fulfill the "minimum standard" of the number of schools. Additionally, Denmark, New Zealand, Norway, Sweden and Latvia did not meet the minimum sample size of 4.500 students.

As a consequence, at least 15 more countries (besides the Netherlands) should have been taken out of the study.

6 The Result of the Study

Most of the results stated in the PISA-study cannot be evaluated as the necessary information about model, sample size, methods, etc. are not made available. Therefore, the following investigations are restricted to the main statement with respect to the reading literacy in the 32 countries under consideration.

As already criticized almost no absolute data are given in the OECD report or in any of the national reports. What is stated in the reports are transformed and exclusively relative data. Neither the mean scores, nor the "standard deviation" SD, nor the "means standard error" SE which are displayed in the reports are absolute numbers. Even worse, in none of the national reports the formulae for these quantities are stated. Therefore, the meaning of any of the given quantities remains vague.

The following quotations from the U.S. report shall illustrate the lack of relevant information in the reports. "Scoring", i.e. the way how to evaluate the test papers, is explained as follows:

> "Scoring
>
> PISA's assessment of reading included 270 minutes of testing time, of which 45 percent was devoted to items requiring open-ended responses. The mathematics and science tests included 60 minutes of testings time, of which 35 percent was assessed through open-ended items. The process of scoring these items was an important step in ensuring the quality and comparability of the PISA data."

The German report contains a list of 11 different tasks within the reading test with their percentages within the overall test, but no hint of the maximum number of credits neither for the single tasks nor for the overall test. These information in conjunction with the number of credits actually obtained by the students would enable to evaluate the tests and the tasks itself. It is a strange fact that these basic information are not given for the PISA study.

Besides scoring, the weighting procedures used in the PISA study are important as they determine to a large extent the final results. Similar to any other relevant information, the used weights are not stated. Instead the following explanation can be found in the U.S. report:

> "Weighting
>
> Students included in the final PISA sample for a given country are not at all equally representative of the full student population, even though random sampling of schools and students is used to select the sample. The use of sampling weights is necessary for the computation of statistically sound, nationally representative estimators. Survey weights help adjust for intentional over- or under-sampling of certain sectors of the population, school or student nonresponse, or errors in estimating size of a school at the time of sampling."

As neither strata nor their number is given in the report the meaning of "representative" is unclear and the weights are not defined. Analogously to any other relevant information, the weights used for calculating the final results are not stated.

Notwithstanding these and other objections, the results for reading literacy as given in the various reports shall be analyzed here. There are slightly different representations of the results in the various national reports, however, the stated numbers seem to be the same.

Table 9: The results with respect to the average reading literacy as given in [1]).

Country	Average M	Means Standard Error SE	Standard Deviation SD
Finland	546	2.6	89
Canada	534	1.6	95
New Zealand	529	2.8	108
Australia	528	3.5	102
Ireland	527	3.2	94
Korea	525	2.4	70
United Kingdom	523	2.6	100
Japan	522	5.2	86
Sweden	516	2.2	92
Austria	507	2.4	93
Belgium	507	3.6	107
Island	507	1.5	92
Norway	507	2.8	104
France	505	2.7	92
United States	504	7.0	105
OECD-average	500	0.6	100
Denmark	487	2.4	98
Switzerland	494	4.2	102
Spain	493	2.7	85
Czech Republic	492	2.4	96
Italy	487	2.9	91
Germany	484	2.5	111
Liechtenstein	484	4.1	96
Hungary	480	4.0	94
Poland	479	4.5	100
Greece	474	5.0	97
Portugal	470	4.5	97
Russia	462	4.2	92
Latvia	458	5.3	102
Luxembourg	441	1.6	100
Mexico	422	3.3	86
Brazil	396	3.1	86

Note that the averages given in the above table represent weighted, transformed and relative data and an interpretation is hardly possible. However, they allow a ranking and this ranking caused a lot of excitement in the countries involved. In order to specify the meaning of any ranking procedure by means of Table 9 consider one student (or equivalently) one country and let him/her/it participate in several reading tests performed independently by different examiners. It is to be expected that the student/country will get in each test a different score and one could list them and use them for ranking. Of course, ranking would make no sense as the reading literacy of the student/country is always the same and the different results only reflect the inherent variability in the

"random experiment" of testing. The same could be true for the results in Table 9. The only reliable statement one can make about the data in Table 9 says that none of the stated figures is equal to the average reading literacy to be determined no matter how it is defined.

To rank Australia with respect to the average reading literacy in front of Ireland, Korea or the United Kingdom is totally unfounded and by no means justified by the results given in Table 9.

Assuming for the time being that an appropriate model was applied and sampling and statistical analysis have been performed in a scientifically satisfactory way and relying on the stated standard errors SE of the average scores of the different countries, then in case of the United States, the following statements would be justified based on standard t-tests with significance level 5% as described in Appendix 1 of [6].

With respect to the average reading literacy of 15-year-olds a pairwise comparison between the United States and the other countries shows the following result:

- Finland, Canada, New Zealand, Australia, Ireland, Korea, United Kingdom and Japan perform better than the United States.
- There is no evidence of a difference between the United States and the following countries: Sweden, Austria, Belgium, Iceland, Norway, France, Denmark, Switzerland, Spain and Czech Republic.
- Italy, Germany, Liechtenstein, Hungary, Greece, Portugal, Luxembourg, Mexico, Russian Federation, Latvia, Brazil perform worse than the United States.

If instead of a pairwise comparison a multiple comparison is made and a procedure with the same reliability, i.e. 95%, is applied, then the following statement would be justified:

With respect to the average reading literacy of 15-year-olds a multiple comparison between the United States and the 30 other countries shows the following result:

- Finland, Canada, New Zealand, Australia and Ireland perform better than the United States.
- There is no evidence of a difference between the United States and the following countries: Korea, United Kingdom, Japan, Sweden, Austria, Belgium, Iceland, Norway, France, Denmark, Switzerland, Spain, Czech Republic, Italy, Germany and Liechtenstein.
- Hungary, Greece, Portugal, Luxembourg, Mexico, Russian Federation, Latvia, Brazil perform worse than the United States.

As mentioned, the above results are obtained following precisely the procedures described in Appendix 1 of [6]. In contrast, in [6] itself the following statement about the U.S. performance is made:

> "U.S. students perform better than students in the OECD countries Greece, Luxembourg, Mexico and Portugal, and the non-OECD nations Brazil, Latvia, and the Russian Federation. Students in Canada, Finland, and New Zealand outperform U.S. students. U.S. students perform at about the same level as the other 19 participating OECD countries and Liechtenstein."

As for the presentation of the results, any ranking 1 to 31 constitutes more or less a pure random order of the countries and should by no means be taken seriously. However, the above statements about the U.S. performance is only valid if the above assumptions are met by the study. In the following sections an attempt is made to check the assumption by deriving a model for the situation aiming at determining an "average literacy" within a country.

7 The Stochastic Model in the PISA Study

As a rule the results of a statistical procedure are meaningful only in conjunction with a stochastic model. Unfortunately, in the official national reports there are almost no hints about the underlying model. There is a short section entitled "Statistical Procedures" in Appendix 1 of the U.S. report from which one can infer that the normal model was assumed for the analysis.

Moreover, in [2] there is a section entitled "How is the accuracy of a sample estimation determined?" which contains some hints concerning the model and basically refers to the *Central Limit Theorem*and describes the way how to analyze the data. It is claimed:

> "However, it is of fundamental importance that the standard deviation (variance) of the means $\sigma^2_{\bar{X}}$ follows a law:
>
> $$\sigma^2_{\bar{X}} = \frac{\sigma^2_X}{n} \tag{7}$$
>
> where σ^2_X is the variance of the characteristic in the population. Thus, the variance of the means gets smaller the larger the sample size is (since n is in the denominator) and the smaller the characteristic varies in the population. Formally, this law is only valid for characteristics which are normally distributed in the population, i.e. follow the Gauss distribution (as displayed on any 10 DM bill). However, the central limit theorem also states that the distribution of the means with increasing sample size n is normally distributed even if the characteristic is not normally distributed in the population. This is in particular important, since in practice the population distribution is unknown. Therefore, whether for instance the test achievements of the students in a study are normally distributed or not is in case of a large sample size (> 1000) almost irrelevant for the distribution of the mean."

From the above quotation it is possible to infer the following:

- The assumed model for the "mean" is the normal one.
- The performance of the students of the population has been modelled by independent and identically distributed random variables.

The second implication holds, since in contrast to the above quotation, formula (7) only holds in the case of independent and identically distributed random variables. With other words neither the central limit theorem nor the normal distribution are necessary for (7).

In order to judge the asumption of independent and identically distribute random variables, imagine the population of eligible 15-year-olds (and as we have seen, this population is not at all identical with the population of 15-year-olds in a country, as claimed in each of the national reports) numbered from 1 to N, and denote by X_j the random achievement of the jth student in the reading literacy test. Then (7) assumes that the random variables $X_1, \ldots, X_N$ are independent and identically distributed, a condition, which, of course, is not at all met, but greatly violated. Therefore, basing data analysis on (7) leads to meaningless results.

In view of (7) there is another strange observation referring to the confidence intervals given in the various national reports. These confidence intervals are calculated as follows:

$$[\bar{X}_1 - 2SE, \bar{X}_1 + 2SE] \tag{8}$$

Assuming that (7) is applied to calculate the standard error of the average $\bar{X}$, we obtain

$$SE = \frac{SD}{\sqrt{n}} \tag{9}$$

If the assumption (9) is correct then countries with a large sample size n and a small value of the standard deviation SD should have a small standard error SE and consequently a short confidence interval, and those countries with small sample size n and with a large value of the standard deviation SD should have a large standard error SE and consequently a longer confidence interval.

Now consider the results of Japan and Germany as given in Table 9:

Table 10: Comparison of Japan and Germany with respect to sampling characteristics.

Country	n	M	SD	SE	$\bar{x} \pm 2SE$
Japan	5.256	522	86	5.2	522 ± 10.4
Germany	4.983	484	111	2.5	484 ± 5.0

Strange enough, a larger sample size and a smaller standard deviation for Japan lead to a confidence interval which is twice as long as the one for Germany.

8 Documentation

For illustrating the problems, the German situation shall be used. In Germany a stratified sampling plan was selected with the following strata:

- Federal states of Germany
- School-types of Germany

There are 16 federal states in Germany and although not clearly expressed six different types of schools. If this is correct then there are in all 96 different strata and it is assumed that the number of eligible 15-year-olds in each stratum is known sufficiently correct.

The study involves three types of random experiments:

1. Random selection of the schools within each stratum.
2. Random selection of the students of each sampled school.
3. Performing the competence test.

8.1 Documentation in the PISA Reports

Any scientifically serious study must be documented in detail. Otherwise, it is scientifically worthless and the results are nor meaningful. The German report shall be taken as an example of the quality of documentation in the PISA-reports.

According to [1] the school sample in Germany consisted of 219 schools and it is claimed that they constitute a representative sample for the stratification in federal states and school-types. From each selected school on an average 23 students of age 15 were selected randomly and asked for their agreement to participate in the tests. The proportion of those who agreed in taking part in the test is not reported in [1]. However, in [11] it is given as about 83%.

In [12] in Section "Allgemeiner Überblick" the German school sample consisted of 211 schools and about 28 students were selected from each participating school. In Section "PISA-Hauptuntersuchung: Ablauf der Datenerhebung" it is stated "In Germany all the 220 schools selected for the international sample have taken part in the study, thus, the school response rate is 100 %.

Finally, the U.S. report gives a number of 213 schools for the German sample.

In Germany the competence tests were identical in the different school types except for the type "Sonderschule" where handicapped students underwent a reduced test of 20 minutes duration. The regular test in Germany took two hours as stated in [1], in contrast to the United States were it took 90 minutes [6].

These few examples should be sufficient to characterize the documentation of studies and and results in the official national PISA reports. In the next section, a model is developed and some hints concerning a documentation are given.

8.2 Necessary Documentation

Consider the eligible population of size N of 15 years old students in spring 2000 arranged according to the Q strata "states & school types". Note that the strata refer to schools and not to students.

Each student is represented by a random vector $\vec{X} = (X_1, X_2, X_3)^T$ describing her/his performance in the sequence of the three PISA-tests.

Let the number of schools in stratum no. i be given by M_i, $i = 1, \ldots, Q$ and the number of eligible students in School no. j of stratum no. i be given by $N_{i,j}$, $j = 1, \ldots, M_i$ and $i = 1, \ldots, Q$. Then the population may be described as follows:

$$\Big((\vec{X}_{1,1,1}, \ldots, \vec{X}_{1,1,N_{1,1}}), \ldots, (\vec{X}_{Q,M_Q,1}, \ldots, \vec{X}_{Q,M_Q,N_{Q,M_Q}})\Big) \tag{10}$$

where $\vec{X}_{i,j,k}$ belongs to the student no. k of school no. j in stratum no. i.

The size of stratum no. i is given by:

$$N_i = \sum_{j=1}^{M_i} N_{i,j} \tag{11}$$

with $N = \sum_{i=1}^{Q} N_i$.

The aim is to determine the students' average capacity or literacy in the three fields reading, mathematics and science. In order to illustrate the model, it is sufficient to restrict it to reading literacy. The expected performance of student No. (i, j, k), i.e. the student no. k in school no. j of stratum no. i, in the reading test is given by

$$E[X_{i,j,k,1}] = \mu_{i,j,k,1} \tag{12}$$

The average performance of eligible 15-year-olds in the country in question is given by

$$\mu_1 = \frac{1}{N} \sum_{i=1}^{Q} \sum_{j=1}^{M_i} \sum_{k=1}^{N_{i,j}} \mu_{i,j,k,1} \tag{13}$$

Thus, μ_1 defines the average reading literacy.

For making the study understandable and reproducible at least the following information are necessary:

1. A clear definition of the population to be assessed including its size N.
2. The definition of the Q disjoint strata and their sizes $N_1, \ldots, N_Q$.
3. Some arguments motivating the stratification and showing that the variation given by the variance $V[X_{i,j,k,1}]$ is smaller in each stratum than in the overall population.
4. The definition of the random variables $X_{i,j,k,1}$.

Remark:
As a matter of fact, none of these information is supplied in the reports. Since for the international study no results concerning the federal states of Germany were of interests, the question should be answered why to select them as one of the stratification characteristics. As mentioned before, the stratification characteristics should be selected in a way that the resulting strata are more homogeneous with respect to the characteristics of interest. Since the educational systems in the different federal states of Germany are rather similar, it is to be expected that there at best only minor effects on the variation of literacy. •

The average μ_1 shall be determined based on a "representative" sample of size n by means of a statistical procedure with reliability given by the confidence level of 95%. According to [2] and the technical standards given in [10] the term "representative" means that a proportional sample is used where the proportion of elements from stratum i in the sample is the same as in the population.

Let

$$\left(a_{i,1}, \ldots, a_{i,m_i}\right) \tag{14}$$

denote the random school sample of size m_i from stratum no. i, $i = 1, \ldots, Q$. Then, the overall number of schools in the sample is

$$m = \sum_{i=1}^{Q} m_i \tag{15}$$

with

$$\frac{m_i}{m} \approx \frac{M_i}{M} \tag{16}$$

Let

$$\left(\vec{X}_{i,a_j,1}, \ldots, \vec{X}_{i,a_j,n_{i,j}}\right) \tag{17}$$

denote the planned random sample of size $n_{i,j}$ from school a_j, $j = 1, \ldots, m_i$ of stratum no. i, $i = 1, \ldots, Q$ referring to reading literacy. The size of the overall sample from stratum no. i is denoted by n_i and the size of the overall sample by n. Then

$$n_i = \sum_{j=1}^{m_i} n_{i,j} \tag{18}$$

$$n = \sum_{i=1}^{Q} n_i \tag{19}$$

The requirement of a proportional sample implies:

$$\frac{n_i}{n} = \frac{N_i}{N} \tag{20}$$

The following information are necessary:

1. The overall sample size n.
2. The sample sizes n_i, $i = 1, \ldots, Q$ of each stratum.
3. The sample sizes $n_{i,j}$, $j = 1, \ldots, m_i$ and $i = 1, \ldots, Q$
4. Some arguments motivating the sample sizes n, n_i and $n_{i,j}$, $j = 1, \ldots, m_i$, $i = 1, \ldots, Q$.
5. A detailed description of sampling schools and students.

Remark:
The information about the sample sizes are partly contained in the reports. However, the data are contradictory and confusing and not complete. There are almost no information about the strata and their weights. Moreover, the sample size determines essentially the accuracy of the statistical procedure and , therefore, some explanations about minimum requirements with respect to sample size and accuracy should have been provided. •

Since the response rates of schools and students are not at all 100% the planned proportional sample will turn into a disproportional sample at the end. Even more serious, however, is the question about the motives of non-response. It is to be expected that the reasons for responding and not responding are dependent of the school's quality on the one hand and the student's literacy on the other hand. If this is true, then the sample cannot be considered as random sample and statistical procedures will produce invalid results.

The non-response of schools and students and subsequent replacement results in a school sample of stratum no. i given by:

$$\left(a_{i,\ell_1}, \ldots, a_{i,\ell_{m'_i}}\right) \quad \text{for } i = 1, \ldots, Q \tag{21}$$

with random sample size m'_i and $m' = \sum_{i=1}^{Q} m'_i$. Moreover,

$$\frac{m'_i}{m'} \neq \frac{M_i}{M} \tag{22}$$

Denote by

$$\left(\vec{X}_{i,\ell_j,1}, \ldots, \vec{X}_{i,\ell_j,n'_{i,j}}\right) \tag{23}$$

the finally used sample of students in school no. ℓ_j of stratum no. i after the non-response of schools and students have taken into account with random sample size $n'_{i,j}$. Then the random sample size of stratum no. i and the overall random sample size n' are given by:

$$n'_i = \sum_{j=1}^{m'_i} n'_{i,j} \tag{24}$$

$$n' = \sum_{i=1}^{Q} n'_i \tag{25}$$

The arithmetic mean $\bar{X}_i$ of the reading literacy in stratum no. i is defined by:

$$\bar{X}_i = \frac{1}{n'_i} \sum_{j=1}^{m'_i} \sum_{k=1}^{n'_{i,j}} X_{i,\ell_j,k,1} \tag{26}$$

For compensating the disproportionality with respect to the stratification, the stratum means have to be weighted accordingly yielding a weighted overall average $\bar{X}$ for the reading literacy.

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{Q} N_i \bar{X}_i \tag{27}$$

Next the most difficult problem has to be solved namely to determine the most relevant distributional properties of $\bar{X}$.

1. The expectation $E[\bar{X}]$.
2. The variance $V[\bar{X}]$ in case $E[\bar{X}] = \mu_1$ holds.
3. The mean squared error $E[(\bar{X} - \mu_1)^2]$ in case $E[\bar{X}] \neq \mu_1$.

The above problems are difficult because

- The sample sizes are random.
- The implications of response and non-response on the distribution of $\bar{X}$ have to be taken into account.
- Each element of one of the sub-samples consists of three dependent random variables.
- The elements of each sub-sample are dependent, particular those coming from one school and having same teachers.
- The marginal distributions of the element $\vec{X}_{i,\ell_j,k}$ are different as well as the marginal distribution of the elements of $(X_{i,k\ell_j,k,1}, X_{i,\ell_j,k,2}, X_{i,\ell_j,k,3})^T$. Particularly, they differ in expectation and variance.

These properties imply that the conditions for (7) are not met and, therefore, it can not be used. Moreover, it also implies that the tests of significance used in the U.S.-study for comparing the averages of other countries with those of the United States as described in Appendix 1 of [6] cannot be used, too, even in the comparatively simple case of reading literacy. The situation for the averages $\bar{X}_2$ and $\bar{X}_3$ gets more complicated because of their dependence on $\bar{X}_1$.

For instance, the conditional variance of $\bar{X}$ under the condition of the sample sizes includes the following variances and correlation coefficients.

$$\sigma^2_{i,j,k} = V[X_{i,\ell_j,k,1}] \quad \text{for } k = 1,\ldots,n'_{i,j},\ j = 1,\ldots,m'_i \text{ and } i = 1,\ldots,Q \tag{28}$$

$$\rho_{i,j,k;u,v,w} = \frac{V[X_{i,\ell_j,k,1}X_{u,k\ell_v,w,1}]}{\sigma_{i,j,k}\sigma_{u,v,w}} \quad \text{for } (i,j,k) \neq (u,v,w) \tag{29}$$

The correlation coefficients and the necessity to determine them are briefly mentioned in [10], but in none of the national reports. Particularly, no hint is given, how the correlation coefficients are estimated and which values have been used for calculating the standard deviation SD given in the reports. In contrast to (7), the existence of correlation implies that it is not possible to measure the average literacy arbitrary accurate by increasing the sample size. For a large sample size, the contribution in variability due to correlation becomes more and more decisive, whereas the contribution of the inherent variations of the individuals gets smaller and smaller. Moreover, generally an increase of the sample size in a finite population also means in increase in the number of correlated sample elements and thus of the correlation effects.

Correlation is particularly high in cases where only few schools participated in the test and consequently many students are from the same school and have the same teachers. Therefore, small countries like Liechtenstein, Luxembourg and Iceland will exhibit a high correlation.

9 Summary

From a stochastic and scientifically point of view the PISA-Study as described in the national reports reveals some major deficiencies.

- General Comments
  1. The national reports are devoted mainly to international comparisons instead of concentrating on the national findings.
  2. The quantitative results and their interpretations are not always clearly separated.
  3. The use of statistical methods without sufficient confirmation seems to be hazardous.
- Stochastic Documentation:
  1. The models used in the different countries are not specified and, therefore, a founded comparison not possible.
  2. The quantity to be determined (average literacy) is not clearly defined.
  3. The procedures for determining the average literacy are not specified.
  4. The sample sizes for the various tests are either not stated or stated contradictory.

- Methodology
  1. The population to be assessed is not clearly defined.
  2. The samples are screened subsequent to drawing.
  3. The standards set are in many cases violated.
  4. The procedures indicated are based on assumptions which are not met.
  5. There are no hints that the stratifications used are appropriate.
  6. The sample size requirements or recommendations seem to be unfounded.
- Presentation of Results
  1. The number of national reports containing more or less the same information is for each country large and confusing.
  2. The presentation of arranged lists according to the mean score is scientifically meaningless and furthers misinterpretations.
  3. Many presentations are merely descriptive and, therefore, have no founded interpretation.
  4. The presentations include almost no absolute numbers, but are restricted to relative numbers making an interpretation at least very difficult.

In all it is to be expected that a thoroughly performed statistical analysis of the data would reveal that the samples are not random samples in a statistical sense and, therefore, it is not possible to draw founded inference from the sample results on the population of the 15-year-olds. Moreover, it would be no surprise at all, if the observed differences in the measured values of the countries' average literacy are to be attributed to the inherent variation in literacy of 15-year-olds and, therefore, cannot be taken as early indicators for improving the national educational systems.

The first impression of the study when trying to solve the mystery of the participating countries to be not carefully planned, performed and presented in the national reports proved to be correct. The insufficient quality of the *2000 Program of International Student Assessment* as reflected in the national documentations does not allow a meaningful comparison between the educational systems of the different countries. Particularly any conclusions to change or not to change an existing system on the basis of the study should be avoided unless the results are confirmed by a thoroughly performed analysis following acknowledged scientific standards.

In any case, however, conclusions should be drawn with respect to the future of PISA, as it is planned to have it cyclically repeated. The next study shall be performed in 2003 focussing on the mathematical literacy of 15-year-olds in the participating countries. The model, the sample design and the stochastic procedures used in the different countries should be revised and harmonized.

Moreover, the documentation should include all items being relevant for understanding the significance of the results obtained as a necessary condition for any scientifically serious

study. The information should necessarily include exact definitions of the population, of the stratification, the model and the model assumptions, the sampling design and the procedures used for analyzing the obtained data.

References

[1] Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K.-J. & Weiß, M. (2001): OECD- PISA - Progamme for International Student Assessment - Schülerleistungen im internationalen Vergleich. http://www.mpib-berlin.mpg.de/pisa

[2] Zur Stichprobenziehung innerhalb der PISA-Erweiterung. (2001). http://www.mpib-berlin.mpg.de/pisa

[3] Die PISA-Stichprobe in Österreich. http://www.sbg.ac.at/assess/pisa/54stichprobe-text.htm

[4] Shiel, G., Cosgrove, J., Sofroniou, N. and Kelly, A. (2001): Ready for Life? The Literacy Achievements of Irish 15-Year Olds with Comaparative International Data. Summary Report. Educational Research Centre, St. Patrick's College, Dublin.

[5] Programme for International Student Assessment (PISA). http://www.minedu.govt.nz/goto/pisa

[6] Lemke, M., Calsyn, Ch., Lippman, L., Jocelyn, L., Kastberg, D., Liu, Y.Y., Roey, St., Williams, T. and Kruger, T. (2001). *Outcomes of Learning Results From the 2000 Porgram for International Student assessment of 15-Year-Olds in Reading, Mathematics, and Science Literacy.* U.S. Department of Education, National Center for Education Statistics, NCES 2002-115. http://nces.ed.gov/surveys/pisa

[7] Bussière, P., Cartwright, F., Crocker, R., Ma, X., Oderkirk, J. and Zhang, Y. (2001): *Measuring up: The Performance of Canada's Youth in Reading, Mathematics and Science.* Human Resources Development Canada, Council of Ministers if Education, Canada and Statistics Canada. http://www.pisa.gc.ca

[8] International Student Assessment. Results for England 2000. National Statistics. http://www.statistics.gov.uk/releases

[9] International Student Assessment. Results for Northern Ireland 2000. National Statistics. http://www.statistics.gov.uk/releases

[10] Pisa Sampling Manual. Main Study Version. (1999). ACER, CITO, ETS, NIER, Westat. http://www.pisa.oecd.org/Docs/Download/SamplingManualUpdated.doc

[11] PISA-Hauptuntersuchung: Ablauf der Datenerhebung. (2001) http://www.mpib-berlin.mpg.de/pisa/html-dt/Testablauf.htm

[12] PISA: Allgemeiner Überblick. (2001) http://www.mpib-berlin.mpg.de/pisa/

Elart von Collani
University of Würzburg
Sanderring 2
D-97970 Würzburg
Germany