OECD/PISA's response to the paper "What do international tests really show about American student performance?" by Martin Carnoy and Richard Rothstein

OECD/PISA shares the view expressed in the paper about the importance of examining countries' performance levels from various perspectives. Indeed, one of the five volumes in which the initial results from PISA 2009 are published, is dedicated to the examination of performance by various background characteristics of students, schools and countries. This is not acknowledged in the paper. For example, Table II.3.2 (http://www.oecd.org/pisa/pisaproducts/pisa2009keyfindings.htm) presents performance scores in reading after accounting for countries' socio-economic backgrounds.

More importantly, the Carnoy/Rothstein paper contains several fundamental misunderstandings and misinterpretations of the PISA data. In particular, the paper claims that there are flaws in PISA samples, which is simply incorrect and unsupported in the paper. Some of the key misunderstandings and misinterpretations are listed below.

The paper compares results from PISA with results from other studies such as NAEP and TIMSS. However, the paper fails to adequately state the important and distinctive features of the respective studies, and to acknowledge the differences in, for example, the target populations and the assessment frameworks. When the results are compared across these different studies, the paper does not carefully interpret the results based on a correct understanding of the data collected in the respective studies but instead tends to immediately conclude that any differences in results that are found are attributable to flaws in one of the studies. (Even though the paper does discuss the differences between the assessments in terms of the target age or grade (pp.61-63), there is insufficient discussion of the implications of these differences on the different results from the respective studies.)

No flaws in the Finish PISA 2000 sample

The paper claims that, for Finland, there is a discrepancy in the social class profile (measured by the number of books at home) between students who responded to reading items and those who responded to mathematics items in PISA 2000.

> p.52 The last paragraph:
> The sampling methodology is complex, and the possibility of sampling flaws is another reason why results should be treated with caution. In 2000, for example, PISA reported separate samples for its reading and mathematics assessments. (In subsequent years, reading and mathematics questions were presented in a single test booklet for all sampled test takers). If the samples were completely accurate, we should expect the social class distribution of test takers to have been the same for the reading and math assessments in 2000. Mostly, this was the case. But not always. The biggest discrepancy was in Finland, where 12 percent of the reading sample came from the highest social class group (more than 500 books in the home), but only 7 percent of the math sample came from this group. Because we know that advantaged test takers score higher, on average, than students from lower social classes, Finland's overall average scores in 2000 cannot have been accurate (i.e., representative) in both reading and mathematics, and perhaps in neither.

However, as is evident from the PISA 2000 compendia, which is available on the PISA public website (at the bottom of http://pisa2000.acer.edu.au/downloads.php), that the proportion of students in the category "More than 500 books in the home" is around 6-7% in both the reading and the mathematics compendia: 6.4% in reading and 6.7% in mathematics. The table below summarises the results from the relevant item (Q37) in the reading and mathematics compendia.

Table 1. Compendia: PISA 2000 Student Questionnaire Q37

		PISA 2000 Q37. How many books are there in your home?							
		None	1 to 10	11 to 50	51 to 100	101 to 250	251 to 500	More than 500	Missing
Reading	Canada	0.9	5.6	17.7	20.1	23.9	18.6	12.5	0.6
	Germany	1.3	7.0	19.6	22.1	20.8	15.1	12.0	2.2
	Finland	0.6	6.6	23.0	24.1	24.1	13.9	6.4	1.4
	France	2.6	8.6	20.8	20.9	20.3	13.0	8.1	5.7
	United Kingdom	1.1	7.3	21.1	20.8	20.6	14.4	12.9	1.8
	Korea	1.1	7.1	18.0	22.6	27.7	15.9	7.5	0.2
	United States	2.7	8.8	18.7	18.4	19.3	13.3	9.2	9.7
Mathematics	Canada	0.9	5.5	17.2	20.6	23.8	18.8	12.6	0.6
	Germany	1.0	7.2	19.7	21.6	21.3	15.0	12.5	1.8
	Finland	0.6	6.6	23.9	24.2	23.1	13.3	6.7	1.6
	France	2.7	8.4	20.5	21.6	20.6	13.3	7.5	5.4
	United Kingdom	0.9	7.8	21.5	19.8	20.4	14.5	13.2	2.0
	Korea	1.2	7.6	18.2	22.6	26.7	15.9	7.6	0.2
	United States	2.8	9.2	19.0	17.7	18.6	13.4	9.8	9.7

In the same paragraph on page 52 of the paper, it is stated that "in subsequent years, reading and mathematics questions were presented in a single test booklet for all sampled test takers", but this is not in fact correct. For example, in PISA 2003, the 167 main study items were allocated to 13 item clusters (seven mathematics clusters and two clusters in each of the other domains), with each cluster representing 30 minutes of test time. The items were presented to students in 13 test booklets, with each booklet being composed of four clusters according to the rotation design shown in a table below (source: Table 2.1 in the PISA 2003 Technical Report

http://www.oecd.org/edu/preschoolandschool/programmeforinternationalstudentassessmentpisa/35188570.pdf reproduced as Table 2 below). M1 to M7 denote the mathematics clusters, R1 and R2 denote the reading clusters, S1 and S2 denote the science clusters, and PS1 and PS2 denote the problem-solving clusters. Each cluster appears in each of the four possible positions within a booklet exactly once. Each test item, therefore, appeared in four of the test booklets. This linked design enabled standard measurement techniques to be applied to the resulting student response data to estimate item difficulties and student abilities.

Table 2. Cluster rotation design used to form test booklets for PISA 2003

Booklet	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	M1	M2	M4	R1
2	M2	M3	M5	R2
3	M3	M4	M6	PS1
4	M4	M5	M7	PS2
5	M5	M6	S1	M1
6	M6	M7	S2	M2
7	M7	S1	R1	M3
8	S1	S2	R2	M4
9	S2	R1	PS1	M5
10	R1	R2	PS2	M6
11	R2	PS1	M1	M7
12	PS1	PS2	M2	S1
13	PS2	M1	M3	S2

No flaws in the US PISA 2009 sample

The paper claims that 40% of the United States PISA sample was drawn from schools where half or more of the students were eligible for free or subsidized lunches, while only 23% of all high school students in the US attend such schools. The paper claims that this discrepancy is the result of an error in the sampling in PISA.

> pp.53-54
> Therefore, for an accurate sample, PISA should not only have a proportion of FRPL-eligible students that is similar to that proportion nationwide, but should have FRPL-eligible students whose distribution among schools with concentrated disadvantage is also similar to the distribution nationwide.
>
> Table 25 compares the distribution of all U.S. high school students nationwide, by share of FRPL-eligible students in their high schools, to the distribution of students in the 2009 PISA sample, by share of FRPL-eligible students in their high schools.
>
> The table shows that the average PISA score of U.S. students in both reading and math decreases dramatically as the share of their schools' students who are FRPL-eligible increases. The table also makes apparent that PISA's FRPL test-takers were heavily concentrated in severely disadvantaged schools, where unusually large proportions of students were FRPL-eligible. Forty (40) percent of the PISA sample was drawn from schools where half or more of the students were eligible for free or subsidized lunches. Only 23 percent of all U.S. students attend such schools. Sixteen (16) percent of the PISA sample was drawn from schools where more than 75 percent of students are FRPL-eligible, yet fewer than half as many, 6 percent of U.S. high school students, actually attend schools that are so seriously impacted by concentrated poverty.
>
> Likewise, students who attend schools where few students are FRPL-eligible, and whose scores tend, on average, to be higher, were undersampled. This oversampling of students who attend schools with high levels of poverty and undersampling of students from schools with less poverty, results in artificially low PISA reports of national average scores. If other countries' PISA samples better reflect the actual spatial distribution of disadvantaged 15 year olds, the real U.S. average performance should rank higher relative to other countries than the reported PISA averages indicate. We have queried officials at the U.S. Department of Education's National Center for Education Statistics (NCES) in an attempt to determine why the PISA sample was skewed in this way, but while these officials acknowledge that there may be a sampling error, they have been unable to provide an explanation. We can only speculate about it. One possibility is that the PISA sampling methodology excluded very small schools, where poverty is less likely to be concentrated. Another possibility is that because participation in PISA is voluntary on the part of schools and districts that are randomly selected for the sample, schools serving more affluent students may be more likely to decline to participate after being selected. Perhaps this is because such schools are generally less supervised by the federal government than schools serving disadvantaged students and feel freer to decline government requests. Whatever the reason, an initial PISA sample that was representative would lose some validity if schools serving higher proportions of more affluent children were more likely to decline to cooperate, and were then replaced in the sample by schools serving lower proportions of affluent students. An underestimation of national average scores is then bound to result.

However, investigation of this claim clearly shows that there is no flaw in the US PISA sample, but rather that two sources of data have been compared (in Table 25 of the paper) that, for one reason or another, are not consistent.

Columns (a) and (b) in Table 3 below are copied from columns (a) and (b) in Table 25 of the paper. The third column, Column (I) shows our contractor Westat's attempt to reproduce the results in Column (b), just using public schools, and including all PISA schools (a very few PISA students are in middle schools). Note that these results are very consistent with the figures in Column (b). In Column (II), Westat again used the PISA public school sample, but rather than using the data that were reported on the PISA school questionnaire, they merged data concerning free and reduced price lunch from the 2007-2008 national public school file (CCD) that is released by NCES. Note that these results use exactly the same sample of students as in Column (I), which presumably very closely resembles the sample used in Column (b) and yet gives results that are very close to those in Column (a).

Note, however, that Column (a) refers to all high school students, rather than PISA students as reported in Column (II). Westat also endeavored to reproduce figures in Column (a), by analysing the whole CCD file, but restricted this to schools that have grade 10, thus giving a proxy for 'high schools' (weighting by the number of students in the school, so as to retain a 'student-centered' analysis). There are about 21,000 such schools and the mean percentage of FRPL-eligible students is 36.2% i.e. very close to the mean for the PISA sample. Also, the distribution is very close to that reported in Column (a). The conclusion from this is that the difference between Columns (a) and (b) is entirely due to systematic differences in the percentage of FRPL-eligible students reported in the school questionnaires in PISA and those reported in the 07-08 CCD data for the same school. This means that the achievement projections that are given in Columns (c) and (d) and Row (g) of the Table 25 in the paper are totally spurious.

Table 3. Percentages of students eligible for FRPL in student's school

	Computed by Carnoy and Rothstein (see Table 25 in the paper)		Computed by Westat		
	(a)	(b)	(I)	(II)	(III)
	Share of all U.S. High School Students, by Share of FRPL-Eligible Students in Student's School, 2007-2008 (percent)	*Share of PISA 2009 Sample in High Schools, by School percent of Students Eligible for FRPL (percent)*	*Share of PISA 2009 Sample in All Public Schools, by School percent of Students Eligible for FRPL (percent) as reported in PISA school Q - Westat*	*Share of PISA 2009 Sample in All Public Schools, by School percent of Students Eligible for FRPL (percent)as reported on 2007-08 CCD file- Westat*	*Share of all U.S. Students, in public schools that offer grade 10, by Share of FRPL-Eligible Students in Student's School, 2007-2008 (percent)- Westat*
75 percent or more	6%	16%	16%	5%	7.4%
50 to 74.9 percent	17%	24%	23%	20%	17.2%
25 to 49.9 percent	33%	36%	35%	34%	33.4%
Less than 25 percent	36%	24%	23%	32%	34.9%
No data available	6%		4%	9%	7.1%
All	99%	100%	100%	100%	
Mean of non-missing			43.6	36.3	
25th percentile			25	18	
75th percentile			64	52.4	

Below is a cross-tabulation of the variables from the two sources, showing their inconsistency. These are weighted PISA results (the same data as reported in Columns (I) and (II) in Table 3). Remarkably, schools were hardly ever reported in PISA as being in a lower category than was recorded in the CCD data, whereas a quarter of the time they were reported in PISA as being in a higher category (Note that in PISA, the schools reported results in terms of whole percentages, which, for this analysis, Westat converted into four categories to be consistent with the data shown in the paper). One thing to keep in mind in viewing

these data is that participation in the NSLP program increased noticeably between 2007-08 and the time in late 2009 when PISA was conducted, due to the changes in the economy during that period.

		PISA School Questionnaire					
		Missing	<25	25-50	50-75	>75	Total
2007-08 CCD	Missing	1	0.5	3.6	1.4	2.8	9.2
	<25	1.6	22.3	7.9	0	0	31.8
	25-50	0.5	0	22.8	9.5	1.3	33.9
	50-75	0.6	0	0.8	11.0	7.4	19.8
	>75	0	0	0	0.7	4.5	5.2
	Total	3.7	22.8	35	22.7	15.9	100

To further examine the relationship between the data from the two sources, Westat ran a regression of the school principals' response to the PISA 2009 School questionnaire (column (I)) on the CCD data for the school (column II). The correlation is 0.93, and both the slope and the intercept are significant. The regression model is:

School principal's response = 1.0828*(07/08CCD) + 3.0127.

This means that the 'typical' school response to the PISA question was 8 percent (not 8 percentage points) higher than the CCD data shows, plus another 3 points. Thus, if the CCD data indicated 36.3% eligibility (the mean of Column (II) in Table 3), the model prediction for the school's response would be 42.3%. This is slightly inconsistent with the mean for Column (I) in Table 3 (which shows the mean school response in PISA as 43.6%, and a linear regression should run through the two means) but this slight difference is no doubt because some schools have missing data for one variable but not the other (and the regression is only run for schools where data are not missing in either source).

If a model is fitted with no intercept (a ratio model), then the coefficient is 1.14.

PISA 2000 mathematics results are not directly comparable to PISA 2003 mathematics results

The paper compares mathematics results in PISA 2000 and PISA 2009 (e.g. Tables 14a-c and 15a-b) and claims that observed changes in scores are different between PISA and other studies. However, mathematics results in PISA 2000 are not directly comparable to those in PISA 2009. As described in detail in PISA 2009 Technical Report (pp.211-213), the primary PISA reporting scales in reading, mathematics and science were established in the year in which the respective domain was the major domain, since in that year the framework for the domain was fully developed and the domain was comprehensively assessed. The primary reporting scale in mathematics was developed in PISA 2003, when mathematics was the major domain. This scale is directly comparable to the mathematics scale in PISA 2006 and PISA 2009, but not to the mathematics scale in PISA 2000. Here is the link to *PISA 2009 Technical Report*:
http://www.oecd.org/edu/preschoolandschool/programmeforinternationalstudentassessmentpisa/pisa2009technicalreport.htm

The paper also claims that there is a "V-shape of the PISA results in Figure 7" (p.59), but it is important to note that the score-point difference in mathematics between PISA 2003 (483 points) and PISA 2006 (474 points) is not statistically significant, after accounting for the standard errors and the link error.