Response from Martin Carnoy and Richard Rothstein to OECD/PISA Comments (by Andreas Schleicher, OECD Deputy Director for Education and Special Advisor on Education Policy to the OECD's Secretary-General, January 14, 2013) regarding our report "What do international tests really show about American student performance?" (Economic Policy Institute 2013)

January 24, 2013.

On December 21, 2012, we sent a draft copy of our report on international test score comparisons to Andreas Schleicher of OECD/PISA. Our report was scheduled for release at noon on January 15, and Mr. Schleicher sent timely comments to us late on January 14. We are grateful to him for getting his comments to us. Our response to his comments follows. A copy of Mr. Schleicher's comments is attached to the end of our response.

Mr. Schleicher's critique does not address the main points of our report which are that there is a social class achievement gap in every country, that the U.S. achievement gap is surprisingly small on international tests, and that the achievement of disadvantaged U.S. students on these tests has been rising very rapidly, while the achievement of similarly disadvantaged students in other countries, including some with which the U.S. educational system is usually unfavorably compared, has either been stagnant or falling rapidly.

Rather, the Schleicher critique raises some relatively small statistical issues, some of which are valid. Where they are valid, we have revised the posted version of our report online. We have preserved the originally posted (January 15) report in our files, and will provide a copy to anyone interested in this controversy.

We appreciate that OECD/PISA has had some interest in the social class background of students in its various national samples and has reported a relationship between average national scores and socioeconomic status. We acknowledge the table that Andreas Schleicher mentions in his comments, Table II.3.2 from PISA's Volume II, and we regret that our report did not do so. We also acknowledge Figure V.2.9 from PISA's Volume V. These aspects of OECD/PISA's documentation are similar to Tables 3A and 3B of our report, showing how each country's score would change in 2009 if countries had similar student social class distributions, and Table 13 of our report, showing how changes within countries over time in their overall social class composition affect these countries' overall average scores. For these estimates, OECD/PISA uses its index of economic, social, and cultural status (ESCS) and, as our report states, its index tracks our "books-in-the-home" measure well, although ESCS is unique to OECD/PISA and cannot be used for comparisons with TIMSS.

That said, neither Andreas Schleicher's comments, nor OECD/PISA's Table II.3.2, Figure V.2.9, nor its surrounding text in PISA 2009 Volumes II and V, permit consumers

of OECD/PISA data to address the main issues raised in our report. Table II.3.2 and Figure V.2.9 do not permit a comparison between social class groups in different countries, and do not permit an analysis of how social class groups within countries perform relative to one another. Consumers of OECD/PISA reports cannot tell from these tables or from OECD/PISA's narrative reports whether a country's relatively good (or poor) overall average scores are attributable to relatively good (or poor) performance of advantaged or disadvantaged students (or both), or whether advantaged (or disadvantaged) students in one country perform better or worse than students with similar social class status in other countries, to an extent not suggested by these countries' relative average scores. When policymakers are unable to compare students' PISA test results across countries by social class groups, the conclusions about relative national performance they draw from national averages alone may be incorrect and may support misguided policy reforms.

As we observe in our report, U.S. policymakers and analysts have become quite sophisticated in their analysis of domestic test scores. Aware of the great influence that social and economic background factors have on student performance, domestic reports of U.S. student achievement always now disaggregate the data by race, ethnicity, and family poverty. Indeed, federal law now requires such disaggregation. Our report expresses puzzlement that the same policymakers and analysts do not feel compelled to apply similar sophistication to international test comparisons.

Our report compares average test scores of disadvantaged, middle class, and advantaged U.S. students with their counterparts in other countries. It offers more nuanced policy conclusions about the relative quality of U.S. education, particularly when it comes to disadvantaged U.S. students, than those offered in the one document that OECD/PISA produced specifically addressing U.S. students' performance: *Strong Performers and Successful Reformers in Education. Lessons from PISA for the United States.* Paris: OECD. http://dx.doi.org/10.1787/9789264096660-en.

For example, OECD/PISA in that document (or elsewhere) fails to show that from 2000 to 2009, U.S. disadvantaged students' scores increased substantially in both reading and math compared with the scores of disadvantaged students in many comparison countries, and especially compared with those in some countries that OECD/PISA holds up as models for the United States. In the most conspicuous case, during the same period that disadvantaged U.S. students' scores increased, disadvantaged students' scores in Finland fell sharply in both reading and math. Understanding these results should have influenced analysts' views of U.S. educational policy and would have raised questions about the success of other models, especially for improving the achievement of disadvantaged students.

Similarly, the U.S. government concluded from the OECD/PISA report that the achievement gap between advantaged and disadvantaged students is a relatively serious,

perhaps even unique problem for the United States. However, the OECD/PISA database that we analyzed does not support this inference. Indeed, we find that the achievement gap in the United States is consistently smaller than it is in similar post-industrial countries, and in some comparisons, is quite similar to that in some top-scoring countries. The OECD/PISA narrative and tables do not address this issue.

Mr. Schleicher's comments address methodological issues related to our report's discussion of the PISA samples and the use of the data.

1. The OECD/PISA comment on our comparison of PISA with NAEP and TIMSS contends that our report

   "…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."

   We reject this particular criticism. While our report observes that different tests may define "mathematics" achievement differently (i.e., assume different implicit mathematics curricula), each of the tests purports to assess a representative national sample of students in "mathematics" at the relevant age or grade.

   We discuss differences in the target populations, as OECD/PISA acknowledges, and state in several places in our report that the tests characterize themselves as measuring different aspects of mathematics. Our main point in comparing the tests is to show that different tests may provide different assessments of whether students in the United States or in any country are adequately learning a subject, in this case, "mathematics." Our conclusion from these comparisons is that judgments about how well U.S. students are doing should never be based on a single international (or any) test at a single point in time. This is borne out starkly by the recent 2011 TIMSS results showing that Finnish 8th grade students scored not significantly higher in mathematics than 8th grade students in the U.S., whereas Finnish mainly 9th grade students score substantially higher than U.S. mainly 9th and 10th grade students on the PISA math test. It is implausible that this very large difference can be attributable mainly to much more rapid achievement growth in Finland than in the United States during a period of about one grade, after performance was nearly identical before that grade began.

2. OECD/PISA comment: "*The 2000 math results are not **directly** comparable with 2009 math results*."

The criticism here is leveled at our comparison of changes in PISA math from 2000–"2007" (a calculated score averaging PISA 2006 and 2009 scores) with TIMSS math from 1999–2007, in a series of tables. Restricting our analysis to adjacent years and to countries in which both PISA and TIMSS scores are available, we display trends in combined PISA/TIMSS performance for the U.S., Korea, and England/UK. We similarly display combined trends in Finland for PISA 2000–2009 and for TIMSS 1999–"2009" (a calculated score averaging TIMSS 2007 and 2011 scores). We demonstrate that the trends in mathematics scores for PISA and TIMSS are quite different. This is particularly true when we break down the comparisons by social class of students. In its comment to us, OECD/PISA claims that the PISA 2000 math scores would be very different if the 2000 test looked like it did in 2003–2009. We reply that differences between PISA and TIMSS score changes are so large that even with possible differences between PISA 2000 and 2009 scaling, trends in TIMSS and PISA would still differ.

Mr. Schleicher's comment refers to Figure "7" in the draft we provided to him in December. The same figure is now labeled Figure G in the published report. The OECD comment suggests that the 9 point difference we show is not statistically significant. However, we estimate that this is more than two standard errors, and is an important difference. Figure G does indeed demonstrate that the point estimates form a V, and the change in the PISA math scores in this period differs from the change during this period in NAEP scores. This is strong evidence of a contradiction between the PISA report of math performance and other reliable measures.

3. OECD/PISA comment: "*No flaws in the Finnish PISA2000 sample.*"

We acknowledge that we made an error in calculating the social class composition of the Finnish sample on the 2000 PISA reading test, and as a result, inaccurately stated that there was a large difference between the social class composition in Finland's reading and math test takers in that year. In fact, the difference for Finland is small, as we reported it was for other countries. We are grateful to Mr. Schleicher for calling attention to this, and regret the error. We have removed references in the report text to this alleged difference, and have substituted corrected Tables 8A and 13 and Figures D1 and D2 where our error affected the data display. The effect of this correction to these tables and figures is very small, does not influence the implications of the tables and figures, and will be barely noticeable even to careful readers of the original and corrected reports.

Without minimizing our embarrassment regarding this error, this is a minor point, as is the OECD/PISA point about the test booklets in 2003. Our chief claim

related to these points was that the 2000 test was conducted differently from later tests concerning how the test scores for math were estimated. We note that the OECD/PISA comments do not contest that point. Nor does OECD/PISA, in its reports, warn consumers of its data not to compare scores from 2000 to 2009 or to make judgments about whether a country's students overall made improvements during this period. Indeed, quite the contrary.

4. OECD/PISA comment: "*No flaw in the 2009 PISA sample for the U.S.*"

This relates to our claim that PISA's 2009 U.S. sample was flawed because 40 percent were students who attended schools where more than half of the students were poor or near poor. We stated that the percentage was actually 23 percent, and estimated what the U.S. score would have been if disadvantaged students in schools with heavy concentrations of such students had not been overrepresented in the sample.

Mr. Schleicher disputed our criticism of his sample. We did make an error, but not to the extent Mr. Schleicher claims. After correcting this error, it remains apparent that disadvantaged students in schools with heavy concentrations of such students were overrepresented in the U.S. PISA 2009 sample.

We spent about two years preparing this report, and while we were in the process of preparing it, new data were released and in this case, we did not revise our calculations before publication to incorporate the most recent data. We should have done so.

When we did this particular estimate shortly after the PISA 2009 database became available, the most recent NCES data on the share of students by school who were eligible for lunch subsidy was for the school year 2007–2008, and showed that 23 percent of U.S. high school students attended schools where more than half the students were in the lunch program. We assumed that this figure could not jump from 23 percent to 40 percent in a two-year period, and that therefore a sample in which 40 percent attended such high poverty schools could not be accurate. And we used the 23 percent figure as the best available number.

In investigating Mr. Schleicher's claim that our conclusion was "totally spurious," we engaged in extensive correspondence with Daniel McGrath, director of the International Activities Program in the National Center for Education Statistics, the branch of the U.S. Department of Education that manages U.S. involvement in international assessments like PISA. We are indebted to Mr. McGrath for his assistance in explaining to us how PISA went about conducting its 2009 assessment in the United States. Although he was not successful in persuading us

to withdraw our claim regarding oversampling, he made every effort to provide us with careful explanations, whether it supported his point of view or not.

From these exchanges, we have now learned that the PISA sampling in the United States was even more questionable than we suspected. This is primarily because the United States deviates from PISA procedure in a crucial respect—although PISA 2009 was established to assess a representative sample of a nation's 15 year olds in the spring of 2009, the United States received permission from OECD (as it had in previous years) to administer PISA in the following academic year, in the fall of 2009.

This raises other very complicated issues about the comparability of U.S. scores with those of other nations. We did not address these in the report and will not attempt to do so now, other than mentioning that they exist, and have been discussed in previous PISA reports.

However, this procedural deviation caused a difficulty that is directly relevant to our conclusion about oversampling of disadvantaged students in schools serving large proportions of such students. NCES was concerned that in the fall of 2009, principals would not yet know what their free and reduced-price lunch (FRPL) enrollment for the current year would be. So when PISA was administered, instead of asking principals for their schools' FRPL percentages, NCES asked them to report their schools' FRPL percentages for the previous year. We consider this a very serious leap. By basing its report in the PISA database of the distribution of schools by their FRPL percentages on principals' reports of their school percentages for the previous year, OECD/PISA is effectively asserting that sampled schools had identical demographic composition in the previous year as in the current (testing) year. This in itself is a dubious proposition. The demographics of many schools change from year to year, not only because of changes in the same students' economic circumstances (e.g., from the recession), but because of boundary changes, choice programs, changes in neighborhood characteristics, etc. In many cases, these changes may be small, but they are frequently large enough to affect the sample's social-class weighted average national score, the subject of our investigation.

The considerations just mentioned assume that principals' reports of previous year FRPL percentages, even if different from the current year's, are accurate. We don't consider it established that principals are more likely to accurately remember the FRPL enrollment of the previous year than they are to know the FRPL enrollment of the current year, even if it is only a month or two into the school year. And another factor is principal turnover. We have not investigated this, but are aware that this turnover could be as high as 20 or even 30 percent on

average. Principals of schools where they did not work the previous year may be less likely to report accurately the previous year's FRPL percentage.

The U. S. practice of administering PISA in the year following the year it is administered elsewhere raises another confounding issue. Governments establish the sampling frame (i.e., pick the schools whose students will take the PISA) the year before they administer the test. We assume this is necessary because full data on the nation's universe of schools for any year may not be available until the following year. Thus, for a test administered in 2008–2009 (the OECD-established year for PISA) the sample was drawn with data for the full universe of schools in 2007–2008. In the United States, the sample was also drawn from the 2007–2008 universe of all schools, but instead of being one school year before the test was administered, the sample year was now two school years before the test was administered. The sample of schools is accurately representative of all schools in 2007–2008. But for all the reasons we stated above that school demographics can change from year to year, they can change even more over two years.

So where does this leave us? When we discovered that the U.S. PISA sample comprised students, 40 percent of whom attended schools where half or more of students were in the FRPL program, we compared this with the actual percentage of students in all such U.S. schools in 2007–2008, the year in which the PISA sample was drawn, or 23 percent. Had we instead compared it with the actual percentage of students in all U.S. schools in the year for which PISA collected its data from its principals' questionnaire (2008–2009), we would instead have used a figure of 28 percent. And had we instead compared it with the actual percentage of students in all U.S. schools during the year in which PISA was actually administered, we would instead have used a figure of 32 percent. In no case, however, does the actual number of students in high-poverty schools approach the 40 percent number that OECD/PISA believes characterizes its sample.

Mr. Schleicher criticizes our 2007–2008 number (23 percent) because he says it is a year out of date and suggests (incorrectly, we think), that we should instead of have used 2008–2009. He is correct in his criticism that 2007–2008 was not the correct year to use. But if his criticism is correct, it can only be because the demographics of schools changes from year to year. If our use of 2007–2008 school universe data for comparison with PISA was problematic because it was a year out of date, then all the more problematic is the PISA sample itself that was two years out of date.

We remain uncertain which is the proper number to use for comparative purposes. If we assume that the demographics of schools were unchanged from 2008–2009 to 2009–2010, then we should use the 2009–2010 number of 32 percent for

Economic Policy Institute

making social class adjustments to the average national U.S. score. If we want to assess the representativeness of the PISA sample by comparing principals' reports to the actual universe of schools for the year they were reporting, then we should use the 2008–2009 number of 28 percent.

The most conservative number to use is 32 percent, and we have adjusted Table 25 and the surrounding text to substitute data on the FRPL percentage of all U.S. schools for 2009–2010 for the 2007–2008 data we initially used. The issue of oversampling remains, though it is not as great as it was in our initial report.

In our January 15 report, we said that adjusting the average U.S. score to account for oversampling of disadvantaged students in high poverty schools, along with other (and much more important) appropriate adjustments, would raise the 2009 U.S. PISA ranking among OECD countries from 14th to 4th in reading and from 25th to 10th in math. We have corrected the posted report to state that the re-adjustment would raise the ranking from 14th to 6th in reading, and from 25th to 13th in math.

While we continue to show that PISA oversampled disadvantaged students in the most disadvantaged schools, we thank Andreas Schleicher for calling attention to the error in our specific claim regarding the magnitude of this oversampling.

At the conclusion of lengthy correspondence with Mr. McGrath, he raised a new issue. He stated that principals' reports of the FRPL data of their schools could not be relied upon. Instead, he said, to calculate Table 25 of our report, we should have applied for a license to analyze PISA's restricted data that would give us the identity of specific schools that participated in PISA. We could then (with a similar license to analyze NCES' Common Core of Data) calculate the share of FRPL students by school in the actual schools that participated in PISA in 2009–2010.

This was a good suggestion, but when we prepared our report, we saw no reason to follow this much more complicated procedure because we relied in good faith on data published by OECD/PISA in its own public database, based on its questionnaire of principals. If OECD/PISA regards these data as unreliable, it should not publish them. It is possible that, as Mr. McGrath suggests, the data on FRPL participation by school is more accurate in the CCD, but we note that these data too are based on reports by principals.

Nonetheless, we suggested that rather than advise us to apply for a license to examine restricted PISA data, Mr. McGrath himself provide us with the summary information needed for Table 25 based on the CCD rather than the PISA database, inasmuch as he could do so in a very few hours. This would provide a check on

whether there was an overestimate in the PISA sample of FRPL students attending schools with high percentages of FRPL students.

As of our preparation of this comment (January 24, 2013), Mr. McGrath has not responded to our request. Therefore, we are proceeding to correct Table 25 using public CCD data for 2009–2010, and comparing it with the information OECD/PISA publishes in its public database on the distribution of schools in the PISA sample by their FRPL percentages. Should a future analysis of the restricted PISA and CCD databases suggest that no oversampling occurred, and if OECD/PISA corrects the data in its public database, we will reflect that in future reports, but we will not be able to revise the January 15 report further.

Ultimately, critics of our methodology will claim that even if all we have said above is correct, FRPL percentages are not a reasonable way to assess whether the PISA sample is representative. This is because the PISA sample was not selected with stratification for FRPL percentages. However, although the PISA sample was not selected with this stratification, the sample's stratification should fairly proxy other critically important background characteristics. Pure random sampling error should not produce as large a difference as we find. As we explain in the report, disadvantaged students in high-poverty schools will tend to achieve at lower levels than similarly disadvantaged students who are dispersed in more heterogeneous schools. In our view, this is one of those critically important background characteristics that the actual stratification should proxy. Thus the unrepresentativeness of the PISA sample in this respect, even if unavoidable, is an essential consideration for the interpretation of average national scores.

From Andreas Schleicher to Martin Carnoy and Richard Rothstein

January 14, 2013

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.

Economic Policy Institute

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 Mean of non-missing 25th percentile 75th percentile	99%	100%	100% 43.6 25 64	100% 36.3 18 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

Economic Policy Institute

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.