

George W. Bush Institute Global Report Card

Technical Appendix

Jay P. Greene
Department of Education Reform
University of Arkansas
201 Graduate Education Building
Fayetteville, AR 72701
(479) 575-3172; jpg@uark.edu

Josh B. McGee
Laura and John Arnold Foundation
1800 Post Oak Blvd., Suite 380
Houston, Texas 77056
(713) 554-1916; josh@arnoldfoundation.org

I. Introduction

This document serves as the technical appendix to the George W. Bush Institute Global Report Card. The goal of the Global Report Card is to allow users to compare the performance of school districts in the United States to the performance of their international economic competitors. In an increasingly global labor market students in the U.S. must not only be competitive locally or even nationally, but must also be prepared to compete internationally. To date it has been very difficult to make comparisons between academic performance at a district level to academic performance internationally.

The Global Report Card developed in this paper seeks to remedy this deficiency by combining state, national, and international testing data into a single measure of student achievement. The remainder of this document is devoted to describing the construction of the Global Report Card.

II. Construction of the Global Report Card

In order to construct the Global Report Card we combine testing information at three separate levels of aggregation: state, national, and international. At each level we use the available testing information to estimate the distribution of student achievement. To allow for direct comparisons across state and national borders, and thus testing instruments, we map all testing data to the standard normal curve.

We must make two assumptions for our methodology to yield valid results. First, mapping to the standard normal requires us to make the assumption that the distribution of student achievement on each of the testing instruments is approximately normal at each level of aggregation (i.e. district, state, national). Second, to compare the distribution of student achievement across testing instruments we assume that standard deviation units are relatively similar across the

testing instruments and across time. In other words we assume that being a certain distance from mean student performance in Arkansas is similar to being the same distance from mean student performance in Massachusetts.

We begin our calculations at the lowest level of aggregation by estimating average district quality within each state. We then estimate each state's average quality using national testing data. And finally, we estimate average national quality using international testing data. We then use our estimates at the national and international level to shift the district level distributions. This re-centers the distribution of district quality based upon the relative performance of the individual state when compared to the nation as a whole as well as the relative performance of the nation when compared to our economic competitors.

For example, in 2007 the average student in the Scarsdale School District in Westchester County, New York scored nearly one standard deviation above the mean for New York on the state's math exam. The average student in New York scored six hundredths of a standard deviation above the national average on the NAEP exam given in the same year, and the average student in the United States scored about as far in the negative direction (-0.055) from the international mean on PISA. Our final Global Report Card score for Scarsdale in 2007 is equal to the sum of the district, state, and national estimates ($1 + 0.06 + -0.055 = 1.005$). Since the final Global Report Card score is expressed in standard deviation units, it can easily be converted to a percentile for easy interpretation. In our example Scarsdale would rank at the seventy seventh percentile internationally in math.

The sections that follow will discuss the calculations necessary to construct our Global Report Card in detail. We will begin at the state level and work our way up to the international level.

State Level Calculations

Since the passage of No Child Left Behind states have been required to implement a testing regiment in grades 3-8 for school accountability. We use student achievement data from these tests to estimate the distribution of district quality within each state. Our data source for this project is the National Longitudinal School-Level State Assessment Score Database (NLSLSASD) maintained by the American Institutes for Research (AIR)¹. This dataset contains information on the percentage of students who reach each cut score (basic, proficient, advanced, etc.) for each school within each state in the U.S. The NLSLSASD contains testing data from 2004, 2005, and 2007, and will continue to be updated on an ongoing basis in odd years (2009, 2011, etc.).

We begin by aggregating to the district level for each grade and exam. For the purpose of this project we use the percent of students who achieve the proficiency cut score or better. We aggregate by calculating the student weighted average of the school level percent of students

¹ <http://www.schooldata.org/>

who score at the proficient level and above. We perform this calculation for both the reading and math exam and for every grade for which data are available.

Next, we standardize the district level percentages to get a measure of how far each district is from mean state performance. If we assume that student performance is approximately normal within each district and state, then we can use an inverse normal transformation to estimate this distance in standard deviation units.

Each district's percent proficient and above can be mapped to a specific point on the cumulative normal distribution using the inverse standard normal transformation. This point, or z-score, represents the number of standard deviation units above the mean on the standard normal curve where the specified percentage of the population would achieve the cut score. However, these z-scores are not comparable across states because cut scores in some states are more difficult to achieve than others. Therefore, we need to shift the districts' z-scores based on how difficult the state's proficiency cut score is to achieve.

We begin this calculation by recovering the mean percent proficient and above for each state by simply calculating the student weighted average at the state level for each exam and grade. We then apply the same inverse standard normal transformation discussed above to the state's proficiency rates. We use the resulting z-score as our measure of how difficult the state's proficiency cut score is to achieve. Each district's z-score is then shifted by simply subtracting the state's z-score. The shifted z-score is our estimate of how far each district is from mean state performance in standard deviation units.

We then aggregate the standardized proficiency percentages (z-scores) across grade to the subject level. We perform this aggregation by calculating the student weighted average of the z-scores. These district level quality scores will serve as the foundation for our subsequent calculations. We are particularly grateful to Martin West at Harvard University for suggesting this technique for estimating the standard deviation of achievement within each state.

Because we want to generate an annual measure of education quality, it is necessary to fill in the gaps in our data. We accomplish this by simply linearly interpolating between the two adjacent year's z-scores for the district. For example, our data does not include 2006, so we assume that district quality evolved in a linear path between 2005 and 2007. This assumption allows us to take the midpoint between the two years as our quality score estimate for 2006.

National Level Calculations

At the national level, we use the National Assessment of Educational Progress (NAEP) exam to estimate the distribution of state education quality. The aim of this calculation is to generate a measure of the relative education quality for each state, and then to use these measures to shift, or re-center, the distributions of district quality within each state. The NAEP exam is given to a

representative sample of 4th and 8th grade students in each state in odd years (2003, 2005, 2007, etc.) in both math and reading.

We begin by standardizing the state average NAEP scale score for each subject using the national student level mean and standard deviation. This yields a z-score for each state in each subject which can be interpreted as the relative position of the average student in each state and subject. This interpretation of the z-score as the mean for the state is the basis for using it to re-center the district quality distributions. Just as with the district level data, we use linear interpolation to fill in data for the even numbered years between NAEP administrations².

After calculating a z-score for each state in each subject and year, we are ready to re-center the district level distributions. We do this by adding the appropriate state level z-score to each district's z-score within the state. This effectively re-centers the quality distribution within the state as our estimate of the state's quality relative to the nation. Given our previous assumptions of normality and the comparability of standard deviation units, re-centering the district quality distributions using the procedure described above allows us to compare district quality scores across state lines.

International Level Calculations

The previous section developed a procedure by which we can compare individual districts across state borders. This section extends this concept by developing comparisons between districts in the United States and their international competitors abroad. We use testing data from the Program for International Student Assessment (PISA) exam to generate international comparisons.

PISA is administered by the Organization for Economic Co-operation and Development (OECD). PISA includes both math and reading exams which are given every three years (beginning in 2000) to a representative sample of 15 year-olds in each of the participating countries. PISA scores will serve as the main basis for our international comparisons.

Just as we did with the national testing data, we start by standardizing using the appropriate student level means and standard deviations. We generate z-scores in math and reading for all PISA participating countries from the 2000 administration through the 2009 administration. We then use linear interpolation to fill in the missing years between test administrations³.

Next, we construct a group of economic competitors based on population and GDP per capita to be our international comparison group. To be included in the comparison group a country must have had a population of at least two million and a GDP per capita of at least \$24,000 (2007 USD) in 2007⁴. We also further limited the comparison group by excluding members of The

² Interpolation was used for NAEP in the years 2004 and 2006.

³ Interpolation was used for PISA in the years: 2001, 2002, 2004, 2005, 2007, and 2009.

⁴ Population and GDP per capita data come from the Penn World Table 6.3. <http://pwt.econ.upenn.edu/>

Organization of the Petroleum Exporting Countries (OPEC). Table 1 provides a list of the 25 countries included in the comparison group.

Table 1: Comparison Group

| Country | Population (000s) | GDP per Capita |
|----------------|--------------------------|-----------------------|
| Australia | 20,750 | \$39,694 |
| Austria | 8,200 | \$38,303 |
| Belgium | 10,392 | \$35,953 |
| Canada | 32,936 | \$39,089 |
| Denmark | 5,468 | \$36,198 |
| Finland | 5,238 | \$33,912 |
| France | 63,682 | \$31,447 |
| Germany | 82,401 | \$33,181 |
| Greece | 10,706 | \$29,483 |
| Hong Kong | 6,980 | \$45,446 |
| Ireland | 4,109 | \$43,351 |
| Israel | 6,990 | \$25,302 |
| Italy | 58,148 | \$30,505 |
| Japan | 127,433 | \$32,063 |
| Korea | 48,250 | \$24,950 |
| Netherlands | 16,571 | \$36,394 |
| New Zealand | 4,132 | \$27,440 |
| Norway | 4,628 | \$53,968 |
| Singapore | 4,553 | \$48,490 |
| Slovenia | 2,009 | \$27,868 |
| Spain | 40,448 | \$33,616 |
| Sweden | 9,031 | \$35,271 |
| Switzerland | 7,555 | \$39,161 |
| Taiwan | 22,829 | \$27,884 |
| United Kingdom | 60,776 | \$34,320 |

Our aim in creating this comparison group is to limit international comparison to countries with whom students are likely to compete in the global labor market. If we were to broaden our definition of competitor to include all countries who took PISA, we would be including countries that are clearly not economic competitors of the United States (e.g. Saudi Arabia, Latvia, Chile, etc.). Our more narrow definition of the comparison group ensures that our international comparisons are not weakened by including too broad a population. It is highly likely that the comparison group will need to change in the future in order to reflect changes in world economic realities, but for this iteration of the Global Report Card our criteria for selecting competitors appears reasonable.

Given our comparison group we would like to generate estimates of the distributional parameters for the student population in the comparison group countries. We already know the distributional

parameters for the entire PISA sample, but the parameters are not known for our specially-constructed comparison group. To address this, we will use generated parameter estimates to re-standardize the PISA testing data to reflect the distribution of student achievement in our more narrowly defined comparison group.

In order to generate parameter estimates we first collect the country mean and within country standard deviation on the PISA exam for each country included in the comparison group across all administrations (2000, 2003, 2006, and 2009). The mean and standard deviation for countries that took PISA in a particular administration is provided in the PISA reports; however, there are a few countries in our comparison group that did not take PISA in all the administrations⁵. In these instances we use the mean and within country standard deviation from their next available PISA result as an estimate of their mean and standard deviation in the previous cases where they did not participate in PISA.

We will use Hong Kong to illustrate this calculation. Hong Kong did not participate in the 2000 PISA administration but did participate in 2003. Our estimation procedure for the mean and within country standard deviation is to look forward to the next PISA administration in which the country participated, and to use the distributional parameters from this later administration as our estimate of the earlier mean and standard deviation. Hong Kong participated in the 2003 administration of the PISA exam and had a mean of 550 and within country standard deviation of 100. We use these values as our estimates for their 2000 mean and within country standard deviation.

After collecting the mean and standard deviation for each country in our comparison group, we use Monte Carlo simulation to estimate the distributional parameters of the overall student population in this comparison group of countries. Specifically, we use the mean and standard deviation for each individual country to generate 500,000 synthetic student observations per country, and calculate the mean and standard deviation of the combined synthetic population. We replicate this procedure 100 times and use the average mean and standard deviation across all replications as our estimates for the comparison group's mean and standard deviation. In doing so we are treating each country as having equal weight in our comparison group. That is, we allow each country to contribute 500,000 synthetic student observations to generate the mean and standard deviation for the international comparison group.

In Table 2 we provide our estimates for the control group's distributional parameters. We use the parameter estimates presented below to calculate z-scores for all countries relative to the international comparison group. Re-standardizing in this way allows us to compare mean student performance in the United States to the performance of students in the comparison group.

⁵ Countries that did not participate in at least one PISA administration include Hong Kong, Israel, Netherlands, Singapore, Slovenia, and Taiwan.

Table 2: Control Group Distributional Parameters

| Year | Math Mean | Math StdDev | Reading Mean | Reading StdDev |
|-------------|------------------|--------------------|---------------------|-----------------------|
| 2000 | 514.89 | 99.63 | 507.12 | 98.60 |
| 2003 | 512.58 | 100.10 | 502.42 | 99.32 |
| 2006 | 509.73 | 97.84 | 500.77 | 101.76 |
| 2009 | 509.92 | 97.78 | 503.04 | 95.72 |

We can interpret the z-score for the United States as the relative position of the mean student in the United States in the international comparison group's student achievement distribution. We use the z-score for the United States to shift the district level distributions once again. By adding the United States' z-score to each district's z-score we are re-centering the distribution of student achievement around the mean performance of the United States relative to the comparison group.

Additionally we can use the mean scores for the United States to calculate the relative position of the mean student in other countries' distribution. For example, we could directly compare the average student in the United States to students in Canada, Switzerland, and Singapore.

Conversion to Percentiles

In the sections above we developed measures of education quality in both math and reading that allow for the comparison of education quality at individual districts in the United States not only across state and national borders, but also to an international comparison group comprised of US economic competitors. To this point we have dealt primarily with standard deviation units or z-scores. For ease of interpretation, the final calculation we make in constructing the Global Report Card is to convert these z-scores to percentiles.

III. Anticipated Criticisms and Rebuttals

We make no claims that this Global Report Card is a perfect reflection of school district student achievement relative to international norms. The question is whether the limitations of the Global Report Card are acceptable for a first attempt. In essence, we want to know whether we have more information with the Global Report Card than we would have were it never developed and publicized.

Critics could rightly highlight a number of defects in the Global Report Card, but we believe that those defects are not fatal. For example, critics might observe that the state, national, and international tests are designed to measure different things, undermining any attempt to compare across them. Of course, this criticism is true, but we believe that there is an underlying quality of student achievement that is imperfectly and indirectly captured by all of the tests. There is information about that underlying quality that we can obtain if we compare across tests that would be lost if we refused to make any comparisons.

In addition, critics might note that the ages at which students take the exams differs across the state, national, and international levels. Our response is essentially the same as to the previous concern. There is an underlying student achievement that will be reflected by students in all grades, which we capture imperfectly by comparing students of different ages. If we were to focus only on the same or closest age students we would reduce the error introduced by comparing different aged students but we would have many fewer observations and much less precise estimates of the underlying school district quality. We think the trade-off of precision of estimate for comparability of age is worth it.

Critics could also note that the results for all of these tests are not always normally distributed, which we assume them to be. Our feeling is that the results are often approximating a normal distribution, making our assumption reasonable even if imperfect. Again, we would rather not make the best the enemy of the good.

Critics could observe that we make no adjustments for student characteristics or other factors and so cannot make claims about the contributions of schools to achievement. That is true, but we are not attempting to make any claims about school contributions to achievement. We are simply trying to gauge student achievement in districts and states relative to an international norm of achievement so that we can have some information of where students are doing well and where they are doing poorly regardless of what caused that higher or lower achievement. Perhaps in subsequent versions of the Global Report Card we will attempt to isolate school quality from student achievement.

We are certain that there are more criticisms, but this should help address some of the major concerns.