v03 – 2014-07-15

George W. Bush Institute Global Report Card 3.0
Technical Appendix

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

Jonathan N. Mills
Department of Education Reform
University of Arkansas
208 Graduate Education Building
Fayetteville, AR 72701
(479) 575-3172; jnm003@uark.edu

Introduction

This document describes the methods used to construct the third version of the George W. Bush Institute's Global Report Card (GRC 3.0). The purpose of the GRC is to place school and district performance on state assessments within the broader national and international testing contexts. While prior versions of the GRC have only been able to focus on district-level performance (Greene & McGee, 2011), the GRC 3.0 uses more detailed information on school-level performance to translate school level performance to national and international scales.

While school-level proficiency scores are available for most public schools across the United States, these measures have rarely been placed in national and international contexts. Without such adjustments, it is hard to recognize that public schools with low reported performance in states with relatively rigorous performance standards are actually doing quite well with their students, while high performing schools in states with low standards might not be as strong as they appear. In short, the GRC 3.0 uses school-level proficiency data to construct measures of school quality that place public schools across the United States on the same scale. To do this, we combine state, national, and international testing data to create a single measure of student achievement. 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.

In the following sections, we describe the process used to construct the GRC 3.0 estimates of school quality. This document builds on the technical appendix associated with prior versions of the GRC (see Greene & McGee, 2011).

Construction of Global Report Card 3.0

In this section, we describe the steps taken to construct GRC 3.0 estimates of school quality. We begin by describing the process we use to combine our school-level data across years and then describe how these averages are used to place both schools and districts in state contexts. In addition, we describe the methods and assumptions required to place school- and district-level performance estimates in national and international contexts.

Averaging Across Years

The goal of GRC 3.0 is to place school-level performance in state, national, and international quality distributions. We use as our estimates of school quality data on school-level proficiency on state assessments compiled through the U.S. Department of Education's ED*Facts* initiative.[1] Through the ED*Facts* initiative, the Department of Education has combined into a single database K-12 performance data provided by the various state education agencies for use by policymakers, practitioners, and researchers. Specifically, the *EDFacts* data includes school-level information on the number of test takers as well as the percentage of students scoring proficient or above on state assessments in grades 3 through 8 and high school for school years 2008-09 through 2011-12. We have obtained these data from the Institute for Education Science's National Center for Education Statistics (NCES).

We begin by first aggregating school-level proficiency data over four years to create a single measure of performance for each school and district. Individual schools and districts can experience moderate fluctuations in performance in any given year, and we therefore take this first step of aggregating across years to provide a more accurate picture of school and district quality. We therefore average the reported proficiency rates for each school and district across school years 2008-09 through 2011-12 while weighting each estimate by the number of students taking the test in a given year. These student-weighted school and district proficiency averages are calculated separately for each grade (3 through 8 and high school) and subject (math and reading).

It is important to note that by averaging proficiency rates across years, we are assuming that states have not dramatically changed their proficiency standards over the four-year time period examined. If this were not the case, our preferred method would be to first standardize the proficiency scores within each subject, year, and grade and then calculate student-weighted averages of these standardized values. Unfortunately, because we do not have within-state student-level standard deviations for each test, we would not be able to place these student-weighted standardized performance calculations on similar scales through an additional standardization step. As such, we are limited to our current process of combining school-level proficiency rates by computing student-weighted school and district proficiency averages between 2008-09 and 2011-12.

[1] For more information, see http://www2.ed.gov/about/inits/ed/edfacts/index.html

State Level Calculations

After computing student-weighted school and district performance averages using the proficiency rates and test-taker counts reported in the ED*Facts* database, we employ the same methodology developed for earlier versions of the GRC (see Greene & McGee, 2011) to compute our GRC 3.0 estimates of school and district quality.

First, using school-level data on test takers and average proficiency rates, we calculate student-weighted proficiency averages for each state. These measures represent estimates of the performance of the average child in each state over the four year period.

Next, we standardize the school- and district-level average proficiency rates described in the proceeding section in order to get a measure of how far each school and district is from the mean performance in the state. Because we do not have access to student-level test distributions for each subject, test year, and grade, we must make an additional assumption in order to calculate these school and district standardized scores.

In particular, if we assume that student performance is approximately normal within each school, district, and state, then we can use an inverse normal transformation to estimate each school and district's distance from the state's mean performance. In doing so, we are effectively using the assumption that each school and district's percent proficient represent specific points on the cumulative normal distribution to translate these percentages to the standard normal probability density function using the inverse normal transformation. This point, or z-score, represents the number of standard deviations units about the mean on the standard normal curve where the specified percentage of the population would achieve the cut score.

Nevertheless, it is important to note that these school- and district-level z-scores are not comparable across states because cut scores in some states are more difficult to achieve than in other states. Thus, in order to create school and district performance measures that are roughly comparable across states, we need to shift the school- and district-level z-scores based on how difficult the state's proficiency cut score is to achieve. We do this by first applying the same inverse normal transformation outlined above to the student-weighted state average percentages computed earlier. We then use the resulting z-score as our measure of how difficult the state's proficiency cut score is to achieve. Finally, we shift each school and district z-score by simply subtracting the state's z-score. The shifted school and district z-scores serve as our estimates of how far each school and district is from the mean state performance in standard deviation units.

Finally, we aggregate the standardized proficiency percentages (shifted z-scores) for each school and district across grade to the subject level. We perform this aggregation by calculating the student-weighted average of the z-scores separately at the school and district levels. The resulting values serve as our primary estimates of school and district quality and form the basis for each of our subsequent calculations. We are particularly grateful to Martin West of Harvard University for suggesting this technique for estimating the standard deviation of achievement within each state.

National Level Calculations

Up to this point, our calculations have produced quality estimates that allow us to compare school and districts within states. In addition, we are interested in how the quality of education varies across the states. At the national level, we use the National Assessment of Educational Progress (NAEP) 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 use these measures to re-center the distributions of state- and district-quality outlined in the prior section within each state. The NAEP provides an ideal means to estimate the distribution of state educational quality as it is administered biennially to a representative sample of 4^{th} and 8^{th} grade students in every state in both reading and math. For our purposes, we use data from the 2009 and 2011 NAEP administrations.

We begin by standardizing the state average NAEP scale scores within each year 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.

After calculating a z-score for each state in each subject and year, we combine the estimated z-scores across years by taking a simple average of the 2009 and 2011 standardized scores. We then use the resulting averages to re-center the district level distributions by adding the appropriate state level z-score to each school-level and district-level quality measure 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 for the 2009 and 2012 test administrations.

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 GPD per capita of at least $24,000 (2007 USD) in 2007.[2] We also further limited the comparison group by excluding members of The 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 (1000s)	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

[2] To maintain consistency across GRC versions, we use the same criteria as was used in prior versions of the GRC. See Greene and McGee (2011).

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. Earlier versions of the GRC relied on Monte Carlo simulations to generate these parameters because the OECD did not make the necessary distributional parameters available for our comparison group (see Greene & McGee 2011). Fortunately, the NCES now calculates both test means and standard deviations for all subsets of PISA participating countries in their Elementary/Secondary Information System (ELSi)[3]. Specifically, we retrieve both the mean and standard deviation for our comparison group countries in 2009 and 2012 in both math and reading. We then average these values across years to create estimates of the comparison group testing distribution. Finally, we shift each school and district's national z-score by adding to it a calculated z-score capturing the position of the United States in the comparison group countries' testing distribution.

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. After computing these z-scores within year, we then collapse these values across years using simple averages. We then use the resulting average z-score for the United States to shift the school- and district-level distributions once again. By adding the United States' z-score to each school and 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 schools and 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.

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.

[3] See http://nces.ed.gov/ccd/elsi/.

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.

Similarly, we do not attempt to estimate the sampling distribution of our GRC estimates which could be used to calculate confidence intervals around individual school performance estimates. While it would be helpful to provide estimates of the variance underlying our school and district performance estimates, the nature of our data do not allow for these calculations. In short, while it is important to realize that the GRC estimates of school and district quality are estimated with error, we are unable to precisely estimate that error. At the same time, it is important to recognize that both the reliability requirements on state tests, as well as our exclusion of schools and districts with fewer than 100 tested students over a four year period, are sufficient to significantly minimize standard error of measurement concerns.

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.

Finally, critics could point out that by initially averaging school- and district-level proficiency percentages across years, we are not accounting for differences in proficiency standards within states between 2008-09 and 2011-12. We agree that the preferred method would be to first standardize these scores within state, subject, year, and grade; however data limitations effectively remove our ability to do this. In particular, because we do not have information test distributions at the student level, we would not be able to re-standardize the resulting averages of standardized values; and would therefore lose our ability to map our measures back to the standard normal curve. Nevertheless, while we agree that these measures will not fully account for changing proficiency standards within states, our estimates still represent a meaningful progress.

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

References

Greene, J. P. & McGee, J. M. (2011). George W. Bush Institute Global Report Card: Technical Appendix. Report prepared for the George W. Bush Institute.