Are International Tests Worth Anything?

Do the U.S. rankings on international achievement tests signal doom for the country's future standing in the world? Mr. Baker sets out to answer that question by looking beyond the test scores to other dimensions on which nations can be compared.

BY KEITH BAKER

THE IDEA THAT America was being harmed because our schools were not keeping up with those in other advanced nations emerged after Sputnik in 1957, took a firm hold on education policy when *A Nation at Risk* appeared in 1983, and continues today. Policy makers justify this concern by pointing to evidence showing that, for individuals within the U.S., higher test scores predict a number of important life advantages, such as going on to college and making more money as an adult. From this they extrapolate that higher *national* test scores correlate with global success. The origins of the notion that education is crucial to the nation date back to the Founding Fathers, especially Jefferson, who held that a well-educated citizenry was the foundation of a nation's, especially a democracy's, success in the world.

Since Sputnik, the evidence driving worries about the performance of U.S. schools has come primarily from a series of international achievement testing programs that started in 1964 with the First International Mathematics Study (FIMS). This was followed by the Second International Mathematics Study (SIMS), the Third International Mathematics and Science Study (TIMSS), and, most recently, the Programme for International Student Assessment (PISA).

In this article I will show that for the U.S. and for the top dozen or so most-advanced nations in the world, standings in the league tables of international tests are worthless. There is no association between test scores and national success, and, contrary to one of the major beliefs driving U.S. education policy for nearly half a century, international test scores are nothing to be concerned about. America's schools are doing just fine on the world scene.

BACKGROUND

When policy makers and politicians infer that the same relationship holds *between* nations as is found *with-*

■ *KEITH BAKER is retired as a researcher for the U.S. Department of Education and now lives in Utah.*

Photo: Liquid Library

in nations, they commit the logical error known as the ecological correlation fallacy. Evidence of the effects of education within nations does not transfer to differences among nations.

To see the ecological fallacy at work, picture fans doing "the wave" at a football stadium. Watching only the up and down movements of individual "citizens" of "Stadium Nation," however, tells us nothing about the direction in which the wave circles the stadium, its "national" movement. If we had two such stadiums side by side, and our view from the Goodyear Blimp showed one wave circling to the left and the other circling to the right, neither wave nor both would tell us how the citizens are moving. Going down into the crowd and watching citizens move up and down tells us nothing about how the wave appears from the blimp — or what is going on in the neighboring stadium. Likewise, the effects of high test scores on the individuals within a nation tell us nothing about the relationship of those test scores to national success.

The mathematics of the ecological correlation fallacy is a proof that generalizing from the relationship between variables at the individual level to larger aggregate levels, such as nations, is indeterminate.[1] That is, maybe the generalization holds, maybe it doesn't. Therefore, when such a generalization is made, we must treat it as a hypothesis, never as established fact, until it has been confirmed at the level of nations. Only then is it wise to act on the hypothesis.

FIMS

To see if the leap from within-nation results to between-nation results is justifiable, I looked at how well test scores on FIMS, the first international comparison study, predicted national success in the first half-decade of the 21st century. FIMS was administered in 1964 to samples of 12-year-olds in 11 nations. Today's world is largely a world created and operated by the now 55-year-old FIMS generation. If there is a connection between high test scores and national success, it will show up in looking at how well the 1964 FIMS scores predicted where nations are today. Among the 11 FIMS nations, the U.S. finished second to last (ahead of Sweden).

Jefferson told us where to look to see if a nation is a success. He did not say to look at test scores. Instead, he said to look at "life, liberty, and the pursuit of happiness." To find out how the FIMS nations are doing on life, liberty, and the pursuit of happiness, I looked at seven indicators of national success. I related them to FIMS scores.

Wealth. First, and perhaps most important to a nation, is the creation of wealth. The best measure of generating wealth is per-capita GDP adjusted for cost of living differences, or purchasing power parity (PPP-GDP). The wealth of nations scoring higher than the U.S. on FIMS averaged 73% of the per-capita income in the U.S. in 2002.[2] FIMS scores in 1964 correlate at r = -0.48 with 2002 PPP-GDP. In short, the higher a nation's test score 40 years ago, the worse its economic performance on this measure of national wealth — the opposite of what the Chicken Littles raising the alarm over the poor test scores of U.S. children claimed would happen.

Jefferson told us where to look to see if a nation is a success. He did not say to look at test scores. Instead, he said to look at "life, liberty, and the pursuit of happiness."

Rate of growth. One can argue that since the U.S. had a big post-WW II economic lead over the rest of the world, the rate of economic growth is at least as important as GDP as an indicator of national achievement. The nations that scored better than the U.S. in 1964 had an average economic growth rate for the decade 1992-2002 of 2.5%; the growth rate for the U.S. during that decade was 3.3%. The average economic growth rate for the decade 1992-2002 correlates with FIMS at r = -0.24. Like the generation of wealth, the rate of economic growth for nations improved as test scores dropped.

Productivity. GDP is a measure of a nation's total economic output. Productivity — GDP per hour worked — might be a better measure of a nation's economic success than GDP, since nations differ in the number of hours a year that the average worker spends at work.[3] There is no relationship between FIMS scores and hourly output, r = -.03. In 2004, the average hourly output of those nations that outscored the U.S. in 1964 was 3.4% lower than U.S. productivity, though the three nations with higher hourly output all had higher test scores than the U.S. However, on the PISA test, which I discuss below, none of these three nations scored higher than the U.S.

Quality of life. Some argue that GDP is too simple a measure of national goals, that there is more to the good life than money. The United Nation's Quality of Life Index addresses this concern. Those who worry about

international test score standings base their worries on an assumption that high-scoring nations are more successful at doing the things nations should be doing, and offering a good quality of life to citizens is one of those things. But again, they are wrong. The average rank on the Quality of Life Index for nations that scored above the U.S. on FIMS was 10.8. The U.S. ranked seventh (lower numbers are better). FIMS scores correlated with Quality of Life at r = -0.57.

Livability. An alternative to the Quality of Life Index, the Most Livable Countries Index, shows that six of the nine countries that scored higher on FIMS than the U.S. are worse places to live. Livability correlates with FIMS scores at r = -.49.

Democracy. Jefferson also held that a well-educated citizenry is necessary for good democratic government. On the Economy Intelligence Unit's Index of Democracy, those nations that scored below the median on FIMS have a higher average rank on achieving democracy (9.8) than do the nations that scored above the median (18). Once again, the U.S. scored higher on attaining democracy than did nations with higher 1964 test scores.

Creativity. A good school system should foster creativity. The number of patents issued in 2004 is one indicator of how creative the generation of students tested in 1964 turned out to be. The average number of patents per million people for the nations with FIMS scores higher than the U.S. is 127. America clobbered the world on creativity, with 326 patents per million people. However, FIMS scores do correlate with the number of patents issued: r = .13 with the U.S. and r = .49 without the U.S.

THE FIMS PREDICTIONS

The hypothesis that low scores on international tests lead to national disaster, or at least inferior performance as a nation, predicts that the nine nations scoring higher than the U.S. on FIMS should outperform the U.S. on measures of national success. If the hypothesis is correct, nations with higher FIMS scores than the U.S. should be doing better than the U.S. on the seven indicators of national success in a world that is now run by the FIMS generation.

What's the bottom line? Altogether, there are 61 possible comparisons between the U.S. and a higher-scoring nation across the seven indicators. According to the hypothesis, 100% of these comparisons — or, at the very least, an impressive majority — will show the U.S. doing a worse job than the higher-scoring nations. In fact, the U.S. comes out on top in 74% of the comparisons.

In the face of such evidence, we can do more than reject the widely held hypothesis that high test scores lead to national success in the future. We can also hypothesize that high test scores are damaging to nations. That the U.S. comes out on top in national success in 74% of the comparisons with higher-scoring nations is statistically significant (p <.0001, binomial test).

Sputnik went up, and America's test scores went down compared to other advanced nations. But there was no need to panic or to proclaim, as so many did, that America's schools were in a crisis of poor performance. In looking at the world four decades after FIMS, the U.S. turned out more than just okay compared to nations with higher test scores. No matter how you look at it, high test scores in 1964 were not positive predictors of how the world would turn out. At best, international test scores are useless and may well be harbingers of failure, rather than success.

The logic of the ecological correlation fallacy warned that jumping to policy conclusions from international tests was a dubious enterprise. Since this logical fallacy was known by 1950, there was no excuse for policy makers at the time of FIMS or at any time since to proclaim the existence of problems in U.S. schools because some other countries posted higher test scores.

PISA

PISA, a second and more recent international testing program, included more than twice as many nations (n = 27) as FIMS (n = 11).[4] Like FIMS, PISA shows

"I'm taking practical language arts. I'm studying medical Latin, restaurant French, musical Italian, and business English."

no connection between high test scores and how well a nation does at achieving wealth, growth, democracy, or quality of life for its people.

On these indicators of success, the nations that scored at the PISA average generally outperformed those scoring either above or below average. For example, per-capita GDP was $22,495 for the 11 nations scoring above average, $34,414 for the five average nations, and $16,375 for the 11 below-average nations. The same pattern holds for quality of life, democracy, and creativity as measured by patents.

International comparisons on many factors show that Norway is the best place in the world to live, and, like the U.S., Norway scored right at the PISA average. Mediocre test scores correlate with better, more successful countries than do top scores (or lower scores). Mediocrity in test scores is, for nations, a good thing! This finding is highly counterintuitive. Why should it be so?

CONCLUSIONS

Among high-scoring nations, a certain level of educational attainment, as reflected in test scores, provides a platform for launching national success, but once that platform is reached, other factors become more important than further gains in test scores. Indeed, once the platform is reached, it may be bad policy to pursue further gains in test scores because focusing on the scores diverts attention, effort, and resources away from other factors that are more important determinants of national success.

The fixation on test scores has so dominated policy that little attention has been paid to finding out what makes America's schools the best in the world with regard to international economic competition. But a recent conversation I had with a Swede now living in Los Angeles seems to point in the right direction. He holds a high position in a bioscience company and has lived in 10 different nations. He told me, "There is no doubt that graduates of European high schools know a lot more than American grads, but I prefer my kids go to school in America because Americans acquire a spirit that the other countries lack." Other anecdotal sources suggest this "spirit" involves ambition, inquisitiveness, independence, and perhaps most important, the absence of a fixation on testing and test scores.

"Imagination is more important than knowledge," Einstein observed, and this principle applies to physics, to science, to what makes a modern economy succeed, and to what schools should teach. As to the relative importance of test scores and that "spirit" that U.S. schools seem to cultivate better than those anywhere else, Einstein again is on the mark: "The true sign of intelligence is not knowledge but imagination." How America's schools beat the rest of the world in developing imagination may not yet be clear, but that — rather than raising test scores — should be the focus of both policy and research.

For more than a quarter of a century, the American public has been barraged by politicians and pundits claiming that America's schools are disaster zones because we are not at or near the top of the league standings in test scores. This claim is flat out wrong. It is wrong in fact, and it is wrong in theory. For almost 40 years, those who believe this fallacious theory have been leading the nation down the wrong path in education policy. It turns out that the elementary teachers who have said all along that there is more to education than what is reflected in test scores were right and the "experts" were wrong.

Trying to raise America's test scores in comparison to those of other nations is worse than pointless. It looks to be harmful, for the only way to do it is to divert time, energy, skill, and resources away from those other factors that propel the U.S. to the top of the heap on everything that matters: life, liberty, and the pursuit of happiness.

The fixation with test scores also harms the nation by diverting time, attention, and resources away from America's real educational problems, such as too few minorities graduating from college, the run-down schools in the nation's inner cites, misdirected parental interference in schools, and the lack of parental and administrative support for teachers. There are more, of course, but nowhere on the list of our educational problems should we ever again find worries over our performance on tests compared to that of other nations.

---

1. Nor, as I showed in my article, "Yes, Throw Money at Schools," *Phi Delta Kappan*, April 1991, pp. 628-32, do relationships found among larger aggregates, such as nations or schools, generalize to individuals.

2. Although I began the analysis with statistical tests of the hypothesis that high scores lead to high national success in the future, these results are not presented, since statistical testing turned out to be unnecessary because the hypothesis underlying the policy concern about America's poor test scores is a directional hypothesis. Being a directional hypothesis, it is sufficient to conclude that the null hypothesis cannot be rejected when we find negative relationships between the independent and dependent variables.

3. Hourly output is available only for the OECD nations, which include nine of the 11 FIMS nations.

4. Using PISA scores to examine the effects of high test scores on national success has both problems and advantages. The main problem is that the students tested in 2000 have not had time to have much effect on their nations. PISA's advantage is that it included many more countries than FIMS. Inspection of the international test score data suggests it is reasonable to assume that national test scores are stable over time, a conclusion confirmed by the fact that similar patterns show up for PISA and for FIMS.