EDUCATION WEEK
Published Online: July 24, 2013

When Bad Things Happen to Good NAEP Data

By Stephen Sawchuk

The National Assessment of Educational Progress is widely viewed as the most accurate and reliable yardstick of U.S. students' academic knowledge.

But when it comes to many of the ways the exam's data are used, researchers have gotten used to gritting their teeth.

Results from the venerable exam are frequently pressed into service to bolster claims about the effect that policies, from test-based accountability to collective bargaining to specific reading and math interventions, have had on student achievement.

While those assertions are compelling, provocative, and possibly even correct, they are also mostly speculative, researchers say. That's because the exam's technical properties make it difficult to use NAEP data to prove cause-and-effect claims about specific policies or instructional interventions.

"It's clearly not NAEP's fault people misuse it, but it happens often enough that I feel compelled to call [such instances] 'misnaepery,' " said **Steven M. Glazerman**, a senior fellow at Mathematica Policy Research, a Princeton, N.J.-based research and policy-evaluation nonprofit.

"NAEP is so tempting, because it has very wide coverage," he said. "But what it tries to do is actually pretty modest, pretty narrow. And that's a good thing."

Contrasting Claims

Often called "the "nation's report card," NAEP represents the achievement of a nationally representative sample of students at three grade levels: 4, 8, and 12. Under the Elementary and Secondary Education Act, each state receiving federal Title I funds also must participate in the exam at the 4th and 8th grade levels in reading and math every two years.

Because of this stipulation, achievement trends across states can be compared, an impossibility using the results of states' own hodgepodge of exams.

Twenty-one urban districts also volunteer to have their students' results reported through the Trial Urban District Assessment, or TUDA.

Parsing Claims

Advocates are fond of making claims about what data from the National Assessment of Educational Progress mean, but not all of them stand up to scrutiny.

NAEP data are generated through a technique known as matrix sampling, in which a portion of exam questions are given to each sample of students; no child takes a "full" exam.

In a sense, what has made NAEP unique in the annals of testing—its commonality and independent administration in an era of cheating scandals—has also rendered it susceptible to misinterpretation and misuse.

"The NAEP exams have good measurement quality and assess subjects other jurisdictions don't have assessment data on," said Sean P. "Jack" Buckley, the commissioner of the U.S. Department of Education's National Center for Education Statistics, which administers the exam. "They are comparable across state lines, which is unusual, and they are well known in the policy world. And unlike trying to negotiate with states and [privacy laws], NAEP data are right there on our website."

Use of Data:

"Public education is supposed to be the great equalizer in America. Yet today the average 12th grade black or Hispanic student has the reading, writing, and math skills of an 8th grade white student."

—From a 2009 *Wall Street Journal* op-ed written by Joel I. Klein, then the chancellor of the New York City school system, and the Rev. Al Sharpton

Problem:

NAEP scales differ by subject and grade.

The downside is that examples of "misnaepery" are legion.

During the height of implementation of the No Child Left Behind Act, the most recent rewrite of the ESEA, dozens of press releases went out from the U.S. Department of Education, then headed by Margaret Spellings, attributing gains on NAEP to the effects of the law.

In the District of Columbia, promoters of the policies instituted by former Chancellor Michelle A. Rhee **have seized on readings of NAEP** as evidence that her aggressive changes to personnel policies boosted student achievement.

On the other hand, a **report recently released by the Broader, Bolder Approach to Education**—a coalition housed in the Education Policy Institute, a left-leaning think tank—drew on the data to support the exact opposite conclusion. And Broader, Bolder's claims that increased access to charter schools, teacher evaluations tied to student test scores, and school closures in Washington and two other cities didn't lead to improvements for poor and minority students were picked up and repeated by influential education figures.

Parsing Claims (Cont.)

Use of Data:

"Among these low-performing students [on 2009 NAEP in reading], 49 percent come from low-income families. Even more alarming is the fact that more than 67 percent of all U.S. 4th graders scored 'below proficient,' meaning they are not reading at grade level..."

—From advocacy organization StudentsFirst's website

Problem:

NAEP's definition of "proficient" is based on "challenging" material and is considered harder than grade-level standards.

"The lesson of the new report: Billions spent on high-stakes testing have had minimal to no effect on test scores," New York University education historian **Diane Ravitch wrote** about the paper. "High-stakes testing has failed."

The practice seems to transcend typical divides: Parties representing both liberal and conservative points of view have drawn on NAEP data to advance an argument.

Teachers' union leaders are fond of pointing out that Massachusetts, among the highest-performing states on NAEP, has **strong collective bargaining laws**. Yet conservative groups point to heavily

unionized cities with poor student performance, like Detroit and Chicago, **to advance an opposing argument**.

Statistics 101

Most such claims suffer, researchers say, from failing to consider that a correlation or relationship between two points of data does not prove causation.

"They're committing the fundamental and almost inexcusable error of leaping to the causal conclusion they prefer, when hundreds of others are possible," said Grover M. "Russ" Whitehurst, the director of the Brown Center for Education Policy at the Brookings Institution and a former director of the U.S. Department of Education's research wing.

Another spurious use: treating NAEP data as though they track the same students as they progress through school. Such longitudinal data generated from state tests are frequently used by statistical researchers, who can take into account students' background characteristics to control for the effect of poverty or family education on scores.

But NAEP data, represents repeated cross-sectional snapshots of achievement, not the progress of individual students, making it much more challenging to institute such controls.

"I can understand why people think if test scores go up, it's because schools get better," said Matthew Di Carlo, a senior fellow who writes about education research for the Albert Shanker Institute, a think-tank affiliated with the American Federation of Teachers. But with NAEP, "you're comparing two different groups of students and assuming they're not changing over time."

Some misuse occurs entirely outside of policy contexts.

"The states see this happening more than even we do nationally," said Cornelia Orr, the executive director of the National Assessment Governing Board, the body that sets policy for NAEP.

For instance, she said, "they're concerned about real estate companies and how they abuse their own state test data, and they're concerned it will happen with NAEP."

New Techniques?

The issue has been sufficiently worrisome that a joint

Parsing Claims (Cont.)

Use of Data:

"When the market-based policies at the center of the reform agenda play out in a comprehensive manner across many years, the results, as captured in reliable data, are not encouraging. ... Reforms that produce a lack of progress on improving test scores or closing achievement gaps are no different from the 'status quo' that they purport to break."

—From "Market-Oriented Reforms' Rhetoric Trumps Reality," by the Broader, Bolder Approach to Education coalition

"In Charlotte, N.C., and Austin, Texas, both cities in right-to-work states where collective bargaining is not required, students in 4th and 8th grade are performing higher than the national average in both reading and math."

—From "Collective Bargaining and Student Academic Achievement," by the American Action Forum

Problem:

Both statements imply that specific policies affected scores, but casual conclusions are difficult to validate using NAEP.

Parsing Claims (Cont.)

Use of Data:

"We subtracted the percentage of students in the state who scored proficient or better from the state NCLB test from the percentage of students in that state who passed the NAEP, and used this difference (or gap) to align each school and district test scores across the nation."

—From real estate website NeighborhoodScout

task force of NAGB and the Council of Chief State School Officers began to catalog it in 2009-10.

Scholars say that is possible to do high-quality research using NAEP data, but doing so appropriately requires research expertise beyond what most lobbyists and policy analysts possess.

Problem:

NAEP cannot be used to generate comparable school results.

Claims compiled by Stephen Sawchuk.

"NAEP is just an outcome measure. It's no different from an IQ test or the number of teachers with advanced degrees," Mr. Whitehurst said. "The ability to draw causal inferences about any education variable depends not on NAEP, but on the quality of the research design for which NAEP is the outcome measure."

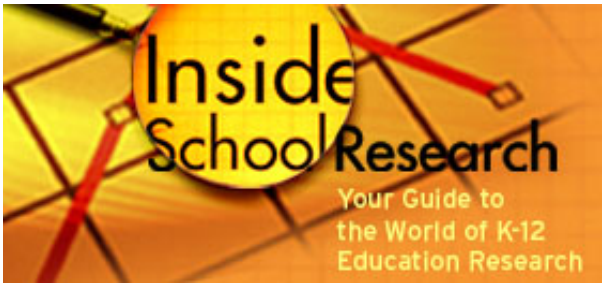
High-quality studies have drawn on NAEP results, for example, to estimate **the impact of Georgia's expanded early education program**, Mr. Whitehurst noted.

Mr. Glazerman cautioned, though, that such studies are few and far between.

Advocates' desire to seek quick confirmation for their policy prescriptions—especially when they are gaining or losing momentum—means that it's unlikely that interest in using NAEP for policy analysis will end anytime soon.

"There is just this unwillingness to accept that policy analysis is difficult, takes a long time, and often fails to come to strong conclusions about individual policies," said Mr. Di Carlo.

RELATED BLOG



Visit this blog.

Over time, the difficulties inherent in interpreting NAEP results have even posed challenges for NCES and NAGB, which must weigh how to report and disseminate data from the exam to minimize misinterpretations.

For example, the NCES itself has on occasion produced reports that include correlations, Mr. Buckley noted. Even when accompanied with caveats, he said, they have been misinterpreted in press accounts.

Still, Mr. Buckley said, the benefits of NAEP data far outweigh the harm that accompanies ill usage.

"We're not the country's education data police," Mr. Buckley said of the NCES. "We want the data to be useful, and we trust that the marketplace of ideas will drive the bad uses out."

Vol. 32, Issue 37