THE RURAL SCHOOL AND COMMUNITY TRUST

It's about Fairness: The Importance of Using "Confidence Intervals" for Making AYP Judgments About Small Schools

[*Note*: This is a summary of a more detailed report by Dr. Theodore Coladarci, entitled *Gallup Goes to School: The Importance of Confidence Intervals for Evaluating "Adequate Yearly Progress" in Small Schools*. The longer report, which assumes little familiarity with statistics, presents a thorough discussion of the issues and includes formulas, charts, and a reference list. *Gallup Goes to School* is available as a downloadable PDF file on The Rural School and Community Trust Web site (http://www.ruraledu.org/docs/nclb/coladarci.pdf).]

Introduction

The ambitious agenda of the No Child Left Behind (NCLB) Act sets challenges for public schools as never seen before. These challenges are particularly troublesome for states that have a sizable rural population, where the demands of the law often are at odds with the reality of rural education.

Among the most ambitious NCLB requirements is that, by 2014, all students must be proficient on "challenging academic content standards and challenging student achievement standards." To monitor movement toward that goal, each state must determine whether its schools are making "adequate yearly progress" (AYP) in reading and mathematics.[1] This is done by testing students annually in certain grades and then determining the percentage of students in each school who score proficient or above. A school "makes AYP" if its proficiency percentage meets the state's target for that year (as in the case where, say, the target is 45% and the school's proficiency percentage is 50%). Targets increase over time so that all students will be proficient by 2014.

So what's the problem?

There are many ways for a school to *not make* AYP: NCLB requires AYP evaluations—for reading and for mathematics—based on all students as well as separate AYP evaluations for the economically disadvantaged, major racial and ethnic groups, students with disabilities, and students with limited English proficiency. A school that does not make AYP is subject to specific sanctions under NCLB. The sanctions get more severe with each year a school fails to make

[1] Actually, AYP applies equally to schools and school districts. Because many of the issues are the same, only *schools* will be referenced in this summary (for ease of presentation).

AYP. After five years, for example, a school may be identified for restructuring, which could mean turning over school operations to the state or private enterprise. Clearly, a lot rides on the comparison of a school's proficiency percentage with the corresponding AYP target.

But how reliable are these percentages?

It turns out that a school's proficiency percentage (like any measure of school performance) varies from year to year in much the same way that the results of a Gallup poll will vary from one random sample to another. And this random variation is much greater for smaller schools than for larger schools. Using Maine data, Coladarci shows that among really small schools, the percentage of proficient students from one year to the next could *decline* from 60% to 13% or *increase* from 11% to 57%. By contrast, among larger schools performance was much less "volatile"—declining or increasing by fewer than 10 percentage points.

What does this mean for AYP? The problem is that when a small school drops below the AYP target one year, it is quite likely that this school had a "bad bounce" rather than a real decline due to weak instruction, poorly aligned curriculum, ineffective leadership, and the like. Given the high-stakes consequences of AYP, the important policy question is this: When a school falls short of the AYP target, how do we know that this school—particularly if it is a small school—*truly* is not making adequate progress?

Confidence intervals

When a pollster asks a random sample of likely voters to weigh in "pro" or "con" on some current event, you know that the percentage falling in either category would be different if a new random sample were selected from this population ("likely voters"). This is why reputable pollsters attach a "margin of error" to their results. If the pollster reports that, say, 55% (±4%) of respondents approve of our foreign policy in the Middle East, we conclude that between 51% and 59% of the population of *all* likely voters feel this way. "55% (±4%)" is a confidence interval: a range of values within which we are reasonably confident the true, or population, value lies.

The same reasoning applies to AYP, and here's the logic: We know that a school's proficiency percentage is subject to random variation from year to year. This means that a school's "observed" proficiency percentage—what you calculate directly from the test scores—is merely an estimate of the school's "true" level of proficiency.[2] A confidence interval addresses the fundamental question, "Where does the true proficiency percentage for this school probably fall?" If you are told that 50% of your school's students are proficient, with a confidence interval of 45% - 55%, you conclude that your schools' true proficiency percentage could be as low as 45% or as high as 55%. To evaluate AYP, simply compare the AYP target to the upper limit of the confidence interval (55% in the present example): If the AYP target falls above the upper limit, the school has not met AYP; if the target is less than or equal to the upper limit, the school meets AYP.

[2] Think of a school's true proficiency percentage as being what we would get if, as Rich Hill says, "we could test an infinite number of students from the school's catchment area an infinite number of times on all the test questions that might be asked."

As you see, then, a school can meet AYP even though the observed proficiency percentage is lower than the AYP target. In this case, the difference between the two is not large enough for you to conclude—with confidence—that the school truly falls short of making adequate yearly progress. Think of it this way: It's like tossing a coin 50 times and getting 20 heads rather than the expected 25. Your conclusion, no doubt, would be that five fewer heads than expected is not a meaningful discrepancy in this instance. There is insufficient evidence that the coin is biased, and the assumption of a fair coin therefore stands. It's the same with AYP.

How school size figures in

Although no sample is free of sampling error, the size of this error is negatively related to sample size: the smaller the sample, the wider the margin of error. Take a school where 50% of the students are proficient. If this school had only five students, the true percentage could be as low as 17% or as high as 83%. With 300 students, however, the interval width reduces to 44% - 56%, and in a humongous school with 5,000 students, the interval width shrinks to 49% - 51%.

It stands to reason that a confidence interval will be relatively narrow when "*n*" is large and, conversely, relatively wide when "*n*" is small. Just as the Gallup Organization can gauge national sentiment more accurately from a larger sample than from a smaller sample, a larger school provides a more accurate estimate of the true level of proficiency than a smaller school can. There simply is greater uncertainty surrounding small-school achievement, and a confidence interval captures the degree of this uncertainty.

The upshot is this: When AYP is evaluated within the context of confidence intervals, small schools are not put at a disadvantage for being small. Because the confidence interval for small schools is wider than that for large schools, a bigger AYP "shortfall" is required before a small school is identified as a failing school. And this is as it should be, given the greater uncertainty associated with achievement in small schools.

Even with wide confidence intervals, however, small schools still can be identified as not making adequate progress and, therefore, in need of improvement. In other words, small schools do not get a "pass" merely because they are small. This, too, is as it should be, for the burden of accountability should not fall only on large schools.

Conclusions

To be sure, NCLB has its positive attributes. But much of this law is troublesome, not least of which is the expectation that all students reach proficiency by 2014—an expectation that seems to defy what's possible. Also, NCLB is problematic for states having many small and rural schools, especially around the provisions regarding school choice, technical assistance, supplemental educational services, and teacher qualifications. For example, consider St. Lawrence Island, Alaska, where school choice or the delivery of supplemental services from a qualified provider would require an airplane ride across the Bering Sea.

Time will tell whether the requirements of NCLB will be modified to make this legislation more realistic, whether for public schools in general or small and rural schools in particular. But as long as NCLB (or any policy) calls for high-stakes evaluation of school performance, the random variation associated with school achievement results must be taken into account. This is particularly true for small schools, where this volatility is more pronounced.

Skeptical readers might conclude that, by using confidence intervals for evaluating AYP, we merely game the system. On the contrary, the use of confidence intervals is a carefully reasoned reply to the NCLB call for the "statistically valid and reliable" determination of AYP. As such, they reduce the likelihood that a school—especially a small school—will be falsely identified as a failing school. It is a matter of fairness.

Coladarci is Professor of Education at the University of Maine. He can be reached by e-mail (theo@maine.edu) or at the address below:

Theodore Coladarci
Shibles Hall
University of Maine
Orono, ME 04469