THE RURAL SCHOOL AND COMMUNITY TRUST

POLICY BRIEF

Gallup Goes to School: The Importance of Confidence Intervals for Evaluating "Adequate Yearly Progress" in Small Schools

By Theodore Coladarci

October 2003

Rural School and Community Trust
1825 K Street NW, Suite 703
Washington, DC 20006
Telephone: (202) 955-7177
Fax: (202) 955-7179
E-mail: info@ruraledu.org
www.ruraledu.org

The Rural School and Community Trust (Rural Trust) is the premier national nonprofit organization addressing the crucial relationship between good schools and thriving rural communities. Working in some of the poorest, most challenging rural places, the Rural Trust involves young people in learning linked to their communities, improves the quality of teaching and school leadership, advocates for appropriate state educational policies, and addresses the critical issue of funding for rural schools.

Gallup Goes to School: The Importance of Confidence Intervals for Evaluating “Adequate Yearly Progress” in Small Schools

Theodore Coladarci
University of Maine

Indicators of school-level achievement, such as the percentage of students who are proficient in a particular content area, are subject to random year-to-year variation in much the same way that the results of an opinion poll will vary from one random sample to another. This random variation, which is more pronounced for a small school, should be taken into account by education officials when evaluating school progress in a policy climate of high stakes. To do otherwise is to unnecessarily risk the false identification of a failing school. In this monograph, I describe the application of confidence intervals to the evaluation of “adequate yearly progress” for No Child Left Behind (NCLB). Throughout, I demonstrate the particular relevance of confidence intervals for small schools. Upon completion, readers will understand why 27 states included confidence intervals in their NCLB accountability plans (and perhaps wonder why the remaining states did not).

The ambitious agenda of the *No Child Left Behind Act of 2001* ([NCLB] 2002) sets unprecedented challenges for public schools in the United States. And these challenges are particularly daunting for states that have a sizable rural population, where the major precepts of NCLB are often at variance with the reality of rural education (e.g., Reeves, 2003; Tompkins, 2003). To be sure, NCLB provisions regarding school choice, teacher qualifications, technical assistance, supplemental education services, and the evaluation of adequate yearly progress will be tough for any school to accommodate. But these provisions will be considerably more difficult for schools that are small and geographically isolated—that is, for the many schools in this country that reside in rural communities.

Among the most ambitious NCLB mandates is that, by 2014, all students must reach proficiency on “challenging academic content standards and challenging student achievement standards.” To monitor progress in this regard, the state determines whether each school is making “adequate yearly progress” (AYP) in reading/language arts and mathematics.[1] This is accomplished by first establishing a proficiency “starting point” for each content area, based on 2001-2002 state test scores. The NCLB starting-point criterion that most states are using is the 20th percentile school for the particular content area. That is, you first rank-order all schools according to percent proficient[2] (e.g., in mathematics) and then count up from the bottom until 20% of the student population is reached. The percentage of students who are proficient in *that* school is the 2001-2002 starting point for the state. Now subtract this percentage from 100% and divide by 12 (years), and you have the average annual gain required for all students to be proficient in 2014.

For illustration, suppose that in your state the mathematics starting point for high schools is 28%. As Figure 1 shows, an annual gain of 6% would be required to reach 100% proficiency in mathematics by 2014.[3] This trajectory provides the basis for evaluating a school’s progress toward that goal. For example, if at least 40% of students in your high school were proficient in, say, 2004, your school would “meet AYP” that year. Figure 1 illustrates a linear trajectory with annual increases, whereas Figure 2

Correspondence concerning this monograph should be addressed to Theodore Coladarci, University of Maine, College of Education & Human Development, 5766 Shibles Hall, Orono, ME 04469-5766. (coladarci@umit.maine.edu)

Dale Carlson, Charlie DePascale, Larry Hedges, Tad Johnston, Scott Marion, and Phoebe Winter provided helpful comments on a draft of this paper. Each is blameless, of course, for any shortcomings in the final product.

[1] Although AYP applies to both schools and districts, I will refer only to schools insofar as (a) the school is where accountability pressure is most acute, (b) *schools* is less awkward than *schools and districts*, and (c) the general argument is the same. In doing so, however, I gloss over several issues. In many rural communities, for example, the school *is* the district. Further, where a multi-school district comprises exceedingly small schools, it is possible that one or more schools in the district are excluded from certain aspects of the accountability system (but not the district as a whole).

[2] To avoid the awkward *proficient and above* or *at least proficient*, I will define *proficient* to also include students who fall in the higher performance category.

[3] (100% - 28%)/12 = 6%

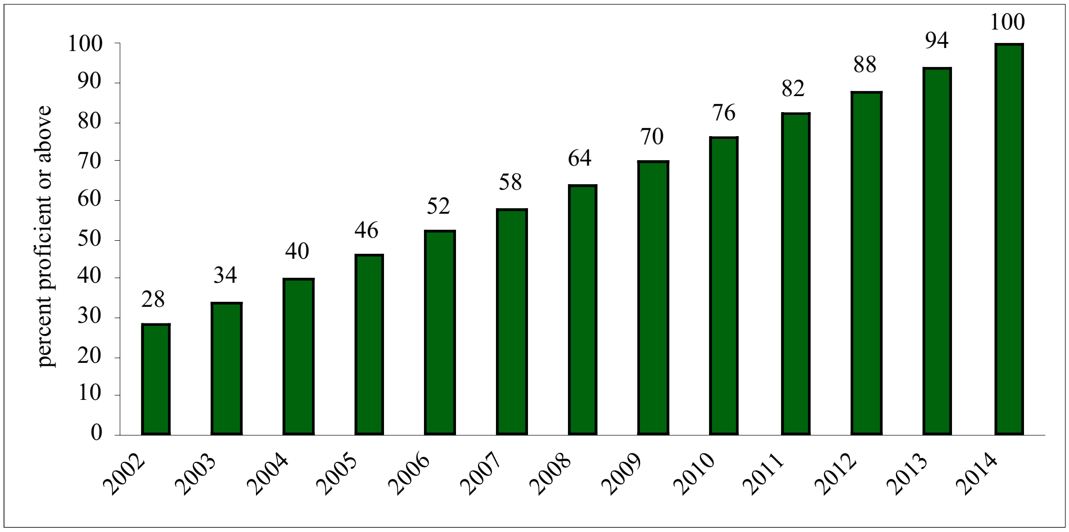

Figure 1. Hypothetical AYP targets for mathematics: Starting point of 28% and annual gain of 6%.

shows a linear trajectory with "intermediate goals." To date, 21 states have adopted intermediate goals but with a nonlinear twist (Olson, 2003). As illustration, consider the trajectory adopted by Maine for high school mathematics (Figure 3), which permits slower growth initially and, in later years, requires more rapid (seemingly Herculean) growth toward 100% proficiency in 2014.

Readers familiar with NCLB and AYP see that I have conveniently sidestepped the "safe harbor" provision as well as the requirement to conduct disaggregated subgroup analyses of AYP. I will comment on both later. Further, although NCLB by 2005-2006 will require annual testing in grades 3-8 and at least once in grades 10-12, AYP judgments ultimately are rendered about *schools*—not separately by grades within schools. Achievement data therefore must be combined to permit a single school-wide judgment regarding AYP (Marion, White, Carlson, Erpenbach, Rabinowitz, & Sheinker, 2002). The importance of these issues notwithstanding, at the center of it all is the initial comparison of an AYP target and a school-level percentage. And it is the nature of this comparison that is my focus.

Volatility of School-Level Achievement

The consequences for a Title 1 school that makes *in*adequate yearly progress are well known. After failing to meet AYP for two consecutive years, the school is identified for school improvement. Among other things, this means allowing for within-district school choice. After three consecutive years of failure, the school must make available "supplemental educational services" from a qualified provider. And after five years, the school is identified for restructuring. In the worst case, this could entail surrendering school operations to the state (which is against state law in Maine). Clearly, a lot rides on the comparison of a school's proficiency percentage with the corresponding AYP target.

But how dependable are such percentages? A school's achievement status (e.g., mean score, percentage or proportion proficient, achievement index) is subject to random year-to-year variation, and this random variation is much greater for smaller schools than for larger schools (Hill & DePascale, 2003; Kane & Staiger, 2001; Kane, Staiger, & Geppert, 2002; Linn & Haug, 2002). Figure 4, for example, shows how the one-year change in a school's "proportion proficient" on the Maine Educational Assessment (MEA) is related to the size of the cohort tested in that school. As you can see for both fourth-grade reading (left) and eighth-grade mathematics (right), the average change from one year to the next hovers around zero for all schools. However, there is considerably greater variability among smaller schools in the amount of this change. For schools having 15 or fewer fourth graders, this change ranges from -.47 (a school declining from 60% proficient to 13% proficient) to +.83 (a school increasing from 17% proficient to 100% proficient). In contrast, the corresponding figures are only -.07 and +.09, respectively, among schools having 150 or more fourth graders.[4]

[4] As the fourth-grade plot shows, the +.83 school (upper left corner) is somewhat of an outlier. The small-school range is -.47 to +.46 with this discrepant case excluded.

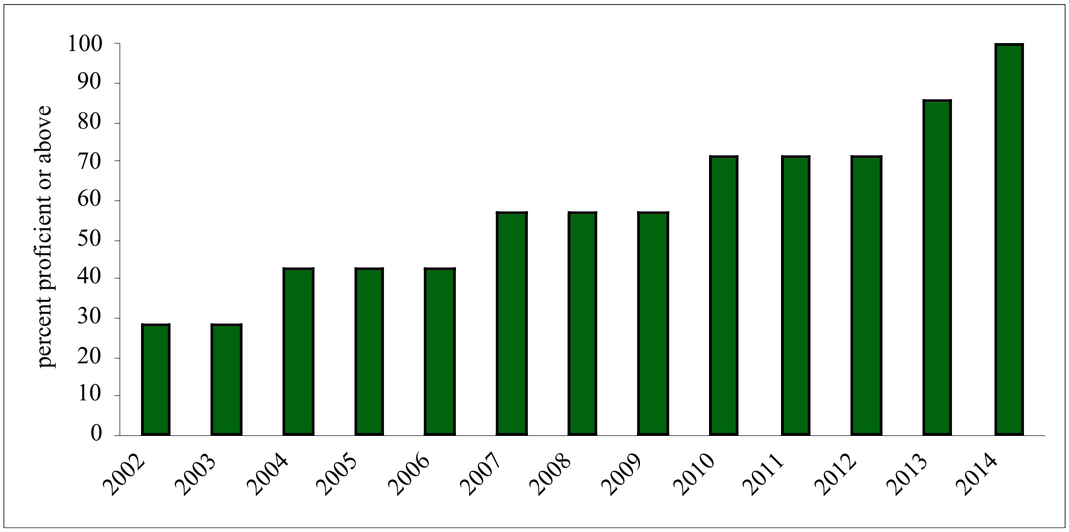

Figure 2. Hypothetical AYP targets for mathematics: Intermediate goals (linear growth).

There simply is greater volatility in achievement in smaller schools. 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. The question, then, is this: When a school falls short of the AYP target, with what warrant can we conclude that the school—particularly a small school—*truly* is not making adequate progress?

School Cohort as "Sample"

The volatility of school-level performance is analogous to the sampling error that attends any opinion poll. If you ask 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 doubtless would change if a new random sample were selected from the same population. This, of course, is why reputable pollsters attach a margin of error (e.g., "±4%") to their results. Although no sample is immune to sampling error, the magnitude of error is inversely related to sample size (*n*): Other things being equal, results based on small *n*s have wider margins of error than those based on large *n*s.

Just as pollsters attach a measure of sampling error to their results, any school-level result (e.g., mean score or proportion proficient) similarly should be considered within the context of sampling error. Perhaps you find it odd to regard school data as *sample* data. In what sense, you may reasonably wonder, do the achievement scores for a school represent a "sample"? After all, these scores are based on the overwhelming majority of—possibly all—students in the school. And what, pray tell, would be the corresponding "population"?

In fact, school data *can* be treated as a sample from a larger population of observations, although this population is decidedly theoretical (Cronbach, Linn, Brennan, & Haertel, 1997). Rich Hill phrased the proposition this way:

> The result for a school for one year is just one observation from which to infer a school's *true score*—what the school's average would be if 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. (Hill, 2002, p. 2; emphasis in original)

A central premise of this proposition is that the desired inference is about the *school*, not the specific cohort of students on whom achievement data were obtained. As Cronbach et al. (1997) argued, "[t]o conclude on the basis of an assessment that a school is effective *as an institution* requires the assumption, implicit or explicit, that the positive outcome would appear with a student body other than the present one, drawn from the same population" (p. 393, emphasis added). Hill and DePascale, with a nod to Dale Carlson, unpack this argument well:

> When the results are reported, they are not attributed to a particular group of students, but to the school as a whole. Since the inference is about the school, not a particular group of students, it is

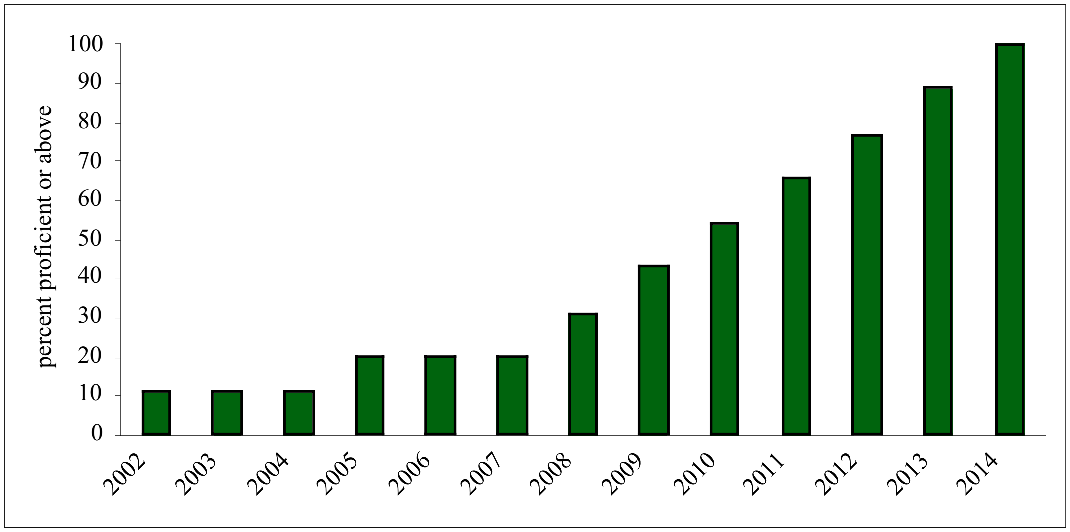

Figure 3. Maine AYP targets for mathematics, high school: Intermediate goals (curvilinear growth).

> important to take into account the fact that the group tested in any particular year might not be representative of students in that school across years. If people were to insist that a particular group of students in, say, 2001, fully represents the school—is the sufficient definition of that school—then when a new group of students is tested in 2002, they actually represent a new school. Under such a belief system, it would be impossible to have any school ever fail to meet AYP in 2 consecutive years, since the population to which the inference was being limited would never be the same across those years. (Hill & DePascale, 2003, pp. 12-13)

We see, then, that school-level achievement indices are subject to random sampling variation from year to year. Further, this variation must be considered when evaluating AYP. But how?

One solution is to reserve AYP judgments only for schools having a sufficient number of students to provide reliable judgments—judgments that would be reasonably similar had a different (random) sample of students been tested that year. But any minimum number (*n*) is arbitrary insofar as there is no single value that separates patently unreliable school results from those for which reliability is unimpeachable (Linn, Baker, & Herman, 2002). Moreover, although a large minimum *n* may allay reliability concerns, it invites an unintended negative consequence, particularly for rural states. As Marion et al. (2002) warn us, "[s]tates with primarily rural and small schools would see these schools excluded from accountability systems, with the full burden of accountability shifted to large schools" (p. 64). In Wyoming, for example, almost half of all schools initially would be excluded if the minimum *n* were set at 30 (Marion et al., 2002, p. 65); in Maine, the figure is roughly 40%.[5]

A more reasonable solution is to relax the minimum *n* for schools to be included in the accountability system yet acknowledge the greater uncertainty that accompanies school-level achievement based on small samples. This is accomplished by interpreting school achievement in light of the "standard error" (analogous to the ±4% in the opinion poll above), which is our best estimate of the error in school-level achievement due to random sampling variation.[6] There are two ways to proceed toward this end: hypothesis testing and interval estimation (e.g., Marion et al., 2002).

[5] Nevertheless, NCLB requires that the progress of excluded schools be evaluated in some manner. In Wyoming and Maine, this is accomplished by drawing on data from local assessment systems.

[6] This form of error is not to be confused with the measurement error inherent in any test score for an individual student. Measurement error reflects an assessment's reliability, which, relative to random sampling variation, is largely inconsequential in the present context (e.g., Arce-Ferrer, Frisbie, & Kolen, 2002; Hill, 2001).

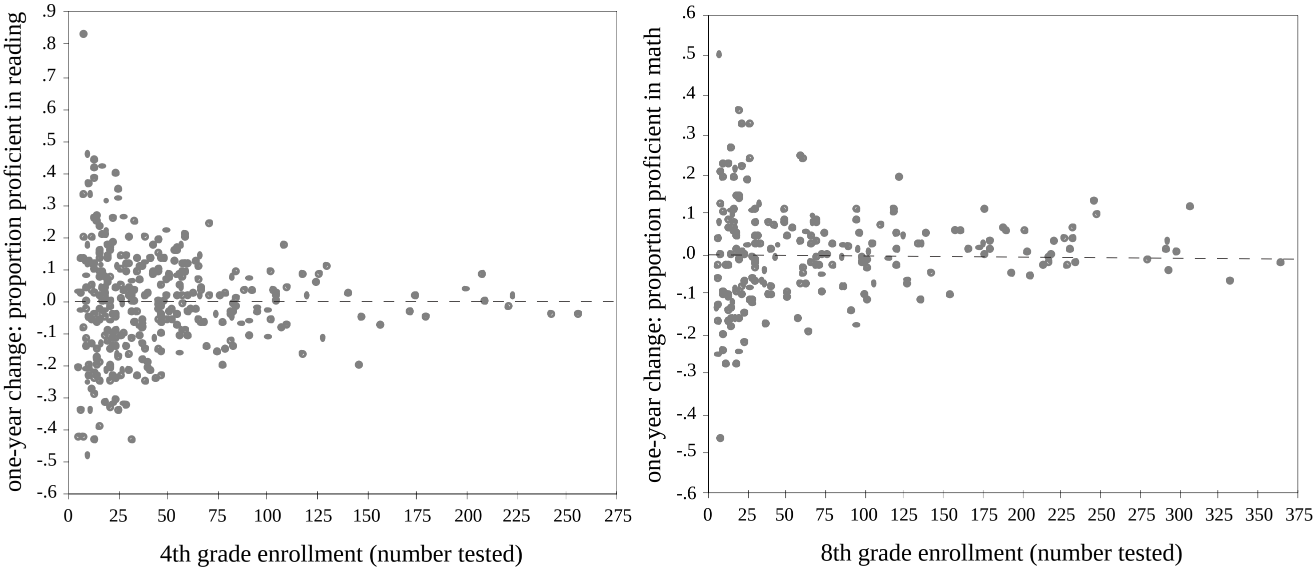

Figure 4. The relationship between (a) the number of students tested in a school and (b) the one-year change in proportion proficient on the Maine Educational Assessment, shown separately for fourth-grade reading (left) and eighth-grade mathematics (right).[7]

Hypothesis Testing

In hypothesis testing, one tests the "statistical significance" of the difference between a school's achievement status and the target value. Say you wish to determine whether your school's proficiency proportion is significantly (read "truly") below the AYP target. (There would be no need for this test, of course, if your proficiency proportion were equal to or greater than the AYP target.)[8] Calculate the *z* ratio,

$$z = \frac{p - \pi}{\sqrt{\frac{p(1-p)}{n}}}$$

where *p* is the proficiency proportion for your school, *n* is the number of students tested in your school, and π (pi) is the AYP target. As you see, the numerator of the *z* ratio is the difference between your school's proportion and the target. What may not be immediately apparent is that the denominator is the *standard error of a proportion*—the aforementioned "best estimate" of random sampling variation inherent in the school-level proportion, *p*.

As a whole, then, the *z* ratio evaluates the difference between *p* and π in reference to the magnitude of random sampling variation. Why is this called hypothesis testing? You are testing the null hypothesis that the school's true proportion is at least equal to π (the AYP target); the alternative hypothesis is that the true proportion is less than π. If your negative *z* ratio is less than, say, the one-tailed critical value of -1.65 (e.g., -2.00), then the null hypothesis is rejected and the alternative hypothesis prevails: You declare the difference between *p* and π statistically significant at the .05 level and, in turn, conclude that your school truly is not making adequate progress. That is, your school does not meet AYP. There is a small probability (.05) that you are wrong in drawing this conclusion—that you have committed a Type 1 error (false positive)—but it is the appropriate conclusion nonetheless, given the data.

On the other hand, a negative *z* ratio that is greater than -1.65 (e.g., -1.00) is *not* statistically significant. The null hypothesis is retained: The difference between *p* and π is no more than what one would expect from random sampling variation alone. Although *p* falls short of π, the magnitude of this difference is insufficient to conclude that your school is not making adequate progress. Indeed, the war-

[7] I omitted schools in which fewer than five students were tested in either year. For both plots, the horizontal axis is mean enrollment across the two years. There are 356 schools represented in the fourth-grade plot and 218 schools in the eighth-grade plot.

[8] But random sampling variation plays no favorites: A school that just meets the AYP target may, in truth, be falling short of making adequate progress. (This school may have had a *good* bounce, as it were.) There is considerable uncertainty for *any* school that is close to the target—whether below or above.

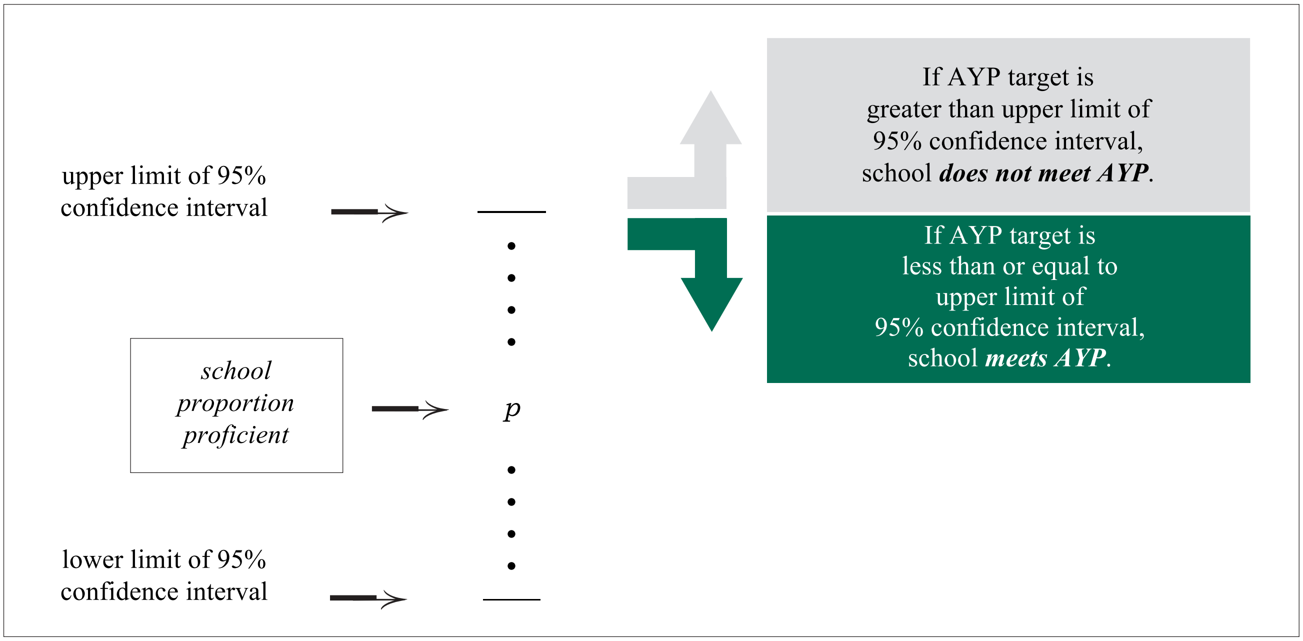

Figure 5. Evaluating AYP within the context of a 95% confidence interval.

ranted conclusion is that your school, in fact, meets AYP. It would be like tossing a coin 50 times and obtaining 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 unbiasedness therefore stands.

Interval Estimation

An alternative to hypothesis testing is interval estimation: the construction of a "confidence interval" around *p*. Like hypothesis testing, the confidence-interval approach to evaluating AYP explicitly takes into account the standard error of a proportion. In the AYP context, however, I find confidence intervals preferable over hypothesis testing for several reasons. First and foremost, a confidence interval is easier to explain, easier to understand, and, because of its common use in opinion polls, largely familiar to the public. That is, as a strategy for evaluating AYP, interval estimation is more transparent than hypothesis testing. If the proverbial person on the street understands the opinion poll result "50% ±4%", this person already understands the logic of a confidence interval (for "50% ±4%" *is* a confidence interval). Second, a confidence interval is sufficient for evaluating AYP. With a confidence interval in hand, we learn nothing more from hypothesis testing. Third, confidence intervals clearly show the relationship between (a) the degree of uncertainty accompanying a school-level proportion and (b) the number of students tested, as you soon will see. Finally, confidence intervals are more responsive to the precipitating question. Let me briefly elaborate on this last point.

We know that a school's proficiency proportion is subject to random sampling variation, analogous to the sampling error that attends the results of any opinion poll. Thus, a school's *observed* proficiency proportion—what you calculate directly from the data—is merely an estimate of the school's *true* proportion, and it arguably is the latter that should be used for making inferences about school performance in general and for evaluating AYP in particular. So, what *is* the true proficiency proportion for a school? This, in my view, is the precipitating question, which a confidence interval addresses head on. Although we can never know a school's true proficiency proportion, or estimate this single value with any semblance of accuracy, we can estimate a *range* of values within which we are 95% confident the true proportion lies. This range, reasonably enough, is called a 95% confidence interval (or, 95% CI), and any value within this interval is a plausible candidate for the school's true proficiency proportion. Unlike hypothesis testing, then, interval estimation addresses the fundamental question, "Where does the true proportion for this school probably fall?" If you calculate $p = .50$ for your school and obtain a confidence interval of, say, .40 to .60, you may conclude with 95% confidence that your schools' true proportion could be as low as .40 or as high as .60.

And what about the evaluation of AYP? Simply compare the AYP target to the confidence interval: If the target falls above the upper limit of the confidence interval, the school has not met AYP; if the target is less than or equal to the upper limit, the school meets AYP (Figure 5). As you see, then, a school can meet AYP even though the observed proportion is lower than the target. In this case (as in hypothesis testing), the discrepancy between *p* and the AYP target would be insufficient to conclude that the school is not making adequate progress.

Constructing a Confidence Interval

So, just how does one construct a 95% CI for a proportion? The traditional method, which relies on the standard error that you earlier encountered in the *z*-ratio formula, is based on the normal distribution. Large samples therefore are required, particularly where *p* is either very low or very high. An alternative formula, which Maine uses, appears below. Based on the binomial distribution and attributed to Ghosh (1979), this formula provides accurate confidence intervals regardless of the magnitude of *n* or *p* (Glass & Hopkins, 1996):

$$P_{\text{L}} = \frac{n}{n+3.84}\left[p + \frac{1.92}{n} - 1.96\sqrt{\frac{p(1-p)}{n} + \frac{.96}{n^2}}\;\right]$$

$$P_{\text{U}} = \frac{n}{n+3.84}\left[p + \frac{1.92}{n} + 1.96\sqrt{\frac{p(1-p)}{n} + \frac{.96}{n^2}}\;\right]$$

The derivation of this formula goes well beyond my scope (and probably your tolerance). The familiar terms, *n* and *p*, have the same meaning as in the *z* ratio. The only new terms are P_{L} and P_{U}, which represent, respectively, the estimated lower and upper bounds of the confidence interval. In the present context, P_{L} and P_{U} are the lower and upper limits of the school's *true* proficiency proportion. Notice the use of uppercase *P* here, which distinguishes the estimated limits from the single *observed* proportion, *p*, which one calculates directly from school data.

As an example, let's take a school in which 30%, or .30, of its 50 students are proficient in mathematics. The calculations[9] are as follows:

$$\begin{aligned} P_{\text{L}} &= \frac{50}{50+3.84}\left[.30 + \frac{1.92}{50} - 1.96\sqrt{\frac{.30(1-.30)}{50} + \frac{.96}{50^2}}\;\right] \\ &= .9287\big(.3384 - .1327\big) \\ &= .19 \end{aligned}$$

$$\begin{aligned} P_{\text{U}} &= \frac{50}{50+3.84}\left[.30 + \frac{1.92}{50} + 1.96\sqrt{\frac{.30(1-.30)}{50} + \frac{.96}{50^2}}\;\right] \\ &= .9287\big(.3384 + .1327\big) \\ &= .44 \end{aligned}$$

Thus, you can state with 95% confidence that the true proportion for this school is anywhere between .19 and .44. If the AYP target for mathematics is .44 or less, then this school meets AYP; if the target is greater than .44, the school does not.

Interval Width and School Size

Table 1 shows the 95% CI for various values of *p* and *n*. Notice that for a given value of *p*, interval width decreases as *n* increases. Take a school in which 50% of the students are proficient (*p* = .50). If this school has only 5 students in the tested grades, the true proficiency proportion could be as low as .17 or as high as .83—quite a range, indeed.[10] With 300 students, however, the interval width is reduced considerably: .44 - .56. And in a decidedly hypothetical school with 5,000 students, the interval width would shrink 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 estimate national sentiment more accurately from a larger sample than from a smaller sample, a larger school provides a more accurate estimate of the true proficiency proportion than a smaller school can. Again, there simply is greater uncertainty surrounding school-level statistics based on small samples, and a confidence interval captures the degree of this uncertainty.

The upshot is that when AYP is evaluated within the context of confidence intervals, small schools are not unwittingly penalized for being small. Because the confidence interval for small schools is wider than that for large schools (for a given *p*), a bigger discrepancy between *p* and the target—the proficiency shortfall—is required before a small school is identified as not meeting AYP. And this is as it

[9] To effortlessly obtain P_{L} and P_{U} for entered values of *n* and *p*, download the Excel file "CI (proportions)" from http://www.umit.maine.edu/~coladarci/. This file allows for either 95% or 99% confidence intervals.

[10] Although a bit clumsy, the fact that 50%, 17%, and 83% of 5 students all yield a fractional individual should not detract from the larger point.

Table 1
95% Confidence Interval for *P*: For Different Values of the Observed Proportion (*p*) and Number of Students Tested (*n*)

observed proportion (*p*) of students who are proficient or above	number of students tested (*n*)								
↓	**5**	**10**	**15**	**20**	**30**	**50**	**75**	**100**	**300**
.10	.01 - .54	.02 - .40	.02 - .34	.03 - .30	.03 - .26	.04 - .21	.05 - .19	.06 - .17	.07 - .14
.25	.05 - .66	.08 - .56	.10 - .50	.11 - .47	.13 - .43	.15 - .38	.17 - .36	.18 - .34	.20 - .30
.50	.17 - .83	.24 - .76	.27 - .73	.30 - .70	.33 - .67	.37 - .63	.39 - .61	.40 - .60	.44 - .56
.75	.34 - .95	.44 - .92	.50 - .90	.53 - .89	.57 - .87	.62 - .85	.64 - .83	.66 - .82	.70 - .80
.90	.46 - .99	.60 - .98	.66 - .98	.70 - .97	.74 - .97	.79 - .96	.81 - .95	.83 - .94	.86 - .93

Note. Calculations are based on Equations 13.8C and 13.8D in Glass and Hopkins (1996, p. 327).

should be, given the greater sampling variation associated with achievement in small schools.

Even with wide confidence intervals, however, small schools nonetheless can be identified as not making adequate progress. In other words, small schools do not necessarily get a "pass" merely because they are small. For example, consider a school having 10 students (an island school in Maine, perhaps), only one student is proficient (*p* = .10), and the AYP target is .50. As Table 1 shows, the true proportion for this exceedingly small school could be as high as .40—the observed proficiency *quadrupled*—which still falls short of the target. The school consequently is judged not to be making adequate progress. This, too, is as it should be. Again, the burden of accountability should not fall only on large schools (Marion et al., 2002).

Related Issues

Subgroups

NCLB requires separate AYP evaluations for the economically disadvantaged, major racial and ethnic groups, students with disabilities, and students with limited English proficiency. The argument of confidence intervals applies equally to subgroup analyses. As you would expect from the discussion above, confidence intervals for disaggregated subgroups are wider (because the *n*s are smaller) than when based on all students in the school.

Minimum n *(of Students)*

Because interval estimation takes into account the greater volatility of small-school achievement results, the minimum *n* for a school to be included in the accountability system can be reduced appreciably when confidence intervals are used, thereby making a more equitable representation of schools participating in the system. Here, the primary concern when specifying a minimum *n* arguably is the protection of student privacy. For example, the accountability plans for 5 of the 27 states using confidence intervals have stipulated minimum *n*s ranging from 5 to 11 (Olson, 2003). However, most of the 27 CI states specify considerably higher minimum *n*s of 20 to 30 (with some higher still), which go beyond what is necessary to protect student privacy. The use of confidence intervals in conjunction with a high minimum *n* is an exceedingly conservative policy for identifying schools in need of improvement.

Confidence Intervals and "Safe Harbor"

I have concentrated on achievement *status*: comparing a school's present achievement level to the AYP target. Some states also are applying confidence intervals to achievement *improvement*. Within the context of NCLB and AYP, this means constructing a confidence interval around *the difference between two proportions*: the difference between the proportion of below-proficient students one year and this same proportion the following year. If a school has not met AYP as described above, the school nevertheless satisfies AYP through the provision of "safe harbor" (as it has come to be called) if the proportion of below-proficient students is reduced by at least 10%.

Clearly, if confidence intervals are important for evaluating status, they are equally important for evaluating improvement. Indeed, achievement *change*, particularly over the short term, is subject to even greater sampling variation than achievement determined at single point in time.

Nevertheless, only five states intend to take sampling variation into account when appraising safe harbor (Erpenbach, Fast, & Potts, 2003).[11]

Level of Confidence

I have focused on the 95% level of confidence. Because of the high stakes associated with AYP judgments, however, 11 of the 27 states that specify confidence intervals in their NCLB accountability plans are adopting the 99% level of confidence (Olson, 2003). Similar in structure to the formula above, the formula for constructing a 99% CI is

$$P_L = \frac{n}{n + 6.66}\left[p + \frac{3.33}{n} - 2.58\sqrt{\frac{p(1-p)}{n} + \frac{1.66}{n^2}}\right]$$

$$P_U = \frac{n}{n + 6.66}\left[p + \frac{3.33}{n} + 2.58\sqrt{\frac{p(1-p)}{n} + \frac{1.66}{n^2}}\right]$$

A 99% CI is wider than its 95% counterpart, so naturally there is a greater chance that a school's true proportion resides somewhere in this interval. Hence, you have greater confidence in the statement that, for an observed value of p, the true proportion falls between P_L and P_U. You also have greater confidence that you have not committed a Type 1 error (false positive): misidentifying a school as having not met AYP when, in truth, this school is making adequate progress. But in exchange for this greater confidence you must accept a higher likelihood of a Type 2 error (false negative): failing to identify a school that, in truth, is not making adequate progress.[12]

Let's return to that small school in which one of the ten students is proficient. With $p = .10$ and $n = 10$, the upper limit of a 99% CI is .51, which surpasses the AYP target of .50. In contrast to the 95% CI scenario, then, this school now meets AYP. There is no way to know whether a Type 1 error (misidentification) was committed using the 95% CI or, alternatively, a Type 2 error (failing to identify) was committed using the 99% CI. Rather, the choice between the two confidence levels is a policy decision, one that is made only after careful deliberation of these two types of errors and their possible consequences.

Inflated Type 1 Error Rate

When we use a 95% (or 99%) CI for a single evaluation of AYP—e.g., one group's performance in one content area on one occasion—we accept a probability of .05 (or .01) that a school will be misclassified as not meeting AYP when, in truth, it is making adequate progress. This is the probability of a Type 1 error. We can never know when we have committed a Type 1 error, but we do know from statistical theory that over an infinite number of schools, all of which are making adequate progress, we nevertheless would misclassify 5% (or 1%) of these schools as *not* making adequate progress. But the problem is that AYP involves multiple evaluations for each school. For example, AYP is evaluated for each subgroup in each content area. While the probability of a Type 1 error is .05 (or .01) for any one AYP evaluation, the probability of at least one Type 1 error *across all AYP evaluations* for a school is considerably greater.

Some measurement specialists are recommending a statistical adjustment in order to maintain a Type 1 error probability of roughly .05 (or .01) across all AYP evaluations for a school—i.e., across the "family" of AYP evaluations. Although *family-wise* adjustments are commonplace in research, their application to the evaluation of AYP raises both technical and policy questions. Regarding policy, for example, a consequence of family-wise adjustments is that many small schools will be removed from the accountability system. This is because family-wise adjustments entail wider confidence intervals. As you saw above, small schools (rightly) have a wider confidence interval to begin with. But to widen this interval further will make it seemingly impossible to identify small schools that are making inadequate progress—except where p is very low and the AYP target is very high. This is particularly true for subgroup AYP, where ns are smaller still.

Conclusion

Like many educators, I have mixed emotions about NCLB. On one hand, I find it difficult to disagree with the call for greater school accountability. Further, I like the focus on measurable objectives (for those objectives that are indeed measurable) and, in particular, the fundamental concerns that this attention necessarily brings to the surface in each state: the integrity of announced content standards, the alignment of curriculum and instruction with these standards, the provision of adequate opportunity to learn for all students, and the validity and reliability of assessments, to mention a few.

[11] With an apparent concern for Type 2 error, the U.S. Department of Education is asking these states to submit "impact data" that speak to the consequences of applying confidence intervals to safe harbor.

[12] For example, I am 100% confident that a school's true proficiency proportion falls between 0 and 1.00. But with this confidence level, it would be impossible to identify any school as not making adequate progress.

But much about NCLB is troublesome, not least of which is the delusional expectation that all students reach proficiency by 2014. In the more tempered words of Robert Linn, "[t]he policymaker expectations almost certainly exceed the ability of schools to make this sort of progress" ("No Child Left Behind," 2002, p. 4). Given the 1990-2000 trend of mathematics scores on the National Assessment of Educational Progress, Linn (in press) observed that it will take 57 years for fourth graders to reach 100% proficiency, 61 years for eighth graders, and 166 years for twelfth graders. Yet the NCLB architects believe that each state, using "challenging academic content standards and challenging student achievement standards," can do it in 12. In fact, the annual gains required to meet this ambitious goal have no empirical precedent—there are no "existence proofs" that this goal is attainable (Linn, in press). Even under the best conditions (e.g., highly motivated educators fully implementing a proven school-wide intervention), achievement gains still fall short of those expected under NCLB.[13]

Further, as I mentioned at the outset, NCLB is problematic for states having many small and rural schools, particularly around 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 entail an airplane ride across the Bering Sea. Comparable (and sea-less) circumstances can be found in countless other rural areas across the country. And while few would disagree with the importance of having highly qualified teachers, recruiting them can prove difficult in rural communities where, as in Winnett, Montana, "the newspaper has stopped printing, the nearest movie theater is 53 miles away, and there are only two stores, one saloon, and the Kozy Korner Cafe" (Dillon, 2003). Even when one is drawn to just such communities, the teaching conditions can be a disincentive nevertheless. Tompkins (2003, p. 31), in reference to a recent school-funding case before the Arkansas Supreme Court, described the circumstance of a particular high school math teacher: "He has two electrical outlets in his classroom, calculators for half the students, and a single blackboard on which he writes exams by hand because there is no photocopier."

In view of these challenges and others, Senators Michael Enzi (R-WY), Susan Collins (R-ME), John Edwards (D-NC), and Kent Conrad (D-ND) began to organize in January 2003 the Rural Education Caucus to address the educational needs of rural communities in the implementation of NCLB. In similar spirit, Secretary of Education Rod Paige announced in April 2003 the formation of a "high-level" Department of Education task force, in part to work with the Rural Education Caucus in finding solutions to the challenges that small, rural schools face.

Time will tell whether the requirements of NCLB will be modified to render this legislation more feasible, whether for public schools in general or small and rural schools in particular. But as long as NCLB (or, indeed, any policy) calls for high-stakes evaluation of school-level student achievement, the random sampling variation associated with school achievement must be taken into account. This is particularly true for small schools, where sampling variation is more pronounced. Toward this end, confidence intervals have the advantage of being easy to explain, easy to understand, and familiar to the public because of their use in opinion polls.

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.

References

Arce-Ferrer, A., Frisbie, D. A., & Kolen, M. J. (2002). Standard errors of proportions used in reporting changes in school performance with achievement levels. *Educational Assessment*, *8*, 59-75.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research*, *73*, 125-230.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

[13] The recent meta-analysis by Borman, Hewes, Overman, and Brown (2003) is perhaps illustrative in this regard. From their synthesis of evidence on the effectiveness of comprehensive school reform (CSR), Borman et al. estimated that the impact of CSR on academic achievement corresponds to an "effect size" of d = .15 when based on all 232 studies identified and retrieved, and d = .09 when restricted to 109 third-party evaluations involving a comparison group. (CSR implementation varied from 1 to 14 years, with a mean of roughly 3 years.) How does one interpret effect size? A popular, if somewhat arbitrary, guideline is to consider d = .20 "small," d = .50 "moderate," and d = .80 "large" (Cohen, 1988), in which case the CSR effect is small at best. A norm-referenced interpretation of effect size also is available. For example, if the true effect of CSR is d = .15, this suggests that schools implementing CSR would outperform, on average, 56% of comparable schools not implementing CSR and, conversely, the average CSR school would be outperformed by 44% of comparable non-CSR schools. Although the findings of Borman et al. do not translate easily into the language of AYP, they nevertheless throw cautioning light on the NCLB expectation that schools can effect 100% proficiency by 2014.

Coladarci, T., Cobb, C. D., Minium, E. W., & Clarke, R. C. (2004). *Fundamentals of statistical reasoning in education*. New York: Wiley.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, *57*, 373-399.

Dillon, S. (2003, June 23). New Federal law may leave many rural teachers behind. *New York Times*, p. A1.

Erpenbach, W. J., Fast, E. F., & Potts, A. (2003). *Statewide educational accountability under NCLB*. Washington, DC: Council of Chief State School Officers.

Ghosh, B. K. (1979). A comparison of some approximate confidence intervals for the binomial parameter. *Journal of the American Statistical Association*, *74*, 894-900.

Glass, G. V, & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Boston: Allyn & Bacon.

Hill, R. K. (2001). *Issues related to the reliability of school accountability systems*. Unpublished manuscript. Portsmouth, NH: The National Center for the Improvement of Educational Assessment. (Available online: http://www.nciea.org/publications/RILS2000Paper_Hill01.pdf)

Hill, R. K. (2002, April). *Examining the reliability of accountability systems*. Paper presented at the 2002 meeting of the American Educational Research Association, New Orleans. (Available online: http://www.nciea.org/publications/NCME_RHCD03.pdf)

Hill, R. K., & DePascale, C. A. (2003). Reliability of No Child Left Behind accountability designs. *Educational Measurement: Issues and Practices, 22*(3), 12-20.

Kane, T. J., & Staiger, D. O. (2001, May). Volatility in school test scores: Implications for test-based accountability systems. *Accountability and its consequences for students: Are children hurt or helped by standards-based reforms?* Symposium presented at the Brookings Institution, Washington, DC.

Kane, T. J., Staiger, D. O., & Geppert, J. (2002). Randomly accountable. *Education Next*, *2*(1), 57-61. (Available online: http://www.educationnext.org/20021/56.html)

Linn, R. L. (in press). Accountability: Responsibility and reasonable expectations. *Educational Researcher*.

Linn, R. L., Baker, E. L., & Herman, J. L. (2002, Fall). Minimum group size for measuring Adequate Yearly Progress. *The CRESST Line*, Fall , pp. 1, 4, 5. (Available online: http://www.cresst.org/products/newsletters/CL2002fall.pdf)

Linn, R. L., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis*, *24*(1), 29-36.

Marion, S. F., White, C., Carlson, D., Erpenbach, W. J., Rabinowitz, S., & Sheinker, J. (2002). *Making valid and reliable decisions in determining adequate yearly progress*. Washington, DC: Council of Chief State School Officers.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002). (Available online: http://www.ed.gov/legislation/ESEA02/)

"No Child Left Behind" test increases unlikely to be met. (2002, Spring). *The CRESST Line*, Spring 2002, pp. 4-5. (Available online: http://www.cresst.org/products/newsletters/CLSpring02final.pdf)

Olson, L. (2003, August 6). 'Approved' is relative term for Ed. Dept.. *Education Week*, pp. 1, 34-36.

Reeves, C. (2003). *Implementing the No Child Left Behind Act: Implications for Rural Schools and Districts*. Naperville (IL): North Central Regional Educational Laboratory. (Available online: http://www.ncrel.org/policy/pubs/html/implicate/)

Tompkins, R. B. (2003, March 26). Leaving rural children behind. *Education Week*, pp. 44, 30, 31.

Appendix

The general formula for a confidence interval based on the binomial distribution, from which the 95% and 99% confidence intervals in this monograph derive, is taken from Glass and Hopkins (1996) and attributed to Ghosh (1979). The Ghosh method "is very accurate and is now the method of choice for all values of *p* and *n*" (Glass & Hopkins, p. 326). The general formula is

$$P_L = \frac{n}{n+z^2}\left[p + \frac{z^2}{2n} - z\sqrt{\frac{p(1-p)}{n} + \frac{z^2}{4n^2}}\right]$$

$$P_U = \frac{n}{n+z^2}\left[p + \frac{z^2}{2n} + z\sqrt{\frac{p(1-p)}{n} + \frac{z^2}{4n^2}}\right]$$

where

- P_L and P_U are the lower and upper limits of the 1-α CI (where α usually is either .05 or .01)
- *p* is the sample proportion
- *n* is the sample size
- *z* is the two-tailed critical value ($z = 1.96$ where $\alpha = .05$, and $z = 2.58$ where $\alpha = .01$)

Consult any introductory statistics textbook for elaboration on the statistical concepts and procedures you encountered in this monograph. Glass and Hopkins (1996) is a superb resource, particularly if you desire a more technical treatment. For a text having somewhat more relaxed prerequisites, you may find Coladarci, Cobb, Minium, and Clarke (2004) helpful.