Computation of Confidence Intervals for Growth Performance in Determination of Safe Harbor Eligibility

Sean W. Mulvenon and Charles E. Stegman
University of Arkansas - Fayetteville

Abstract
As part of No Child Left Behind (NCLB) legislation, many states are using confidence intervals to determine a range of scores for evaluating a school system. More specifically, the states are employing confidence intervals to help minimize measurement error in determining a school system's performance. The methodology and techniques employed in these NCLB calculations for confidence intervals have raised several questions with regard to appropriateness, methods, and the transfer to educational policy. The purpose of this paper is to review the methodology, application, and impact of the various methods in regard to educational policy. Additionally, simulations that examine variations in sample size and proportions were completed in order to examine how inconsistency can impact the determination of a school's performance relative to the achievement goals.

Background

The No Child Left Behind (NCLB) legislation, implemented in 2002, mandated that schools and districts be evaluated relative to state performance standards. Further, their performance is assigned a "grade" or designation of "Meets Standard," "Alert," or "School Improvement." A school is assigned the designation of "Meets Standard" if overall student performance on achievement tests attain the designated criteria established by the individual state. A school is assigned the status of "Alert" if it fails to meet designated performance standards for the current year, but has attained the status "Meets Standard" in the previous year. If a school fails to meet designated performance standards for two consecutive years, the school is placed in "School Improvement" and by statute in NCLB, parents have to be provided

the opportunity to transfer their child to alternative schools which have met the performance standard. Additionally, schools can be required to provide tutoring or other student support mechanisms which translates into increased financial costs for districts.

The assignment of a school to "School Improvement" can be avoided through a "Safe Harbor" provision within NCLB (NCLB, 2002). Safe Harbor is a flexible provision within NCLB regulations that allows for consideration of a school system's academic improvement during the most recent year or other time period as deemed appropriate. A school can be deemed as "Meets Standard" by exceeding the annual performance goals or by improving performance by a predetermined amount. If School A makes "adequate yearly progress," the standard for growth during the assigned time period, it is deemed as "Meets Standard" for performance consistent with NCLB legislation.

For example, Safe Harbor in Arkansas is attained if a school met attendance, percentage of students tested, and a 10% growth in achievement standards during the current year. The attendance rate and percent tested criteria are static at 91.13% and 95.0%, respectively. The 10% growth, however, is based on each school's previous year's performance. The amount of expected growth is very simple and straightforward to compute. For example: a school had 20% of students proficient on the achievement test last year, and this year must increase the percent of students proficient on the exam by 10% of the difference between 20% and 100% (i.e. 100 - 20 divided by 10). Thus, the performance growth goal for Safe Harbor determination for this school is 8%.

It is also acknowledged that a certain amount of measurement error will exist in this process, so to provide a best case scenario for schools, the use of confidence intervals has been proposed to develop lower versus upper bound values in the system. Numerous statistical issues have been raised in regard to the development and implementation of confidence intervals in this process, from inaccurate determination, discrepancy in sample sizes, and one-tail or two-tailed intervals.

Example of a School System Appeal

Suppose School System A has appealed the designation of the performance category of "School Improvement" and applied for "Safe Harbor" in part based on the calculation of confidence intervals by the Arkansas Department of

Education. As stated in the Arkansas Consolidated State Accountability Plan, to invoke the "Safe Harbor" provision schools must meet three conditions: (a) they must have tested 95% of their students; (b) if a school does not have a high school graduation rate, they must meet the 91.13% attendance rate; if a school does have high school graduation, they must attain a graduation rate of 73.9%; and (c) they must have a 10% reduction in the difference between last year's performance and the attainment of 100% of students proficient on the achievement test as described above from 20 to 28%–this is referred to as a 10% growth, but should not be confused with a 10% improvement from last year's performance, or in the example above, from 20% to 22%. Additionally, School System A has raised the issue of performance against the state standards for Literacy and Mathematics (see Table 1 on next page). Table 1 provides the performance goals for schools and the lower bound values. A school is expected to meet the performance goal, but if the school meets the lower bound value for the confidence interval it is considered to have met performance standards for the academic year.

Methodology

Computation of Confidence Intervals

Various methods can be used for computing confidence intervals. A comparison of the differences in the two most widely used methods will be included in this review for the School System A appeal. The first method is the traditional method for computing confidence intervals for proportions (Glass & Hopkins, 1996). The second method is the Ghosh method (1979), which addresses distributional and sample size issues which can be problematic in the more traditional method.

Method 1: The Traditional Method

The confidence intervals were computed using standard statistical methodology for computing these ranges (Glass & Hopkins, 1996). Both, two-sided and one-sided confidence intervals were computed using a 75% confidence band. In layman's terms, this means that over repeated samples, one would expect the "true" percentages of students for a school would reside in 75% of the intervals. A 75% confidence interval was employed in lieu of the more traditional 68% or 95% intervals due to language used in the approval of a statewide school

Table 1
Literacy and Mathematics Performance and Standards for 2003

AYP Group	Literacy Goal	Expected Gain	*SE*	Lower Bound Literacy	Mathematics Goal	Expected Gain	*SE*	Lower Bound Mathematics
K - 5	37.48	6.3	.71	36.66	34.18	6.6	.75	33.32
6 - 8	24.93	7.5	.93	23.86	22.36	7.8	1.05	21.15
9 - 12	26.21	7.4	.66	25.45	17.87	8.2	.69	17.08

Note: Literacy 9 - 12 was 26.21 - (.66)(1.15) producing 25.45.

improvement plan. Originally, if a school met 75% of their performance goal (i.e., if the goal was 10% growth and the school obtained 7.5% or greater) they were consider to have "MET" their growth expectations. Given the language submitted and approved by the U.S. Department of Education, 75% confidence intervals were employed.

One-Tail versus Two-Tailed Confidence Intervals

A two-tailed confidence interval equally divides the 75% confidence interval around a school's obtained percentage of students proficient. Thus, for a 75% confidence interval, 37.5% of this band is below the obtained score and 37.5% is above their score. Using the normal approximation, a *z*-value is obtained to identify 37.5% of the area from the center of a standard normal curve, in this case $z \pm 1.15$, and is used to multiply the standard error and create the confidence interval as demonstrated in the provided example.

If the hypothesis or direction of a percentage is known *a priori* you can also calculate a one-tailed confidence interval using the 75% criteria. Using this method, 75% of the distribution is identified as resting below or above an identified value, predicated on the directional hypothesis for performance. The *z*-value for the standard normal table is identified, in this case $z = .674$, and is used to compute the confidence interval (See Tables 2 - 3 on the following pages).

Table 2
Two-Tailed 75% Confidence Intervals Using Both Traditional and Ghosh Methods for Literacy and Mathematics for Selected School Groups

Condition	2002 Prolit	2003 Prolit	10% Growth Goal	Traditional UB Lit	Ghosh UB Lit	2002 Promath	2003 Promath	10% Growth Goal	Traditional UB Math	Ghosh UB Math
School Total:										
School A	19.14	28.07	27.23	30.56	30.63	17.13	29.47	25.42	31.99	32.05
School B	24.32	18.10	31.89	21.08	21.26	6.93	7.94	16.24	10.71	11.16
School C	13.60	28.29	22.24	31.52	31.62	10.66	21.63	19.60	23.91	24.00
FRLP Total:										
School A	14.34	20.77	22.91	23.66	23.81	10.57	17.31	19.51	20.00	20.17
School B	12.82	6.90	21.54	10.72	11.76	0.00	4.88	10.00	8.75	10.35
School C	8.77	16.00	17.89	20.22	20.66	8.47	12.06	17.63	15.21	15.57
Special Ed.:										
School A	1.75	4.55	11.58	8.16	9.67	0.00	4.55	10.00	8.16	9.67
School B	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
School C	0.00	15.00	10.00	24.18	26.32	0.00	15.00	10.00	24.18	26.32
Black Total:										
School A	12.50	23.25	21.25	26.46	25.18	9.38	19.74	18.44	22.77	22.94
School B	12.82	6.42	21.54	9.12	8.19	0.00	2.63	10.00	4.74	5.69
School C	9.73	24.51	18.76	29.41	27.49	7.32	15.03	16.59	18.15	18.42
White Total:										
School A	28.06	36.63	35.26	40.85	40.94	27.55	47.09	34.80	51.47	51.48
School B	40.00	33.33	46.00	38.87	39.06	19.79	16.28	27.81	22.75	23.74
School C	17.86	36.36	26.07	41.39	41.52	16.28	28.97	24.65	32.54	32.66

Note: Prolit = percent proficient in literatcy; Promath = percent proficient in mathematics.

Table 3

One-Tailed 75% Confidence Intervals Using Both Traditional and Ghosh Methods for Literacy and Mathematics for Selected School Groups

Condition	2002 Prolit	2003 Prolit	10% Growth Goal	Traditional UB Lit	Ghosh UB Lit	2002 Promath	2003 Promath	10% Growth Goal	Traditional UB Math	Ghosh UB Math
School Total:										
School A	19.14	28.07	27.23	29.55	29.56	17.13	29.47	25.42	30.96	30.97
School B	24.32	18.10	31.89	19.86	19.91	6.93	7.94	16.24	9.57	9.71
School C	13.60	28.29	22.24	22.24	30.22	10.66	21.63	19.60	22.98	23.00
FRLP Total:										
School A	14.34	20.77	22.91	22.48	22.52	10.57	17.31	19.51	18.90	18.95
School B	12.82	6.90	21.54	9.16	9.49	0.00	4.88	10.00	7.17	7.68
School C	8.77	16.00	17.89	18.49	18.62	8.47	12.06	17.63	13.92	14.03
Special Ed.:										
School A	1.75	4.55	11.58	6.68	7.17	0.00	4.55	10.00	6.68	7.17
School B	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
School C	0.00	15.00	10.00	20.43	21.16	0.00	15.00	10.00	20.43	21.16
Black Total:										
School A	12.50	23.25	21.25	25.15	25.18	9.38	19.74	18.44	21.53	21.57
School B	12.82	6.42	21.54	8.02	8.19	0.00	2.63	10.00	3.88	4.18
School C	9.73	24.51	18.76	27.41	27.49	7.32	15.03	16.59	16.88	16.95
White Total:										
School A	28.06	36.63	35.26	39.13	39.14	27.55	47.09	34.80	49.68	49.66
School B	40.00	33.33	46.00	36.61	36.65	19.79	16.28	27.81	20.11	20.42
School C	17.86	36.36	26.07	39.34	39.36	16.28	28.97	24.65	31.08	31.10

Note: Prolit = percent proficient in literacy; Promath = percent proficient in mathematics.

Equation for both One-Tailed and Two-Tailed Confidence Intervals

The confidence intervals were calculated using the following formula:

$$.75C.I. = p \pm (z - value) \cdot \sigma_{se} \text{ with,}$$

$$\sigma_{se} = \sqrt{\frac{p \cdot (1 - p)}{N}}$$

and a *z*-value of 1.15, "*p*" is the percent of students proficient, and *N* is the sample size. For the one-tail confidence intervals a *z*-value of .674 is used and the C.I. has the form p + (*z*-value)σ_{se}.

Example of Use of Confidence Intervals

Using the School Total data for School A from Table 2, School A obtained the scores of 19.14 percent proficient for 2002 Literacy and 28.07 percent for 2003 Literacy, with an improvement of 8.93 percent. The goal for growth was (100 - 19.14)/10 = 8.086 or the target for 2003 was 19.14 + 8.086 = 27.23 percent of students proficient. The elementary school met this goal with 28.07 percent of their students proficient. The upper bounds for the confidence intervals are:

one-tailed: 28.07 + (.674)(2.16) = 29.53

two-tailed: 28.07 + (1.15)(2.16) = 30.55

The values demonstrate the inherent value and added statistical power issues associated with using one-tailed versus two-tailed confidence intervals. In practice, smaller confidence intervals are desirable. A common use of these intervals is in hypothesis testing with a distributional hypothesized value, determining statistical significance identified. Failure to have this hypothesized value within the confidence interval indicates a "statistically significant" difference between the obtained value and the hypothesized value. Typically, a difference of this magnitude is important for researchers. The goal is to be 75% confident that the true proportion is less than or equal to 29.53 and 75% confident that the true proportion is between 25.59 and 30.55. In practice, however, if the school had a performance goal of 30% of students proficient it would have been judged to "meet" this standard using the two-tailed interval provided a margin for error in consideration of the school's "true" performance. In the context of attempting to obtain an easier standard to assess

if a school was within a 75% confidence interval in meeting the performance growth, it actually creates a more rigid standard with 30.55 "wider" or a greater upper bound obtained for the two-tailed case over the 29.53 used for a one-tailed confidence interval.

Method 2: The Ghosh Method

The Ghosh method uses the binomial distribution, in contrast to the normal distribution, and has been demonstrated to be more accurate over other procedures (Ghosh, 1979; Glass & Hopkins, 1996). Next, the Ghosh method and equations will be introduced and applied to the same conditions as the traditional method for computing confidence intervals.

Equations for Two-Tailed Confidence Intervals

$$\pi_L = \frac{n}{n+z^2}\left[\hat{p}+\frac{z^2}{2\cdot n}-z\cdot\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}+\frac{z^2}{4n^2}}\right]$$

$$= \frac{n}{n+(1.15)^2}\left[\hat{p}+\frac{(1.15)^2}{2\cdot n}-(1.15)\cdot\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}+\frac{(1.15)^2}{4n^2}}\right]$$

$$= \frac{n}{n+1.3225}\left[\hat{p}+\frac{.6613}{n}-(1.15)\cdot\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}+\frac{.3306}{n^2}}\right]$$

Similarly, the upper bound is calculated using

$$\pi_U = \frac{n}{n+z^2}\left[\hat{p}+\frac{z^2}{2\cdot n}+z\cdot\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}+\frac{z^2}{4n^2}}\right]$$

$$= \frac{n}{n+1.3225}\left[\hat{p}+\frac{.6613}{n}+(1.15)\cdot\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}+\frac{.3306}{n^2}}\right]$$

Thus, one can be 75% confident that the value of π is with the range [π_L, π_U]. The actual probability that π is within any specific interval is either 0 or 1.

Equation for One-Tailed Confidence Intervals

A one-tailed confidence interval that sets an upper bound, i.e., we are 75% confident that π is less than or equal to π_U is as follows:

$$\pi_U = \frac{n}{n+z^2}\left[\hat{p}+\frac{z^2}{2n}+z\cdot\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}+\frac{z^2}{4n^2}}\right]$$

For the 75% one-tail C.I. for π less than or equal to π_U we have:

$$\pi_U = \frac{n}{n+(.674)^2}\left[\hat{p}+\frac{(.674)^2}{2n}+(.674)\cdot\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}+\frac{(.674)^2}{4n^2}}\right]$$

$$= \frac{n}{n+.4543}\left[\hat{p}+\frac{.2271}{n}+(.674)\cdot\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}+\frac{.1136}{n}}\right]$$

Comparison of the Results of the Two Methods

Effectively, the two methods produce similar results for larger sample sizes (see Table 4 on the next page). However, given the NCLB issues associated with smaller sample sizes, and the incredible discrepant nature of school sizes in rural states such as Arkansas, the Ghosh method appears to be more equitable. For example, if you adjust the sample size to represent a very small school of 40 students versus a very large school of 500 students, the values for the Ghosh confidence interval are slightly smaller. You can also see that as the value of π deviates from .5, the Ghosh method makes a greater adjustment from the traditional method. In all cases, the decision using the traditional or Ghosh methods are the same (see Tables 2 and 3), but the Ghosh method provides additional protection against sample size variation and is recognized as the more efficient method.

Importance for Education

The implementation of NCLB legislation has created many interesting measurement and statistical questions. Further, given the bipartisan support in Congress for NCLB, it is unlikely there will be any significant changes

Table 4
Comparing Three Samples Size for Ghosh Method

		Traditional		Ghosh	
Sample Size	Proportion	LB	UB	LB	UB
40	.1	.045	.155	.058	.168
	.2	.127	.273	.137	.281
	.3	.217	.383	.224	.389
	.4	.311	.489	.316	.491
	.5	.409	.591	.411	.589
	.6	.511	.689	.509	.685
	.7	.617	.783	.611	.776
	.8	.727	.873	.718	.863
	.9	.845	.955	.832	.942
100	.1	.065	.135	.071	.140
	.2	.154	.246	.158	.250
	.3	.247	.353	.250	.355
	.4	.344	.456	.345	.457
	.5	.443	.558	.443	.557
	.6	.544	.656	.543	.655
	.7	.647	.753	.645	.750
	.8	.754	.846	.750	.842
	.9	.866	.935	.860	.929
500	.1	.085	.115	.086	.117
	.2	.179	.221	.180	.221
	.3	.276	.324	.277	.324
	.4	.375	.425	.375	.425
	.5	.474	.526	.474	.526
	.6	.575	.625	.575	.625
	.7	.676	.724	.676	.723
	.8	.779	.821	.779	.820
	.9	.885	.915	.884	.914

Note: LB = lower bound; UB = upper bound.

in this legislation until the reauthorization of the Elementary and Secondary Education Act (ESEA). Given that the consequences of NCLB legislation are very real for schools, educators, and students, it is paramount that groups such as educational statisticians complete studies and provide insight on methods and assessment practices that are appropriate.

Recommendations for NCLB and Implications for State Educational Agencies

A reality of NCLB was that the operationalizing of this legislation, aside from some very broad guidelines, was left to SEAs. The use of confidence intervals is understood and appreciated, but some specific recommendations for NCLB include:

1) Allow states to adjust scores using one standard deviation with the standard error of measurement. If the upper limit of a student's interval includes the "passing score," they can report the student as provisionally passing. This would indicate the student's score was below the "passing score" but within measurement error.

2) If use of confidence intervals at the school level is continued, it is recommended the Ghosh method be employed. Additionally, it is recommended the width of the intervals be limited to 68%. Given the large percentage of students tested from the school's "population," it is expected there will be limited measurement error in the "true" score for the school system.

From a policy perspective, it is important that SEAs embrace the intent of NCLB to measure the performance of school systems and ensure that all students are receiving access to a quality education. The use of statistical approaches that are only positively biased, such as how confidence intervals have been applied, represents only one area where the policies of NCLB have been inconsistent with sound mathematical and statistical methodologies. Given the actual NCLB legislation included the term "scientifically based" over 100 times, it seems reasonable to expect, say demand, the measurement and statistical models employed to evaluate school systems be held to a standard that is beyond what is politically expedient or legally within the confines of the law. The use of suspect use of statistical applications tends to detract from the otherwise laudable efforts to improve the K-12 system nationally via the implementation of NCLB. The intent of this research is to help provide improved assessment of school performance, which should be the initial step in any reform model.

References

Arkansas Department of Education, Consolidated State Application Accountability Plan, Section 3.2b. Amended April 18, 2003. Retrieved May 15, 2005, from http://www.ed.gov/admins/lead/account/stateplans03/arcsa.pdf

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.

No Child Left Behind Act of 2001, Pub. L. No. 107-110 (2002). § 1001.