Why NCLB is a Statistical Sham
Part I: How the Confidence Interval (margin of error) Procedures Destroy the Credibility of State NCLB Plans

David Rogosa
Stanford University
November 2003
rag@stat.stanford.edu

Introduction

In presidential politics, there are the color-coded maps of the blue states versus red states. In educational accountability, the implementations of No Child Left Behind have spawned a division based on statistical credibility: the "confidence interval" (nee margin of error) states versus states that avoid such machinations. The purpose of this technical note is to explain the rather ludicrous properties of the NCLB confidence interval procedures, which stem from basic misunderstandings of material from introductory statistics courses.

The genesis of these confidence interval procedures appears to be the December 2002 report from the Council Of Chief State School Officers "Making Valid And Reliable Decisions In Determining Adequate Yearly Progress." Deeply involved in these developments is the Center for Assessment (aka, The National Center for the Improvement of Educational Assessment, NCIEA) as seen, for example, in Marion and Gong (2003).

As with most follies, there is a kernel of credible concern at the core of the confidence interval NCLB schemes. NCLB requires that a school and all its eligible subgroups meet a specified performance goal--e.g., proportion of students achieving the "proficient" designation must be at least .19. (In NCLB-speak this performance goal has the designation of Annual Measurable Objective or AMO.) That would seem a simple enough criterion.

Things get complicated and go astray because of clumsy attempts to deal with statistical uncertainty in the school and subgroup scores. The proportion of students proficient does indeed have some associated statistical uncertainty, commonly described as resulting from the facets: sampling variability in drawing the specific students tested and measurement or classification error in the NCLB subject test instruments. How (badly) the statistical uncertainty is handled in the state NCLB plans is the concern herein. (Whether the NCLB structure represents a wise accountability system is a separate issue not dealt with here.)

One way of expressing the issue confronted by these margin of error schemes is in terms of the burden of proof arising from statistical uncertainty in the school and subgroup scores. Certainly, it's unreasonable to expect

schools to prove beyond a reasonable doubt that they have met the performance goal. On the other hand, it's also indefensible to award schools a passing designation even though the probability is measured in parts per million that the scores stripped of their statistical uncertainty would meet the performance goal. There's a case to be made for awarding the benefit of the doubt in favor of schools, but there has got to be a limit.

An analogy for the confidence interval procedures is an abusive tax shelter, with CCSSO and the Center for Assessment having the role of the sponsoring accountancies and the state education agencies as the subscribing taxpayers. Certainly, NCLB with its strictures, complexity, contingencies (e.g., safe harbor), and spewing of acronyms has obvious similarities to the Tax Code. There are legitimate shelters (or incentives) included in the tax code, but at some point of excess legitimacy fades, and the tax shelter is clearly abusive. Similarly, effects of statistical variability are a legitimate concern in the NCLB accountability process, but the proposed (and approved) state plans for these confidence interval adjustments result in properties so excessive as to be clearly illegitimate.

The calculations in this note represent preliminary statistical work examining these NCLB plans. The particular form of these calculations are directed toward the confidence interval procedures. Many other calculations are necessary to augment this first pass, including how to do NCLB accountability well (i.e. with defensible statistical properties), or whether it can be done well at all. In that spirit, the calculations in this note should be seen as examples of some of the calculations that could/should be done in understanding any state NCLB plan. And the content herein is far more suggestive as to what should be done rather than representing a comprehensive set of results.

I want to congratulate news reports, such as in the Chicago Tribune (September 28, 2003 "Schools Toying with Test Results: Some States Meet Standards with Art of Statistics", D. Rado and D. Little) for calling public attention to the confidence interval scam. Also Redelman (2003) properly criticizes the Indiana use of 99% confidence ("Confidence game in the Hoosier State"). I am happy to add my initial calculations to these justified protests.

The 99% plans.
These NCLB confidence interval procedures employ elementary statistical procedures for inferences about proportions, stating the use of "99% confidence" as if that were a good thing. One version is to construct a crude large-sample Normal approximation confidence interval about the observed proportion proficient, as in Kentucky. Other variants of the 99% plans are couched in terms of hypothesis testing, as in Indiana and Utah. The details of the state procedures are something of moving targets, changing almost weekly in some instances, and varying in unstated ways among states. The analyses here are based on some modal template versions of the procedures. In particular, it's best to think of the calculations here as illustrating useful calculations that should be applied in order to understand a specific state plan.

For the 99% confidence interval, a NCLB procedure is that the school or subgroup's proportion proficient meets the performance goal if the

$$\text{adjusted proportion} = \text{observed proportion proficient} + 2.58 * \text{Sqrt}[pS*(1 - pS)/n]$$

meets or exceeds the stated performance goal (e.g. AMO of .19). Here n is the number of students in the school or subgroup and pS denotes the statewide proportion proficient (or in some versions the AMO or in others the obtained empirical proportion).

Alternative 99% plans in Indiana and Utah make reference to one-sided hypothesis testing which can be represented as

$$\text{adjusted proportion} = \text{observed proportion proficient} + 2.33 * \text{Sqrt}[pS*(1 - pS)/n]$$

where, again, various substitutions for pS can be made.
The more general form for the adjusted proportion proficient kludge:

$$\text{adjusted proportion}[k] = \text{observed proportion proficient} + k * \text{Sqrt}[pS*(1 - pS)/n]$$

where the 99% confidence plans have k = 2.58 or k = 2.33.

A minor issue (mentioned in the CCSSO report p.66 but seemingly ignored in state NCLB activities) is that even introductory educational statistics texts such as Glass and Hopkins (1996, Ch. 13 esp Figs 13.4, 13.5) inform the student that this form of inference for proportions (normal approximation using standard error Sqrt[p*(1 - p)/n]) is crude. Agresti (1990, p.76-7; 1996, p.15) provides exact forms for inference, as do standard statistical computing packages such as SAS.

In the details that follow it is easy to lose the important message. What is shown in this note is that these NCLB confidence interval procedures represent an egregious instance of "if it could be, it is" to an extent that sacrifices credibility of the system.

Issue #1.
A very few proficient students is good enough.

Before turning to the statistical properties of the confidence interval procedures a simple display of their consequences should raise concerns. Table 1 is one of many possible displays of how few proficient students become "good enough"; that is, "close enough" to the AMO via the confidence interval adjustment to satisfy these NCLB plans.

Start with the top frame of Table 1, which considers the performance goal proportion proficient at least .19. Under the k = 2.58 confidence interval procedures, if the statewide proportion proficient for this subgroup is .2, then groups smaller than 30 receive a "free ride", meaning that 0 students proficient is enough to satisfy the NCLB criteria. If the statewide proportion proficient for this subgroup is .5 then groups smaller than 47 obtain the free ride. If the statewide proportion proficient for the subgroup is .35, then groups smaller than 42 obtain the free ride, and a single proficient student out of a group of 51 is good enough. These free-ride numbers are slightly smaller with k = 2.33; groups smaller than 25 with pS = .2 and groups smaller than 38 with pS = .5. Clearly, states employing the margin of error procedure with low AMO and claiming a small minimum group size (Kentucky uses n=10) are not being forthcoming. In reality, those small groups get a free ride which is no different in consequence to the use of a larger minimum group size. (California, NOT a margin of error state, uses minimum n=50).

The lower frame of Table 1 repeats these calculations for a performance goal (AMO) of proportion proficient at least .45. For example, with k = 2.58 the lower frame shows that 10 proficient students out of 40 is close enough to .45 to satisfy NCLB for statewide proportion proficient .5. For k = 2.33 this changes slightly to 10 out of 38 or 11 out of 41. For smaller group size 5 out of 25 is close enough with k = 2.58 and 6 out of 26 with k = 2.33 (pS = .5).

Insert Table 1

The margin of error states have had difficulties even in representing the functioning of their own NCLB plans. For example, Indiana (IDOE, 2003) produced charts (widely used in expositions of these confidence interval approaches) showing their calculations for "Minimum Number Pass to meet AYP". Unfortunately, their own numbers describing their procedures do not appear to be correct. For example in the n=200 entry with ELA AMO .588 proportion proficient, the Indiana chart claims 105 proficient students out of 200 is close enough to the actual minimum number of students required, which is 118. However, a binomial cdf calculation (i.e., the level of a high-school statistics student) shows that 101 or 102 (depending on how the Indiana one-sided .01 test is constructed) proficient students would satisfy the .01 one-sided claim. Not such a large discrepancy, but instructive as to the statistical competence being employed in these NCLB enterprises.

Table 1
Size of subgroup for which minimum number proficient
satisfies performance goal for NCLB 99% confidence plans

	k = 2.58			k = 2.33		
	AMO .19 Proportion Proficient					
	statewide proportion proficient			statewide proportion proficient		
	.2	.35	.5	.2	.35	.5
Minimum Number Proficient	Maximum Number Students in Subgroup			Maximum Number Students in Subgroup		
0	29	41	46	24	34	37
1	39	51	56	33	44	47
2	48	61	65	42	53	56
3	56	69	74	50	61	65
4	64	78	82	58	69	73
5	72	86	91	66	77	81

	AMO .45 Proportion Proficient			
	statewide proportion proficient		statewide proportion proficient	
	.5	.65	.5	.65
Minimum Number Proficient	Maximum Number Students in Subgroup		Maximum Number Students in Subgroup	
0	8	7	6	6
1	12	11	10	10
2	15	15	14	13
3	19	18	17	16
4	22	21	20	19
5	25	24	23	22
6	28	27	26	25
7	31	30	29	28
8	34	33	32	31
9	37	36	35	34
10	40	39	38	37
11	43	42	41	40
12	46	45	43	42
13	48	47	46	45
14	51	50	49	48
15	54	53	52	50
16	57	56	54	53
17	59	58	57	56
18	62	61	60	58
19	65	64	62	61
20	68	66	65	64

Issue #2
Some probability calculations for "if it could be, it is"

As expressed above, the starting point for these NCLB procedures is reasonable--that the observed proportion proficient can be thought of as a version of "true" proportion proficient that is obscured by statistical variability. (If students could be drawn from the school population repeatedly and given very long tests, the average of these proportions proficient would converge to this "true" value.) And the CCSSO (2002) NCLB report speaks consistently in terms of inferences for this “true” percent proficient (e.g., pp. 65, 66, 81). Therefore there would seem to be wide agreement that a key quantity for understanding these NCLB plans is the probability that this (unknown) true proportion proficient for a school meets or exceeds the performance goal (AMO).

The main results of this paper are probability calculations for true proportion proficient, which are shown in the frames of Figure 1, also represented in Tables A-F, with additional summary in Tables 2, 3, 4. Given the data--observed number of proficient students--what can be said about the (unobserved) true proportion proficient? These calculations use the familiar beta-binomial formulation to calculate the posterior probability that the true proportion proficient meets or exceeds the performance goal [see Technical Appendix].

Insert Figure 1

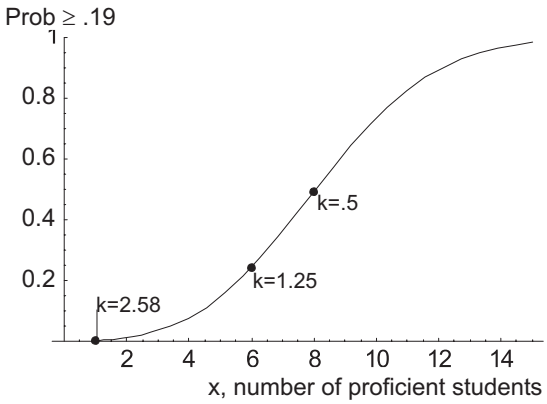
The base component of the various calculations is for a single subject and a single subgroup. Table 2 provides some examples of the probability calculations. The import of these calculations is to demonstrate that the number of proficient students deemed good enough to satisfy the performance goal in the 99% confidence NCLB procedures does not pass the laugh test.

Table 2
Probability "true" proportion proficient meets Performance Goal
at minimum number proficient for 99% margin of error procedures

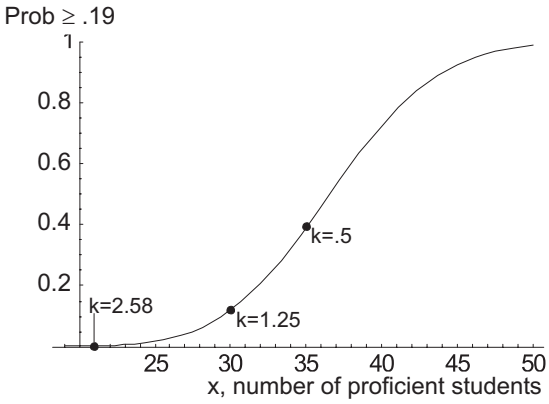
	Probability meets Performance Goal (minimum number proficient)		
	n=30	n=50	n=200
99% "confidence", pS = .35, AMO = .19			
k=2.58	.0287	.0046	.0018
	(0)	(1)	(21)
k=2.33	.0287	.0144	.0060
	(0)	(2)	(23)
99% "confidence", pS = .5, AMO = .45			
k=2.58	.0152	.0116	.0060
	(7)	(14)	(72)
k=2.33	.0354	.0234	.0131
	(8)	(15)	(74)

Figure 1. Consequences of the Confidence Interval Kludge

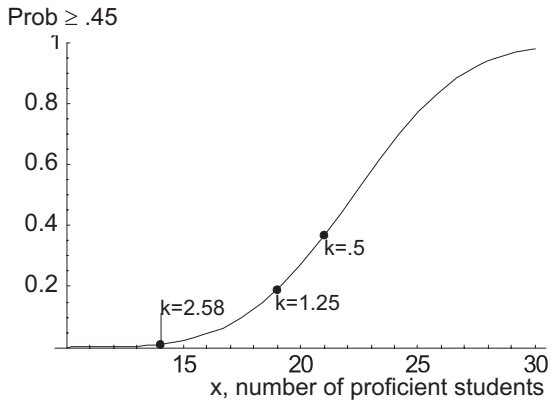
Plot of posterior probability for "true" proportion proficient;
n = 50, statewide proportion .35, performance goal .19



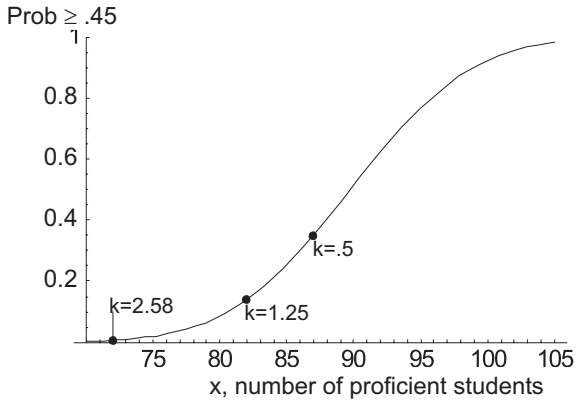
Plot of posterior probability for "true" proportion proficient;
n = 200, statewide proportion .35, performance goal .19



Plot of posterior probability for "true" proportion proficient;
n = 50, statewide proportion .5, performance goal .45

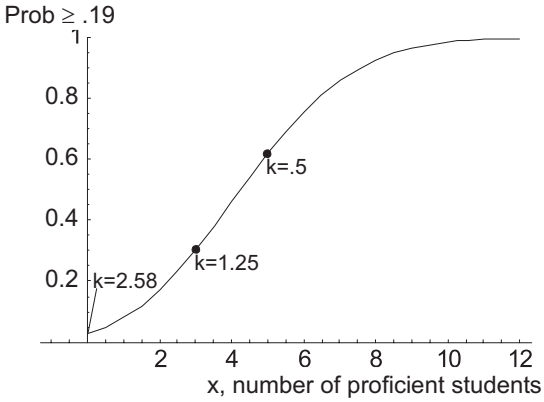


Plot of posterior probability for "true" proportion proficient;
n = 200, statewide proportion .5, performance goal .45

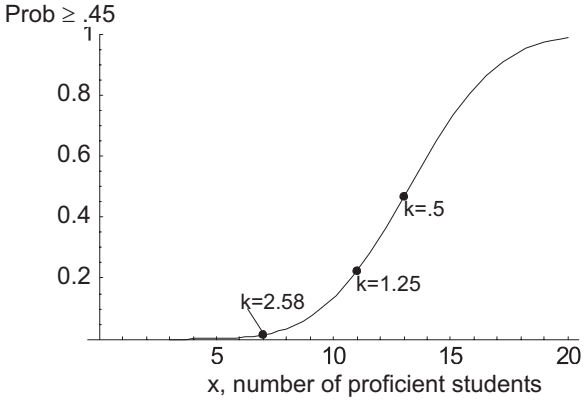


Displays for (very) small groups n=30

Plot of posterior probability for "true" proportion proficient;
n = 30, statewide proportion .35, performance goal .19



Plot of posterior probability for "true" proportion proficient;
n = 30, statewide proportion .5, performance goal .45



For example, in Table 2 the top right entry for a group of 200 students with AMO .19 and k=2.58 indicates that 21 proficient students out of 200 is close enough for the NCLB procedure. Yet the probability that the true proportion proficient meets or exceeds .19, given the data of 21 proficient students out of 200, is only .0018. To calibrate, a probability of .0018 is slightly less than the probability of flipping a fair coin 9 times and obtaining a head each time. It might happen, but is it serious national educational policy to deem that result to be close enough? This example is depicted in more detail in Table B and in the top right frame of Figure 1. For k=2.33 that same scenario requires 23 proficient students out of 200, and for 23 proficient students the probability that the true proportion proficient meets or exceeds .19 is .0060, which is slightly less than the probability of flipping a fair coin 7 times and obtaining a head each time.

It would be counter-productive to try to present analyses of the properties of all possible or even existing variants of these NCLB confidence interval procedures, and state plans are evolving in these details. The purpose here is really to illustrate the kinds of calculations that should be done to understand a proposed NCLB procedure. That said, one variant of the adjusted proportion is worthy of note, in which the AMO is used as the pS yielding:

adjusted proportion[k] = observed proportion proficient + k*Sqrt[AMO*(1 - AMO)/n]

where the 99% confidence plans have k = 2.58 or k = 2.33. This alteration has no effect on the results for the AMO = .45, pS = .5 entries in Table 2. The effect for the pS = .35, AMO = .19 entries is to require 2 or 3 more proficient students to be close enough. That effect is enough to increase the small probabilities in upper half of Table 2 by a factor of 3 or 4.

As expressed in the abusive tax shelter analogy: at some point of excess legitimacy fades. How unlikely is too unlikely? Just from this single test for a single subgroup calculation, even before the compounding effects of the multiple tests and multiple subgroups, the confidence interval procedures appear to have lost all legitimacy. Certainly under the NCLB system, there is a strong case to be made for applying the benefit of the doubt towards the schools; probabilities of .5 that the true proportion proficient met the AMO would clearly be OK (i.e. rounding .5 up to 1). And there may be cogent arguments to be made for .25, and possibly maybe .1, being enough positive evidence (and that's really bending over backwards in benefit of doubt). But as can be seen in Table 2 the 99% confidence procedures confirm results with probabilities measured in parts per hundred or even parts per thousand (for a single group on a single test).

Satisfying Both English and Math AMO.

NCLB AYP requires meeting both English and math AMO, and thus the computed probabilities for a single subject as in Table 2 would greatly overstate the probability of the true proportion proficient meeting both math and English AMO. If math and English results were uncorrelated, then the joint probability would just be the product of the two computed probabilities. Because the same students take both the math and English tests, the two tests are not independent, but also not redundant (math ability and English

ability are not matched perfectly over students and measurement variability in the two tests is regarded as independent). For discussion, take the simplest case of math and English having the same AMO and pS. The probability of true proportion proficient on both subjects meeting the AMO is in-between the probability for one subject and the product of the two probabilities. For the small probabilities seen in these calculations a decent (conservative) approximation is provided by a very simple form: p/3, where p is the computed probability in Table 2 for a single subject. Thus for the k=2.33 99% confidence and n=200 group example above (which indicates a single subject probability .006), the probability that the true proportions proficient for both English and math meet or exceed .19 is computed to be around .0015 (and .006/3 yields .002). The crude p/3 approximation will be used in later displays. The .0015 probability is slightly less than the probability of flipping a fair coin 9 times and obtaining a head each time. Do such probabilities represent close enough?

Schools and Multiple Subgroups.

The NCLB structure requires that the school-wide scores on English and math meet the AMOs and also these subject scores meet the AMOs for each of the included subgroups. The calculations here are for an artificial school composed of three non-overlapping subgroups whose union comprises the school population. One example is a school of size 600, composed of 3 subgroups each of size 200. A second example is a school of 350 students with three subgroups of size 200, 100 and 50 students.

Results for the probability that the true proportions proficient meet the stated AMO for both school and subgroups on both subjects are displayed in Table 3. The probabilities are expressed in millionths; note for reference the probability of flipping a fair coin 20 times and obtaining a head each time is just slightly less than 1 millionth. Two versions of NCLB 99% confidence are shown: k = 2.58 and k = 2.33.

Start with School A, an artificial school of size 600 composed of 3 subgroups each with 200 students, each subgroup having statewide proportion proficient .35 (pS). With k=2.58, AMO .19, and pS = .35 each of these 3 subgroups can satisfy the close enough criteria of NCLB 99% confidence margin of error with 21 proficient students. But NCLB also requires the school-wide score to meet the AMO, and for this close enough according to the margin of error is 84 proficient students out of 600 (not 21*3 = 63). Thus a configuration of proficient students that would satisfy both subgroup and school criteria is 21, 21, and 42 proficient students in the three subgroups respectively. The 99% confidence procedure using k = 2.33 changes this acceptable configuration of proficient students to {23, 23, 41}. For k = 2.33, a configuration of students meeting the AMOs under the NCLB 99% confidence interval procedure has estimated probability of one-millionth that the true proportions proficient meet the stated AMO. Yet this probability is large enough to pass NCLB scrutiny. If schools really were this adept at winning longshots, then school funding would not be an issue.

Table 3
"99% Confidence" Posterior Probabilities for True Proportions Proficient Satisfying AMO for School and Three Subgroups and Two Subjects Expressed in Parts per Million

	AMO = .19	
	k = 2.33	k = 2.58
School A	1	.1
600 students		
groups {200,200,200}	{23,23,41}	{21,21,42}
all pS = .35		
School B	6	.3
350 students		
groups {200, 50,100}	{23, 3,20}	{21, 1,21}
pS = {.35,.35, .5}		
	AMO = .45	
	k = 2.33	k = 2.58
School C	5	1
600 students		
groups {200,200,200}	{74,74,94}	{72,72,95}
all pS = .5		

Artificial School B is a school composed of 350 students who belong to one of 3 subgroups: group 1 with n = 200 and pS =.35, group 2 with n = 50 and pS = .35, and (a typically higher achieving subgroup) group 3 with n = 100 and pS = .5. For the school, 46 (k=2.33) or 43 (k=2.58) proficient students out of 350 is "close enough" to satisfy the performance goal of .19 proportion proficient. Configurations of number of proficient students in each subgroup meeting the "99% confidence" procedures are shown below the probabilities that the true proportions proficient meet the AMO = .19.

School C, an artificial school of size 600 composed of 3 subgroups each with 200 students and each subgroup having statewide proportion proficient .5 (pS) has AMO = .45. For 99% confidence k = 2.33 a configuration of proficient students in the three subgroups of {74,74,94} is considered close enough to the AMO. For this configuration, the probability that the true proportions proficient meet the stated AMO for both school and subgroups on both subjects is 5 millionths.

Issue #3
Is there a more reasonable adjustment setting for "close enough"?

Consider the more general form for the adjusted proportion proficient kludge:

$$\text{adjusted proportion}[k] = \text{observed proportion proficient} + k*\text{Sqrt}[pS*(1 - pS)/n]$$

Clearly, the demonstrations in Issue #2 (see also Figure 1 and Tables A-F) establish that "99% confidence" NCLB state plans induce laughter, not educational excellence. Would a smaller value of k be more defensible, or at least less ludicrous? Alternative values of k explicitly considered in Figure 1 and Tables A-F are k = 1.25 and k = .5. Reducing k does increase the number of proficient students required to be considered "close enough" to the performance goal. Consequently, reducing k will also increase the probability that true proportion proficient meets the performance goal. Examining Tables A-F and Figure 1 will provide some intuitions for the increase in number of required proficient students and the corresponding increases in the probabilities as k is reduced. Table 4 assembles some of those values for single subgroup, single subject. With k = 1.25 these posterior probabilities are around one-quarter for n = 50, and with k = .5 (a modest quantity that might be acceptable) that probability is close to one-half.

Insert Table 4

Taking the school examples with 3 subgroups used in Issue #2 above provides an additional quick look at the consequences of a smaller k. First example is a school of 600 students and AMO = .19, consisting of three subgroups of size 200 each with pS = .35. For k = 1.25, the minimum number of proficient students for the school is 100 (out of 600) and for each subgroup 30 (out of 200). Consequently, a configuration of number of proficient students for the three subgroups that satisfies both the subgroup and the full school performance goals is {30, 30, 40}. This {30, 30, 40} configuration has probability "true" proportions proficient meets Performance Goal of .0103, nearly 4000 times larger than the k=2.58 value. More plausibly "close enough"? Reducing k further to .5, the minimum number of proficient students for the school is 109 (out of 600), and for each subgroup 35 (out of 200). Consequently, number of proficient students in the three subgroups {35, 35, 39} satisfies the subgroup and full school performance goals. This {35,35,39} configuration has probability "true" proportions proficient meets Performance Goal of .1024, 10 times larger than the k=1.25 value and nearly 40000 times larger than the k=2.58 value. Eliminating the margin of error altogether would result in number of proficient students {38, 38, 38} satisfying the subgroup and full school performance goals with probability .217 that "true" proportions proficient meets AMOs. Table 5 provides summary of this example as well as the second school example in Issue #2 (school size 350 with subgroups n= 200, 50, 100). The pattern and indications of the results are similar for both examples. The policy question is: What would reasonable persons deem as "close enough"?

Insert Table 5

Table 4
Probability "true" proportion proficient meets Performance Goal at minimum number proficient for various margin of error procedures for single subgroup, single subject

pS = .35, AMO = .19

	Probability meets Performance Goal (minimum number proficient)		
	n=30	n=50	n=200
k = 2.58	.0287 (0)	.0046 (1)	.0018 (21)
k = 1.25	.303 (3)	.244 (6)	.119 (30)
k = .5	.619 (5)	.494 (8)	.392 (35)
k = 0	.756 (6)	.737 (10)	.601 (38)

pS = .5, AMO = .45

	Probability meets Performance Goal (minimum number proficient)		
	n=30	n=50	n=200
k = 2.58	.0152 (7)	.0116 (14)	.0060 (72)
k = 1.25	.223 (11)	.190 (19)	.138 (82)
k = .5	.469 (13)	.369 (21)	.351 (87)
k = 0	.602 (14)	.582 (23)	.515 (90)

Table 5
Probability "true" proportions proficient for single subject meets Performance Goal at minimum number proficient for margin of error procedures: Schools with 3 subgroups and AMO = .19

	Probability meets Performance Goal (AMO = .19) {minimum number proficient configuration}	
	School Example A	School Example B
k = 2.58	.00000274 {21, 21, 42}	.00000681 {21, 1, 21}
k = 1.25	.0103 (30, 30, 40}	.0216 (30, 6, 20}
k = .5	.102 (35, 35, 39}	.127 (35, 8, 19}
k = 0	.217 (38, 38, 38}	.291 (38, 10, 19}

School A: 600 students in 3 subgroups n = 200 with pS = .35
School B: 350 students in 3 subgroups: group 1 with n = 200 and pS =.35, group 2 with n = 50 and pS = .35, group 3 with n = 100 and pS = .5.

Issue #4
Going the other way: Investigating failures to meet Performance Goals

What has been shown so far is that the NCLB 99% confidence procedures provide remarkable (excessive?) levels of benefit of the doubt for schools meeting performance goals. Configurations with probabilities of success measured in parts per million are deemed as "close enough". The motivation for the margin of error procedures may be the desire to ensure (as much as possible) that schools are not falsely labeled as "needs improvement". But good motivations can create excessive overreaction. Perhaps a more positive view of these margin of error procedures would arise from investigating a different scenario: schools that fail to meet the performance goal either by subgroup or total group scores falling short.

Just miss scenarios.
The worst case scenario for "false fails" would be represented by a school just missing, either by virtue of a school score or subgroup score, the number proficient demanded by the 99% confidence procedure. Under this "just miss" scenario, what's the probability that the school deserved to pass? I.e., calculate the probability that the school and subgroups have true proportions proficient meeting the performance goal. One set of calculations can be based on jiggling the scenarios in Tables 3 or 5 (considering school and subgroups for a single subject) to produce "just-make/just-miss" configurations. Table 6 gives configurations of number of proficient students such that all subgroups are close enough under 99% confidence but the school is one proficient student short. Certainly, Table 6 indicates that the 99% confidence NCLB procedures do not fail schools with reasonable probability of having achieved the requisite educational performance (the small probabilities in Table 6 would be considerably smaller if adjusted for two subjects tested). Also in Table 6 the entries for AMO - .5 show that these "just fail" schools really aren't close to meeting the AMOs, in that the probabilities (even for a single subject) are not large for meeting a reduced performance goal (AMO - .05).

Insert Table 6

A more favorable set of configurations would be two of three subgroups meeting the performance goal without any margin of error kludge but one subgroup just missing the number proficient required to be "close enough" under 99% confidence, k = 2.33. For artificial School A in Table 6 consider a different testing outcome in which the numbers of proficient students in the three subgroups is {38, 38, 22}: a configuration such that two of the three subgroups meet the performance goal of .19 proportion proficient without the benefit of any margin of error adjustment, school proportion proficient .1633 is close enough under the margin of error, but the third subgroup is just one proficient student short. But even here, the probability that all three subgroups have true proportion proficient of at least .19 is about than one-part in a thousand: .00122. Is there harm done by labeling this school as "needs improvement"?

Table 6
"99% Confidence" Posterior Probabilities for True Proportions Proficient Under Just-Miss Scenarios for School and Three Subgroups, Single Subject

	AMO = .19	
	k = 2.33	k = 2.58
School A 600 students with groups {200,200,200} all pS = .35		
True Proportions Proficient meet AMO	.00002635	.00000258
True Proportions Proficient meet AMO - .05	.0715	.0226
Proficient Students configuration	{23,23,40}	{21,21,41}
School B 350 students with groups {200, 50,100} and pS = {.35,.35, .5}		
True Proportions Proficient meet AMO	.0001451	.00000621
True Proportions Proficient meet AMO - .05	.0548	.0072
Proficient Students configuration	{23, 3,19}	{21, 1,20}
	AMO = .45	
	k = 2.33	k = 2.58
School A 600 students with groups {200,200,200} all pS = .5		
True Proportions Proficient meet AMO	.0001156	.0000264
True Proportions Proficient meet AMO - .05	.0458	.0198
Proficient Students configuration	{74,74,93}	{72,72,94}

An even more extreme just-miss scenario for artificial School A would be the number proficient in the three subgroups {70, 70, 22}; two of the three subgroups blow past the performance goal of .19 proportion proficient in meeting their statewide proportion proficient .35, school proportion proficient is a strong .27, but the third subgroup is just one proficient student short of close enough under k = 2.33 99% confidence. But even here, the probability that all three subgroups have true proportions proficient of at least .19 on a single subject is only 1 part in three hundred, .00338.

It's left for a policy decision to judge whether these very small probabilities represent a prudent guard against falsely "failing" a school or depict extreme overkill. If indeed the goal of these margin of error procedures advertising 99% confidence is to control the probability that true probabilities meet the performance goal under a worst case scenario of a "just miss", then perhaps k = 1.25 would appear appropriate. For these school examples, just-make/just-miss scenarios using k = 1.25 will result in probabilities for a single subject that all three subgroups have true proportions proficient of at least .19 around one or two parts in a hundred.

Issue #5
What is a reasonable minimum group size?

There are many additional calculations not taken up here that would more directly bear on this frequent question. Clearly, states using the 99% confidence procedures with a small AMO and claiming to include very small groups (e.g. n=10) are being silly and deceptive, because even sizable groups would be given a free ride by the margin of error. From Table 1 the free ride consequences of the 99% confidence interval with AMO .19 would indicate a minimum group size of 40 or 50. Reducing k reduces the amount of free ride, such that reducing k to .5 eliminates the free ride for most reasonable AMO and pS values.

Comparing results for the n=30, n=50, n=200 examples in this note gives at least some initial impression on group size questions. Consider the performance goal .45 examples from Figure 1 and Tables F (n=30), C (n=50) and D (n=200). The summary in Table 2 shows that the properties of the confidence interval (margin of error) procedures are actually less attractive for the larger groups than the smallest. The probability that a subgroup satisfying the NCLB margin of error adjustment actually deserved it decreases with group size. Table 3 shows this effect of group size for values of decreasing values of k: {2.58, 1.25, .5, .0}. Thus Tables 2 and 3 illustrates yet another deficiency of the margin of error strategies--the form of the confidence interval adjustment does not function well in taking group size into account. And this effect merits notice because a major motivation for the confidence interval procedures expressed in the CCSSO report (see CCSSO, 2002, Ch.3 "Sample Cell Size Issues") is the effect of group size. Note that the relative disparity for larger groups seen in Table 3 diminishes as k is reduced (i.e. as the confidence interval fudge factor disappears).

Concluding Comments

A plan or procedure is properly judged by its consequences (here statistical properties), not by its good intentions. The intention in these NCLB plans to take into account uncertainty in the proportion proficient scores is reasonable; the mechanism developed (margin of error, confidence intervals) by CCSSO and friends is ludicrous because it leads to ludicrous results. It is ironic that the CCSSO report proclaims great concern about "public confidence in educational accountability" (e.g., Executive Summary p.10). Yet the state NCLB plans based on these CCSSO procedures have already served to create press and public derision towards NCLB. With helpers like these, NCLB does not need opponents to fail.

Accountability is not a bad thing, but it can be done badly. And a bad result is assured when the advice to states (from CCSSO, Center for Assessment) consists of uninformed statistical direction from non-statisticians. Moreover, it is especially distressing for the CCSSO report to describe their immensely flawed procedures as "Statistically-Based Approaches". The procedures put forth in CCSSO (2002) and Marion and Gong (2003) should not be confused with good statistical practice.

It is clear that states should not be allowed to pursue these confidence interval scams. Certification of these schemes as "reliable and valid" under the NCLB statute should be revoked.

On the other hand, NCLB AYP requirements are so poorly thought out and overreach so egregiously that it is unclear how a defensible plan should be constructed. Many more relevant calculations are needed to guide useful policy. California and Florida at least have clean plans, but those are not without severe difficulties (see Rogosa, 2003 for some properties of the California NCLB plan). One can applaud with unending vigor the statements and sentiments of Education Secretary Paige "Only if we hold schools and school districts accountable for the improved achievement of all students will we meet the goal of leaving no child behind," while decrying the details of NCLB.

More broadly, the NCLB tragedy is that a state like California is forced to replace a functioning and defensible accountability system with NCLB, which is neither (c.f., Washington Post 1/2/03 for similar sentiments on Kentucky and North Carolina). The lesson from previous work with the California API and the associated award programs is that statistical variability in the school and subgroup scores makes growth targets far more formidable than they might appear, in large part because of the subgroup requirements (as each of the subgroups has larger uncertainty than the school index). In the API award context, to have high probability that school and all subgroup scores meet the improvement criteria requires underlying improvement that far exceeds (blows through) the seemingly modest growth targets (Rogosa, 2002a, 2002b, 2002c). Therefore, a useful accountability strategy is to set modest improvement goals in order for successful schools and their subgroups to have high probability of meeting these conjunctive standards (c.f., the herding cats metaphor in Rogosa, 2002a,b,c API reports). Sadly, the federal mandates of NCLB ignore these important lessons.

Appendix: Technical Formulation and Notes

The intent of this Appendix is to make more explicit the various calculations presented in the body of this paper. As part of the purpose of this work is to encourage, by example, others to apply these calculations to other proposed NCLB plans, the value of this presentation lies in good part in its broader use. At the same time, it's important to make clear the modest scope--nothing hard was done in any of this work--and the limitations of the formulation and calculations.

CCSSO, NCLB Use of Statistical Inference.

The appeal to basic statistical theory by CCSSO and Center for Assessment is to speak in terms of inferences for a true proportion (or percent) proficient, here denoted by π (e.g., CCSSO, 2002, pp. 65, 66, 81) and to recall the elementary facts for large sample normal approximations for the distribution of an observed proportion. The observed proportion proficient for a group is $p = x/n$, where x is the observed number of proficient students and n is the size of the group. The large-sample Normal theory approximation is to say that the sampling distribution of the observed proportion proficient follows $p \sim N[\pi, \pi(1-\pi)/n]$. Then x is regarded as a *close enough* outcome if a confidence interval centered on x/n includes the AMO or if x/n is not in the rejection region from a hypothesis test (typically one-sided) for a null hypothesis that π = AMO.

Whether the inferences for π be based on confidence interval approximations or versions of hypothesis testing leads to the various (but quite similar) forms for the adjusted proportion proficient in the state NCLB plans. The problem with the CCSSO report and state plans is not whether π is a good quantity to know about, rather the issue is the proper use of statistical ideas and methodology.

Beta-Binomial Model and Results for $\pi|x$

The observed number of proficient students, x, in a school or subgroup of size n is assumed to be binomial with parameter π written as *Bin*$[n, \pi]$. Both π and n are linked to the specific school and group and could be subscripted accordingly. (The use of binomial here is an oversimplification and slight understatement of the statistical variability due to measurement error and sampling of students, but keeps this formulation as close as possible to the CCSSO treatment in order to highlight where the consequential failings lie.) In addition, the true proportion proficient π, whether it represents a school attribute or a subgroup within a school, has a distribution over the schools in a state. For convenience, take π to have a beta distribution, written as $B(\alpha, \beta)$. In the calculations, the parameters of the beta distribution for a specific group or subgroup are chosen to correspond to the specified state-wide mean, pS,; for that group or subgroup, $pS = [\alpha/(\alpha + \beta)]$. Values of pS .35 and .5 are used for the artificial school and subgroup examples.

Our knowledge about π is provided by the conditional distribution $\pi|X=x$. Based on the data (number of proficient students) what can be said about the object of inference, the unobserved "true" proportion proficient π?

The standard result can be found in Lehmann and Casella (1998, section 4.1). For π having a beta distribution (see Figure A1 below) written as $B(\alpha, \beta)$ and the observed number of proficient students, x, in a group of size n having binomial distribution *Bin*$[n, \pi]$ then the conditional density of π given $X=x$ is $B[\alpha + x, \beta + n - x]$. This distribution is a combination of the prior (group information) and the data such that the mean of the distribution of $\pi|X=x$ can be written in the familiar form, as the weighted combination of the mean of the statewide information and the observed proportion proficient:

$$(\alpha + x)/(\alpha + \beta + n) = [(\alpha + \beta)/(\alpha + \beta + n)]\cdot[\alpha/(\alpha + \beta)] + [1 - (\alpha + \beta)/(\alpha + \beta + n)]\cdot[x/n]$$

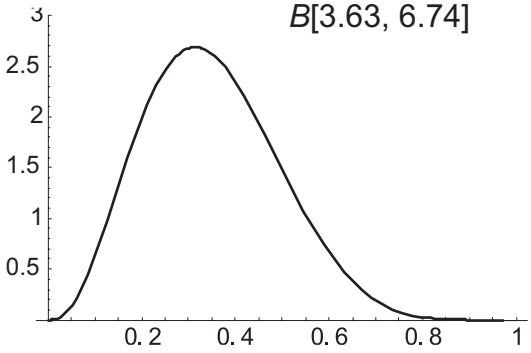
Probability Calculations for State NCLB Plans

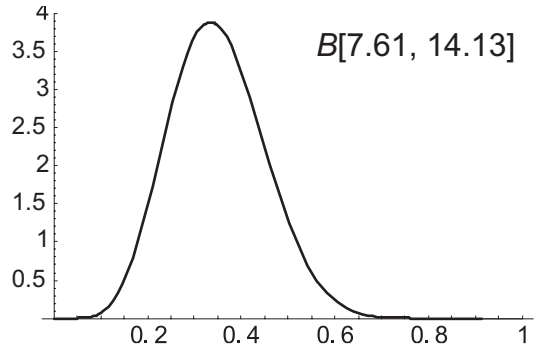
Using the conditional distribution $\pi|X=x$, the probability that π meets or exceeds the AMO can be computed for any specified level of x. The state NCLB plans serve to identify interesting values of x, such as x values so that x/n is deemed just close enough to the AMO (or alternatively x-values that result in a just-miss). Thus the details and variations amongst these state plans are not that consequential for the results here, as the variations in the plans (for a stated "confidence" level) don't much affect the values of interesting choices of x.

The other component of the calculations is the state-wide distribution of π. In the calculations two values for the state-wide proportion proficient are used: pS = .35 and pS = .5. Below in Figure A1 two forms of the beta density with mean .35 and with mean .50 are shown. In the calculations the upper two densities were used: $B[3.63, 6.74]$ for pS = .35 and $B[3, 3]$ for pS = .5; an alternative more peaked density is shown below each for reference. The lower densiities have variance about one-half as large. When pS is a good deal greater than the AMO (as in AMO = .19, pS = .35) the peakedness of the chosen beta distribution will be consequential, so calculations that incorporate realistic shapes are more useful. For example, even with n=200, the value of the probability that $\pi \geq .19$ given X = 23 is .006 using $B[3.63, 6.74]$ (see Table 2, k=2.33 entry), but substituting $B[7.61, 14.13]$ increases that probability to .018. The effect is much less for pS = .5, AMO = .45 calculations; with n=50, the value of the probability that $\pi \geq .19$ given X = 15 is .0234 using $B[3, 3]$ (see Table 2, k=2.33 entry), but substituting $B[6, 6]$ increases that probability to .0356.

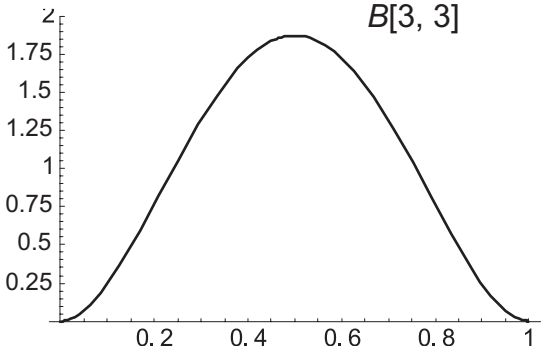
Figure A1. Choice of densities for true proportion proficient (priors).

Priors corresponding to pS = .35





Priors corresponding to pS = .5



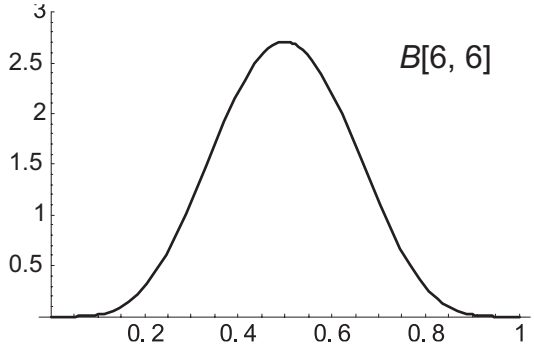


Table A
Posterior probability for "true" proportion proficient for observed number and proportion proficient and for indicated "margin of error" confidence interval kludges:
n = 50, statewide proportion .35, performance goal .19

observed number proficient	observed proportion proficient	probability true proportion proficient at least .19	
0.	0.	0.001	
			------> 1/50 + 2.58*Sqrt[.35*.65/50] = .194
1.	0.02	0.005	
			------> 2/50 + 2.33*Sqrt[.35*.65/50] = .197
2.	0.04	0.014	
3.	0.06	0.037	
4.	0.08	0.079	
5.	0.1	0.147	
			------> 6/50 + 1.25*Sqrt[.35*.65/50] = .204
6.	0.12	0.244	
7.	0.14	0.363	
			------> 8/50 + .5*Sqrt[.35*.65/50] = .194
8.	0.16	0.494	
9.	0.18	0.623	
10.	0.2	0.737	
11.	0.22	0.828	
12.	0.24	0.896	
13.	0.26	0.941	
14.	0.28	0.969	
15.	0.3	0.985	
16.	0.32	0.993	
17.	0.34	0.997	
18.	0.36	0.999	
19.	0.38	1.	
20.	0.4	1.	

Table B
Posterior probability for "true" proportion proficient for observed number and proportion proficient and for indicated "margin of error" confidence interval kludges:
n = 200, statewide proportion .35, performance goal .19

observed number proficient	observed proportion proficient	probability true proportion proficient at least .19	
0.	0.	0.	
1.	0.005	0.	
2.	0.01	0.	
3.	0.015	0.	
4.	0.02	0.	
5.	0.025	0.	
6.	0.03	0.	
7.	0.035	0.	
8.	0.04	0.	
9.	0.045	0.	
10.	0.05	0.	
11.	0.055	0.	
12.	0.06	0.	
13.	0.065	0.	
14.	0.07	0.	
15.	0.075	0.	
16.	0.08	0.	
17.	0.085	0.	
18.	0.09	0.	
19.	0.095	0.	
20.	0.1	0.001	
---------->			21/200 + 2.58*Sqrt[.35*.65/200] = .1
21.	0.105	0.002	
22.	0.11	0.003	
---------->			23/200 + 2.33*Sqrt[.35*.65/200] = .1
23.	0.115	0.006	
24.	0.12	0.01	
25.	0.125	0.017	
26.	0.13	0.027	
27.	0.135	0.041	
28.	0.14	0.06	
29.	0.145	0.086	
---------->			30/200 + 1.25*Sqrt[.35*.65/200] = .1
30.	0.15	0.119	
31.	0.155	0.16	
32.	0.16	0.208	
33.	0.165	0.264	
34.	0.17	0.326	
---------->			35/50 + .5*Sqrt[.35*.65/50] = .192
35.	0.175	0.392	
36.	0.18	0.462	
37.	0.185	0.532	

38.	0.19	0.601
39.	0.195	0.666
40.	0.2	0.726
41.	0.205	0.78
42.	0.21	0.828
43.	0.215	0.868
44.	0.22	0.9
45.	0.225	0.927
46.	0.23	0.947
47.	0.235	0.963
48.	0.24	0.975
49.	0.245	0.983
50.	0.25	0.989

=-

Table C
Posterior probability for "true" proportion proficient for observed number and proportion proficient and for indicated "margin of error" confidence interval kludges:
n = 50, statewide proportion .5, performance goal .45

observed number proficient	observed proportion proficient	probability true proportion proficient at least .19
0.	0.	0.
1.	0.02	0.
2.	0.04	0.
3.	0.06	0.
4.	0.08	0.
5.	0.1	0.
6.	0.12	0.
7.	0.14	0.
8.	0.16	0.
9.	0.18	0.
10.	0.2	0.
11.	0.22	0.001
12.	0.24	0.002
13.	0.26	0.005
----------------------------------> 14/50 + 2.58*Sqrt[.5*.5/50] = .462		
14.	0.28	0.012
----------------------------------> 15/50 + 2.33*Sqrt[.5*.5/50] = .465		
15.	0.3	0.023
16.	0.32	0.044
17.	0.34	0.076
18.	0.36	0.124
----------------------------------> 19/50 + 1.25*Sqrt[.5*.5/50] = .468		
19.	0.38	0.19
20.	0.4	0.272
----------------------------------> 21/50 + .5*Sqrt[.5*.5/50] = .455		
21.	0.42	0.369
22.	0.44	0.475
23.	0.46	0.582
24.	0.48	0.683
25.	0.5	0.772
26.	0.52	0.845
27.	0.54	0.901
28.	0.56	0.94
29.	0.58	0.966
30.	0.6	0.982

=-

Table D
Posterior probability for "true" proportion proficient for observed number and proportion proficient and for indicated "margin of error" confidence interval kludges:
n = 200, statewide proportion .5, performance goal .45

observed number proficient	observed proportion proficient	probability true proportion proficient at least .19
1.	0.005	0.
10.	0.05	0.
20.	0.1	0.
30.	0.15	0.
40.	0.2	0.
50.	0.25	0.
60.	0.3	0.
65.	0.325	0.
66.	0.33	0.
67.	0.335	0.001
68.	0.34	0.001
69.	0.345	0.002
70.	0.35	0.003
71.	0.355	0.004
----------> 72/200 + 2.58*Sqrt[.5*.5/200] = .451		
72.	0.36	0.006
73.	0.365	0.009
----------> 74/200 + 2.33*Sqrt[.5*.5/200] = .452		
74.	0.37	0.013
75.	0.375	0.019
76.	0.38	0.026
77.	0.385	0.036
78.	0.39	0.049
79.	0.395	0.065
80.	0.4	0.085
81.	0.405	0.109
----------> 82/200 + 1.25*Sqrt[.5*.5/50] = .454		
82.	0.41	0.138
83.	0.415	0.172
84.	0.42	0.21
85.	0.425	0.253
86.	0.43	0.3
----------> 87/200 + .5*Sqrt[.5*.5/50] = .453		
87.	0.435	0.351
88.	0.44	0.404
89.	0.445	0.459
90.	0.45	0.515
91.	0.455	0.571
92.	0.46	0.625
93.	0.465	0.677
94.	0.47	0.725
95.	0.475	0.77

96.	0.48	0.81
97.	0.485	0.846
98.	0.49	0.876
99.	0.495	0.903
100.	0.5	0.925
101.	0.505	0.943
102.	0.51	0.957
103.	0.515	0.968
104.	0.52	0.977
105.	0.525	0.984
106.	0.53	0.989
107.	0.535	0.992
108.	0.54	0.995
109.	0.545	0.996
110.	0.55	0.998

=-

Table E
Posterior probability for "true" proportion proficient for observed number and proportion proficient and for indicated "margin of error" confidence interval kludges:
n = 30, statewide proportion .35, performance goal .19

observed number proficient	observed proportion proficient	probability true proportion proficient at least .19
----------------------------------> 0/30 + 2.58*Sqrt[.35*.65/30] = .225		
----------------------------------> 0/30 + 2.33*Sqrt[.35*.65/30] = .203		
0.	0.	0.029
1.	0.033	0.079
2.	0.067	0.17
----------------------------------> 3/30 + 1.25*Sqrt[.35*.65/30] = .209		
3.	0.1	0.303
4.	0.133	0.46
----------------------------------> 5/30 + .5*Sqrt[.35*.65/30] = .21		
5.	0.167	0.619
6.	0.2	0.756
7.	0.233	0.858
8.	0.267	0.926
9.	0.3	0.965
10.	0.333	0.985
11.	0.367	0.994
12.	0.4	0.998
13.	0.433	0.999
14.	0.467	1.
15.	0.5	1.
16.	0.533	1.
17.	0.567	1.
18.	0.6	1.
19.	0.633	1.
20.	0.667	1.

=-

Table F
Posterior probability for "true" proportion proficient for observed number and proportion proficient and for indicated "margin of error" confidence interval kludges:
n = 30, statewide proportion .5, performance goal .45

observed number proficient	observed proportion proficient	probability true proportion proficient at least .45
0.	0.	0.
1.	0.033	0.
2.	0.067	0.
3.	0.1	0.
4.	0.133	0.001
5.	0.167	0.002
6.	0.2	0.006
----------> 7/30 + 2.58*Sqrt[.5*.5/30] = .469		
7.	0.233	0.015
----------> 8/30 + 2.33*Sqrt[.5*.5/30] = .479		
8.	0.267	0.035
9.	0.3	0.073
10.	0.333	0.134
----------> 11/30 + 1.25*Sqrt[.5*.5/30] = .481		
11.	0.367	0.223
12.	0.4	0.338
----------> 13/30 + .5*Sqrt[.5*.5/30] = .479		
13.	0.433	0.469
14.	0.467	0.602
15.	0.5	0.725
16.	0.533	0.825
17.	0.567	0.898
18.	0.6	0.946
19.	0.633	0.974
20.	0.667	0.989
21.	0.7	0.996
22.	0.733	0.999

References

Reference Note.
Specific State NCLB Plans (Accountability Workbooks) are available from http://www.ed.gov/admins/lead/account/stateplans03/index.html

Agresti, A. (1996). An Introduction to categorical data analysis. Wiley-Interscience.

Agresti, A. (1990). Categorical data analysis. Wiley-Interscience.

Chicago Tribune, September 28, 2003 "Schools Toying with Test Results: Some States Meet Standards with Art of Statistics", D. Rado and D. Little).

Council Of Chief State School Officers (2002). Making Valid And Reliable Decisions In Determining Adequate Yearly Progress. A Paper In The Series: Implementing The State Accountability System Requirements Under The No Child Left Behind Act Of 2001. ASR-CAS Joint Study Group on Adequate Yearly Progress, Scott Marion and Carole White, Co-Chairs.
available at
http://www.ccsso.org/content/pdfs/AYPpaper.pdf
("confidence interval approach" in Chapter 3)
Executive summary available at
http://www.ccsso.org/content/pdfs/AYPpapersummary.pdf

Glass, G. V. & Hopkins, K. D. (1996). Statistical Methods in Education and Psychology. Third Edition. Boston: Allyn & Bacon.

Indiana Department of Education (2003). School Accountability in Indiana: Public Law 221-1999 and the No Child Left Behind Act of 2001. Indiana Department of Education 8/14/2003
available at
http://ideanet.doe.state.in.us/esea/pdf/Accountability081403.pdf

Lehman, E. L. and Casella, G. (1998). Theory of point estimation, Second Edition. New York: Springer-Verlag.

Marion, S and Gong, B. (2003). Evaluating the Validity of State Accountability Sytems. Center for Assessment Presentation at CCSSO, St. Louis MO, September 11, 2003. Slides available at
http://www.ccsso.org/content/pdfs/CCSSO_Validity_BGSM03.pdf

Redelman, D. (2003). Confidence game in the Hoosier State. The Education Gadfly: A Weekly Bulletin of News and Analysis from the Thomas B. Fordham Foundation August 7, 2003, Volume 3, Number 28.
available from
http://www.edexcellence.net/foundation/gadfly/issue.cfm?id=111#1392

Rogosa, D.R. (2002a). Plan and Preview for API Accuracy Reports. California Department of Education, Policy and Evaluation Division July 2002.
available from http://www.cde.ca.gov/psaa/apiresearch.htm

Rogosa, D.R. (2002b). Commentaries on the Orange County Register Series: What's the Magnitude of False Positives in GPA Award Programs? and Application of OCR "margin of error" to API Award Programs. California Department of Education, Policy and Evaluation Division. September 2002.
available from http://www.cde.ca.gov/psaa/apiresearch.htm

Rogosa, D.R.. (2002c). Irrelevance of Reliability Coefficients to Accountability Systems: Statistical Disconnect in Kane-Staiger "Volatility in School Test Scores" CRESST deliverable, October 2002.
available from: http://www-stat.stanford.edu/~rag/api/kscresst.pdf

Rogosa, D.R.. (2003). California's AMOs Are More Formidable Than They Appear. California Department of Education, Policy and Evaluation Division. October 2003.
available from http://www.cde.ca.gov/psaa/apiresearch.htm

Washington Post, Thursday, January 2, 2003; Page A01. States Worry New Law Sets Schools Up to Fail, By Michael A. Fletcher.