The NCLB "99% confidence" scam: Utah-style calculations

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

Prepared for the use of The Salt Lake Tribune.
see
Federal mandate: A complex statistical formula allowed them to meet their targets, by Ronnie Lynn, in *The Salt Lake Tribune*, 12-18-03

This research was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences, U.S. Department of Education.

The NCLB "99% confidence" scam: Utah-style calculations

David Rogosa
Stanford University
November 15, 2003
[prepared for the use of Salt Lake Tribune]

Utah, like a number of states, has adopted a NCLB implementation based on what is often called a "confidence interval" adjustment to the observed proportion of proficient students. From the Utah Accountability Workbook:

> In calculating AYP for LEAs, schools, and subgroups, Utah will employ a test of statistical significance with a one-tailed alpha of 0.01. This will allow schools with small subgroup populations to be held accountable without falsely identifying a school. This creates a balance between validity (holding schools accountable for all students) and reliability (assuring that those subgroups identified have not been so identified simply on the basis of random fluctuation of scores). For AYP determination based on the annual measurable objective, a test of statistical significance will be applied for subgroups [of at least] 10. The null hypothesis is that the observed percent of students proficient in any subgroup is equal to the required percent proficient defined by the annual measurable objective. The test of statistical significance is a z-score with the distribution of school mean scores (in terms of percent correct) around the null hypothesis. A school or LEA makes AYP if the null hypothesis is not rejected.

From p.20, Utah State Office Of Education,
State of Utah Consolidated State Application Accountability Workbook,
Plan Approved by U.S. Department of Education June 10, 2003

The genesis of these NCLB 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." In that report these procedures, whether they be based on confidence intervals or hypothesis tests, are unfortunately described as "Statistically-Based Approaches". Regrettably, these well-meaning attempts to deal with statistical uncertainty in the school and subgroup scores are based on serious misunderstandings of material from introductory statistics courses. 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.

The present note describes the often ludicrous results of these NCLB implementations, using calculations based on some of the summary data from Utah testing and on the Utah NCLB procedures as described in the Accountability Workbook. The data used in the construction of the examples was provided by reporters from The Salt Lake Tribune, for whose use these examples were originally prepared. Utah is less diverse and has smaller schools than many of the other NCLB Confidence Interval states. And it would be useful to repeat these calculations for settings representing additional states. But the message that these NCLB 99% confidence interval plans are indefensible will pertain in other settings, as larger and more diverse schools will produce even more extreme results than seen in these Utah-style examples. Note that no attention is given here to the additional set of safe harbor provisions and contingencies in the NCLB plan, as these additional factors would only serve to exacerbate the message of the probability calculations--that schools are granted remarkable, perhaps fanciful, benefit of the doubt by virtue of faulty logic couched in statistical terminology.

I do 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. A companion report (Rogosa, current draft November 2003, How the Confidence Interval Procedures Destroy the Credibility of State NCLB Plans) provides additional generic example of the probability calculations and more extensive discussion of the statistical issues.

1. Utah procedures.

For Utah grades 3-8, the following goals for proportion proficient students are set: for Language Arts, proportion proficient .65 and for mathematics, proportion proficient .57. In NCLB-speak the performance goal has the designation of Annual Measurable Objective or AMO. For a group of 100 students, the AMOs would seem to require that at least 65 of the 100 students be proficient in language arts and at least 57 of the 100 students be proficient in mathematics. The "99% confidence" procedures serve to reduce the required numbers of proficient students that are deemed as "close enough". For example, with a group of 25 students the mathematics AMO of .57 would seem to require 15 proficient students out of the 25. But the one-sided hypothesis test described in the Utah NCLB plan would deem only 8 proficient students as representing close enough.

Table 1 displays the number of proficient students needed to meet the stated AMO (the n*AMO column) and the number of proficient students the Utah procedure would deem as close enough (minimum number proficient) for group sizes from 10 to 500. (These minimum number proficient values were obtained from the binomial cdf, locating the smallest number of proficient students that did not fall in the rejection region of the statistical test, "one-tailed alpha of 0.01", specified in the Utah NCLB plan.)

Table 1
Tabulations of "Close Enough" for Utah 99% Confidence NCLB

	Language Arts (AMO = .65)		Mathematics (AMO = .57)	
group size	min number proficient	n*AMO	min number proficient	n*AMO
10	3	7	2	6
25	11	17	8	15
50	24	33	20	29
75	39	49	33	43
100	54	65	45	57
125	69	82	58	72
150	84	98	71	86
175	99	114	84	100
200	114	130	98	114
225	129	147	111	129
250	145	163	124	143
275	160	179	138	157
300	176	195	151	171
500	300	325	259	285

---

2. Consequences of the "Close Enough" Approach--Some probability calculations for "if it could be, it is"

The logical flaw in the Utah NCLB plan and all similar "margin of error" approaches is a misunderstanding of the failure to reject a statistical hypothesis. 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 using Utah-style scenarios. 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 Appendix). The calculations provide some quantification for the question, Is Utah's close enough really good enough?

School proportion proficient.

The first set of calculations examines the simplest situation: a school with no subgroups. This simple setting is useful in introducing the calculations; in Utah this scenario would pertain for a school with one subgroup constituting the entire school population. That is, the subgroup is the school, as in a small all-white school (with no disadvantaged or EL subgroups). However, the scenario of schools without additional subgroups is not widespread because Utah uses such a small minimum subgroup size, 10 students.

Table 2 shows results for language arts and mathematics for hypothetical schools of sizes: 50, 100, and 200 included students. The Utah statewide proportions proficient are denoted by pS, with values shown in Table 2. To explain the contents of the table, focus on the math row for a school with 100 students (middle column). The raw AMO for math of .57 would seem to require 57 proficient students out of the group of 100. But the "99% confidence" procedure deems 45 proficient students as satisfying the .57 AMO. Is 45 proficient students close enough? The answer provided in Table 2 is the value .0132 for the calculated probability that the true proportion proficient meets or exceeds .57, given the data of 45 proficient students out of 100. That is, the chances are slightly less than 1 in 75 (i.e. odds 75 to 1 against) that the true proportion proficient (think of it as the observed proportion stripped of its statistical uncertainty) in mathematics would meet the AMO, even though the Utah NCLB procedure would certify the AMO is satisfied. Now the probability .0132 is not zero, therefore it might be the case that the true proportion proficient meets the AMO, but following "if it might be, it is" would seem to be a poor basis for state or national educational policy.

Table 2
Probability school "true" proportion proficient meets Performance Goal (AMO) at minimum number proficient for Utah 99% confidence procedure

	Probability meets Performance Goal (minimum number proficient)		
	n=50	n=100	n=200
English/Lang Arts			
pS = .7374, AMO = .65	.0162	.0195	.0137
	(24)	(54)	(114)
Mathematics			
pS = .6729, AMO = .57	.0170	.0132	.0157
	(20)	(45)	(98)

The other entries of Table 2 are similar in magnitude to the math result for n=100. An additional aspect, even for a school without subgroups, is the NCLB AYP requirement that both the math and language arts AMO be satisfied. For the schools and numbers of proficient students in Table 2, a rough approximation for the joint probability of the true proportions proficient for both English and math meeting the respective AMOs is around .005, 1/200 (or a bit greater than probability of obtaining 8 heads in a row from tossing a fair coin). Yet the Utah plan would deem these numbers of proficient students as close enough. Do these very small probabilities represent too much benefit of the doubt?

Schools with Two Non-overlapping subgroups.

Proceeding to slightly less unrealistic configurations, the calculations in Table 3 are based on an artificial school composed of two subgroups, Caucasian and Hispanic Students. The Utah statewide proportions proficient for the two subgroups are denoted by pScauc and pShisp, with values shown in Table 3. Results for schools with 75, 150, and 300 included students are shown in Table 3. For each school size (column) and each subject (row), the numbers of proficient Caucasian and Hispanic students deemed by the Utah procedure to satisfy the AMO school-wide and for each of the two subgroups are shown in parens. Above those numbers is the main entry: the calculated probability, given the numbers of proficient students, that the true proportions proficient all meet or exceed the AMO (details on these numbers given below). For each of the two subjects separately, these probabilities are typically less than one chance in a thousand. Yet these numbers of proficient students satisfy the Utah plan implementation of NCLB. Moreover, in regard to the NCLB AYP requirement that both the math and language arts AMO be satisfied, the rough approximation for the joint probability of the true proportions proficient for both English and math meeting the respective AMOs for both subgroups and school is around 1 in 15,000 (just about probability of obtaining 14 heads in a row from tossing a fair coin).

Table 3
School with Two Non-overlapping subgroups:
Caucasian and Hispanic Students

	Probability meets Performance Goal (minimum number Caucasian, Hispanic proficient)		
	School with 75 students; 50 Caucasian, 25 Hispanic	School with 150 students; 100 Caucasian, 50 Hispanic	School with 300 students; 200 Caucasian, 100 Hispanic
English/Lang Arts: AMO = .65 pScauc = .78, pShisp = .46	.00104 {28, 11}	.00059 (60, 24)	.00084 {122, 54}
Math: AMO = .57 pScauc = .71, pShisp = .40	.00076 {25, 8}	.00063 (51, 20)	.00068 {106, 45}

Furthermore, examples with two non-overlapping subgroups showing less diversity (e.g., 85% white instead of 67% white) produce very similar results. If the Table 3 school with 150 students were instead composed of 125 Caucasian and 25 Hispanic students, a configuration of 73 Caucasian and 11 Hispanic proficient students would be good enough under the Utah plan to satisfy the language arts AMO for the school and the two subgroups. Then the calculated probability, given those numbers of proficient students, that the true proportions proficient all meet or exceed the AMO would be .000598, identical to the .000594 obtained for the subgroup sizes used in Table 3. For mathematics in this less diverse school, a configuration of 63 Caucasian and 8 Hispanic proficient students would be good enough under the Utah plan to satisfy the mathematics AMO for the school and the two subgroups. The calculated probability, given these numbers of proficient students, that the true proportions proficient all meet or exceed the math AMO would be .00031, half as large as the .00063 obtained for the configuration in Table 3. The same conclusion--configurations of proficient students that satisfy the Utah plan represent remarkable benefit of the doubt being awarded to the Utah schools--pertains for this additional example.

One interpretation for the small probabilities that true proportion proficient meets AMO is derived from the approximate fact that PGA golfers score a hole-in-one on par 3 holes about 1 in 2000 (probability .0005). The remarkable level of benefit of the doubt being accorded to schools having this configuration is akin to assuming that a tee shot will make the hole. Conceding an 18 inch putt may be good manners, but assuming that a 180 yard drive will make the hole, just because it might happen, is not rational. Yet this is the magnitude of the leeway (in each of math and English) being provided to Utah schools with similar 2 subgroup configurations.

Details on the entries in Table 3.

For completeness in exposition, the paragraphs below explain, for each of the school examples in Table 3, the specification for number of proficient students used in the calculations. Details on the calculation of probability that true proportion proficient meets the AMO are given in the Appendix.

First, consider the school with 75 included students, 50 Caucasian, 25 Hispanic. School-wide, according to Table 1, 39 proficient students in language arts and 33 in mathematics are sufficient to satisfy the AMOs under the Utah plan (whereas the actual AMOs would require 49 and 43). For language arts, 11 Hispanic and 24 Caucasian proficient students are close enough (to 17 and 33) under the Utah plan, but 11 + 24 is four students short of the 39 total specified for the school. Padding the Caucasian total by an additional 4 proficient students to reach 28, produces the listed {28,11} configuration that would be deemed in the Utah plan as meeting both subgroup and school AMO for language arts. (The 4 students are added to the Caucasian total because that subgroup has a much higher statewide proportion; 28/50 still is smaller than the language AMO and the statewide proportion for Caucasians). For mathematics, 8 Hispanic and 20 Caucasian proficient students are close enough (to 15 and 29) under the Utah plan, but 8 + 20 is five students short of the 33 total specified for the school. Padding the Caucasian total by an additional 5 proficient students to reach 25, produces the listed {25,8} configuration that would be deemed in the Utah plan as meeting both subgroup and school AMO for mathematics. (The 5 students are added to the Caucasian total because that subgroup has a much higher statewide proportion; 25/50 still is smaller than the math AMO and the statewide proportion for Caucasians).

Next, consider the Table 3 school with 150 included students, 100 Caucasian, 50 Hispanic. School-wide, according to Table 1, 84 proficient students in language arts and 71 in mathematics are sufficient to satisfy the AMOs under the Utah plan (whereas the actual AMOs would require 98 and 86). For language arts, 24 Hispanic and 54 Caucasian proficient students are close enough (to 33 and 65) under the Utah plan, but 24 + 54 is 6 students short of the 84 total specified for the school. Padding the Caucasian total by an additional 6 proficient students to reach 60, produces the listed {60,24} configuration that would be deemed in the Utah plan as meeting both subgroup and school AMO for language arts. (The 6 students are added to the Caucasian total because that subgroup has a much higher statewide proportion; 60/100 still is smaller than the language AMO and the statewide proportion for Caucasians). For mathematics, 20 Hispanic and 45 Caucasian proficient students are close enough (to 29 and 57) under the Utah plan, but 20 + 45 is six students short of the 71 total specified for the school as close enough. Padding the Caucasian total by an additional 6 proficient students to reach 51, produces the listed {51,20} configuration that would be deemed in the Utah plan as meeting both subgroup and school AMO for mathematics. (The 6 students are added to the Caucasian total because that subgroup has a much higher statewide proportion; 51/100 still is smaller than the math AMO and the statewide proportion for Caucasians).

Finally, consider the Table 3 school with 300 included students, 200 Caucasian, 100 Hispanic. School-wide, according to Table 1, 176 proficient students in language arts and 151 in mathematics are sufficient to satisfy the AMOs under the Utah plan (whereas the actual AMOs would require 195 and 171). For language arts, 54 Hispanic and 114 Caucasian proficient students are close enough (to 65 and 130) under the Utah plan, but 54 + 114 is 8 students short of the 176 total specified for the school. Padding the Caucasian total by an additional 8 proficient students to reach 122, produces the listed {122,54} configuration that would be deemed in the Utah plan as meeting both subgroup and school AMO for language arts. (The 8 students are added to the Caucasian total because that subgroup has a much higher statewide proportion; 122/200 still is smaller than the language AMO and the statewide proportion for Caucasians). For mathematics, 45 Hispanic and 98 Caucasian proficient students are close enough (to 57 and 114) under the Utah plan, but 45 + 98 is 8 students short of the 151 total specified for the school as close enough. Padding the Caucasian total by an additional 8 proficient students to reach 106, produces the listed {106,45} configuration that would be deemed in the Utah plan as meeting both subgroup and school AMO for mathematics. (The 8 students are added to the Caucasian total because that subgroup has a much higher statewide proportion; 106/200 still is smaller than the math AMO and the statewide proportion for Caucasians).

Schools with Three Non-overlapping subgroups.

A slight extension of the examples in Table 3 is to add a third subgroup, based on the Utah African-American data. This more diverse configuration (3 non-overlapping subgroups) will pertain to relatively few Utah schools; configurations represented in Table 2 are probably far more common. The calculations for Table 4 are based on an artificial school composed of three subgroups, Caucasian, Hispanic, and African-American students. The Utah statewide proportions proficient for the subgroups are denoted by pScauc, pShisp, and pSafam with the values shown in Table 4. Results for schools with 125, 250, and 500 included students are shown in Table 4. For each school size (column) and each subject (row), the numbers of proficient Caucasian, Hispanic, and African-American students deemed by the Utah procedure to satisfy the AMO school-wide and for each of the three subgroups are shown in parens. The number of proficient students needed to satisfy the Utah NCLB plan are obtained in the same manner as depicted in detail for the Table 3 examples; in particular, any additional proficient students needed to satisfy the school-wide number proficient are included in the highest-scoring group, Caucasians.

Above the listing of the minimum number of students is the main entry: the calculated probability, given the numbers of proficient students, that the true proportions proficient meet or exceed the AMO. For each of the two subjects separately, the probabilities that the true proportions proficient meet or exceed the AMO can be expressed in parts per million (i.e. probabilities for the n=250 schools are 8 and 9 millionths). Yet these same numbers of proficient students satisfy the Utah plan implementation of NCLB. Moreover, in regard to the NCLB AYP requirement that both the math and language arts AMO be satisfied, the rough approximation for the joint probability of the true proportions proficient for both English and math meeting the respective AMOs for both subgroups and school would yield a result of one-millionth or less (for comparison, the probability of obtaining 20 heads in a row from tossing a fair coin is about one-millionth).

Insert Table 4

Table 4
School with Three Non-overlapping subgroups:
Caucasian, Hispanic, and African-American Students

	Probability meets Performance Goal (minimum number Caucasian, Hispanic African-American proficient)		
	School with 125 students; 75 Caucasian, 25 Hispanic, 25 Afr-Amer	School with 250 students; 150 Caucasian, 50 Hispanic, 50 Afr-Amer	School with 500 students; 300 Caucasian, 100 Hispanic, 100 Afr-Amer
English/Lang Arts:			
AMO = .65	.000037	.0000084	.0000196
pScauc = .78,	{47,	(97,	{192,
pShisp = .46	11,	24,	54,
pSafam = .53	11}	24}	54}
Mathematics			
AMO = .57	.0000068	.0000089	.0000098
pScauc = .71,	(42,	(84,	(169,
pShisp = .40	8,	20,	45,
pSafam = .42	8}	20}	45}

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.

Utah NCLB Use of Statistical Inference.

Although the wording in the Utah Accountability Workbook is imprecise and often confused, the statistical procedure appears to be a form of inference for the true proportion proficient, here denoted by π . Tests of the null hypothesis that π = AMO against the one-sided alternative $\pi <$ AMO are employed with Type I error rate .01. 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. Then x is regarded as a *close enough* outcome if x/n is not in the rejection region for the hypothesis test. *Close enough* values in Table 1 and throughout are determined by the minimum value of x for which the cumulative distribution function (cdf) of the binomial distribution with n trials and parameter AMO, written as *Bin*[n, AMO], is at least .01. (I.e., determine the smallest x value in the non-rejection region of the one-sided hypothesis test.)

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, π] . Both π and n are linked to the specific school and group and could be subscripted accordingly. 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, in Tables 2-4 the parameters of the beta distribution for a specific group or subgroup are chosen to correspond to the specified state-wide mean proportion proficient, pS, for that group or subgroup, pS = $[\alpha/(\alpha + \beta)]$. (Statewide priors were specified a bit more diffuse than would match the AMO at the 20th percentile.)

Prior Distributions in Table 2-4 Calculations

	English/ Language Arts	Mathematics
All Students		
pS	.7374	.6729
prior	BetaDistribution[5, 1.781]	BetaDistribution[4, 1.944]
Caucasian		
pS	.78	.71
prior	BetaDistribution[5, 1.410]	BetaDistribution[4, 1.634]
Hispanic		
pS	.46	.40
prior	BetaDistribution[5, 5.87]	BetaDistribution[4, 6.]
African-American		
pS	.53	.42
prior	BetaDistribution[5, 4.434]	BetaDistribution[4, 5.524]

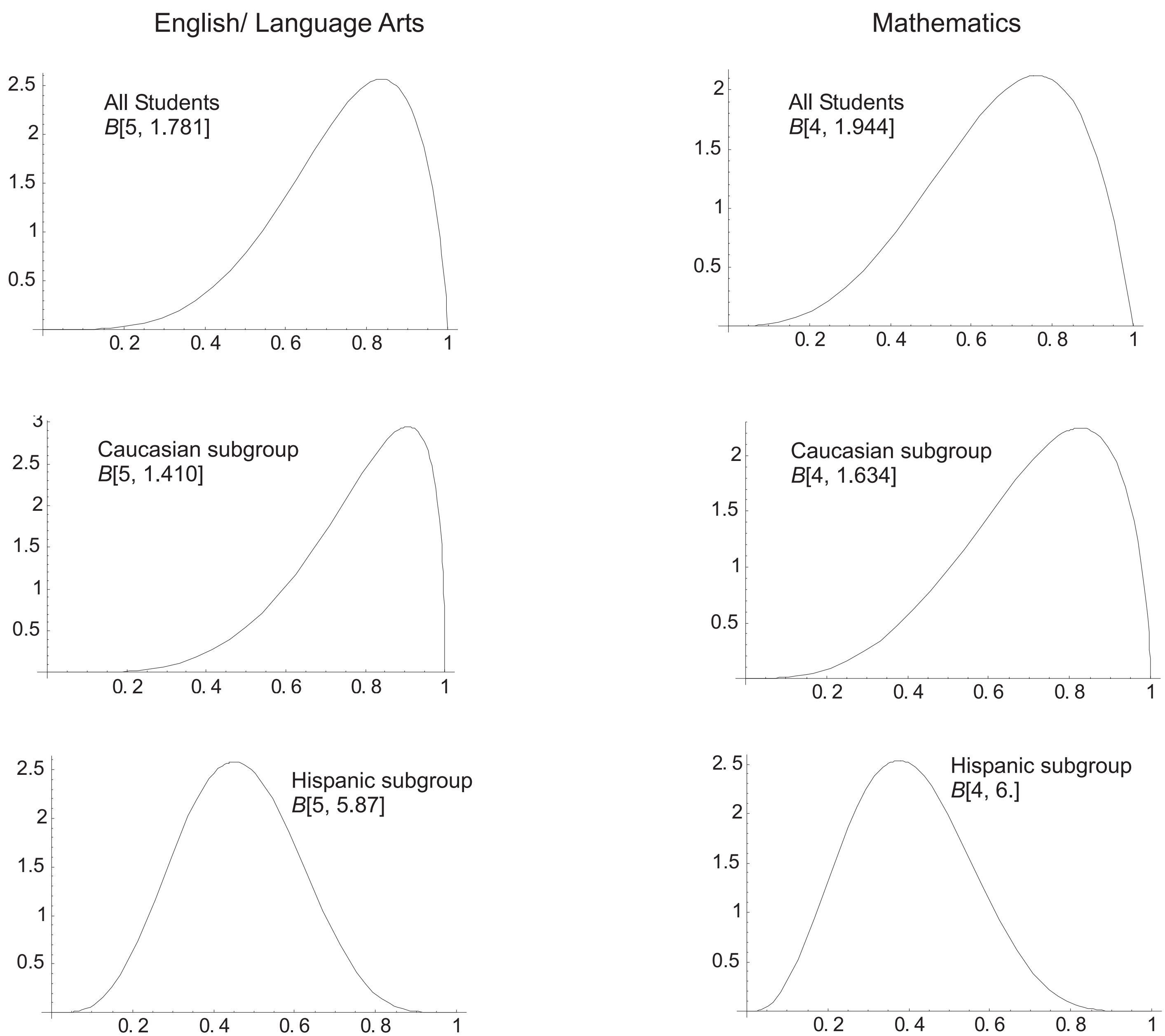
Our knowledge about π is provided by the conditional distribution $\pi|X=x$. Based on the data (number of proficient students) what can be said about the object of inference, the unobserved "true" proportion proficient π?

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: $(\alpha + x)/(\alpha + \beta + n) =$

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

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, 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). These probabilities comprise the entries in Tables 2-4.

Figure A1. Plots of densities for true proportion proficient (priors) for Tables 2 and 3.



References

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

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

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. (curent draft November 2003). How the Confidence Interval (margin of error) Procedures Destroy the Credibility of State NCLB Plans
available from
http://www-stat.stanford.edu/~rag/draft/mssham1107.pdf

Utah State Office Of Education (2003). State of Utah Consolidated State Application Accountability Workbook. Plan Approved by U.S. Department of Education June 10, 2003
Specific State NCLB Plans (Accountability Workbooks) are available from
http://www.ed.gov/admins/lead/account/stateplans03/index.html