California's AMOs Are More Formidable Than They Appear

David Rogosa
Stanford University
October 2003

This note attempts to illustrate some aspects of the effects of statistical uncertainty in NCLB accountability. NCLB requires that a school and all its eligible subgroups meet a specified performance goal, expressed in terms of proportion of students achieving the "proficient" designation. This performance goal has the designation of Annual Measurable Objective (AMO); in California for 2003-2005 the AMO values for grades 2-8 are set to .136 for English/Language Arts and to .16 for mathematics (scores obtained from the California Standards Tests, CST).

The AMO values for California are honest, unambiguous standards. In particular, unlike a majority of states, California does not apply a confidence interval (margin of error) fudge factor to the AMO values. (In separate work, the NCLB state plans that invoke this "confidence interval" adjustment are shown to be laughably indefensible as statistical procedures; similar sentiments have been expressed in many press reports, such as Chicago Tribune, September 28, 2003 "Schools Toying with Test Results: Some States Meet Standards with Art of Statistics", D. Rado and D. Little).

Statistical properties of school and subgroup scores are important in understanding accountability systems. 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).

The proportion of students proficient measure used in NCLB does indeed have some associated statistical uncertainty, commonly described as resulting from sampling variability in drawing the specific students tested and from measurement or classification error in the NCLB subject test instruments (here the CST). The underlying educational achievement (e.g., "true" proportion proficient) can be thought of as obscured by the statistical uncertainty. The purpose of the initial basic calculations presented below is to calibrate some of the consequences of the statistical variability.

Single subgroup example.
Turning to NCLB calculations, first consider a single subgroup. Table 1 below gives separate results for English and math for a subgroup ranging in size from 50 to 200 students. The columns (labeled "pass probability") indicate the specified probability (column values .9, .95, .99) that the

observed proportion proficient is at least as large at the AMO (English .136, math .16). Note that 99% confidence is often cited in state NCLB plans. The entries in each part of Table 1 show the required "true" proportion proficient (loosely speaking, the observed proportion proficient scores hypothetically stripped of the statistical uncertainty). That is, for a subgroup of size 50, the smallest value of true proportion proficient that provides probability .99 of meeting the English AMO of .136 is .267, a value almost double the AMO. A subgroup of size 100 whose real educational attainment in mathematics corresponds to proportion proficient .252 does have probability .99 of satisfying the math AMO of .16. Of course, the discrepancy between the AMO and the required true proportion proficient decreases as group size increases (a result of statistical variability in the proportion proficient score decreasing with larger n).

Table 1
True Proportion Proficient Required with a Single Group of Size n to Meet Subject Performance Goal (AMO) with Stated Pass Probability

English/Language Arts (AMO .136)				Mathematics (AMO .16)			
	"Pass Probability"				"Pass Probability"		
n	.90	.95	.99	n	.90	.95	.99
50	0.201	0.223	0.267	50	0.224	0.247	0.292
75	0.198	0.216	0.251	75	0.213	0.231	0.267
100	0.184	0.199	0.228	100	0.206	0.222	0.252
125	0.175	0.188	0.213	125	0.201	0.215	0.242
150	0.176	0.188	0.211	150	0.198	0.21	0.234
175	0.172	0.181	0.202	175	0.195	0.206	0.228
200	0.171	0.181	0.201	200	0.193	0.203	0.224

Satisfying Both English and Math AMO.
Because NCLB AYP requires meeting both English and math AMO, the Table 1 pass probabilities require some further discussion, as the values in Table 1 would overstate the probability of meeting both standards. That is, the probability that a subgroup of size 100 with true proportion proficient in math .222 and in English .199 meets both AMO criteria is less than .95. How much less? For two independent trials, the probability of success on both trials would be .81, .90, .98 for trials with probability of success .9, .95, .99. 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). Taking typical within-school correlations of English and math around .7 indicates the joint pass probabilities are approximately .85, .92, .98 for single trial probabilities .9, .95, .99 respectively. A version of Table 1 can be recomputed for joint pass probabilities set to .9, .95, .99, and the slight increase in the required true proportion proficient values is shown in Table 1A.

Table 1A
True Proportion Proficient Required with a Single Group of Size n to Meet Both Performance Goals (AMO) with Stated Joint Pass Probability

English/Language Arts (AMO .136)				Mathematics (AMO .16)			
	"Joint Pass Probability"				"Joint Pass Probability"		
n	.90	.95	.99	n	.90	.95	.99
50	0.216	0.238	0.281	50	0.239	0.262	0.306
75	0.21	0.227	0.262	75	0.225	0.243	0.278
100	0.194	0.208	0.237	100	0.216	0.232	0.261
125	0.184	0.196	0.221	125	0.211	0.224	0.25
150	0.184	0.196	0.218	150	0.206	0.218	0.242
175	0.177	0.188	0.209	175	0.203	0.214	0.235
200	0.176	0.188	0.207	200	0.2	0.21	0.23

School Composed of Three Non-overlapping Subgroups.
The single subgroup calculations are extended to the artificial case of a school composed of three subgroups. The simplest scenario is subgroups of equal size, so that a school of 300 students is composed of 3 subgroups each with 100 students. Table 2 (English) and Table 3 (math) present calculations for the three subgroups of equal size scenario. The entry in the tables is the school-wide true true proportion proficient needed to achieve the stated pass probability for that single subject (English or math). (Obvious additional complications that could be tacked on to these simple scenarios include varying the size of the subgroups.)

Tables 2 and 3 each have two frames. In the left frame all three subgroups are constrained to have the same true proportion proficient, so that the school-wide true proportion proficient also pertains to each of the three subgroups. In the right frame the three subgroups have "laddered" true proportion proficient--one subgroup has the school-wide value, while the two other subgroups have values .1 above and .1 below the school-wide value. For example, to obtain probability .99 of meeting the English AMO of .136 for school and subgroup scores, a school of size 150 students (i.e., a school composed of 3 subgroups each with n=50) requires a school wide true proportion proficient of .367 (nearly 3 times the AMO), configured as subgroup true proportion proficient values: .267, .367, .467.

Calculations for the joint probabilities of passing both English and math as were done in Table 1A can also be done for Tables 2 and 3. Table 1A gives reasonable guidance to the size of that effect. Take the n=50 laddered subgroups true proportion proficient entries. Values for single test pass probability .90 of .302 and .325 in Tables 2 and 3 increase to .316 and .34 in order to have (joint) probability .90 of meeting both English and math AMO for school and subgroup scores.

Table 2
School-level True Proportion Proficient Required for Three Subgroups each of Size n to Meet English Performance Goal .136 with Stated Pass Probability

Subgroups constrained to have same true proportion proficient

n	"Pass Probability" .90	.95	.99
50	0.234	0.254	0.293
75	0.224	0.24	0.271
100	0.206	0.219	0.245
125	0.194	0.206	0.229
150	0.194	0.204	0.225
175	0.186	0.196	0.215
200	0.186	0.195	0.213

Subgroups with laddered true proportion proficient

n	"Pass Probability" .90	.95	.99
50	0.302	0.324	0.367
75	0.298	0.316	0.351
100	0.284	0.299	0.328
125	0.275	0.288	0.313
150	0.274	0.288	0.311
175	0.271	0.281	0.302
200	0.271	0.281	0.301

note: the three equal sized subgroups have true proportion proficient p - .1, p, p + .1

Table 3
School-level True Proportion Proficient Required for Three Subgroups Each of Size n to Meet Mathematics Performance Goal .16 with Stated Pass Probability

Subgroups constrained to have same true proportion proficient

n	"Pass Probability" .90	.95	.99
50	0.258	0.278	0.319
75	0.24	0.256	0.288
100	0.229	0.242	0.27
125	0.222	0.233	0.257
150	0.216	0.227	0.248
175	0.212	0.222	0.241
200	0.209	0.218	0.236

Subgroups with laddered true proportion proficient

n	"Pass Probability" .90	.95	.99
50	0.325	0.347	0.392
75	0.313	0.331	0.367
100	0.306	0.322	0.352
125	0.301	0.315	0.342
150	0.298	0.31	0.334
175	0.295	0.306	0.328
200	0.293	0.303	0.324

note: the three equal sized subgroups have true proportion proficient p - .1, p, p + .1

These calculations represent one small glimpse of the statistical issues that arise in an accountability system such as NCLB. An important message which was featured in discussions of API award programs (see Rogosa 2002a,b) is that just having the standard error of the school index does not inform much about the properties of the accountability system.

The calculations in these tables do indicate that schools with what may appear to be reasonable educational performance may not satisfy NCLB. For example, consider a school with 300 students, configured as 3 subgroups of 100 students. If that school had underlying educational achievement corresponding to true proportion proficient .206 English and .229 math, school-wide and for each subgroup, then that school has probability of about .15 of failing to meet both AMO.

A small caution against one misreading of these examples. It may appear that smaller schools and groups are disadvantaged, as these appear to require larger levels of educational attainment to produce a given probability of meeting the AMO. As with most stories there are two sides. A small school with very low real educational attainment has statistical variability as its friend, in that a false positive result may occur more often than for a large school. But a small school that has relatively strong real attainment has statistical uncertainty as its foe, in that a false negative result is more likely than for a large school. The latter situation of false negatives is what's visible in these tables. For relevant discussion in the API context see Rogosa (2002a,b,c), especially Rogosa (2002c, section 4).

Technical Notes

1. Use of Binomial Distribution
The probability calculations using the binomial distribution represent an exceptionally crude approximation to the processes that generate these data. However, for the descriptive purposes here, that approximation does not serve too badly. As these simple methods make the point, simpler seems better in this instance. Moreover, even more crude large-sample normal theory approximations to binomial variability are at the core of the state NCLB plans that use the (unfortunate) confidence interval adjustments to the AMO (CCSSO, 2002). Put briefly, the binomial calculations do represent in a common form the statistical uncertainty resulting from the sampling of students from the school population. In addition, the observed proportion proficient is affected by measurement error in the CST. Harcourt (2002) indicates score reliabilities CST tests at or above .9. For these reliability values the effect of measurement error is not large. The expected value of the observed proportion proficient will slighty exceed the true proportion proficient (being pulled toward .5). The solutions for proportion proficient in Tables 1-3 could be regarded as solutions for the value of expected observed proportion proficient; decreasing those tabled values by approximately .01 would account for the possible additional distortion of measurement error.

2. Joint Probability Calculation.
Clearly the probability of passing both math and English is less than the single test probability (values .9, .95, .99 used in the tables) but greater than the probabilities resulting from two independent trials (.81, .9025, .9801). The intermediate values, .8474, .9199, .9827 were obtained from assuming a bivariate normal distribution for math and English with correlation .71 (1/Sqrt[2]). A reasonable approximation to the joint proability is p^1.5, the geometric mean of p and p^2, (where p is the probability for a single trial). The calculations in Table 1A used the reverse calculation with this bivariate normal distribution; values for a single test of {.9366, .9696, .9943} were used to yield joint probabilities {.9, .95, .99}. For serious discussion of this kind of joint probability calculation, see Kotz et. al. (2000, section 46.4).

For completeness, below are code snippets for the basic functions used to generate the tabled values.

```
table1[amo_] :=
Table[pi /.
  FindRoot[(
    1 - CDF[BinomialDistribution[
            n, pi], Floor[(amo-.001)*n]]) == conflist[[ii]],
            {pi, {.2, .4}}] , {n, 50, 200, 25}, {ii, 1, 3}]
conflist = {0.9, 0.95, 0.99}
```

```
table1A[amo_] :=
Table[pi /.
  FindRoot[(
    1 - CDF[BinomialDistribution[
            n, pi], Floor[(amo-.001)*n]]) == conflistduo[[ii]],
            {pi, {.2, .4}}] , {n, 50, 200, 25}, {ii, 1, 3}]
(r={{1,1/Sqrt[2]},{1/Sqrt[2],1}};
  ndist=MultinormalDistribution[{0,0},r])
conflistduo =
  Table[pduo /.
      FindRoot[CDF[
          ndist, {Quantile[zdist, pduo],Quantile[zdist, pduo]}]\[Equal]
          conflist[[ii]], {pduo, {.9, .98}}], {ii,1,3}]
   = {0.936612,0.969633,0.994315}

table23pvary[amo_] :=
  Table[pi /.
    FindRoot[(
    1 - CDF[BinomialDistribution[n, pi - .1], Floor[(amo-.001)*n]])*
            (1 - CDF[BinomialDistribution[n, pi], Floor[(amo-.001)*n]])*
            (1 - CDF[BinomialDistribution[n , pi + .1], Floor[(amo-.001)*
              n]]) \[Equal] conflist[[ii]], {pi,.3}] , {n, 50, 200,25},
              {ii,1,3}]

table23p[amo_] :=
  Table[pi /.
    FindRoot[(
    1 - CDF[BinomialDistribution[n, pi], Floor[(amo-.001)*n]])*
            (1 - CDF[BinomialDistribution[n, pi], Floor[(amo-.001)*n]])*
            (1 - CDF[BinomialDistribution[n , pi ], Floor[(amo-.001)*n]])
            \[Equal] conflist[[ii]], {pi,{.2,.4}}] ,
            {n, 50, 200,25}, {ii,1,3}]
```

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

Harcourt Educational Measurement. California STAR Technical Report, Fall 2002. December 2002.

Kotz, S., Balakrishnan, N., and Johnson, N. L. (2000). Continuous Multivariate Distributions, Second Edition. New York: Wiley-Interscience.

Linn, R. L., Baker, E. L., & Herman J. L. (2002, Fall). Minimum group size for measuring adequate yearly progress. The CRESST Line, 1, 4-5.

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

Appendix [added 11/15/03]
Impact of Minimum Group Size: Calculations on the Margin

For NCLB accountability, California has adopted a minimum group size of 50. (In API accountability the minimum group size had been 30, subject to additional contingencies.) Some states, in concert with the margin of error adjustments, have set minimum group sizes as low as 10 students. The question of how to set the minimum group size has been highly visible in NCLB discussions (see CCSSO 2002), but good sense has been absent. Attention has often focused on the standard error of a single proportion, and as expressed earlier, examination of just one component often provides little useful information on the properties of the accountability system. A particularly inexplicable analysis is provided by Linn et al (2002), in which the function Sqrt[2*p*(1 - p)] for p = .47 is tabled for n between 10, 100 (their table 1) leading to the recommendation: "In our judgment, a reasonable compromise between the competing goals of more disaggregated reporting and greater statistical reliability would be to set the minimum number of students at 25.(p.5)

The common sense approach to investigating minimum group size is to compare the properties of the accountability system as constituted with the properties resulting from setting a smaller minimum n. That's the logic of an "on the margin" calculation. If the minimum group size is 50, then only groups of 50 or more are included; if instead the minimum group size is lowered to 30, then, for example, a sub-group of size 30, previously not included, is now part of the accountability computation for that school. The brief treatment here shows some examples of that common sense approach, within the limited context of the cartoon-level analyses of this note.

The left frame of Table A1 repeats the 3 group example in Table 3 in order to compare those results with the addition of a fourth non-overlapping sub-group of size nS = 30 in the right frame. As would be expected, the required school-level true proportion proficient increases with the addition of the small nS = 30 group. (Note that this calculation assumes the added small group has the same true proportion proficient as the school- wide population; in many NCLB scenarios the small group may well have lower educational performance which would produce a much larger effect.) For calibration, if the fourth group were of size 50 instead of the 30 used in Table A1 the required increase in the school-level true proportion proficient effect would be about the midpoint of the entries in the two tables.

In particular, notice from Table A1 that the effect of adding the nS = 30 subgroup depends on the size of the other subgroups and on the stated pass probability. The effect is much smaller for n = 75 pass probability .9 (.24 vs .265 for true proportion proficient) than for n = 200 pass probability .99 (.236 vs .34 for true proportion proficient). These calculations illustrate the inadequacy of most investigations of minimum group size, which examine the small group in isolation. The statistical properties of an nS = 30 group, considered in isolation, are far from the full story. What value for understanding these accountability questions is knowing that

the standard error is about .08 for an observed proportion proficient of .25 in a subgroup of size 30? Instead, the effect on the properties of the accountability system resulting from the addition of that additional small subgroup is the relevant question.

Table A1
Marginal effects of adding a fourth subgroup of size 30 on School-level True Proportion Proficient Required for Three Subgroups Each of Size n to Meet Mathematics Performance Goal .16 with Stated Pass Probability

Three Subgroups constrained to have same true proportion proficient				Added fourth subgroup, nS = 30, all subgroups constrained to have same true proportion proficient			
	"Pass Probability"				"Pass Probability"		
n	.90	.95	.99	n	.90	.95	.99
50	0.258	0.278	0.319	50	0.278	0.301	0.35
75	0.24	0.256	0.288	75	0.265	0.288	0.342
100	0.229	0.242	0.27	100	0.258	0.283	0.34
125	0.222	0.233	0.257	125	0.254	0.281	0.34
150	0.216	0.227	0.248	150	0.252	0.28	0.34
175	0.212	0.222	0.241	175	0.251	0.28	0.34
200	0.209	0.218	0.236	200	0.25	0.28	0.34

One interpretation of these effects of including the smaller groups in the NCLB accountability is that is that in order to maintain the same school probability of not failing, everyone else in the school has to work harder (do better). For example, for n = 50 school-wide true proportion proficient .278 corresponds to pass probability .95 with the three subgroups, but the addition of the fourth group nS = 30 reduces the pass probability to .90. Also, for n = 75 school-wide true proportion proficient .288 corresponds to pass probability .99 with the three subgroups, but the addition of the fourth group nS = 30 reduces the pass probability to .95. More vividly, for n = 150 school-wide true proportion proficient .248 corresponds to pass probability .99 with the three subgroups, but the addition of the fourth group nS = 30 reduces the pass probability to less than .90. Small groups can be thought of as a "probability anchor" (in terms of the probability of meeting conjunctive AMO). From the perspective of the schools the question might be, How much drag can the system take and still move forward (as required by NCLB)?

When does the smaller group dominate?
On the other hand, there are configurations where the properties of the small group dominate--i.e., n large, nS very small. Table A2 shows for n=50 and n=200 separately the school-wide true proportion proficient for the four-group configuration, with the fourth sub-group of size nS (nS, 10,..., 50). With nS = 10 the size of the other groups has little consequence. For further comparison, Table A3 shows the calculation for the required true proportion proficient treating the single group of size nS in isolation. For any of the nS values with n = 200 or for

the smallest nS values with n = 50, the properties of the smallest group dominate.

Table A2
School-level True Proportion Proficient Required for Three Subgroups Each of Size n = 200 (left frame) or Size n = 50 (right frame) and a fourth subgroup of size nS to all Meet Mathematics Performance Goal .16 with Stated Pass Probability

	n = 200 "Pass Probability"				n = 50 "Pass Probability"		
nS	.90	.95	.99	nS	.90	.95	.99
10	0.337	0.394	0.504	10	0.34	0.395	0.504
15	0.317	0.363	0.453	15	0.323	0.365	0.453
20	0.304	0.344	0.421	20	0.312	0.347	0.421
25	0.249	0.282	0.349	25	0.279	0.303	0.356
30	0.25	0.28	0.34	30	0.278	0.301	0.35
35	0.25	0.277	0.333	35	0.277	0.299	0.345
40	0.249	0.275	0.327	40	0.276	0.297	0.34
45	0.249	0.273	0.321	45	0.275	0.295	0.337
50	0.23	0.249	0.292	50	0.266	0.286	0.325

Table A3
True Proportion Proficient Required for Single Subgroup of Size nS to Meet Mathematics Performance Goal .16 with Stated Pass Probability

	"Pass Probability"		
nS	.90	.95	.99
10	0.337	0.394	0.504
15	0.317	0.363	0.453
20	0.304	0.344	0.421
25	0.248	0.282	0.349
30	0.249	0.28	0.34
35	0.249	0.277	0.333
40	0.248	0.275	0.327
45	0.248	0.272	0.321
50	0.224	0.247	0.292