Assessing the Effects of Multiple Subgroups: A Rebuttal to the PACE Policy Brief December 2003 "Penalizing Diverse Schools? Similar test scores, but different students, bring federal sanctions" by John R. Novak and Bruce Fuller.

David Rogosa
Stanford University
January 5, 2004

Assessing the Effects of Multiple Subgroups: A Rebuttal to the
PACE Policy Brief December 2003 "Penalizing Diverse Schools?
Similar test scores, but different students, bring federal
sanctions" by John R. Novak and Bruce Fuller.

David Rogosa
Stanford University
January 2004

SUMMARY
Through a series of misunderstandings and mistakes,
the PACE report by Novak and Fuller badly overstates
the effects of subgroups (diversity) on the probability
of meeting the NCLB requirements. Examples of useful
probability calculations on this important question are
presented to correct the misinformation in the PACE
Policy Brief and to guide further investigations on
statistical properties of these accountability systems.

This rebuttal demonstrates that the empirical claims in the PACE Policy
Brief 03-4 December 2003 "Penalizing Diverse Schools? Similar test scores,
but different students, bring federal sanctions" by John R. Novak and Bruce
Fuller do not stand up to scrutiny. Unfortunately, the claims and
conclusions of this PACE report did receive widespread press attention,
with favorable articles spanning the San Francisco Chronicle to the New
York Times (both December 23, 2003).

The PACE report (available at pace.berkeley.edu/pace_publications.html)
asserts under "Key Findings" that
     schools serving diverse students in California are less likely
     to achieve their growth targets....
     even when students display almost identical average test
     scores schools with more subgroups are more likely to miss
     their growth targets under federal rules set by the No Child
     Left Behind Act. Schools serving middle-class children, for
     example, are 28 percent more likely to be labeled "needs
     improvement" by the feds when serving five student subgroups
     than schools serving one group. (PACE page 1)
Such claims are important to examine because, clearly, if statements like
those in the PACE report were credible, then the NCLB accountability system
would not be defensible.

The PACE Policy Brief does raise an absolutely legitimate question about
the statistical properties of California NCLB accountability; what are the
effects of subgroups? Novak and Fuller focus on questions of "fairness":
     Is it fair or motivating to label a school as failing simply
     because it serves more diverse students, not because it's [sic]
     overall achievement level is lower? (PACE Key findings page 1)

> Does the adequate yearly progress (AYP) strategy—with its
> emphasis on the performance of student subgroups—fairly label
> schools as failing, or are diverse schools penalized for
> serving a wider range of subgroups? (PACE page 3)

And the various data tabulations of NCLB outcomes in the PACE report are used to support a conclusion of "unfairness". For example, in their "Tale of Two Schools: Dinged for Diversity?" vignette, Novak and Fuller assert: "This seems unfair, simply an artifact of facing stiffer odds when serving more subgroups." (PACE p.5)

My attempt to explain the logical fallacy in presentations like that of Novak and Fuller is to appeal to the important distinction between "equality of results" and "equality of opportunity". The tabulated accountability results are empirical outcomes from complex confounded factors, as schools differ in many ways other than number of reported (overlapping) subgroups. (Granted, Novak and Fuller at least make a gesture to the complexity with some cross-classifications.) But disparity in results does not imply unfairness (unequal opportunity). In particular, the magnitudes of any observed disparity in results cannot be taken as estimates of unfairness (inequality of opportunity).

A very similar and equally flawed analysis of the effect of subgroups on the properties of a school accountability system is the presentation in Kane and Staiger (2002, esp. Table 4), using results from the California API award programs. Kane and Staiger assert as their Lesson 2: "Incentive systems establishing separate thresholds for each racial or ethnic subgroup present a disadvantage to racially integrated schools"(p. 258), and Kane-Staiger further state in their conclusion: "rules making any rewards contingent on improvement in each racial group present a great disadvantage to integrated schools and generate a number of perverse incentives that may harm rather than help minority students" (p.269). For refutation of these specific claims in Kane-Staiger see Rogosa (2002a, section 4) which also provides a catalog of the many blunders in the Kane-Staiger assertions of volatility in test scores (c.f., Rogosa 2003b).

A major purpose of this note is to present simple probability calculations which provide some quantification for questions about the effects of subgroups. All indications are that such effects are far far less than that claimed by the PACE report. Two sets of calculations use an "on the margin" logic--what is the effect on probability of award or probability of meeting the AMO if another subgroup is added? The first set of artificial binomial calculations for NCLB follow the development in Rogosa (2003a); the advantage of those rudimentary calculations is their accessibility to say a high school student with a knowledge of the binomial distribution and basic spreadsheet computing skills. The second set of calculations use actual California school test data to provide examples of the effects of subgroups on properties of the California GPA award program. What's important is style and logic of these calculations, rather than their details; additional examples with California data or from other states would be desirable.

Small Note #1. How far apart is "similar" achievement?

The full statement in the PACE Key Findings
    Schools serving middle-class children, for example, are 28
    percent more likely to be labeled "needs improvement" by the
    feds when serving five student subgroups than schools serving
    one group. This disparity exists even though average test
    scores are just five percentile points apart between schools.
Examination of PACE Table 2 "Similar Student Achievement Levels across
California Schools" shows these numbers are drawn from the collection of
schools with "less than 25%" students economically disadvantaged. For these
schools having 1 subgroup, 83% (of 616 schools) met AYP, whereas for these
schools with 5 subgroups, 55% (of 107) met AYP. Novak and Fuller draw
conclusions about the effects of subgroups by claiming these two sets of
schools do not differ much in their overall educational performance--
"average test scores are just five percentile points apart between
schools". In the PACE report text and in their Table 2 test scores are
expressed in the NCE (Normal Curve Equivalents) metric. Regrettably,
throughout their report, Novak and Fuller treat and refer to NCE and
percentiles as interchangeable--a serious mistake which turns out to be
consequential. The small table below shows the reported NCE scores and the
equivalents of those NCE score in percentiles (expressed here on a 0,1
scale).

	CAT/6 NCE from PACE Table 2		Equivalent Percentile	
	Math	Reading	Math	Reading
schools with 1 subgroup	64	59	.747	.665
schools with 5 subgroups	58	54	.648	.575

This table shows that the differential of 5 or 6 points in the NCE metric
turns out to be 9 or 10 percentile points. Translating that differential
into the California API scale (see Rogosa 2000, for such calibrations of
the API scale) this 9 or 10 percentile point differential is equivalent to
between 80 and 90 API points, a differential greater than the width of two
full state deciles! By no reasonable standard are these groups of schools
"similar". Lower scoring schools will indeed tend to fare less well in NCLB
than higher scoring schools.

Many other details in the PACE report merit negative comment, but the
intent of this rebuttal is to discredit the general approach (in order to
guide others toward better methodology) rather than to dissect the specific
fumbles of Novak and Fuller. That said, two other specifics on the test
score displays require clarification. It is regrettable from a mathematics
education standpoint that Novak and Fuller use odds and proportions
interchangeably throughout the text of the PACE report. Also the PACE
report is inconsistent in describing the CAT/6 tests employed. In Tables
2,3,4 the CAT/6 test is labeled Reading, but the text of the PACE report
refers to language arts. Reading and Language Arts are separately reported
scores of the CAT/6 battery in STAR. [Note: I did not attempt to recreate
from the state data the test score results shown in the PACE report nor did

I audit the tabled cross-classifications of percent disadvantaged and
number of subgroups.]

Small Note #2. Counting (often) redundant subgroups.

As noted in various spots in the PACE report, subgroup membership overlaps
to an extent that some subgroups are nearly redundant. Therefore, counting
up the reported number of NCLB student subgroups as is done in PACE Tables
1,2,3,4 can be a misleading indicator of the statistical context and
properties. One canonical example would be a school say of 390 students
with 387 of those students classified in the Hispanic subgroup and 386 of
those students also classified as SocioEconomically Disadvantaged. Such a
school would be deemed to have 2 subgroups in the PACE report tables (or
even three subgroups if the disadvantaged Hispanic students were also
classified EL). But clearly, if the school succeeds then all the subgroups
succeed and vice versa. So the properties of this 2 or 3 subgroup school
are actually equivalent to a school having no subgroups (or one subgroup if
the school is counted as its own subgroup). A different school might well
be composed of 3 distinct, non-overlapping, subgroups, yet the two schools
would be considered equivalent in the PACE tabulations. Many of the
probability calculations that follow are expressed in terms of number of
non-overlapping subgroups. And it's useful I believe to think of many of
the California schools with reported 5 or 6 subgroups really being
stochastically equivalent to 3 or 4 non-overlapping subgroups. Very few
California schools are equivalent to 5 or more non-overlapping subgroups.

CALCULATIONS PART I
Binomial Probabilities in NCLB setting:
Comparing artificial schools with identical educational attainment

The goal of these calculations is to do what Novak and Fuller seek--
compare probabilities of NCLB success for schools that are identical except
for the number of subgroups (diversity). But here, fairness is represented
in terms of equality of opportunity (probability of success) rather than
the Novak-Fuller equality of results (i.e., unfairness in terms of
disparity of outcomes). If indeed the claim in the PACE report that
"schools serving diverse students in California are less likely to achieve
their growth targets" is true, it matters how much less likely. For
example, what's the difference between two schools with equal school-wide
educational attainment, where one school is homogeneous, composed of a
single subgroup of size 300 and the other school is diverse, composed of 3
non-overlapping subgroups each of size 100? Calculations accessible to a
high school student with a knowledge of the binomial distribution and basic
spreadsheet computing skills actually can provide useful information on
these questions.

Construction of an artificial school.
The rough approximation to reality used in these calculations is that
schools are entirely composed of non-overlapping subgroups. (Thus if each
subgroup meets AMO, then school does also.) A school of size 100 could be
composed of a single group of 100 students (in this case the subgroup is
the school) or composed of two non-overlapping subgroups each of the
California minimum size of 50. A school of size 600 could be composed of 4
non-overlapping subgroups of 150 students or composed of two non-
overlapping subgroups each of size 300 (subgroup sizes kept equal merely
for convenience in exposition). For purposes of statistical calculation
this artificial school construction isn't as unrealistic as it may appear.
Even though in reality subgroups do have overlapping membership, the
overlapping subgroups are often partially or largely redundant, and
therefore not consequential for this initial statistical analysis. Consider
a school with two racial subgroups, African-American and Hispanic. If all
African-American and Hispanic students in the school are also disadvantaged
(and represent all the disadvantaged students), then success by the
African-American and Hispanic subgroups in meeting AMO guarantees success
by the disadvantaged subgroup. Therefore, the school really acts as having
2 nonoverlapping subgroups rather than the reported 3 subgroups.

Educational attainment via true proportion proficient.
For purposes here, two schools are said to have the same educational
attainment or achievement by having the same value of true proportion
proficient. (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.) The observed proportion
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)
and can be thought of as a version of "true" proportion proficient that is
obscured by statistical variability. Conversely, the underlying educational
achievement (e.g., "true" proportion proficient) can be thought of as the
observed proportion stripped of its statistical uncertainty.

Binomial calculations for probability of meeting AMO.
The core of the calculations use the binomial distribution to compute for
any group or subgroup of size n and true proportion proficient pi the
probability of meeting the AMO. Or, inversely, to solve for the true
proportion proficient necessary to meet the AMO with high probability (e.g.
probability .99). Of course if two schools have similar probability of
success, the concept of fairness, as stated in terms of equality of
opportunity, is satisfied. 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. More
complex versions of these calculations (incorporating exact results for the
bias and sampling variability of the observed proportion proficient) can be
constructed, but that would seem to be technical overkill for the
relatively simple task of getting some handle on the effects of subgroups.

RESULTS

The starting point is Table 1, "Required School-level True Proportion
Proficient for Subgroups of Equal Size and Equal Attainment". Separate
results are given for English and math with the performance goal being the
NCLB 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 entries of Table 1 are values of true proportion proficient that are
large enough to give probability .99 that the observed proportion
proficient will meet the AMO. The point of the previous report
"California's AMOs Are More Formidable Than They Appear"(Rogosa, 2003a) was
to quantify the effects of statistical uncertainty in the proportion
proficient scores in these terms: to have high probability that school and
all subgroup scores meet the criteria (AMO) requires underlying educational
performance that exceeds (blows through) the seemingly modest AMO values.
For example, in Table 1, a school of 200 students, comprised of two
subgroups each of size 100, requires true proportion proficient in English
.239 and math .263, values that exceed the respective AMOs by .103 and
which translate into about 20 additional truly proficient students in each
subject. Note that built into the structure of Table 1 is the requirement
that all subgroups in a school have the same true proportion proficient,
the school-wide value.

These same calculations allow examination of the effect of subgroups by asking how much does the required true proportion proficient increase as the number of subgroups increase. Going across rows in Table 1 would indicate for both English and math an increase of .02 to .025 in true proportion proficient moving from the 0/1 (single subgroup is the school) column to the 5 non-overlapping subgroup column. Given that Novak and Fuller are happy to consider 6 percentile point differences in test score outcomes as within the meaning "similar", a difference of 2 points in percent proficient over that very wide range of non-overlapping subgroups would seem to be a rather puny effect.

Insert Table 1

Another way to express the results in Table 1 is to keep true proportion proficient at the value in the 0/1 subgroup column and ask how much decrease in probability of meeting AMO results from additional subgroups. For Table 1 setting the answer is a decrease in probability of .01 for each additional subgroup. Thus a level of true educational attainment that gives probability .99 of meeting AMO for a completely homogeneous school (comprised of one subgroup, subgroup is the school) would give probability .951 of meeting AMO for an extremely diverse school comprised of 5 non-overlapping subgroups (and very few schools in California are that diverse). That result holds for both subjects and is constant over subgroup size as shown below. Many other versions of this calculation can be done;

       Effects on probability of meeting AMO of additional subgroups
               Number of Non-overlapping Subgroups
               0/1   2   3   4   5

    n
    50   0.98   0.97   0.961   0.951
    100   0.99   0.98   0.97   0.961   0.951
    200   0.99   0.98   0.97   0.961   0.951
----
for example starting with the level of true proportion proficient required to have probability .95 of meeting AMO for a school comprised of two non-overlapping subgroups will indicate that each additional subgroup will reduce the probability of meeting AMO by almost .025.

AYP Requires Meeting Both AMOs.
The results above treat English and math separately. To speak in terms of satisfying AYP requires calculations for the joint probabilities of meeting both the English and math AMO. 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 two independent trials, the probability of success on both trials would be .985 for two trials each with probability of success .99 and .903 for two trials each with probability of success .95. Taking typical within-school correlations of English and math around .7 indicates the joint pass probabilities are approximately .985 for single trial probability .99 and .926 for .95. Thus the AYP version (below) of the AMO table (above) shows a decrease in AYP probability of about .015 for each

Table 1
Required Educational Attainment (School-level True Proportion Proficient)
for Subgroups of Equal Size and Equal Attainment

School-level True Proportion Proficient Required for All
Subgroups each of Size n to Meet English Performance Goal
.136 with Probability .99

	Number of Non-overlapping Subgroups				
	0/1	2	3	4	5
n					
50		0.284	0.293	0.3	0.305
75		0.264	0.271	0.277	0.28
100	0.228	0.239	0.245	0.25	0.253
125	0.213	0.223	0.229	0.232	0.235
150	0.211	0.22	0.225	0.228	0.231
175	0.202	0.21	0.215	0.218	0.22
200	0.201	0.208	0.213	0.215	0.218

School-level True Proportion Proficient Required for All
Subgroups Each of Size n to Meet Mathematics Performance Goal
.16 with Probability .99

	Number of Non-overlapping Subgroups				
	0/1	2	3	4	5
n					
50		0.309	0.319	0.325	0.33
75		0.28	0.288	0.293	0.297
100	0.252	0.263	0.27	0.274	0.277
125	0.242	0.252	0.257	0.261	0.264
150	0.234	0.243	0.248	0.252	0.254
175	0.228	0.237	0.241	0.245	0.247
200	0.224	0.232	0.236	0.239	0.241

additional subgroup (i.e., .97 for two subgroups decreases to .94 for four non- overlapping subgroups).

Effects on probability of meeting AYP of additional subgroups

Number of Non-overlapping Subgroups

n	0/1	2	3	4	5
50		0.970	0.956	0.941	0.927
100	0.985	0.970	0.956	0.941	0.927
200	0.985	0.970	0.956	0.941	0.927

What are the most obvious weaknesses of the Table 1 formulation? One is that although schools with more subgroups tend to be larger, the size differential (even if calibrated for non-overlapping groups) is smaller than proportional to the number of subgroups as in the Table 1 structure. That's why there is a Table 3. Second, not all subgroups have equal educational ability (underlying performance) as is required in the Table 1 structure. That's why Table 2 introduces laddered subgroup true proportion proficient, allowing school-wide true proportion proficient to remain constant even though subgroups differ in their own true proportion proficient. Also, the heterogeneity in ability is constant across the various subgroup configurations. The calculations using the Table 1 specification emphasize settings in which the true educational attainment provides high probability of meeting AMO. Even so, the true proportion proficient values are modest (.2 to .3). Of course, if the school-wide true proportion proficient is much nearer the AMO, then failure to meet the AMO in school or subgroup scores due to statistical variability in the observed proportion proficient is likely, and the effect of subgroups will appear larger. For example, if English true proportion proficient is .156, the probability of school with subgroups of 100 students meeting the AMO of .136 is only .71 for a single subgroup and reduces to .26 with four non-overlapping subgroups. This sort of result fits in with the previously expressed theme that statistical variability makes the AMO more formidable than it might appear.

Subgroups with heterogeneous true proportions proficient.

In Table 1 all subgroups are constrained to have the same true proportion proficient, so that the school-wide true proportion proficient also pertains to each of the subgroups. Calculations for Table 2 are somewhat less unrealistic in relaxing that constraint. Specifically, Table 2 calculations employ a laddered set of true proportion proficient values constructed by deviation of a set amount from the school-wide proportion proficient. These displacement values allow the same school-wide true proportion proficient for artificial schools with differing numbers (2,3,4,5) of non-overlapping subgroups as each of the set of displacements have mean 0. In addition these values are chosen to equalize the heterogeneity by each set of displacement values having variance 2/300.

Laddering of Subgroup True Proportion Proficient

number of subgroups	Displacement from school wide true proportion proficient
2	$\{$-Sqrt[2/3] , Sqrt[2/3]$\}$/10
3	$\{$-1, 0, 1$\}$/10
4	$\{$-1, -Sqrt[1/3], Sqrt[1/3], 1$\}$/10
5	$\{$-1, -Sqrt[2/3], 0, Sqrt[2/3], 1$\}$/10

where Sqrt[1/3] = .577, Sqrt[2/3] = .816.

Table 2 displays the school-wide true proportion proficient required for probability .99 that the school and all subgroups meet the AMO. Thus the school comprised of 3 subgroups of size 100, 300 students total, has in Table 2 school-wide true English proportion proficient .328, with the three subgroups having the values for true proportion proficient for English {.228, .328, .428}. Values of school-wide true proportion proficient in Table 2 are of course larger than corresponding values in Table 1 because failures to meet AMO will be driven by the subgroup with the lowest true proportion proficient. The differential is less than .1, the span of the displacement. Of most interest here is the increase in required school-wide true proportion proficient is around .02 moving from schools with two non-overlapping subgroups to schools with 5 non-overlapping subgroups. That small increase is about twice as large as was seen under the homogeneity constraints in Table 1. Calibrating the effect of subgroups in terms of reduction of the probability of meeting AMO, as was done above, a school with five subgroups having the school-wide true proportion proficient value required for a school with two subgroups to meet the AMO with probability .99 has probability .962 of meeting the AMO. Thus in this setting the "diversity penalty" measured as effect of two subgroups versus five non-overlapping subgroups for schools with identical school-wide educational attainment is a reduction in the probability of meeting the AMO of less than .03.

Insert Table 2

```
Table 2
Required Educational Attainment (School-level True Proportion Proficient)
for Subgroups of Equal Size and Laddered Attainment

---------------------------------------------------------------------
Laddering of Subgroup True Proportion Proficient

number of           Displacement from school wide true proportion proficient
subgroups
    2                       {-Sqrt[2/3] , Sqrt[2/3]}/10
    3                       {-1,   0,     1}/10
    4                       {-1, -Sqrt[1/3], Sqrt[1/3], 1}/10
    5                       {-1, -Sqrt[2/3], 0,   Sqrt[2/3], 1}/10

where Sqrt[1/3] = .577, Sqrt[2/3] = .816.
---------------------------------------------------------------------
```

School-level True Proportion Proficient Required for All
Subgroups each of Size n to Meet English Performance Goal
.136 with Probability .99

	Number of Non-overlapping Subgroups with Laddered Attainment			
n	2	3	4	5
50	0.349	0.367	0.371	0.377
100	0.31	0.328	0.329	0.332
150	0.293	0.311	0.311	0.314
200	0.283	0.301	0.301	0.303
250	0.271	0.29	0.29	0.291
300	0.267	0.286	0.286	0.286

School-level True Proportion Proficient Required for All
Subgroups Each of Size n to Meet Mathematics Performance Goal
.16 with Probability .99

	Number of Non-overlapping Subgroups with Laddered Attainment			
n	2	3	4	5
50	0.374	0.392	0.396	0.402
100	0.333	0.352	0.353	0.356
150	0.316	0.334	0.335	0.337
200	0.305	0.324	0.324	0.326
250	0.298	0.317	0.317	0.318
300	0.293	0.312	0.312	0.313

Constant School Size, Increasing Number of Subgroups

Table 3 presents, in some sense, the worst-case scenario for the effects of subgroups. School size is held constant as the number of subgroups increases such that a computation for a school with 300 students compares two subgroups of size 150 with five subgroups of size 60. (True proportion proficient for all subgroups is constrained to be the same as the school-wide value as in Table 1.) Table 3 is a bookend to Table 1 in bracketing reality. Schools with larger numbers of subgroups do not necessarily have proportionately smaller subgroups. Even so, the increase in the required true proportion proficient in schools of size 600 from 0/1 column to the 5 non-overlapping column is less than .08. Slightly larger effects are driven not by the diversity but by the smaller size (i.e., larger statistical variability) of the multiple subgroups (e.g., 5 subgroups in a school of size 300 have 60 students each).

Insert Table 3

Table 3
School-level True Proportion Proficient Required for All
Subgroups of equal size to Meet AMO: Constant School Size

English, School Size 150

probability of meeting AMO	Number of Non-overlapping Subgroups				
	0/1	2	3	4	5
.90	0.176	0.215	0.234		
.95	0.188	0.231	0.254		
.99	0.211	0.264	0.293		

English, School Size 300

probability of meeting AMO	Number of Non-overlapping Subgroups				
	0/1	2	3	4	5
.90	0.162	0.187	0.206	0.231	0.25
.95	0.17	0.198	0.219	0.246	0.267
.99	0.186	0.22	0.245	0.277	0.302

English, School Size 600

probability of meeting AMO	Number of Non-overlapping Subgroups				
	0/1	2	3	4	5
.90	0.155	0.17	0.186	0.198	0.21
.95	0.16	0.177	0.195	0.208	0.222
.99	0.171	0.191	0.213	0.228	0.244

Math, School Size 150

probability of meeting AMO	Number of Non-overlapping Subgroups				
	0/1	2	3	4	5
.90	0.198	0.23	0.258		
.95	0.21	0.247	0.278		
.99	0.234	0.28	0.319		

Math, School Size 300

probability of meeting AMO	Number of Non-overlapping Subgroups				
	0/1	2	3	4	5
.90	0.187	0.21	0.229	0.246	0.27
.95	0.195	0.221	0.242	0.262	0.288
.99	0.212	0.243	0.27	0.293	0.323

Math, School Size 600

probability of meeting AMO	Number of Non-overlapping Subgroups				
	0/1	2	3	4	5
.90	0.179	0.195	0.209	0.221	0.239
.95	0.185	0.203	0.218	0.231	0.251
.99	0.196	0.218	0.236	0.252	0.274

CALCULATIONS PART II:
Probabilities for API Awards using California School Data

Unlike NCLB, the California award programs, such as the GPA, were based on
year to year improvement. Calculations on the effect of subgroups on the
probability of award start with data from some California elementary
schools, increment that data by a specified amount of true improvement and
use bootstrap resampling to calculate the probability of meeting the award
criteria (school and all significant subgroups meet their growth targets).
Such calculations were previously conducted as part of the refutation of
the claims in Kane and Staiger (2002), who assert "great disadvantage"
(p.269) to schools with more subgroups: "at any given level of overall
improvement, a racially integrated school is much less likely to win an
award than a racially homogeneous school" (p.258) (c.f., Rogosa 2002a,
section 4). In Rogosa (2002a,b) the subgroup criteria are seen to be very
important in reducing false positives and in making the growth targets more
formidable than the small numerical values would indicate.

Start by considering three elementary schools of median school size and
Decile 5 (approximately median) performance which contrast on number of
significant subgroups: two, three and four. (Note: 97% of decile 5 schools
have 2, 3 or 4 subgroups, 85% of all Elementary schools have 2,3 or 4
subgroups). This setting of constant school size with increasing number of
subgroups resembles the Table 3 formulation, which was considered as a
worst case setting in demonstrating the largest effect of subgroups. The
three elementary schools are:

CDS 56726036055735	CDS 19643376011951	CDS 19648816021653
n= 358, se(API)=12.7	n= 350, se(API)=13.7	n= 341, se(API) = 14.9
2 Sig Subgroups:	3 Sig Subgroups:	4 Sig Subgroups:
Hispanic(124),	Soc Dis (221),	Soc Dis (192), AfAm(134)
White(209)	Hispanic(189), White(119)	Hispanic(123), White(53)

Table 4 shows the probabilities of GPA Award Eligibility for three schools
used for this example, all Decile 5 and medium size (~50th percentile).
Explaining the structure of Table 4 requires introduction of some new
quantities. The award probabilities, both for the existing CDE GPA rules
and for the margin of error modifications, are expressed as a function of
the Incrementation (rows). This representation of school improvement has
two forms: Integer Incrementation (Ik) and Partial Incrementation (Pk). In
Integer Incrementation (Ik) every student increases k percentile points on
each test. Partial Incrementation (Pk) provides an intermediate improvement
between the levels of the Integer Incrementation. For grades 2-8, each
student increases k percentile points on math and k-1 on the other 3 tests
(Reading, Lang., Spell). In Table 4 the form of incrementation (Ik, Pk,
k=0,...,6) is shown in the Incrementation column, and the school API score
resulting from the Incrementation is given in the API column (note: "Base"
is I0). (In Section 2 of Rogosa 2000 these forms of incrementation, and
their consequences for API scores, are covered in detail.)

The calculation of probability of award for a specific incrementation is

done by bootstrap resampling because subgroups overlap with each other (i.e., SD subgroup) and with the full school. The calculation starts with the actual 1999 data for the school. First increase all student scores according to the incrementation protocol; then resampling (e.g. 10000 bootstrap resamples) is used to estimate the probability of award for that specified true improvement. These calculations address: What is the probability of award for a specified true improvement? The "Base (I0)" row provides information about false positives: the probability of achieving award status due to statistical variability alone (no real improvement).

Insert Table 4

The comparison of interest here is whether the award probabilities decrease markedly with greater number of subgroups. Figure 1 displays a smooth curve for these probabilities plotted against real API improvement. "Great disadvantage" is awfully hard to discern, but a small effect for subgroups, as is to be expected, is present. A numerical measure for these examples that match the schools on school size and performance is that schools with 3 versus 4 subgroups are separated overall by probability of award differential .004, and schools with 2 versus 4 subgroups are separated overall by probability of award differential .04 . (Different examples may show a larger effect for 3 versus 4 subgroups with two schools of the same size.)

Insert Figure 1

Schools with more subgroups do tend to have lower API scores. In the table of descriptive statistics below each row is the number of subgroups (1 to 5), then the number of 1999 elementary schools having that number of subgroups, then summary statistics for the corresponding API scores. Almost all of the 14% of Elementary schools having only one numerically significant subgroup are highly successful (deciles 8-10) small schools drawing from relatively advantaged populations. This confounding will badly warp any "equality/disparity of results" analysis to detect the effects of subgroup on award probabilities.

Descriptive Statistics: API by NSig: 1999 Elem Schools

NSig	N	Mean	Median	Q1	Q3
1	691	773.13	797.00	736.00	844.00
2	1660	602.60	581.00	466.25	730.75
3	1941	608.81	608.00	535.50	684.00
4	496	617.61	616.00	536.00	689.75
5	41	626.6	633.0	527.5	708.0

A different comparison consists of two decile 5 Elementary schools with two significant subgroups and approximately 150 students, and a decile 5 Elementary school with three significant subgroups and approximately 250 students. The table below shows that a somewhat larger school with 3 significant subgroups has a slightly greater overall probability of GPA award than the smaller schools with 2 significant subgroups. This example

Probability of California GPA award for equal sized schools with 2,3,4 numerically significant subgroups

Table 4. Probabilities of California GPA Award Eligibility for Decile 5 Elementary Schools
of Medium Size (~50th percentile) and Varying Number of Numerically Significant Subgroups

```
                    CDS  56726036055735      CDS  19643376011951      CDS  19648816021653
                    n= 358, se(API) = 12.7   n= 350, se(API) = 13.7   n= 341, se(API) = 14.9
                    2 Sig Subgroups:         3 Sig Subgroups:         4 Sig Subgroups:
                    Hispanic(124),           Soc Dis (221),           Soc Dis (192), AfAm(134)
                    White(209)               Hispanic(189), White(119) Hispanic(123), White(53)

Incrementation      API  PrAPI&Subgr>Targ1   API PrAPI&Subgr>Targ1    API   PrAPI&Subgr>Targ1

P0                  610      0.0511           610     0.0655           617        0.0697
Base                615      0.1080           613     0.1010           620        0.0994
P1                  620      0.1731           615     0.1275           621        0.1137
I1                  624      0.2792           621     0.2446           626        0.1950
P2                  627      0.3294           624     0.3111           631        0.2791
I2                  633      0.4753           630     0.4590           635        0.3798
P3                  635      0.5520           634     0.5321           637        0.4290
I3                  641      0.7375           640     0.6515           641        0.5108
P4                  645      0.8153           642     0.7136           645        0.6159
I4                  650      0.8911           647     0.7927           649        0.6990
P5                  655      0.9439           651     0.8639           651        0.7555
I5                  659      0.9711           658     0.9299           657        0.8679
P6                  662      0.9784           661     0.9564           660        0.8964
I6                  667      0.9919           668     0.9832           665        0.9400
```
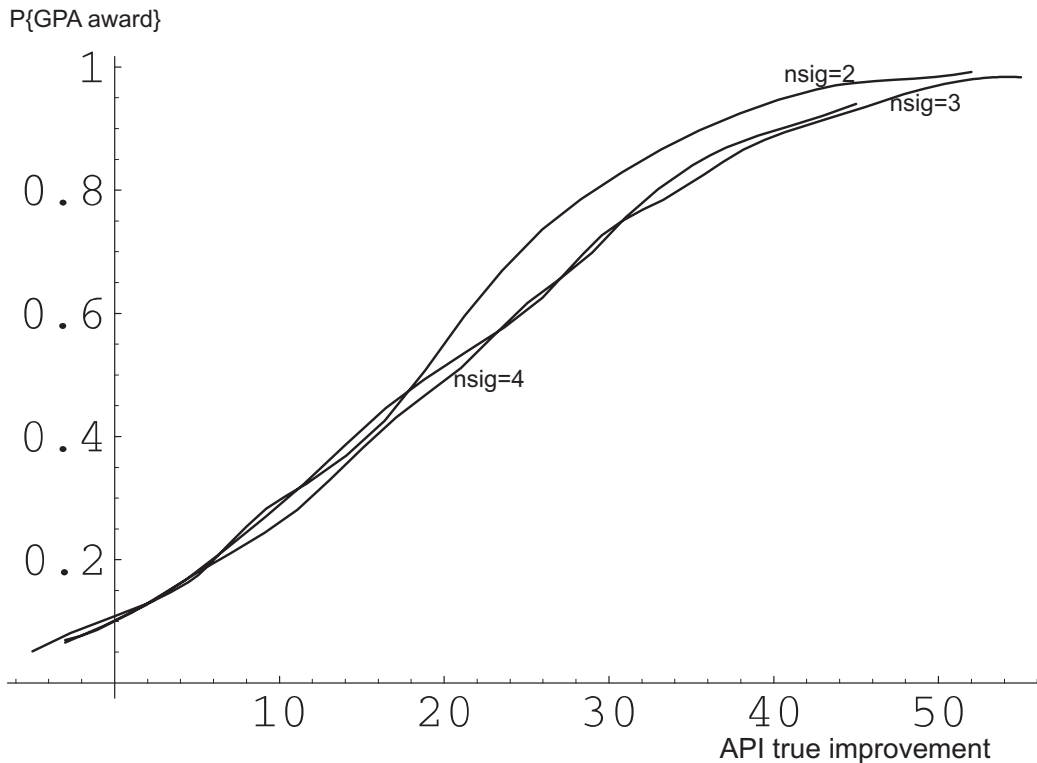

Figure 1. Plot of probability of GPA award as a function of true API improvement, resulting
from Ik, Pk incrementation (k=0,...,6) for three decile 5 elementary schools of median
size with 2,3,4 numerically significant subgroups.

of probability calculations represents another reasonable counterexample to
the Kane-Staiger and Novak-Fuller assertions both of which are based on
(highly confounded) equality of results criteria.

                          Decile 5 Elementary Schools

CDS 37682136038681   CDS 41689736044226   CDS 36679596037410
n=143, se(API)=19.5   n=174, se(API)=18.7   n=244, se(API)=14.3
2 Sig Subgroups: 2 Sig Subgroups: 3 Sig Subgroups:
Soc Dis(80)   Hispanic(66)   Soc Dis(44)
White(85)   White(47)   Hispanic(69) White(161)

        Probability of GPA Award Averaged over Ik, Pk Incrementation
    0.513   0.484   0.524

Final Thoughts

It seems useful to conclude on a note of concordance with the PACE report. In their final recommendations Novak and Fuller state: "Washington could respect states' own methods for determining achievement growth within schools" (page 9). That's an important statement worthy of further comment.

This rebuttal is properly thought of as a defense of good statistical reasoning and practice rather than a defense of NCLB. Regrettably, NCLB AYP requirements are poorly thought out and egregiously overreach. 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.

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). Sadly, the federal mandates of NCLB ignore these important lessons.

In particular, NCLB differs from California PSAA accountability by adding on additional subgroup categories. There can be too much of a good thing, and there is a point where the resulting statistical properties lead to a dysfunctional accountability system. The consequences of these additional NCLB subgroup categories deserve scrutiny, and a good PACE study would empirically demonstrate the undesirable effects of all the added subgroups in NCLB. That is, determine whether a school would have satisfied AYP if NCLB subgroups were restricted to the PSAA/API subgroups? Novak and Fuller do discuss the role of EL subgroups, and that would be a useful direction to pursue.

References

Kane, T. J., and Staiger, D. O. (2002) "Volatility in School Test Scores:
Implications for Test-Based Accountability Systems." Brookings Papers on
Education Policy, 2002 (Washington, DC: Brookings Institution), 235-269.

New York Times, December 23, 2003. New Education Law is Faulted in Gauging
Performance, by Sam Dillon.

Novak, John R. and Fuller, Bruce. PACE Policy Brief 03-4 , "Penalizing
Diverse Schools? Similar test scores, but different students, bring federal
sanctions" December 2003.
http://pace.berkeley.edu/policy_brief_03-4_Pen.Div.pdf

Rogosa, D.R.(2000). Interpretive Notes for the Academic Performance Index
California Department of Education, Policy and Evaluation Division November
20, 2000. California Department of Education website:
http://www.cde.ca.gov/psaa/apiresearch.htm

Rogosa, D.R.. (2002a). 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. (2002b). 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. (2002c). 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.. (2003a). 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

Rogosa, D.R. (2003b). Confusions about Consistency in Improvement. CRESST
deliverable, June 2003.
available from: http://www-stat.stanford.edu/~rag/api/consist.pdf

San Francisco Chronicle, December 23, 2003. Standards tougher on diverse
schools, More 'subgroups' mean more hurdles by Nanette Asimov.

APPENDIX [added Jan 25, 2004]

                I. Heterogeneity Trumps Number of Subgroups

A related calculation that I probably should have included in the main
rebuttal illustrates the obvious--in NCLB the probability of failing to
meet AMO is driven by the lowest achieving subgroup (regardless of whether
there are few or many subgroups). In NCLB, a school is evaluated on the
basis of its lowest-scoring subgroup not on the average achievement in the
school (min not mean). Consider two schools of equal size, each with
school-wide true proportion proficient .5 and each school having two
equally sized subgroups. In the first school the two subgroups are of equal
ability (therefore that school is close to assured of meeting the AMO which
is below .2). The second school is extremely heterogeneous with one
subgroup of very high academic achievement (true proportion proficient .9)
and the other subgroup of low academic achievement (true proportion
proficient .1) so that this school is unlikely to satisfy the California
AMO of .16. Because NCLB focuses on the performance of the lowest
performing subgroup rather than on the school mean, by design, two schools
with the same mean can have different probabilities of satisfying NCLB.

The interesting result of the calculations is how much more sensitive the
NCLB criterion is to small heterogeneity in the subgroups rather than
to the number of subgroups. Table A1 considers the comparison of schools
with two equally sized subgroups, with the subgroups having educational
attainment represented by $p + k$ and $p - k$, where p is the school-wide
true proportion proficient. For $k > 0$ there exists some heterogeneity
in that the subgroups differ (by $2k$) in their educational attainment.

Table A1
School-level True Proportion Proficient Required for both
Subgroups Each of Size n to Meet Math AMO .16 with Probability
.99 in the Presence of Heterogeneity k

	Heterogeneity k				
	0	.025	.05	.075	.01
n					
50	0.309	0.32	0.342	0.367	0.392
100	0.263	0.277	0.302	0.327	0.352
150	0.243	0.259	0.284	0.309	0.334
200	0.232	0.249	0.274	0.299	0.324
250	0.224	0.242	0.267	0.292	0.317
300	0.218	0.237	0.262	0.287	0.312

The first column in Table A1 (heterogeneity 0) reproduces entries the 2
subgroup column in Table 1. Take subgroups of size 100 and see that the
required school-wide true proportion proficient increases to .277 if the
two subgroups are .05 different (the k=.025 column). Refer to Table 1 and
note that this required increase to .277 is identical to the effect of
moving from a school of 2 subgroups to 5 subgroups (all of size 100). Close
equivalence are also seen for the other sub-group sizes. A small amount of

heterogeneity between the groups has the same effect as a large change in number of subgroups.

The calculations for Table A2 repeat this comparison employing a lower probability of satisfying NCLB Math requirements, probability .9 instead of .99. The top frame of Table A2 is a display like Table 1 showing the effect of increasing the number of equivalent subgroups. The lower frame repeats the display of Table A1, showing the effect of increasing heterogeneity between two equal sized subgroups. Again, the effect of small heterogeneity is equivalent to a large increase in the number of subgroups. For subgroups of size 200 the increase in the required school-wide true proportion proficient .203 to .218 resulting from heterogeneity .025 is larger that that from increasing the number of subgroups from 2 to 5 (.203 to .215). For small subgroups of size 50 the increase in the required school-wide true proportion proficient .246 to .275 resulting from heterogeneity .05 is larger than that from increasing the number of subgroups from 2 to 5 (.246 to .272).

Table A2
Comparison of Number of Subgroups and Heterogeneity for Probability .9 of Meeting Math AMO .16 with Probability .90

School-level True Proportion Proficient Required for All Subgroups Each of Size n to Meet Mathematics Performance Goal .16 with Probability .90

	Number of Non-overlapping Subgroups			
	2	3	4	5
n				
50	0.246	0.258	0.266	0.272
75	0.23	0.24	0.246	0.251
100	0.221	0.229	0.235	0.239
125	0.215	0.222	0.226	0.23
150	0.21	0.216	0.221	0.224
175	0.206	0.212	0.216	0.219
200	0.203	0.209	0.212	0.215

School-level True Proportion Proficient Required for both Subgroups Each of Size n to Meet Math AMO .16 with Probability .90 in the Presence of Heterogeneity k

	Heterogeneity k				
	0	.025	.05	.075	.01
n					
50	0.246	0.255	0.275	0.299	0.324
100	0.221	0.233	0.256	0.281	0.306
150	0.21	0.223	0.248	0.273	0.298
200	0.203	0.218	0.243	0.268	0.293
250	0.198	0.215	0.239	0.264	0.289
300	0.195	0.212	0.237	0.262	0.287

II. Computational Functions

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

```
Table 1
readTab1 =
  Table[pi /.
      FindRoot[(1 - CDF[BinomialDistribution[n, pi], Floor[.135999*n]])^
            i == .99, {pi, .3}] , {n, 50, 200, 25}, {i, 1, 5}]
mathTab1 =
  Table[pi /.
      FindRoot[(1 - CDF[BinomialDistribution[n, pi], Floor[.15999*n]])^
            i == .99, {pi, .3}] , {n, 50, 200, 25}, {i, 1, 5}]



Table 2
k = Sqrt[2/3.]; pList2 = {-k, k}/10; pList3 = {-1, 0, 1}/
    10. ; pList4 = {-1, -Sqrt[1/3.], Sqrt[1/3.], 1}/10 ;
    pList5 = {-1, -k, 0, k, 1}/10;
Table[pi /.FindRoot[
      Product[1 -
            CDF[BinomialDistribution[n, pi + pList2[[i]]],
              Floor[.135999*n]], {i,1,
            Length[pList2]}] \[Equal] .99, {pi,.3}] , {n, 50, 300,50}]
Table[pi /.FindRoot[
      Product[1 -
            CDF[BinomialDistribution[n, pi + pList3[[i]]],
              Floor[.135999*n]], {i,1,
            Length[pList3]}] \[Equal] .99, {pi,.3}] , {n, 50, 300,50}]
Table[pi /.FindRoot[
      Product[1 -
            CDF[BinomialDistribution[n, pi + pList4[[i]]],
              Floor[.135999*n]], {i,1,
            Length[pList4]}] \[Equal] .99, {pi,.3}] , {n, 50, 300,50}]
Table[pi /.FindRoot[
      Product[1 -
            CDF[BinomialDistribution[n, pi + pList5[[i]]],
              Floor[.135999*n]], {i,1,
            Length[pList5]}] \[Equal] .99, {pi,.3}] , {n, 50, 300,50}]
Table[pi /.FindRoot[
      Product[1 -
            CDF[BinomialDistribution[n, pi + pList2[[i]]],
              Floor[.15999*n]], {i,1,
            Length[pList2]}] \[Equal] .99, {pi,.3}] , {n, 50, 300,50}]
Table[pi /.FindRoot[
      Product[1 -
            CDF[BinomialDistribution[n, pi + pList3[[i]]],
              Floor[.15999*n]], {i,1,
            Length[pList3]}] \[Equal] .99, {pi,.3}] , {n, 50, 300,50}]
```

```
Table[pi /.FindRoot[
     Product[1 -
          CDF[BinomialDistribution[n, pi + pList4[[i]]],
            Floor[.15999*n]], {i,1,
          Length[pList4]}] \[Equal] .99, {pi,.3}] , {n, 50, 300,50}]
Table[pi /.FindRoot[
     Product[1 -
          CDF[BinomialDistribution[n, pi + pList5[[i]]],
            Floor[.15999*n]], {i,1,
          Length[pList5]}] \[Equal] .99, {pi,.3}] , {n, 50, 300,50}]


Table 3
conflist = {.9,.95,.99}
readTabNS150 =
  Table[pi /.FindRoot[(1 -
             CDF[BinomialDistribution[150/i, pi], Floor[.135999*150/i]])^
          i \[Equal] conflist[[jj]], {pi,.3}] ,  {jj,1,3}, {i,1,3}]
readTabNS300 =
  Table[pi /.FindRoot[(1 -
             CDF[BinomialDistribution[300/i, pi], Floor[.135999*300/i]])^
          i \[Equal] conflist[[jj]], {pi,.3}] ,  {jj,1,3}, {i,1,5}]
readTabNS600 =
  Table[pi /.FindRoot[(1 -
             CDF[BinomialDistribution[600/i, pi], Floor[.135999*600/i]])^
          i \[Equal] conflist[[jj]], {pi,.3}] ,  {jj,1,3}, {i,1,5}]
mathTabNS150 =
  Table[pi /.FindRoot[(1 -
             CDF[BinomialDistribution[150/i, pi], Floor[.15999*150/i]])^
          i \[Equal] conflist[[jj]], {pi,.3}] ,  {jj,1,3}, {i,1,3}]
mathTabNS300 =
  Table[pi /.FindRoot[(1 -
             CDF[BinomialDistribution[300/i, pi], Floor[.15999*300/i]])^
          i \[Equal] conflist[[jj]], {pi,.3}] ,  {jj,1,3}, {i,1,5}]
mathTabNS600 =
  Table[pi /.FindRoot[(1 -
             CDF[BinomialDistribution[600/i, pi], Floor[.15999*600/i]])^
          i \[Equal] conflist[[jj]], {pi,.3}] ,  {jj,1,3}, {i,1,5}]


Table A1
kList2 = {-k, k}
Table[pi /.FindRoot[
     Product[1 -
          CDF[BinomialDistribution[n, pi + kList2[[i]]],
            Floor[.15999*n]], {i,1,
          Length[kList2]}] \[Equal] .99, {pi,.3}] , {n, 50, 300,50}, {k,
   0, .1, .025}]
```