

A School Accountability Case Study: California API Awards
and the *Orange County Register* Margin of Error Folly

David Rogosa
Stanford University

To appear in Richard Phelps, Ed. *Defending Standardized Testing*.
Mahwah, NJ: Lawrence Erlbaum, 2004.

Running Head: A California School Accountability Case Study

Acknowledgments

Preparation of this chapter 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 findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Institute of Education Sciences, or the U.S. Department of Education.

This school accountability episode has the following timeline.

August 2002. The *Orange County Register* (Orange County, California), after months of preparation, launches a week-long series of attack pieces against the state of California school accountability system, which offered rewards for improvements in the Academic Performance Index (API). The main assertions by the *Orange County Register* to be examined in this case study:

California's \$1 billion school testing system is so riddled with flaws that the state has no idea whether one-third of the schools receiving cash awards actually earned the money (8/11/02)

the Register's findings, which showed about one-third of the award-winners gains were within the error margin making it impossible to tell if the school really improved (8/16/02)

That error rate means that of the \$67.3 million the state plans to give to schools as rewards this year 35 percent is slated for schools where improvement had little or no statistical significance. (8/11/02)

These claims by the newspaper garnered considerable press attention and serious governmental concern (OCR 8/14/02). For this series, the Orange County Register received from the Education Writers Association the 2002 National Award for Education Reporting in the Investigative Reporting category. Obviously, if these claims were at all correct, the California school accountability system would not be credible or defensible.

The *OCRegister* reporters repeatedly cite the advice and contributions to their analysis by Richard Hill, Center for Assessment (e.g., Hill, 2001). Additional experts cited include Thomas Kane of UCLA.

September 2002. David Rogosa (9/9/02) distributes "Commentaries on the Orange County Register Series" (Rogosa 2002a) on the California Department of Education (CDE) website demonstrating that instead of the 1/3 or 35% claimed by the *OCRegister*, the correct answers were 2% of schools, 1% of funds. The lead from (Rogosa, 2002a):

The Orange County Register analysis (8/11-8/16/02) of the California API awards is so riddled with statistical blunders and misstatements that credibility should not be given to their numerical assertions.

In addition, Rogosa (2002b) addressed the Education Writers

Association Meetings on September 5, 2002 including a panel discussion with the *OCRegister* reporters.

October 2002. Unabashed, the *OCRegister* reporters repeat their assertions and stress the importance of the *margin of error*:

in the first three years, the state handed out about \$1 billion in awards and money for remedial programs based on API scores that sometimes fell within the margin of error, so it was unclear if the school really improved or dropped. (10/17/02)

September 2003. The *OCRegister* assertions receive national exposure by Malcolm Gladwell in the *New Yorker* (9/15/2003):

But the average margin of error on the A.P.I. is something like twenty points, and for a small school it can be as much as fifty points. In a recent investigation, the Orange County Register concluded that, as a result, about a third of the money given out by the state might have been awarded to schools that simply got lucky.

The major lesson demonstrated in this chapter is that the *OCRegister* use of the *margin of error*, which they define as 1.96 times the standard error of the school score, represents a serious misunderstanding of basic material from introductory statistics courses. Regrettably, the statistical blunders involved in the *OCRegister* margin of error calculations also arise in slightly different contexts in many state accountability plans for *No Child Left Behind* (CCSSO, 2002). More broadly, the calculations mustered to refute the *OCRegister* claims also provide useful approaches and tools for understanding the properties of accountability systems.

I. California API Award Programs and the ORegister Calculations

The California Academic Performance Index (API) is a weighted index of student national percentile ranks on the Stanford 9 battery, producing a possible range for school (or subgroup) scores of 200 to 1000 (see Appendix A for computational details). To provide some calibration of the scale it's useful to note that a school with about half of its students above the national 50th percentile on the tests will have an API score around 650; also, a one percentile point improvement by each student on each test translates into a 8 to 10 point improvement in the school API (Rogosa, 2000). Data analysis for student and school progress in the California testing programs is provided in Rogosa (2003).

The focus of the *ORegister* series, and thus this analysis, is the California Governor's Performance Awards (GPA). For most schools (schools below API 800) the API growth target for GPA awards is an arithmetic form of 5% toward the interim state goal of API 800. To receive a GPA award, targets for the numerically significant subgroups in the school must also be met (see Appendix A). The dollar awards to schools and teachers for 1999-2000 API improvement totaled \$227 million from GPA (plus another \$350 million from the Schoolsite Employee Performance Bonus program to GPA schools); for 2000-2001 API improvement GPA awards totaled \$144 million (disbursed in two fiscal years).

ORegister calculations. The margin of error tallies that produce the claimed 1/3 or 35% start out by counting the schools receiving GPA awards for 1999-2000 and for 2000-2001. The calculation that best matches the *ORegister* description frames the calculation in terms of year-to-year improvement in the school API. The *ORegister* margin of error for improvement is given as: $1.3 * 1.96 * (\text{standard error for that school's API score})$. Tag a school with the designation "state has no idea if it really improved or earned the money" whenever the school's second year API score does not exceed the first year API score by more than this margin of error. The proportion of GPA award schools in 1999-2000 and 2000-2001 passing that criteria for is $(3357 + 1895) / (4526 + 3200) = .68$; i.e., 32% of the GPA award schools fail these criteria.

Appendix A considers three related versions of the margin of error

calculations, all of which produce numbers reasonably consistent with the *OCRegister* claims. As all the margin of error calculations are seen to be statistically unsound, the exact form doesn't much matter. The following sections demonstrate how the *OCRegister* application of the margin of error produces such misleading results, a folly which can be summarized as piling might-have's upon could-be's and counting all those as certainties.

II. The Margin of Error Folly: If it Could be, it is

The margin of error, the main *OCRegister* tool for analysis of the California API awards, is shown to have no value and to lead to preposterous results. A series of examples show California schools with year-to-year improvement in the API falling within this margin of error (and therefore no real improvement according to *OCRegister*) also having probability .98 and larger that true change is greater than 0.

Blood pressure parable

As a lead-in to API data demonstrations, consider an artificial setting, using diastolic blood pressure (DBP) for hypertension diagnosis. Like educational standards, standards for blood pressure diagnosis are subject to revision (Brody, 2003). The example uses part of the hypertension standard: DBP, 90 or above. For this statistical demonstration consider the distribution of DBP for adult males to be represented as normally distributed with mean 82, standard deviation 11.5, which would indicate about 25% of DPB at or above 90. (Hypertension diagnoses can also be based on elevated systolic pressure, leading to a hypertension rate of 1/3 or more.) Also, for purposes of this example assume DPB has measurement uncertainty indicated by standard error of measurement of 6.12 (due to time of day, patient factors etc.) Consequently, the *margin of error* for DPB is taken to be 12.

A patient is measured to have $DPB = 101$. The margin of error logic indicates that the physician has "no idea" whether the patient is indeed hypertensive (as by the margin of error a reading of 101 is indistinguishable from 89). To the contrary, this single DPB measurement does provide information, and the statistical question is: What does a DPB reading of 101 indicate about hypertension status? That is, calculate the conditional probability that true DBP is at least 90 given a DPB measurement of 101. As shown in Appendix B, this probability is .912. That is, the probability is better than nine out of ten, or odds of ten to one, that the true DPB is at least 90. Figure 1 provides a depiction of this example. Does the application of the margin of error appear to promote good health (or good educational policy)?

Insert Figure 1

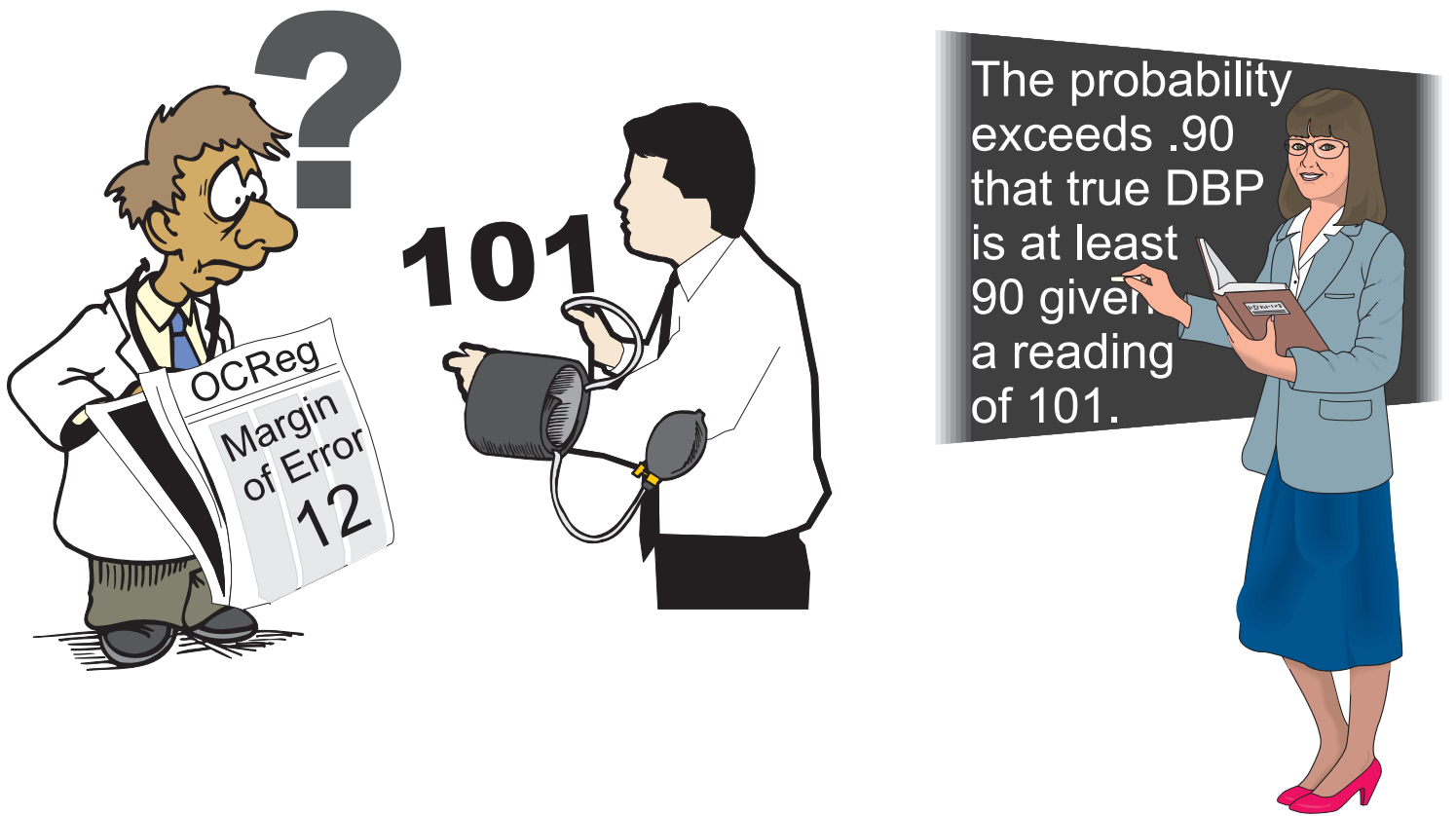


Figure 1. Non-verbal depiction of blood pressure parable. Man (center) has diastolic blood pressure reading 101. Doctor (left) heeding the Orange County Register margin of error "has no idea" of whether the man is hypertensive. Statistician (right) explains the probability is above .9 that the patient's perfectly measured DBP is at least 90.

California school examples: margin of error vs actual probability calculations

In the API context, the disconnect between the margin of error logic and any reasonable probability statement is even larger than that seen in the Blood Pressure Parable. Three school examples, all telling the same story, are displayed in Table 1. Bottom line: application of the *OCRegister* margin of error to API scores produces nonsense.

School Example 1. The *OCRegister* hypothetical elementary school (from *OCRegister* 8/11/02, "API's error margin leaves a lot to chance"). has API score 620 in year 2000 implying a growth target of 9 points. The growth target is surpassed by the 640 score in year 2001. However the *OCRegister* margin of error for improvement is 21.2, exceeding the observed improvement. The *OCRegister* piece, by Ronald Campbell, poses the question "How do you know if the score really increased"? His answer is that the margin of error for improvement (here 21.2 points) indicates "how much the score could have risen by chance alone." His conclusion is "Bottom line: Chance, not improved academics, might be behind that 20 point increase."

Anything "might" or "could" happen; probability calculations

Table 1
 Three School Examples: Probability Calculations for API Improvement

	School Example 1		School Example 2		School Example 3	
	yr1	yr2	yr1	yr2	yr1	yr2
API	620	640	685	701	616	647
n	900	900	900	1002	349	355
se(API)	8.32	8.32	7.5	6.95	14.2	13.2
margin of error						
API	16.3	16.3	14.7	13.6	27.8	25.9
Improvement		21.2		18.4		34.9
P{true change ≤ 0 observed data}						
		.0189		.0341		.00798

(formally or informally) are useful to quantify the "mights" and "coulds" of life. The relevant calculations (see Appendix B), using this *OCRegister* hypothetical school data and the actual California 2000-2001 Elementary school data, are that $P\{\text{true change} \leq 0 \mid \text{observed data}\} = .0189$ and $P\{\text{true change} < 9 \mid \text{observed data}\} = .125$. The probability exceeds .98 that the true change is positive for this school (and the odds are 7 to 1 that the true improvement meets the growth target of 9 points).

But according to the *OCRegister*, because the 20 point improvement does not exceed their 21.2 margin of error, this hypothetical school would be a school for which it's "impossible to tell if the school really improved". Most importantly, this school would be included in the *OCRegister* tally (approximately 35% total) of schools for which the state has no idea of whether they really improved. In effect, the *OCRegister* would regard this school as having probability 1 (i.e. certainty) of no real change. What this amounts to is rounding the .0189 probability up to 1.0 (in effect multiplying the actual probability by more than a factor of 50). Pile on enough "might haves" upon "could bes" and then count all of those as certainties, and it's not hard to obtain a dramatically inflated number such as the 1/3 or 35% trumpeted by the *OCRegister*.

School Example 2. This California middle school (CDS code 19644516057616) for years 2000-2001 was chosen for similarity to the *OCRegister* hypothetical school in Example 1. This school has 3 significant subgroups, Socioeconomically Disadvantaged (SD), Hispanic and White and received a GPA award of \$44,404 for year 2000 to 2001 improvement. The year-to-year improvement of 16 points for the school (see Table 1) is 10 points above the growth target of 6 points, but still less than the margin of error for improvement calculated to be 18.4 points.

Again the probability that true (real) change is positive is very large, above .96, even though the school's improvement is less than the stated margin of error and, therefore, not real according to the *OCRegister*. The *OCRegister* calculation for the award programs counts this school as one for which "it's impossible to tell if the school really improved (OCR 8/16/02)" and includes this school in the tally that yields their 35% figure; in effect the *OCRegister* is taking the small probability .0386 and rounding it up to 1.0. Also note that the $P\{\text{true}$

change $< 6 \mid$ observed data} = .133 and thus the odds are better than 6 to 1 that this school's true change met its growth target of 6.

School Example 3. This elementary school (CDS code 19647336018253) for years 1999-2000 has 2 significant subgroups, Socioeconomically Disadvantaged (SD) and Hispanic, and this school received a GPA award of \$40,262 for year 1999 to 2000 improvement. The year-to-year improvement of 31 points for the school is 22 points above the growth target of 9 points, but still less than the margin of error for improvement, calculated to be 34.9 points.

The probability is .992 that the true improvement is greater than 0. Yet, because the observed improvement of 31 points is less than the margin of error for improvement of 34.9, according to the *OCRegister* we are to have "no idea" whether this school actually improved and this school is included in the 35% tally. In effect *OCRegister* rounds .008 up to 1.0 in their tabulation of "impossible to tell" schools. Moreover, $P\{\text{true change} < 9 \mid \text{observed data}\} = .0386$ and thus the odds are 25 to 1 that the true improvement exceeded the year 2000 growth target.

The intent of these examples is to expunge the margin of error from any discourse about the California API awards or any other educational assessment or accountability program. Understanding the accuracy (statistical properties) of all aspects of an assessment program is essential; botching the important task of understanding accuracy must be avoided in both policy research work and press reporting.

III. Aggregate Results from the School Probability Calculations

The statistician would calculate for each of the GPA award schools the quantity: $P\{\text{true improvement} \leq 0 \mid \text{data}\}$. This empirical Bayes calculation is described in some detail in Appendix B. The calculation is carried out separately by school type (elementary, middle, high) and by award cycle (1999-2000, 2000-2001), because improvement differs over school type and between year-to-year comparisons. The California Data set used here had 4545 GPA award schools out of 6630 in 1999-2000 and 3167 award schools out of 6570 in 2000-2001. The aggregate results of this collection of six analyses are shown in Table 2.

The expected number of schools in each cell (schools having no real improvement and given GPA award) is simply the sum of the probabilities of all the schools (or the mean probability times the number of GPA award schools). The 1.25% result for 1999-2000 award cycle is obtained from $(35 + 12.67 + 8.95)/4545 = 0.01246$ and the 3% result for 2000-2001 is obtained from $(74.53 + 14.21 + 7.94)/3167 = 0.03053$. The cumulative 2% of schools is $(35 + 12.67 + 8.95 + 74.53 + 14.21 + 7.94)/(4545 + 3167) = 0.01988$.

The total funds associated with the $P\{\text{true improvement} \leq 0 \mid \text{data}\}$

Table 2.
Probability Calculations: No Real Improvement for Schools Given Awards

Award Cycle	School type		
	Elementary	Middle	High
1999-2000	.0098 35.0 (3585/4717)	.0199 12.67 (637/1100)	.0277 8.95 (323/813)
2000-2001	.0296 74.53 (2522/4667)	.0304 14.21 (468/1100)	.0448 7.94 (177/803)

Each cell contains
the average probability of no improvement for GPA award schools
the expected number of schools having no real improvement and given GPA award
(number of GPA awards / number of schools)

calculations are closer to 1% of the award monies than to the overall 2% of schools. Two factors make the expected amount of funds relatively smaller than the expected number of schools; GPA awards in the 1999-2000 cycle were about twice as large as the awards in the 2000-2001 cycle, and because these funds are per student, the small schools which tend to have the higher false positive probabilities receive less total funds.

Most GPA award schools have very large calculated values of $P\{\text{true change} > 0 \mid \text{observed data}\}$. As displayed in Appendix B (Table B3), for the 4545 elementary school GPA award winners in 1999-2000, the median value of $P\{\text{true change} > 0 \mid \text{observed data}\}$ was .9988, over 75% of the schools had probabilities above .99, and over 90% of these schools had probabilities above .97. Yet the *OCRegister* applies their margin of error to assert "no idea" of whether many of the schools really improved. Moreover, results for the probability that true change exceeded the growth target, $P\{\text{true change} > \text{growth target} \mid \text{observed data}\}$, further refute the *OCRegister* claims. For elementary schools, the median probability for those schools receiving GPA awards is .9926 for 1999-2000 and .951 for 2000-2001. Percentiles for the distributions of these probabilities are given in Appendix B, Table B3.

IV. How a High School Student Could Have Set the Orange County Register Straight

The fable of "The High School Intern and the API Dollars" is used to demonstrate a simple plausibility check on the *OCRegister* claims, one that yields results remarkably close to the more complex probability calculations. This discussion also introduces a familiar framework for discussing and evaluating properties of accountability systems.

Medical Diagnostic Test Context

The statistical approach to the accuracy of award programs follows standard ideas from medical diagnostic and screening tests. The accuracy of the award programs is expressed in terms of false positive and false negative events, which are depicted in the chart in Table 3. Commonly accepted medical tests have less than perfect accuracy. For example, prostate cancer screening (PSA) produces considerable false positives and in tuberculosis screening, false negatives (sending an infected patient into the general population) are of considerable concern. In the context of API awards, false positives describe events where statistical variability alone (no real improvement) produces award eligibility. False negatives describe events for which award

Table 3
 2x2 diagnostic accuracy table with joint probabilities a,b,c,d

	Good Real Improvement	NO Real Improvement
GPA Award	TRUE POSITIVE {a}	FALSE POSITIVE {b}
NO GPA Award	FALSE NEGATIVE {c}	TRUE NEGATIVE {d}

status is denied due to statistical variability in the scores, despite a (stated) level of underlying ("real") improvement.

The common derived quantities from Table 3 are *sensitivity*, $a/(a+c)$, which determines the proportion of false-negative results, and *specificity*, $d/(b+d)$ which determines the proportion of false-positive results. Note especially that the *OCRegister* quantity of interest, $P\{\text{no real improvement} | \text{award}\} = b/(a + b)$, which would be termed 1 – predictive value positive.

The High School Intern and the API \$\$\$

The short version of this fable is expressed in the equation:

$$\begin{aligned} &\text{Smart High School Statistics Student} + \text{Publicly Available Information} \\ &= \text{Correct Answer} \end{aligned}$$

The setting for the fable is California, July 2002. A newspaper preparing a series on the API has a summer intern who has recently completed a high school statistics course. The intern is asked by supervisors: "Do you think our finding is reasonable that a third or more of the GPA award schools made no real improvement? That is, can you confirm the statement $P\{\text{no real improvement} | \text{GPA award}\} > .3$ "

The High School statistics student makes the following presentation to the newspaper's reporters:

"In my class we learned about false positives and false negatives, for disease diagnosis and stuff [see Figure 2]. To get the information I needed for the API, I did the following:

a. From reports on the CDE website I can get $P\{\text{award} | \text{no improvement}\}$ for two examples, a typical elementary school with a .1 probability, and a typical high school with a .01 probability. Middle schools will be in between elementary and high, and since there are more elementary schools, that probability might average out over California schools to .07 or .08.

b. But $P\{\text{award} | \text{no improvement}\}$ is not the probability I was asked about. From my statistics course, I know a little about conditional probability; namely that $P\{\text{no improvement} | \text{award}\} = P\{\text{award} | \text{no improvement}\} * P\{\text{no improvement}\} / P\{\text{award}\}$.

c. From newspapers or CDE site I see that the GPA award rate for

1999-2000 was just about 2/3, which is my $P\{\text{award}\}$.

d. From reports on the CDE website [Rogosa, 2001a,b], I can get the observed distribution of year-to-year change in the API, and I calculate that proportion of schools with observed improvement less than or equal to 0 (which overstates proportion no true improvement) is approximately .1.

So now I can plug into my conditional probability formula and get a guesstimate for the 1999-2000 GPA awards that $P\{\text{no real improvement} \mid \text{GPA award}\}$ is approximately $.07 \cdot .1 / .67 = .01$. For 2000-2001 awards, less awards were given, and the CDE reports (Rogosa, 2001b) tell me that at least twice as many schools showed no improvement compared to 1999-2000. Combine those factors, and for 2000-2001 I can compute that $P\{\text{no real improvement} \mid \text{GPA award}\}$ is at least .03. The two results average out to an overall .02. As 1/50 is a whole lot less than 1/3, I can't confirm the reporters' story.”

V. Applying Margin of Error to Award Qualification

Using their margin of error the *OCRegister* claimed that in 35% of schools statistical uncertainty is too great to justify an award; although that headline is seen to be far off the mark, it does serve to suggest a constructive exercise of considering the properties of alternative award rules. A legitimate award, indicated by the *OCRegister* logic and rhetoric, would be given if and only if the school exceeded the growth target by the margin of error; then the award would not be due to chance and not subject to *OCRegister* criticism. In his justified protest of the *OCRegister* series, Haertel (2002) states: "Your reporting suggests that only schools that exceeded their target by 20 points should get rewards." Properties of such an award system can be examined, and some examples are examined here.

For example, suppose a more stringent GPA award rule withheld awards from those schools that are currently eligible but do not exceed their school API targets by at least 20 points. Another alternative award rule would (conservatively) apply that school 20 point surcharge also to the subgroups. Table 4 provides probability calculations on the properties of these amended award rules (in the rightmost columns "Applying *OCRegister* MOE") to illustrate the tradeoffs between false positives and false negatives for a typical elementary school (upper portion) and typical high school (lower portion).

Insert Table 4

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), and for grades 9-11 each student increases k percentile points on Math and Reading and k-1 percentile points on the other 3 tests (Lang, Science, Social Science). 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; Rogosa, 2002c does the CDE GPA award calculations for these two schools.)

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 (e.g. no improvement, "moderate" improvement, "large" 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). Applying a margin of error adjustment to the award criteria does markedly lower that (already often small) probability. But because most schools are making good improvement (c.f. Rogosa, 2003 for documentation of the continuing improvement in California student performance) the consequences of non-zero false positive probabilities are minimal.

The middle school in example 2 from the "margin of error folly" section provides an additional note on false positives. For this school the probability that statistical variability alone producing an award, $P\{\text{GPA award} \mid \text{no real improvement}\}$, is .077. To illustrate the role of the subgroups criteria (and the minimal information about awards from just the standard error of the school API), note that this probability would be .226 if awards were based solely on the school API, without the additional subgroup criteria.

The marked consequences of adding the margin of error adjustments to the award criteria are seen from considering false negatives. For calibration, consider the P3 row to indicate "moderate" real improvement (incrementation corresponding to about 20 API points) and the I4 row to indicate "stronger" real improvement (incrementation corresponding to over 30 API points). For the moderate improvement scenario the false negative probabilities are a

little less than .5 under the existing GPA award but soar under the margin of error adjustments to as high as .985 using subgroups and even to .90 just applying the adjustment to the school score. For strong improvement (I4), the false negative probabilities are small for GPA: .05 for the high school and .2 for the elementary school. These probabilities increase by .2 if the school margin of error adjustment is applied and exceed .6 if the adjustment is also applied to the subgroups.

One important issue relating to false negatives is the claim that “small schools [have] an advantage in winning awards” (OCR 8/11/02, c.f. OCR 8/12/02). The fallacy of a small school advantage (also claimed by Kane & Staiger, 2002) lies in the neglect of false negatives. A small school having made no real improvement 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 made substantial real improvement (which so far has been the more likely event) has statistical uncertainty as its foe, in that a false negative result may occur more often than for a large school. An imperfect illustration using Table 2 compares the elementary school example (n=350) versus the high school example (n=1115). For true improvement 29 points the false negative probability $P\{\text{no award} \mid \text{strong real improvement}\}$ is more than twice as large for the smaller school, .13 versus .29. (c.f., Rogosa 2002d demonstrations of even larger differences in false negatives between the two school sizes from cleaner comparisons between elementary schools of similar subgroup configuration.)

The short summary is that false positives in GPA awards can be lowered further by more stringent rules, but the cost is a large increase in false negatives (i.e. lower probability of award for a school that really improved). If false positives aren't much of a problem (which is the case if most schools are making good improvement), then measures to reduce those further constitute an unwise policy option.

Table 4.
Comparing Award Rules: Probabilities of Award Eligibility

Elementary School Example. CDS 19643376011951
n= 350, CA Rank = 5, API = 613, growth target = 9, s.e.(API) = 13.7,
Significant Subgroups: SD, Hispanic, White

Incrementation (real improvement)	API	CDE GPA P{API&Subgr> Target}	Applying ORegister MOE	
			P{API-20& Subgr>Target}	P{API-20& Subgr-20>Target}
P0	610	0.0655	0.0080	0.0028
Base (I0)	613	0.1002	0.0169	0.0036
P1	615	0.1275	0.0234	0.0054
I1	621	0.2446	0.0597	0.0196
P2	624	0.3111	0.0849	0.0309
I2	630	0.4590	0.1857	0.0774
P3	634	0.5321	0.2602	0.1180
I3	640	0.6515	0.3995	0.1963
P4	642	0.7136	0.4766	0.2588
I4	647	0.7927	0.5992	0.3639
P5	651	0.8639	0.7105	0.4752
I5	658	0.9299	0.8566	0.6345
P6	661	0.9564	0.9017	0.7275
I6	668	0.9832	0.9665	0.8647

High School Example CDS 15635291530708
n= 1115, CA Rank = 5, API = 609, growth target = 10, s.e.(API) = 7.8,
Significant Subgroups: SD, African-American, Hispanic, White

Incrementation (real improvement)	API	CDE GPA P{API&Subgr> Target}	Applying ORegister MOE	
			P{API-20& Subgr>Target}	P{API-20& Subgr-20>Target}
P0	605	0.0015	0.0000	0.0000
Base (I0)	609	0.0097	0.0002	0.0000
P1	613	0.0307	0.0002	0.0000
I1	618	0.1457	0.0025	0.0000
P2	622	0.2700	0.0145	0.0002
I2	626	0.4480	0.0525	0.0052
P3	629	0.5737	0.1047	0.0150
I3	634	0.7207	0.2432	0.0512
P4	638	0.8717	0.4515	0.1532
I4	644	0.9555	0.7725	0.3917
P5	648	0.9792	0.8825	0.5647
I5	653	0.9935	0.9690	0.7792
P6	655	0.9932	0.9830	0.8152
I6	662	0.9982	0.9987	0.9405

References

Brody, J. E. (2003). 'Normal' blood pressure: health watchdogs are resetting the risk. *New York Times* August 12, 2003.

Carlin, B. P., & Louis, T. A. (2000). Bayes and empirical Bayes methods for data analysis (2nd ed.). New York: Chapman & Hall.

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>

Haertel, E. H. (2002). Letter to Editor, Orange County Register, published August 18, 2002, with title: State test program among most reliable in nation.

Hill, R. (2001). The Reliability of California's API. Center for Assessment, February 2001.
available at www.nciea.org

Gladwell, M. (2003). Making the Grade. In *The New Yorker*, Talk of the Town, September 15, 2003.

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

Lehman, E. L., & Casella, G. (1998). Theory of point estimation (2nd ed.). New York: Springer-Verlag.

Morris, C. N. (1983). Parametric empirical Bayes inference: theory and applications. Journal of the American Statistical Association, 78, 47-55.

The Orange County Register:

August 2002 series available from

http://www.ocregister.com/features/api/text_version/index.shtml

Sunday August 11, 2002

Test scores unreliable: Error margin means state can't precisely measure how schools are doing, but the cash still flows. By Keith Sharon, Sarah Tully Tapia and Ronald Campbell.

API's error margin leaves a lot to chance: Mathematical imprecision could lead to inaccurate interpretations. By Ronald Campbell.

Monday August 12, 2002

Rules hurt diverse schools: Groupings create more hurdles than chances for educators in the hunt for state money. By Keith Sharon, Sarah Tully Tapia and Ronald Campbell.

Wednesday August 14, 2002

Lawmakers urge changes in school testing law: Flaws uncovered in Register probe prompt calls for reform, but governor's office defends system. By Keith Sharon, Maria Sacchetti, Sarah Tully Tapia And Kimberly Kindy

Friday August 16, 2002

State testing expert says API margin of error is insignificant Leader who helped design index calls it as accurate as possible. By Sarah Tully Tapia and Keith Sharon.

Thursday, October 17, 2002.

State will not cite error margin in API scores: Missing data on results to be released today would reveal numbers' precision, By Keith Sharon and Sarah Tully.

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

Rogosa, D. R. (2001a). Year 2000 Update: Interpretive Notes for the Academic Performance Index. California Department of Education, Policy and Evaluation Division, October 2001.
available from <http://www.cde.ca.gov/psaa/apiresearch.htm>

Rogosa, D. R. (2001b). Year 2001 Growth Update: Interpretive Notes for the Academic Performance Index. California Department of

Education, Policy and Evaluation Division, December 2001.
available from <http://www.cde.ca.gov/psaa/apiresearch.htm>

Rogosa, D.R. (2002a). 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. (2002b). Accuracy Is Important, in Testing Programs and in Reporting. Education Writers Association Meetings (Is School Reform Working in California?), Stanford Calif., September 5, 2002.

Rogosa, D. R. (2002c). 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. (2002d). 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/ksresst.pdf>

Rogosa, D. R. (2002e). Accuracy of API Index and School Base Report Elements. California Department of Education, Policy and Evaluation Division. December 2002.
available from <http://www.cde.ca.gov/psaa/apiresearch.htm>

Rogosa, D. R. (2002f). Year 2000 Update: Accuracy of API Index and School Base Report Elements. California Department of Education, Policy and Evaluation Division. December 2002.
available from <http://www.cde.ca.gov/psaa/apiresearch.htm>

Rogosa, D. R. (2003). Four-peat: Data Analysis Results from Uncharacteristic Continuity in California Student Testing Programs California Department of Education, Policy and Evaluation Division. September 2003.
available from <http://www.cde.ca.gov/psaa/apiresearch.htm>

Appendix A: Computational Details for California API index, GPA awards, and ORegister Margin of Error

To compute the API, start with a Stanford 9 test and transform the national percentile rank into quintiles: 1-19, 20-39, 40-59, 60-79, 80-99. The quintiles are assigned values 200, 500, 700, 875, 1000; an individual's API score on a single test is the value for the attained quintile. For any collection of students, the API component score for a single test (e.g. Reading) is the average, over the individuals, of these values (any missing test scores are imputed by the mean of the group). The resulting scores are combined across tests; API scores in grades 2-8 are a combination of Math (.4), Reading (.3), Language (.15), and Spelling (.15) whereas API scores in grades 9-11 are a combination of Math (.2), Reading (.2), Language (.2), Science (.2) and Social Science (.2).

The school API growth target for GPA awards is a rounded version of $40 - \text{API}/20$, an arithmetic form of 5% toward the interim state goal of API 800 (for schools below API 800). For example, the school-wide yearly growth target for a school with a 600 API is 10 points. The API target is simply the previous year API plus the growth target. To receive a GPA award, targets for the numerically significant subgroups in the school must also be met; for subgroups the growth target is $4/5$ of the school-wide improvement target. In addition, for the 2000-2001 award cycle minimum growth targets of 5 points for the school and 4 points for subgroups were imposed.

ORegister calculations. For completeness, consider the variants of the tallies based on the margin of error that indicate the claimed $1/3$ or 35% numbers.

Version 1. Set the requirement to be 20 points (the *ORegister* mean margin of error) above the API GPA award target. Then the proportion of GPA award schools passing (combining 1999-2000, 2000-2001 award cycles) is $(3263 + 1785)/(4526 + 3200) = .653$.

Version 2. Almost equivalently, do the calculation school-by-school, using the criteria $1.96 \times (\text{standard error for that school's API score})$ above the API target. Then the proportion of GPA award schools passing is $(3230 + 1726)/(4526 + 3200) = .641$.

Version 3. Framing the calculation in terms of improvement rather than moving past the API school target for GPA award, then the requirement would be $1.3 \times 1.96 \times (\text{standard error for that school's API score})$ above the previous year API, and the proportion of GPA award schools passing is $(3357 + 1895)/(4526 + 3200) = .68$.

Appendix B: Technical Details for Probability Calculations

1. Posterior Distribution for Gaussian variables.

The basic statistical facts for what is termed the normal/normal model can be found in Carlin and Louis (2000, sec. 1.5.1, eqs. 1.6, 1.7 and sec. 3.3.1) and Lehman and Casella, (1998, sec. 4.2). The likelihood specification (following notation in Morris, 1983) is that $Y_i | \theta_i \sim N(\theta_i, V_i)$, a Gaussian distribution with mean θ_i and variance V_i for units (e.g., schools) $i = 1, \dots, k$. Specifying the (prior) distribution of θ over units as $\theta_i | \mu \sim N(\mu, A)$ yields the distribution of the unknown parameter given the data:

$$\theta_i | y_i \sim N(B\mu + (1 - B)y_i, (1 - B)V_i), \quad (B1)$$

where $B = V_i / (V_i + A)$.

2. Blood pressure example

The DPB example serves to illustrate the simplest form of calculations based on Equation B1. In this example θ_i is the perfectly measured DBP for individual i , and for the DPB measurement $V_i = (6.12)^2$. The prior distribution of true DPB in the adult male population has $\mu = 82$, $A = (11.5)^2$. Consequently, $B = .221$, and for the DBP observation of 101, the posterior distribution is $N(96.8, 29.21)$ and $P\{\text{true DPB} > 89.5 | \text{DBP} = 101\} = .9117$.

3. Details on the three school examples

Start with the quantities in Table 1: standard errors and margins of error. For school example 1, the *OCRegister* hypothetical school, the margin of error for the yearly API is stated to be 16.3 (OCR 8/11/02), implying that the standard error of the school API, $se(\text{API})$, is 8.32 (= 16.3/1.96). To obtain the *OCRegister* margin of error for year-to-year improvement, the method (credited to Richard Hill) is to multiply the yearly margin of error by 1.3 to obtain 21.2 (1.3*16.3 = 21.2). For the California schools in examples 2 and 3, the $se(\text{API})$ values for each year, $se(\text{API}_1)$ and $se(\text{API}_2)$, are obtained for each school each year from bootstrap calculations described in Rogosa (2002e,f). The margin of error values shown for each of the school scores are simply $1.96*se(\text{API}_i)$. The margin of error for improvement is calculated as $1.3*1.96*[(se(\text{API}_1)^2 + se(\text{API}_2)^2)/2]^{1/2}$.

For the probability calculations shown for each school example, the parameterization is in terms of year-to-year improvement; thus for school i y_i is the observed year1, year2 improvement, θ_i is the true improvement and V_i is the (error) variance for improvement. The prior distributions are calculated separately by school-type and award cycle, from the improvement data over all schools, with the values shown below (part 4). Special attention needs to be given to the calculation of the V_i ; in Table B1 results for the posterior probabilities are shown for three versions of V_i , described by the column headings: overlap, OCR 1.3, and independence.

The probabilities reported in the text and Table 1 are taken from the column labeled OCR 1.3. For this column $V_i = 1.69 * [(se(API_1)^2 + se(API_2)^2) / 2]$, following the *OCR Register* method of multiplying $se(API)$ by 1.3 to obtain standard error of improvement. The independence column, using $V_i = [(se(API_1)^2 + se(API_2)^2)]$ has larger mistake probabilities than in OCR 1.3, and represents an upper bound for these calculations. The column labeled overlap takes into account the within-school correlation between year1 and year2 scores and the partial overlap of student populations (e.g. grades 2-6) in year1 and year2. Setting the score correlation at .75 and the proportion of students present both years at 2/3 would indicate $V_i \approx [(se(API_1)^2 + se(API_2)^2) / 2]$. This quantity is used in the column labeled overlap, in which the mistake probabilities are the smallest.

Table B1

Probability calculations for school examples with alternative calculations of V_i

	P{true change ≤ 0 observed data}		
	overlap	OCR 1.3	independence
Ex1	.0049	.0189	.0259
Ex2	.0105	.0341	.0449
Ex3	.0025	.0080	.0104

	P{true change < target observed data}		
	overlap	OCR 1.3	independence
Ex1	.0768	.125	.1404
Ex2	.0777	.1328	.1511
Ex3	.0207	.0386	.0445

4. Calculations for Table 2.

The prior distributions used for the school examples and the Table 2 results are calculated separately by schooltype and award cycle, using API data from all California schools. Values of μ and A for the prior distributions are calculated from the improvement data over all schools. The value used for μ is the mean improvement, and the value of A is the variance of observed improvement corrected for the error variance in the API scores. Table B2 shows those prior distributions in the form NormalDistribution[mean, sd] .

Table B2: Distributions for true API year1-year2 improvement.

	Elementary	Middle	High
1999-2000	[38.23, 22.67]	[21.58, 18.82]	[15.19, 18.95]
2000-2001	[20.94, 22.15]	[12.88, 19.25]	[2.388, 18.12]

The values in Table 2 accumulate the $P\{\text{true improvement} \leq 0 \mid \text{data}\}$ computed for each school, using the independence calculation for V_i . This upper bound for the mistake probabilities overstates the error rate for awards in a manner most favorable to the *OCRegister* claims, and the school examples above would indicate these are a factor of four larger than probabilities computed with the partial overlap specification.

Data description for calculated probabilities. Two probabilities are calculated for each GPA award school at each award cycle: no improvement, and failure to meet growth target. Over the collection of schools, distributions of these probabilities have extreme shapes, with only a relatively few schools having sizeable probabilities. Percentiles of these distributions (expressed in parts per thousand) are given in Table B3. As with Table 2 quantities, these probabilities use the independence assumption in calculation of variance improvement, therefore representing an upper bound.

Table B3. Percentiles of school distributions for probability calculations

		percentiles of $1000 * P\{\text{true change} \leq 0 \mid \text{observed data}\}$								
		10	20	30	40	50	60	70	80	90
Elem	1999-2000	.0013	.022	.131	.436	1.19	2.68	5.84	12.5	29.9
	2000-2001	.025	.413	1.91	5.22	10.9	19.1	32.0	51.8	93.8
Middle	1999-2000	.0008	.022	.226	.813	2.44	5.44	14.6	31.4	65.5
	2000-2001	.0029	.196	1.16	3.08	7.48	13.8	27.9	49.9	102
High	1999-2000	.00003	.0002	.038	.352	2.20	6.37	14.7	36.0	94.6
	2000-2001	.024	.334	1.64	4.54	11.8	23.8	46.0	83.3	148

		percentiles of $1000 * P\{\text{true change} < \text{growth target} \mid \text{observed data}\}$								
		10	20	30	40	50	60	70	80	90
Elem	1999-2000	.039	.383	1.35	3.60	7.43	14.8	26.6	47.7	88.8
	2000-2001	.811	4.98	15.5	29.8	49.5	73.6	113	170	242
Middle	1999-2000	.089	.839	2.96	8.43	22.1	42.3	76	119	197
	2000-2001	.582	5.53	17.6	39.1	63.9	96.6	15.5	212	327
High	1999-2000	.032	.305	2.71	10.5	27.8	54.0	114	187	243
	2000-2001	.583	5.53	17.6	39.1	63.9	96.6	15.5	212	327

Figure captions

Figure 1. Non-verbal depiction of blood pressure parable. Man (center) has diastolic blood pressure reading 101. Doctor (left) heeding the *Orange County Register* margin of error "has no idea" of whether the man is hypertensive. Statistician (right) explains the probability is above .9 that the patient's perfectly measured DBP is at least 90.