Educational Accountability: Fair and Balanced

Resource Index for presentation
September 7, *Seminar on Testing,*
Hechinger Institute and Center on Education Policy

David Rogosa
Stanford University
rag_at_stat.stanford.edu

Premise:

Most "experts" in the educational research community that you as journalists would reasonably rely upon for expertise in assessment and accountability issues cannot supply such. Arising from this dearth of knowledge on statistical issues key to accountability systems (or even large scale assessments) is the opportunity for many leading figures in educational research to substitute their own ideological (anti-testing) biases for the facts or to bash testing programs for self-promotional purposes. All educational researchers best left behind?

Accountability is not a bad thing, but it can be done badly. And that's where statisticians (should) come in, to insure that the policy directives are implemented in a defensible form.

Policy Research and Journalism Vignettes

- **The Volatility Scam**
  Claims of "volatility" in the school-level scores from testing programs by Linn and Haug (2002) and by Kane and Staiger (2002) represent a serious threat to defensible policy uses of test scores in school accountability systems. However, such claims are based on blunders at the level of high school statistics instruction. See Confusions about Consistency in Improvement (especially intro examples in section 1.2); additional, perhaps less accessible, material specific to Kane-Staiger in Irrelevance of Reliability Coefficients to Accountability Systems
- **"Margin of Error" Nonsense and the Orange County Register Debacle**
  The *margin of error* is a misunderstanding of elementary statistical concepts that leads to hilarious assertions. Sadly, last August the *Orange County Register* based their series of attacks on the California API on this nonsense: chief experts/charlatans Richard Hill and Thomas Kane. See the "Commentaries on the Orange County Register Series" at the API Research Page ; in particular the "High School Intern and the API Dollars" in What's the Magnitude of False Positives in GPA Award Programs? and the "Blood Pressure Parable" in Application of OCR "margin of error" to API Award Programs
- **Sanctions are Not the Flip-side of Awards**
  Award programs, such as California API GPA have false positives and false negatives, and these are not symmetric. Basing sanctions on a failure to reach award criteria is undesirable. In other words, where should the "the benefit of the doubt" be applied? Properties of award programs are discussed in various documents on the API Research Page
- **Accuracy of Individual Scores**
  Properties of individual student scores, such student percentile rank scores from standardized tests that go to parents and schools and which are also sometimes used for high stakes decisions, are typically described by test reliability coefficients. Unfortunately, reliability coefficients are one of the dumbest ideas ever and provide little useful information. Various documents and analyses for the accuracy of individual scores (including analyses of the CAT/6 and Stanford 9) are provided on the Accuracy Guide page
  Simplest place to start is the Shoe-Shopping Example.
- **Demographics are far from Deterministic**
  The California Teachers Association (and other critics of testing programs) seek to undermine the credibility of assessment programs with slogans such as "It's All Zip Codes" and renaming the API as the "affluent parent index". Many policy researchers (e.g. California Budget Project) feed this misrepresentation with unthoughtful correlational and multiple regression analyses. Reasonable data analysis shows that schools (and students) with similar demographic composition have very different educational performance. See the analyses in the *Interpretive Notes* series on the API Research Page.
- **NCLB, Where Accountability Came to Die?**
  A teaser for forthcoming work, should the question mark should be removed?
  Wise man statement:
  "It is a bad system to punish people when you set standards they can't possibly make," said Roy Romer, superintendent of the Los Angeles Unified School District, the largest school system in the state. (*Los Angeles Times*, Aug 16)

Discussion Item. Rebutting bad research, Process and policy?
Is *null set* the best and only answer? What do and what should journalists do after reporting in good faith on demonstrably incompetent research? Contrast education with reporting on medical research (e.g. *New York Times* Tuesday Health section).
Education examples: charter schools, teacher credentialling
Public Forum on School Accountability "*A Better Student Data System for California*"

Acknowledgements Support for the research reported here has been provided by

- the California Department of Education, Policy and Evaluation Division.
- the Educational Research and Development Centers Program, PR/Award Number R305B60002 and Award Number R305B960002-01 as administered by the Office of Educational Research and Improvement, 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,

THE VOLATILITY SCAM (from Rogosa *Confusions about Consistency in Improvement*) brought to you by Linn and Haug (2002) by Kane and Staiger (2002) and others

Example A: Extreme Volatility? Consider a set of 5 schools with four years of test data (scores are in the California API scale so year-to-year improvement of 50 points is very strong). The four years of API data produce the following results for year-to-year improvement.

	Improvement		
School	Yr1toYr2	Yr2toYr3	Yr3toYr4
A	40	50	50
B	40	40	50
C	50	40	40
D	59	41	49
E	45	45	45

Each school has healthy improvement each year of approximately the same amount. School officials and parents in these schools would cheer. Yet, remarkably, LH would determine 0% stability, 100% volatility for these school scores ($r_{LH} = 0$). KS, whose methods use the first three years of data, would determine $p_{KS} = .062$, indicating 94% of change nonpersistent, again extreme volatility! A lingering question from this example is: If the above represents "volatile", then can consistent, stable improvement ever be identified?

Example D: Up-and-Down, LH stability. LH draw conclusions regarding patterns of scores in which schools initially improve and then decline, thus giving up the gains. In the example below, schools initially improve, then flatten out, then decline, such that the overall improvement over the four years is exactly 0.

	Improvement		
School	Yr1toYr2	Yr2toYr3	Yr3toYr4
A	39	7	−47
B	39	−3	−54
C	49	−3	−41
D	49	−1	−36
E	44	2	−44

Yet for these schools $r_{LH} = .977$, and thus LH would determine great stability for these year-to-year improvements school scores! An up-and-down example for the KS procedure

	Improvement	
School	Yr1toYr2	Yr2toYr3
A	40	−40
B	40	−50
C	50	−50
D	50	−40
E	45	−45

KS would obtain 0% volatility

Even though the initial improvements in year 1 to year 2 are erased by the declines year 2 to year 3, leaving overall improvement from year 1 to year 3 of exactly 0,; $r_{KS} = 0$ and thus $p_{KS} = 1$.

Contrary to TESTTALK CEP October 2002, good consistency in improvement

Table 2.4

Consecutive Improvement for SD Subgroups in High SD Elementary Schools (n=2045)

Three year Improvement 1999-2001

ImpLevel 1999-2000	Number exceeding ImpLevel	Proportion of those improving 2000-2001	Amount of Improvement 2000-2001 {lowest decile lower quartile median upper quartile}
0	1922	0.816	-11.9 7.6 26.8 48.2
25	1516	0.807	-13. 6. 25.3 46.
50	874	0.767	-18.4 2.1 21.3 43.8
75	375	0.72	-22. -3.7 17.4 42.1
100	129	0.667	-27.5 -6.9 18.6 42.2

Fourth-year Improvement, 1999-2002

ImpLevel 1999-2000 and 2000-2001	Number exceeding both ImpLevels	Proportion of those improving 2001-2002	Amount of Improvement 2001-2002 {lowest decile lower quartile median upper quartile}
0	1569	0.829	-10.1 7.7 27.2 45.3
25	763	0.81	-12.9 5.6 26. 43.9
50	174	0.77	-18.1 3.2 25. 47.1

"Margin of Error" Nonsense *if it could be, it is*

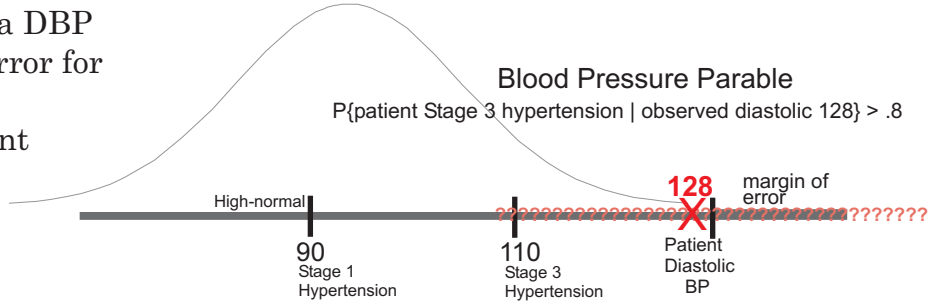
references: OCRegister Series, CDE Sept 2002, The Misunderstanding of Confidence Intervals (forthcoming)

The blood pressure parable reveals 'margin of error' folly
CA schools falling within their 'margin of error' have probability .97 and larger of positive true improvement--should we 'round' .03 up to 1.0?
Different award rules will give different results; Do we care about False Negatives?

Blood Pressure Parable

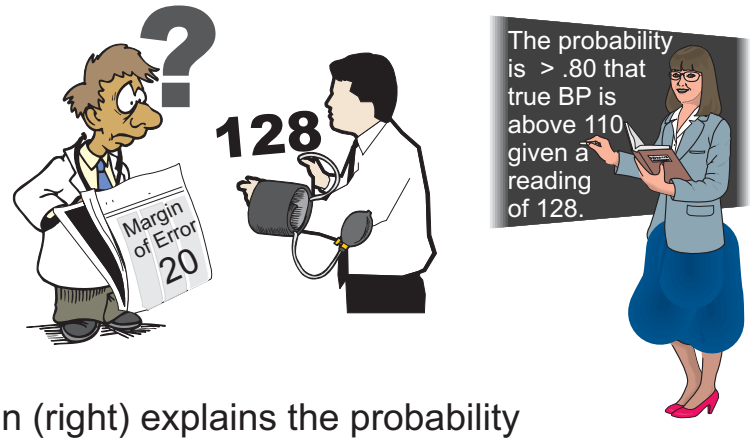
Consider this artificial setting, using diastolic blood pressure (DBP) for hypertension diagnosis. The error of measurement in diastolic blood pressure is assumed to have standard deviation 10.2 to yield a margin of error of 20 points.
A diagnosis of Stage 3 hypertension, DPB 110 and above, will lead to drug intervention with beta-blockers, diuretics etc. Following the margin of error 'logic,' no DPB reading below 130 is *really* Stage 3 hypertension. The headline would be: *Billions of Dollars wasted on diuretics and beta blockers,* and the lead would read: "Doctors have **no idea** whether millions of diagnosed patients are really in Stage 3 hypertension, yet the drugs keep flowing..."

This parable considers a patient with a DBP reading of 128, within the margin of error for Stage 3 hypertension. Does the doctor really have *no idea* whether this patient



is properly diagnosed in Stage 3 hypertension? The relevant calculation is :
Probability{true DBP at least 110 given measurement of 128}. That probability is greater than .8 (depending on the details of the population distribution etc). As the patient, would odds of 4 to 1 that you are in Stage 3 hypertension provide useful information for therapy or does the 20 point margin of error make it "impossible to tell"? [Small reality check: the 10.2 value is set high for convenience, and real diagnosis is based on multiple BP measurements.]

A numerical result closer to that seen for California schools is obtained for a diagnosis of Stage 1 hypertension (90 or above) and a patient DPB reading of 108.. Again the reading is within a margin of error (20 points) of the target but Probability{true DBP at least 90 given measurement of 108} = .94. Are odds of 15 to 1 consistent with "it's impossible to tell"?



Non-verbal depiction of blood pressure parable. Man (center) has diastolic blood pressure reading 128; doctor (left) heeding the margin of error "has no idea" of whether the man is Stage 3 hypertensive; statistician (right) explains the probability is about .80 or more (depending on the details of the artificial scenario) that the patient's perfectly measured diastolic BP is above 110.

School Examples: Margin of Error vs Probability Calculations

Back to the API context where it is shown that the disconnect between the margin of error and any reasonable probability statement is considerably larger than that shown in the Blood Pressure Parable. Examples, all telling the same story, follow below. Bottom line: application of the margin of error to API scores produces nonsense.

Example 1. Elementary School 19647336018253 1999-2000
Year-to-year improvement of 31 points is 22 points above it's growth target. Margin of error for improvement is 34.9 points.

	1999	2000		Margin of error
API	616	647	Growth target for 2000 was 9	API 27.8, 25.9
se(API)	14.21	13.2	Received GPA award $40,262	improvement 34.9
n	349	355		

Calculations:

P{true change ≤ 0| observed data} < .01
P{true change ≤ 9| observed data} < .04

Thus the probability exceeds 99% that the true improvement is greater than 0, and the odds are better than 25 to 1 that the true improvement exceeded the year 2000 growth target. Yet because the observed improvement of 31 points is less than the margin of error for improvement of 34.9, we are to have "no idea" whether this school actually improved. True, .992 is not 1.0, but it is horribly wrong to round this probability down to zero (or more exactly to round .01 up to 1.0). To take a current event analogy, I believe that if one were told that the probability was above .96 that an equity would beat it's target, many would clamor for the ticker symbol.

Example 2. CDS 19644516057616 This Middle School has 3 significant subgroups (SD Hisp Wht). Year-to-year improvement of 16 points is well within the margin of error for improvement of 19.1 points. Improvement is 10 points above the growth target.

	2000	2001		Margin of error
API	685	701	Growth target for 2000 was 6	API 14.7
se(API)	7.496	----	Received GPA award $44,404	improvement 19.1
n	900	1002		

Calculations:

P{true change ≤ 0| observed data} = .0386
P{true change ≤ 6| observed data} = .14

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 . Also note that the odds are better than 6 to 1 that this school's true change exceeded its growth target of 6.

One additional aggregate comment . As an example consider the set of the 1999-2000 elementary school GPA award winners. For these 4545 schools over 75% had P{true change > 0| observed data} above .99, and over 90% had P{true change > 0| observed data} above .97. Yet the margin of error acolytes assert "no idea" of whether many of the schools really improved.

Awards and Sanctions: False Positives and False Negatives

Medical Diagnostic Test Context

2x2 diagnostic accuracy table (see also CDC site http://www.cdc.gov/hiv/pubs/rt/sensitivity.htm)

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 on the following page (adapted from the exposition on the CDC web page). 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 status is denied due to statistical variability in the scores, despite a (specified) level of underlying ("real") improvement. The tradeoff between false positives and false negatives is the important policy decision in the formulation of an award or sanction system.

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

Note that P{no real improvement| award} = b/(a + b) , 1 - predictive value positive

In discussing accuracy of the award programs, a main factor is the subgroup criteria-- the descriptive phrases that I have used in prior discussions are “saved by the subgroups” and “herding cats”. The herding cats metaphor is that it’s unlikely that a set of cats will all move in the same direction (past the growth target) by accident, but a strong enough probe (real improvement) may persuade all the cats to move in unison. The number of significant subgroups is an important factor: having many subgroups in a school tends to make false positives less likely and make false negatives more likely (the more cats, the tougher to herd them). Furthermore, statistical variability in the school and subgroup scores makes growth targets far more formidable than these might appear because of the subgroup requirements (as each of the subgroups has larger uncertainty than the school index). To have high probability that all subgroup scores will meet the criteria requires underlying improvement that far exceeds (blows through) the seemingly modest growth target.

Richard Rothstein, *New York Times* column January 24, 2001, “Flaws in Annual Testing” http://www.nytimes.com/2001/01/24/national/24LESS.html citing my work, covered false positives,false negatives, the saved-by-the-subgroups message and the effects of multiple subgroups on false negatives for the API awards.

> Professor Rogosa estimated that if awards were based on school averages alone, over one-fourth of schools with no gains would still qualify. But California avoids this problem by insisting that schools succeed for low-income students and each minority group, as well as schoolwide. This reduces the chances of undeserved awards, because simultaneous false gains in each group are unlikely. But the reverse is also true. Schools deserving rewards will be more likely to lose them because if any group fails as a result of random events or sampling error, the school will be disqualified. Diverse schools will fail more often; they have more subgroups where false declines can occur.

Probability of award given no improvement.
Four Elementary School Examples. Schools All Decile 5 with 3 Numerically Significant Subgroups: Socioeconomically Disadvantaged, Hispanic, White. Four school sizes
The entry PrAPI&Subgr>Targ1 in the "Base" row for each school indicates the probability that statistical variability alone (i.e., null improvement I0) will result in school eligibility for GPA. The derived display extracts those quantities for the four examples.

		Base (I0)	
	se(API)	PrAPI&Subgr>Targ1	PrAPI>Targ1
n=148	20.2	.131	.313
n=244	14.26	.094	.274
n=350	13.7	.101	.273
n=491	11.1	.084	.226

KS Small School Advantage??? Probability GPA Award: Elementary School Examples

	True Improvement				
	0	29	41	Comp1	Comp2
smallest n	0.131	0.447	0.596	0.342	0.441
(~5th percentile n=148)					
small n	0.094	0.620	0.797	0.445	0.563
(~20th percentile n=244)					
medium n	0.101	0.714	0.890	0.509	0.627
(~50th percentile n=350)					
large n	0.084	0.822	0.961	0.576	0.669
(~80th percentile n=491)					

Comp1: Prob{true improvement = 0} = 1/3, Prob{true improvement = 29} = 2/3
Comp2: Prob{true improvement = 0} = 1/3, Prob{true improvement = 41} = 2/3

The High School Intern and the API $$$

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

Smart High School Statistics Student + Publicly Available Information = correct answer

The setting is California, July 2002. A newspaper preparing a series on the API has a summer intern who has recently completed one of the fine (one semester) high school statistics courses. The intern is asked: "Do you think the finding that a third or more of the GPA award schools made no real improvement--P{no real improvement | GPA award} > .3-- is reasonable?"

The High School statistics student makes the following presentation to the newspaper's reporters: "In my class we learned about about false positives and false negatives, like the chart on the CDC website. To get the information I needed for the API, I did the following:

a. I went to Rogosa's "Plan and Preview" report on the CDE site. The information I get there is P{award | no improvement} for two examples, a typical elementary school with a .1 probability, and a typical high school with a .01 probability. So the best I can do is reason that middle schools will be in between elementary and high, and since there are more elementary schools, that probability might average out to .07 or .08.
b. But I'm not done because that's not the probability I was asked about. From my statistics course, I know a little about conditional probability
   P{no improvement | award} = P{award | no improvement}*P{no improvement}/P{award} .
   So I need a bit more information.
c. From newspapers or CDE site I see that the GPA award rate for 1999-2000 was just about 2/3.
d. From Rogosa's Year 2000 Interpretive Notes on the CDE site, 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 overestimates 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 P{no real improvement | GPA award} ≈ .07*.1/.67 = .01 . For 2000-2001 awards, less awards were given, and from Rogosa's Year 2001 growth Interpretive Notes I see that at least twice as many schools showed no improvement compared to 1999-2000. Combine those factors, and for 2000-2001 I get that P{no real improvement | GPA award} is at least .03. Average them out to an overall .02, and 1/50 is a whole lot less than 1/3."

ACCURACY OF INDIVIDUAL SCORES

1.SHOE SHOPPING and the RELIABILITY COEFFICIENT

Dedicated to Al Bundy

A man who cares as much about good measurement as he does about his own children.

Try this on

1. A population of male and female shoe-shoppers who have *true* shoe sizes between size 5 and size 15 (e.g. the small sizes are female feet translated to the male shoe-size scale).

2. Mr. Bundy measures each shopper's shoe size as either too large or too small with equal probability.

- On a good day Mr. Bundy misses the correct shoe size by one-half size too big or one-half size too small.
- On other days Mr. Bundy misses the correct shoe size by a full size too big or a full size too small.

In each case the shoe size measurement error has mean 0 (overall and at each level of shoe size) and is uncorrelated with actual shoe size. (Standardized testing analogs might be to the full battery and abbreviated battery versions of testing.)

3. The accuracy of shoe fitting on the *good day* is poor (as most wearers would notice a half-size misfitting), and on the other days the accuracy is totally unacceptable (as a full-size misfitting would presumably be unwearable).

4. The reliability coefficient for Al Bundy on the *good day* is .973 (better than any standardized test, even though accuracy is poor). The reliability coefficient for Al Bundy making errors of a full shoe size is .902 (comparable to many standardized tests, even though accuracy is unacceptable). {Pub version .94, .81, .871 for error processes, emp. dist}

---

2. ACCURACY OF INDIVIDUAL NATIONAL PERCENTILE RANK SCORES

How Accurate Are the STAR National Percentile Rank Scores for Individual Students?–An Interpretive Guide: Versions 1.0 and 2.0

Some Teasers: Stanford 9

♦What are the chances that a student who "really belongs" at the 50th percentile in the national norms obtains a score more than 5 percentile points away from the 50th percentile?

For Math grade 9 it's 70%, for Reading grade 4 it's 58%.

♦What are the chances that two students with "identical real achievement" obtain scores more than 10 percentile points apart?

For two students really at the 45th percentile for Math grade 9, it's 57%.

For two students really at the 45th percentile for Reading grade 4, it's 42%.

Some Teasers: CAT/6 Survey

- What are the chances that a student who "really belongs" at the 50th percentile in the national norms obtains a score more than 10 percentile points away from the 50th percentile?

For Math grade 2 it's 57%, for Reading grade 2 it's 47%.

- What are the chances that two students with "identical real achievement" (at the 50th percentile in the national norms) obtain scores more than 20 percentile points apart?

For Math grade 2 and Language grade 2 it's 40%, for Reading grade 2 it's 29%.

Version 2.0, CAT/6 Survey

The specific accuracy quantities shown below are computed for an "average" student, a student who under perfect measurement would score at the national 50th percentile:

Hit-rate Accuracy(10) . Probability that the student score is within 10 points of the true value (here, 50th percentile).

Retest Accuracy(20). Probability that two scores (retest, or scores from identical students) are within 20 points of each other.

	Hit-rate Accuracy(10)	Retest Accuracy(20)
Math		
grade 2	.43	.60
grade 4	.57	.74
grade 8	.59	.76
Language		
grade 2	.42	.60
grade 4	.53	.72
grade 8	.60	.77
Reading		
grade 2	.53	.71
grade 4	.61	.79
grade 8	.68	.84

"Retention" Calculations (or HSEE)

Textbook example: Passing Standard set at 30th percentile of Observed Norms

1. False Fails: Probability below 30th percentile for a student with true score at the stated percentile of the observed norms distribution.

	Percentile						
	35	37.5	40	42.5	45	47.5	50
test reliability							
0.6	0.413	0.372	0.334	0.298	0.264	0.233	0.204
0.8	0.378	0.323	0.272	0.227	0.186	0.151	0.12
0.9	0.33	0.258	0.196	0.145	0.104	0.072	0.049
0.95	0.267	0.179	0.113	0.067	0.037	0.019	0.01

2. False Pass: Probability above 30th percentile for a student with true score at the stated percentile of the observed norms distribution.

	Percentile				
	15	17.5	20	22.5	25
test reliability					
0.6	0.209	0.258	0.308	0.357	0.406
0.8	0.126	0.18	0.239	0.303	0.369
0.9	0.053	0.097	0.158	0.233	0.318
0.95	0.011	0.033	0.078	0.151	0.251

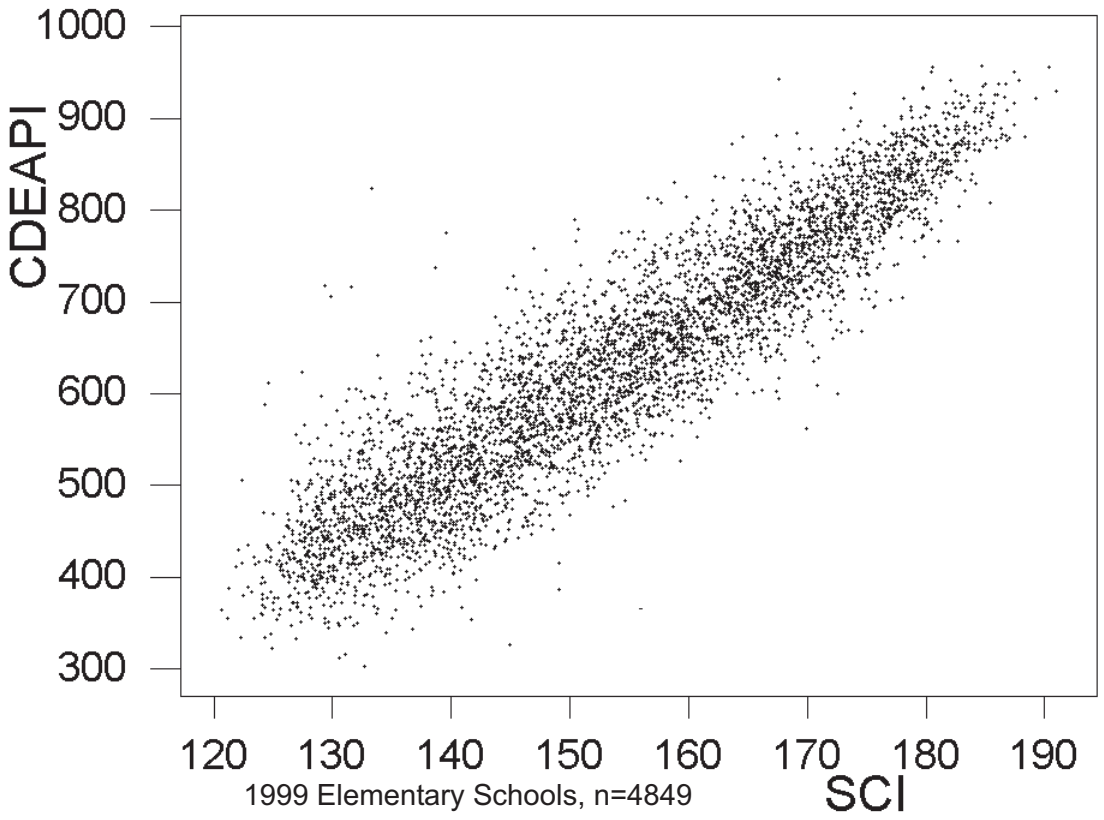
Passing Standard set at 30th percentile of Observed Norms

False Fails:Probability below 30th percentile for a student with true score at the stated percentile of the observed norms distribution.

	Percentile						
	35	38	40	43	45	48	50
ReadGr4	.191	.106	.067	.048	.029	.013	.007
MathGr9	.325	.234	.19	.138	.11	.075	.059

Are Demographics Determinant?

	Elem	Middle	High	
Correlation of SCI and API =	**0.924**	**0.951**	**0.946**	=> R^2 = .85, .90



Range of Similar Schools

Range Similar School API for all Elementary Schools [from Interpretive Notes... 11/00]

Variable	N	Mean	Median	Q1	Q3	Minimum	Maximum
RangeSimSAPI	4849	**281.50**	**277.00**	**243.00**	**304.00**	154.00	522.00

Range Similar School API for all Elementary Schools at each State Decile

	Range Similar School API					
CA Decile	N	Mean	Median	Q1	Q3	Minimum	Maximum
1	478	326.24	**294.00**	**279.75**	374.00	209.00	522.00
2	490	322.36	**301.00**	**276.00**	374.00	209.00	522.00
3	477	307.44	**290.00**	**260.50**	354.00	200.00	522.00
4	488	295.78	**286.00**	**253.00**	317.00	205.00	522.00
5	480	284.57	**279.00**	**249.00**	303.75	198.00	522.00
6	487	271.97	**272.00**	**247.00**	292.00	203.00	464.00
7	485	270.79	**265.00**	**246.00**	288.00	181.00	407.00
8	491	270.81	**265.00**	**243.00**	290.00	182.00	389.00
9	480	252.38	**258.00**	**217.00**	280.00	154.00	349.00
10	493	214.22	**208.00**	**192.00**	220.00	165.00	349.00

The Statewide result at the top of the table says that half the Elementary Schools show a range of their Similar Schools API scores of at least 277 points, and 75 percent of elementary have a range of their Similar Schools API scores of at least 243 points. As for elementary schools the statewide decile categories typically span 40-45 API points, the median range 277 represents a span of 6 (or more) statewide deciles.

The second part of the table breaks down the Range Similar School API for each State Decile. For the 490 elementary schools placed in the second state decile, half of those schools have Range Similar School API of over 300 points, and 75 percent of those schools have Range Similar School API of over 275 points.

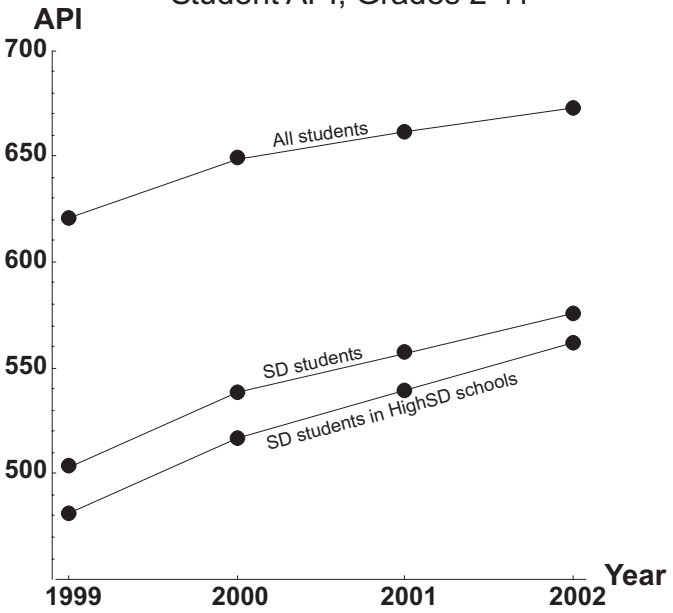
From forthcoming
"Four-peat: Results from Uncharacteristic Continuity in California Student Testing Programs"
D Rogosa 9/03

API Scores 1999-2002, All California Students

Four Years of California API Scores for Indicated Grade Ranges and Student Groups

Grades Included	1999API	2000API	2001API	2002API	Improvement '99-02
Grades 2-6					
All Students	619.85	657.57	676.26	693.45	73.6
SD Students	505.73	550.1	576.03	602.08	96.35
SD Students in HighSD Schools	483.39	527.64	555.84	585.	101.6
Grades 2-8					
All Students	622.09	655.03	671.08	684.93	62.84
SD Students	505.38	544.75	567.55	589.41	84.03
SD Students in HighSD Schools	482.6	521.79	547.06	571.94	89.34
Grades 9-11					
All Students	617.15	631.34	634.39	636.87	19.72
SD Students	494.34	510.79	512.58	519.15	24.81
SD Students in HighSD Schools	475.2	489.1	493.58	501.15	25.95
Grades 2-11					
All Students	620.84	648.88	661.56	672.47	51.63
SD Students	503.27	538.06	556.96	575.86	72.6
SD Students in HighSD Schools	481.52	517.01	539.41	561.94	80.42

Student API, Grades 2-11



Improvement 1999-2002,
Grade-by-grade for Student Groupings

Grade	Students All	Students SD	Students SD in HighSD Schools
2	77.88	104.38	110.31
3	83.62	109.75	116.31
4	83.75	107.	112.75
5	67.75	87.81	91.5
6	54.75	70.88	72.88
7	40.62	52.75	52.44
8	28.62	38.56	38.5
9	24.25	29.69	30.31
10	16.38	19.12	20.19
11	17.38	21.75	24.56

Socioeconomically Disadvantaged (SD) Student Comparisons

SD versus non-SD students

Grades Included	1999API	2000API	2001API	2002API	Improvement '99-02
Grades 2-6					
SD Students	505.73	550.1	576.03	602.08	96.35
non-SD Students	721.24	789.51	802.84	816.12	94.87
Grades 2-11					
SD Students	503.27	538.06	556.96	575.86	72.6
non-SD Students	704.29	753.95	761.51	769.44	65.15

SD in HighSD schools comparison (the economic integration question)

Grades Included	1999API	2000API	2001API	2002API	Improvement '99-02
Grades 2-6					
SD Students in HighSD Schools	483.39	527.64	555.84	585.	101.6
SD Students in non-HighSD Schools	600.17	639.97	659.95	672.98	72.81
SD Students	505.73	550.1	576.03	602.08	96.35
Grades 2-11					
SD Students in HighSD Schools	481.52	517.01	539.41	561.94	80.42
SD Students in non-HighSD Schools	565.78	594.42	605.94	614.07	48.29
SD Students	503.27	538.06	556.96	575.86	72.6

Artificial Cohort Comparisons

Cohort formed from Grade 2 and 3 Students in 1999

Grades 2-3 In 1999	1999API	2000API	2001API	2002API	Improvement '99-02
All Students	622.81	655.39	663.84	685.55	62.74
SD Students	514.43	549.45	558.36	586.06	71.62
SD Students in HighSD Schools	493.8	527.15	537.12	565.55	71.75

Improvement 1999-02, Cohorts formed Grade-by-grade for Student Groups

Grade in 1999	All Students	SD Students	SD Students in HighSD Schools
2	50.12	57.75	58.69
3	75.38	85.88	85.56
4	60.12	67.5	65.5
5	50.75	55.44	54.12
6	2.5	11.75	14.62
7	-14.75	-19.38	-12.44
8	29.62	34.19	45.38