Handouts and Overheads for

Statistical Blunders by the Proponents and Opponents of Accountability

Presentation April 16, 2004
Education Writers Association 2004 Seminar

David Rogosa
Stanford University

Margin of Error and March Employment Report

The March Employment Report showed a payroll increase of 308,000 jobs (4/2/2004).

On that strong number markets were up big, weak dollar was up 5%, even Kerry was saying nice things.

But..... What about statistical uncertainty?

Press reports [CNBC, Leesman] showed standard error for job growth 176,800.
[display 90% CI for 0 jobs growth would be -290,000 to 290,000]

increase appears wobbly because it's relatively small compared to 130,000,000 total jobs
(Compare year-to-year school achievement)

apply 95% margin of error criteria and discover margin of error for jobs growth is about 346,000.
Thus the margin of error would dictate that: "the feds have NO IDEA whether any real job growth occurred"

Burden of proof required by 99% NCLB confidence intervals would imply any jobs growth less than 410,200 would be indistinguishable from zero.
[converse of 'close enough is good enough']

So who has it right--those moving billions of dollars on these job numbers or the margin of error piety?

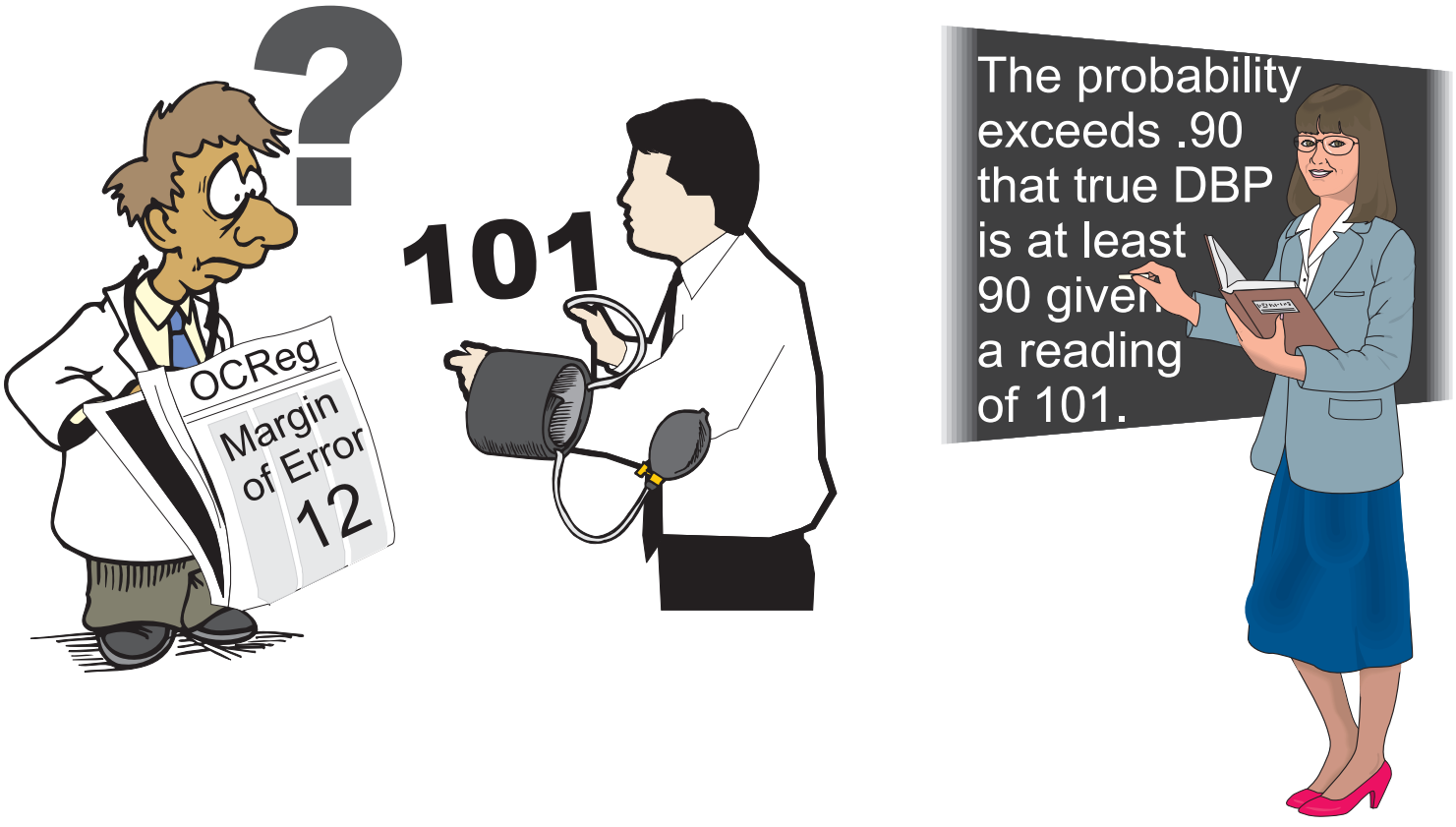


Figure 1. Non-verbal depiction of blood pressure parable. Man (center) has diastolic blood pressure reading 101. Doctor (left) heeding the Orange County Register margin of error “has no idea” of whether the man is hypertensive. Statistician (right) explains the probability is above .9 that the patient’s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

The Focus of the Reforms

The second aspect of the statement of purposes is a declaration of the types of students and schools that are the object of attention or reform. This is the essence of the defining of the construct—including the definition of the schools that merit improvement status. This might seem like an unnecessary step—the focus is on schools where the students aren't learning and the test scores reflect it, what's to talk about? It turns out that this is only true if (1) all results were based on several years of data; (2) if there were no student mobility, and (3) if all schools drew from the same types of communities. Since these three conditions are hardly ever met, it does make a difference how one looks at the scores for a school; different—and equally rational—ways of looking at the results yield quite different results for many schools. Hanushek and Raymond (2002) confirm the need to think carefully about this topic; as they see it, this is just one of the ways that accountability systems fail to show any evidence of a rational design or plan.

One conception uses a simple two-by-two table to look at four main models or approaches (Carlson, 2001). Two other documents (Gong et al., 2001; Hill, 2001) expand on that conception by drawing out the underlying assumptions about the nature of schooling and school reform, showing how each would identify schools of different types and presenting what is known at this point about their technical characteristics. Figure 4 presents an abbreviated description of the three most popular models.

FIGURE 4. THREE COMMON WAYS OF REPRESENTING PROGRESS OF SCHOOL ACHIEVEMENT RESULTS

School-level Model	Example of the Type of Data Used	Advantages	Disadvantages
A. Status	68% of the 4th graders meet the Proficiency level in reading in 2001.	Easy to understand; school scores are quite reliable.	Scores indicate the level of current achievement only.
B. Change, Successive Groups, or Improvement	Difference between percent meeting the Proficient level for two years for students at the same grade level (e.g., change of 4% between 2001 and 2002 for 3rd graders in Lincoln School).	Easy to compute.	Differences between cohorts are large and random, leading to wide variability and lack of reliability (Kane & Staiger, 2001, 2002; Linn & Haug, 2002).
C. Growth or Longitudinal[9]	Average progress from one grade to the next (or from fall to spring) for a set of students (e.g., average growth from 3rd to 4th grade in Lincoln School is 33 scale score points).	Allows inferences about school effectiveness.	Requires assessments at adjacent grade levels; scores are also quite unstable from year to year; and some believe that this method requires vertically-scaled assessments.

All accountability systems have to define the construct by answering the question, "What kinds of schools should be identified for improvement?" It is an obvious exaggeration, but one might say that good schools are very similar, but bad schools are bad in different ways. This is why it is important for the accountability system to clearly communicate the definition of a school identified for improvement. There are two main philosophical positions, both of which are rational and defensible, but lead to the identification of different schools:

[9] The distinction is usually made between true-longitudinal (where the same students are tracked over a year, or more), and quasi-longitudinal (where a whole group of students is tracked over time, but not individuals are present in all calculations, e.g., where the third grade for a school in one year is compared with the fourth grade for the next year).

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

LH 0% stability, KS 94% volatility

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

LH 98% stability

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

From: The NCLB "99% confidence" scam: Utah-style calculations
David Rogosa Stanford University

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

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

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

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

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

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

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

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

federal officials staunchly defend these "close enough is good enough" NCLB confidence interval procedures.

From: Federal mandate: A complex statistical formula allowed them to meet their targets
The Salt Lake Tribune December 18, 2003
By Ronnie Lynn

> "There's a requirement in the law that accountability decisions be statistically valid and reliable," said Celia Sims, special assistant in the U.S. Department of Education's Office of Elementary and Secondary Education. "Confidence interval is a common statistical technique, and states are allowed to use them."

Deconstruction of this statement is important. It is true that interval estimation (confidence intervals) is a basic tool of statistical inference, and (crude) confidence intervals are prominently taught in elementary statistics courses. But useful tools can be egregiously misused, and the use of the confidence interval notions in these state NCLB plans is indefensible, violating any possible interpretation of "statistically valid and reliable".

A toothpick metaphor illustrates my point that useful tools can be misused and misapplied:
A sharp toothpick does a fine job of cleaning teeth. However, the sharp toothpick is less well-suited to picking one's nose, is dangerous if used to clean ears, and possibly catastrophic if used to remove a fallen eyelash from the cornea (epithelium).

To sum up, confidence intervals have an important role in statistical inference but are not appropriately applied in the NCLB "close enough is good enough" state plans.

Table 3
"99% Confidence" Posterior Probabilities for True Proportions Proficient Satisfying AMO for School and Three Subgroups and Two Subjects Expressed in Parts per Million

	AMO = .19	
	k = 2.33	k = 2.58
School A	1	.1
600 students		
groups {200,200,200}	{23,23,41}	{21,21,42}
all pS = .35		
School B	6	.3
350 students		
groups {200, 50,100}	{23, 3,20}	{21, 1,21}
pS = {.35,.35, .5}		
	AMO = .45	
	k = 2.33	k = 2.58
School C	5	1
600 students		
groups {200,200,200}	{74,74,94}	{72,72,95}
all pS = .5		

Artificial School B is a school composed of 350 students who belong to one of 3 subgroups: group 1 with n = 200 and pS =.35, group 2 with n = 50 and pS = .35, and (a typically higher achieving subgroup) group 3 with n = 100 and pS = .5. For the school, 46 (k=2.33) or 43 (k=2.58) proficient students out of 350 is "close enough" to satisfy the performance goal of .19 proportion proficient. Configurations of number of proficient students in each subgroup meeting the "99% confidence" procedures are shown below the probabilities that the true proportions proficient meet the AMO = .19.

School C, an artificial school of size 600 composed of 3 subgroups each with 200 students and each subgroup having statewide proportion proficient .5 (pS) has AMO = .45. For 99% confidence k = 2.33 a configuration of proficient students in the three subgroups of {74,74,94} is considered close enough to the AMO. For this configuration, the probability that the true proportions proficient meet the stated AMO for both school and subgroups on both subjects is 5 millionths.

From:
A School Accountability Case Study: California API Awards and the *Orange County Register* Margin of Error Folly
To appear in Richard Phelps, Ed. *Defending Standardized Testing*.
Mahwah, NJ: Lawrence Erlbaum, 2004.

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	

DIVERSITY PENALTY????

From
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.

Effects on probability of meeting AMO of additional subgroups

	Number of Non-overlapping Subgroups				
n	0/1	2	3	4	5
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

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 Mathematics Performance Goal .16 with Probability .99

	Number of Non-overlapping Subgroups				
n	0/1	2	3	4	5
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

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 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.

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			
	2	3	4	5
n				
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

NCLB FULLY QUALIFIED TEACHER MANDATE
From: Teacher Credentials and Student Progress: What do the data say?

Spurious Correlation Table
Table 4C. Static School-Level Tables (Traditional) 2001 Scores

Mean ECrED_01 by State Decile and Schooltype

CARank01	E	H	M
1	18.985	20.811	25.468
2	14.996	16.872	19.500
3	12.515	15.691	17.615
4	11.088	12.099	14.000
5	8.973	13.229	11.211
6	6.574	10.297	9.802
7	5.695	7.740	9.180
8	4.566	8.038	7.018
9	4.099	7.085	7.472
10	3.511	6.663	5.784

Table 1:
Comparison of Student Scores in High ECRED Schools versus no ECRED Schools

	Improvement, 2000-2001 School Year	
	Students in Schools with ECRED > 15 in AY 2000-2001	Students in Schools with ECRED = 0 in AY 2000-2001
All Students, Grades 2-6		
API2k	552.7	747.7
API01	582.3	764.1
Improvement	29.7	16.4
SD Students, Grades 2-6		
API2k	509.4	617.9
API01	542.6	639.1
Improvement	33.2	21.3
SD Students in High SD Schools, Grades 2-6		
API2k	503.4	582.6
API01	537.4	606.2
Improvement	34.0	23.6

Table 2A:
Deciles Breakdown for 2000-2001, All Students Grades 2 - 6

All Students Grades 2 - 6
Improvement by Decile 2000-2001 for High and Low Emergency Credential

State Decile	ECRED > 15 API2k	API01	Imp	ECRED = 0 API2k	API01	Imp
1	444.3	484.1	39.8	454.4	478.3	23.9
2	517.3	548.1	30.8	520.6	554.6	34.
3	567.7	595.5	27.8	571.8	600.1	28.3
4	611.2	634.4	23.2	613.6	640.3	26.7
5	649.8	671.4	21.6	653.	672.8	19.8
6	691.1	704.9	13.9	692.2	710.9	18.7
7	730.7	748.2	17.5	732.7	749.7	17.
8	769.2	783.4	14.2	774.3	790.3	16.
9	820.	829.2	9.2	821.	832.6	11.7
10	879.3	892.4	13.1	881.6	889.3	7.7

=-=

Table 2C:
Deciles Breakdown for 2000-2001, SD Students in High SD Schools Grades 2-6

SD Students in High SD Schools Grades 2 - 6
Improvement by Decile 2000-2001 for High and Low Emergency Credential

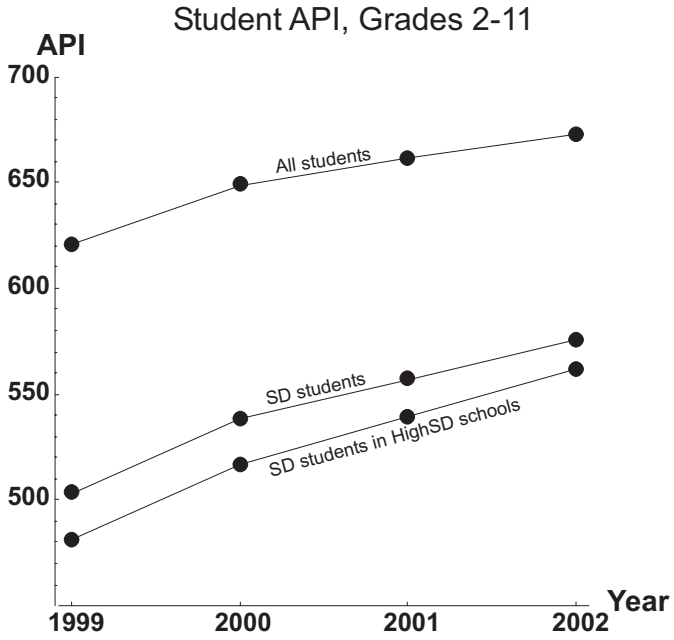
State Decile	ECRED > 15 API2k	API01	Imp	ECRED = 0 API2k	API01	Imp
1	440.3	481.5	41.1	448.4	472.5	24.1
2	505.	537.4	32.5	503.8	538.	34.2
3	548.5	577.1	28.6	541.8	571.2	29.4
4	581.3	609.9	28.5	569.8	602.2	32.4
5	614.2	642.6	28.4	601.7	621.	19.4
6	654.2	670.2	16.	648.2	662.8	14.6
7	710.	728.4	18.4	689.6	695.9	6.3
8	758.2	761.	2.8	741.5	753.	11.6
9	807.9	790.	-17.9	797.2	801.3	4.

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

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



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

Class not Race?

Table 2. Comparisons Across Subgroups

API Students in Grades 2-6

Subgroup	1999API	2000API	2001API	2002API	Improvement '99-02
All	619.85	657.57	676.26	693.45	73.6
Afam	520.35	562.97	587.54	610.02	89.66
Asian	749.41	782.79	806.	827.03	77.61
Hisp	498.28	542.67	572.23	599.74	101.46
White	744.13	778.44	794.31	805.33	61.19

SD API Students in Grades 2-6

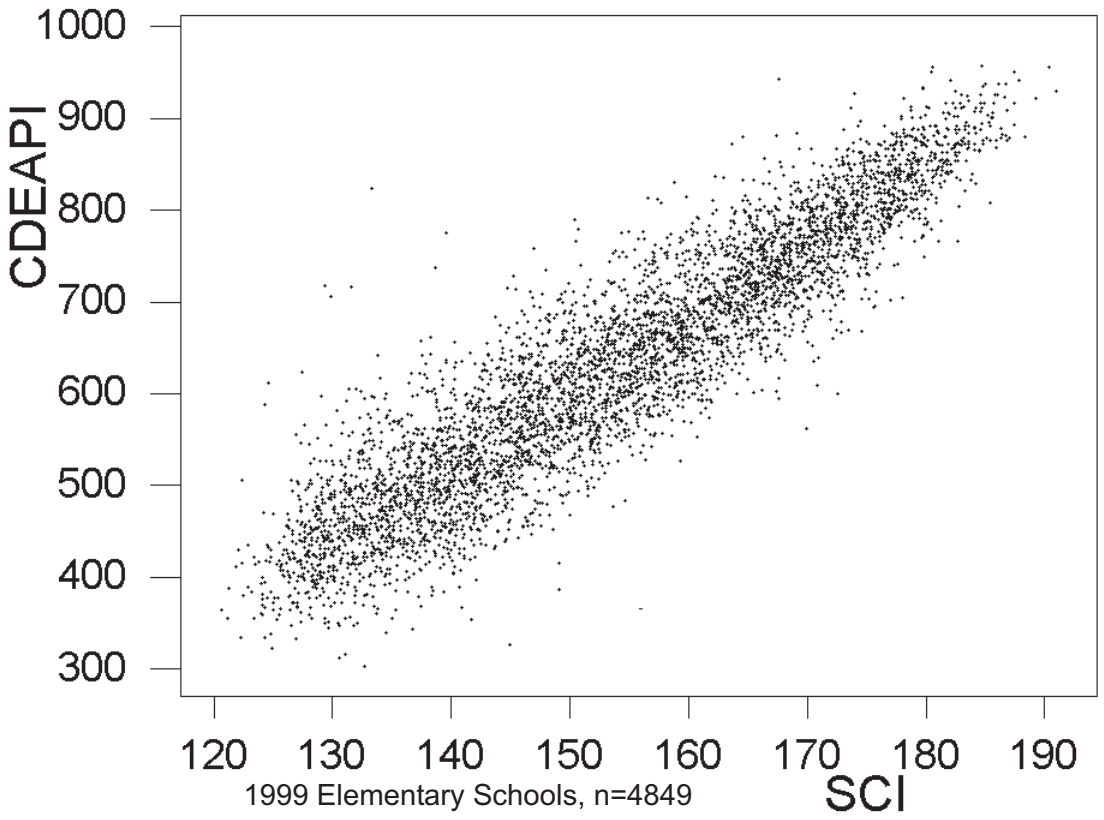
Subgroup	1999API	2000API	2001API	2002API	Improvement '99-02
All	505.73	550.1	576.03	602.08	96.35
Afam	479.56	520.75	544.35	568.8	89.24
Asian	633.82	673.02	704.85	734.67	100.85
Hisp	465.21	512.34	543.99	574.48	109.27
White	612.15	652.64	671.93	687.78	75.63

SD API Students in HighSD Schools, Grades 2-6

Subgroup	1999API	2000API	2001API	2002API	Improvement '99-02
All	483.39	527.64	555.84	585.	101.6
Afam	465.96	505.81	530.85	556.59	90.63
Asian	608.11	646.84	680.21	711.91	103.79
Hisp	454.72	501.48	533.82	566.3	111.58
White	578.04	619.28	640.91	660.12	82.07

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.