Plan and Preview for API Accuracy Reports

David Rogosa
Stanford University
July 3, 2002

Preamble

About two years ago I began work on the statistical properties of the Academic Performance Index (API) and the associated award programs (e.g., GPA and AB1114). To date, the reports that have been released consist primarily of the *Interpretive Notes* series. Those data analysis reports (available from this page) were originally envisioned as the opening chapters for the forthcoming Accuracy Reports, following the simple logic that before explaining the statistical properties (e.g., accuracy) of the API, it's useful to first understand and explain a bit about the basic index and its reporting. The initial Interpretive Notes covered the 1999 API data, with the additional update reports covering the year 2000 and the year 2001(growth) data. These reports provide descriptions and interpretation for the API scores and the improvement in the API (including subgroups). Additional topics in those reports include demonstrations that the link between demographics and school performance is much weaker than is asserted by various interest groups and that school size has almost no relation with school performance.

Thus the *Interpretive Notes* series represents a first step (and a bit of a detour) on the road to the Accuracy Reports. The plan for releases during this summer of 2002 has the structure of two main reports with various updates and add-ons:

Accuracy of API Index and Base Report Elements

This first report will present the statistical properties of the API and the accuracy resulting from the use of the API index to determine the Statewide (decile) Rank and the Similar Schools Rank. The main report is based on the 1999 data, with a separate update report using the year 2000 data.

Accuracy of Award Programs

The second report examines the accuracy resulting from the use of the API index to determine awards for the Governor's Performance Awards (GPA) and the eligibility for the (now discontinued) Certificated Staff Performance Incentive (AB 1114) programs. The main report is based on the 1999 data, with a separate update report using the year 2000 data.

Also, additional specialized reports on statistical topics arising in these Accuracy Reports will appear on the CRESST website.

Overview/Preview of Results

The purpose of the body of this note is to provide a "quick tour" of the type of results that will be contained in these forthcoming Accuracy Reports. Previous press reports in major newspapers have contained brief treatments of some of these accuracy topics. In the January 17, 2001 *Los Angeles Times* (pp. A9-10), Martha Groves writes in some detail about the accuracy of the quantities in the API reports. And Richard Rothstein's column in the January 24, 2001 *New York Times* (p. A16) discusses my results on the accuracy of the award programs.

Accuracy of API Index and Base Report Elements

Standard errors of school API. The first topic is the standard error (a measure of statistical uncertainty) of the school-level API index. The table on the next page first shows descriptive statistics for the standard error of the API-- s.e.(API) -- separately for each school type and below that the median standard error for each state decile. Further display of s.e.(API) is provided by the plots for Elementary and High Schools. Regardless of school type, schools have a wide range of values for s.e.(API), One calibration for these s.e.(API) values, which may be helpful to readers familiar with educational testing, is to speak in terms of a reliability coefficient; even for a small elementary school (having s.e.(API) of nearly 20), the reliability of the API score exceeds .98.

A major feature of s.e.(API) is the dependence on the number of students (denote by n) contributing to the school's API index. In California, Middle Schools have about twice the number of API students as Elementary Schools, and High Schools have about three times the number as Elementary Schools. The table shows that the median standard errors for each school type follow quite closely the ratio indicated by relative school sizes (proportional to square root of relative sizes). Furthermore, the plots of s.e.(API) versus 1/Sqrt(n) for Elementary and High Schools show the strong dependence of the standard error on the number of students. (To calibrate those plots note that axis points .1, .05, .025 correspond to n = 100, 400, 1600.)

Although the dependence of s.e.(API), on the number of students is strong, the plots also show some sizable differences for schools of the same size, mainly a result of the additional dependence of s.e.(API) on the school's API score. The plots of s.e.(API) versus API show a pattern of larger s.e.(API) for API scores in the middle of the distribution, a pattern readers with an introductory statistics course will recognize as characteristic of a proportion score. (And readers of the *Interpretive Notes* series will recall the demonstrated correspondences between the API and proportion of students above the 50th and 25th national percentile ranks.) The tables displaying the median s.e.(API) by state decile similarly show larger values for schools in the middle state deciles for each school type.

Standard Error of API (bootstrap resampling)

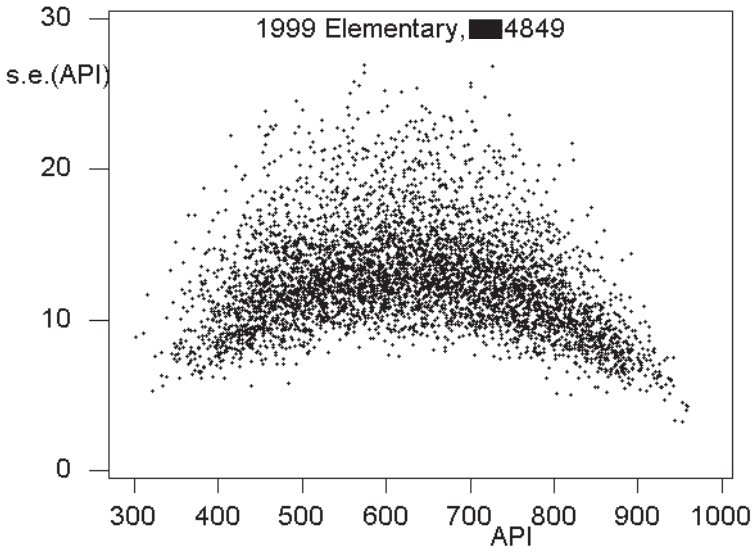
Descriptive Statistics: s.e.(API)

	N	Median	Q1	Q3	Minimum	Maximum
Elem	4849	**12.217**	10.329	14.338	3.244	27.411
Mid	1118	**8.491**	7.1906	10.005	3.687	24.975
High	837	**6.931**	5.831	8.863	2.014	23.149

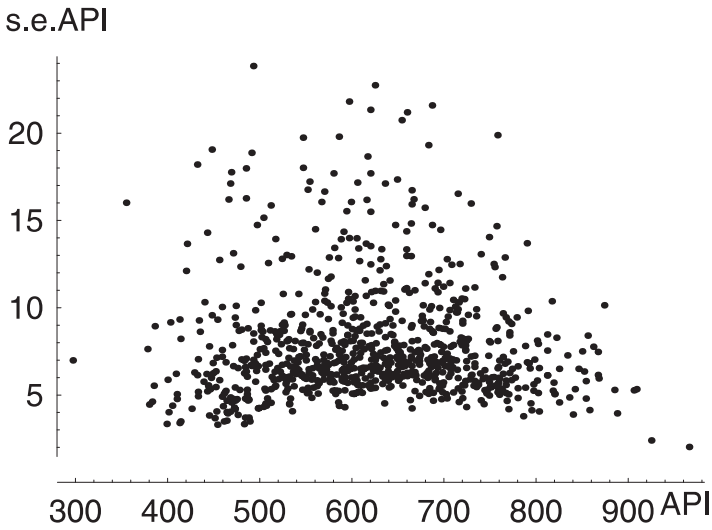
Median s.e.(API) by CARank (state decile)

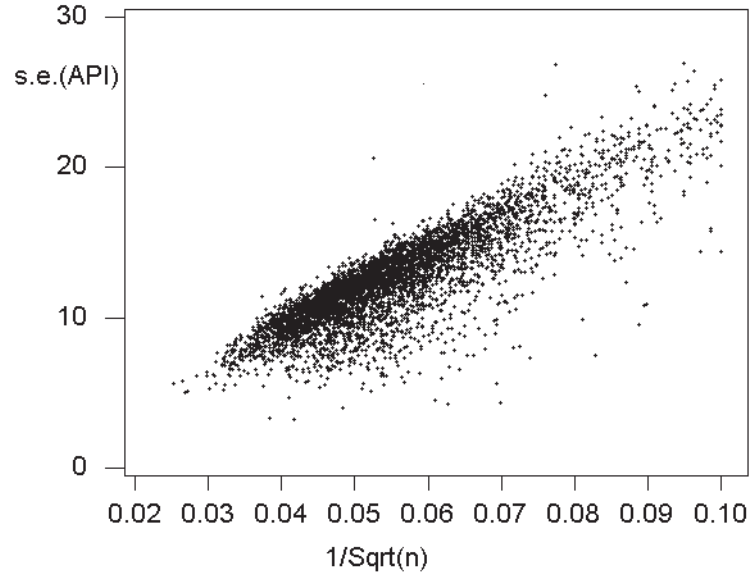
	Elem		**Middle**		**High**	
CARank	N	Median	N	Median	N	Median
1	478	10.242	110	7.627	85	5.845
2	490	11.994	111	8.103	84	6.749
3	477	12.744	110	9.032	84	7.048
4	488	13.241	115	9.219	82	6.968
5	480	13.554	111	9.295	78	7.880
6	487	13.674	110	9.145	89	7.357
7	485	13.152	111	8.690	83	7.208
8	491	12.401	115	8.489	84	7.192
9	480	11.350	110	8.223	82	6.692
10	493	8.760	115	6.485	86	6.043

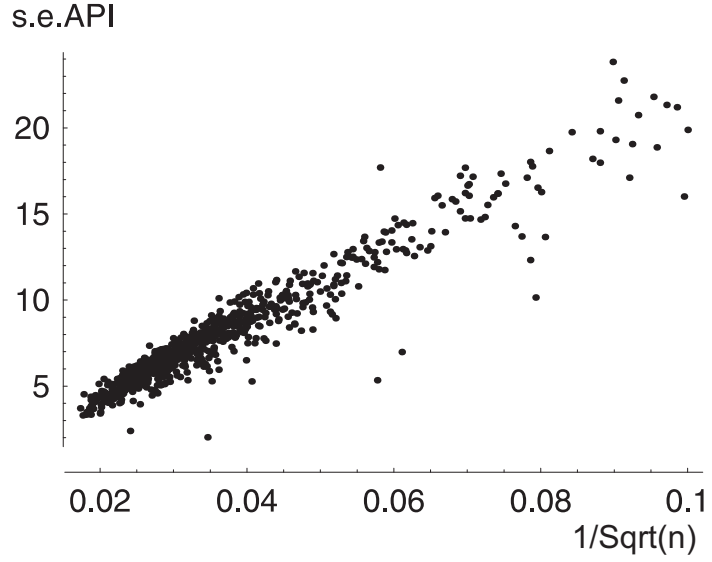
Plots for s.e.(API): Elementary Schools



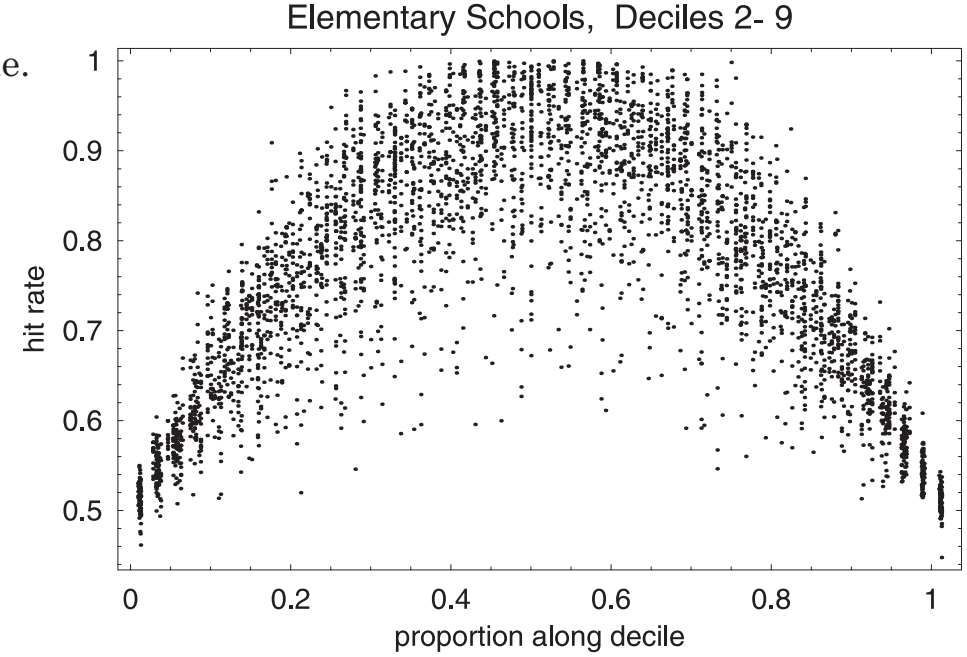
High Schools







Accuracy of Statewide Decile Ranks. The accuracy of the use of the school API score to determine the reported statewide rank is quantified by the *hit-rate* which is 1 - Prob{sampling variability in API score moves the school out of its assigned decile}. The plot below shows the hit rate for 1999 Elementary schools in statewide deciles 2 through 9; the hit-rates are estimated from a bootstrap resampling. Of course, a school with an API score near a decile boundary will have a much larger probability of statistical variability moving it's API score into a different decile; that's what motivates plotting hit-rate versus position in the decile. Almost all schools are contained within two adjoining deciles; the median hit-rate for Elementary schools is above .75.



Accuracy of Similar Schools Rank. The accuracy of the use of the school API score to determine the reported similar schools rank depends upon both the uncertainty in the target school API score and the uncertainty in API scores of the 100 similar schools. A double bootstrap procedure used to obtain category probabilities for the similar school ranks shows the effects of statistical variability on the observed similar schools rank. The examples below illustrate that the similar schools rank is best thought of as a +/- 1 decile "smear" from assigned rank. Even the relatively large High School shows considerable uncertainty in the similar schools rank.

Examples of Similar Schools Ranks Accuracy

	Elementary Schools		High School
CDS	01612596001978	30666216029805	15635291530708
CDEAPI	480	587	609
SEAPI	15.98	11.47	7.93
CDEN	241	502	1115
CDECARnk	2	5	5
CDESmRnk	5	7	3
SM01	0.0005	0	0.0065
SM02	0.01975	0	0.1125
SM03	0.13275	0.00025	0.3283
SM04	0.26425	0.00875	0.3568
SM05	0.2585	0.10425	0.1631
SM06	0.2135	0.309	0.0308
SM07	0.0805	0.39075	0.002
SM08	0.02675	0.17575	0
SM09	0.0035	0.01125	0
SM10	0	0	0

Statistical Properties of Award Programs: False Positives and False Negatives

My mantra (and the main reason I'm doing this work) is the critical importance of understanding statistical uncertainty in school accountability indices (e.g., API) and, in particular, understanding the effects of statistical uncertainty on rewards and/or sanctions based on these accountability indices. Analyses of the properties of the award programs (GPA, AB1114) are to be organized into a separate report in part because of the complexity of the required analysis, but also in large part to emphasize (or respect) the disconnect between describing the properties of the API index and the analysis of the properties of decisions made using the API index.

Most important, intuitions formed by examining the standard error of the API score are not easily transformed into conclusions about the award programs. The main factor is the subgroup criteria employed in the award programs-- in regard to the award programs the descriptive phrases that I have used in prior discussions are "saved by the subgroups" and "herding cats". The herding cats metaphor (to keep in mind for the technical development to follow) is that it is unlikely that a set of cats will all move in the same direction (past the growth target) by accident, but a strong enough probe (incrementation) may persuade all the cats to move in unison. Furthermore, statistical variability in the school and subgroup scores makes growth targets far more formidable than they 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.

The statistical approach to discussing the award programs follows ideas from medical diagnosis or screening tests. The accuracy of the award programs is quantified in terms of False Positive and False Negative probabilities. Commonly accepted medical tests have less than perfect accuracy. For example, prostate cancer screening (PSA) has a considerable False Positive rate and in tuberculosis screening, False Negatives (sending an infected patient into the general population) are of considerable concern. In the context of PSAA awards, False Positives describe the chance that sampling variability alone (no real improvement) produces award eligibility. False Negatives are calculated by inducing an artificial improvement and then observing the effects of statistical variability on award eligibility. The number of significant subgroups is an important factor: having many subgroups in a school tend to make False Positives smaller and make False Negatives larger (the more cats, the tougher to herd them).

Statistical calculation of the False Positives and False Negatives is not straightforward because subgroups overlap with each other (i.e., SD subgroup) and with the full school. Bootstrap calculations provide the most direct approach, and there are many forms these calculations could take. In the reports, accuracy properties for the statewide collection will be described using the statistical methodology of "False Discovery Rates," methods commonly used in genomics for gene-chip data. In this note the displays are

limited to analyses of award accuracy for a single exemplar school at a time.

Representing School Improvement. The device used for the calculation of False Negatives is to augment the school data by forms of individual score incrementation. The two forms used in the school examples on the final page of this note are:

Integer Incrementation (**Ik**). Every student increases k percentile points on each test.

Partial Incrementation (**Pk**). This 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). 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 Section 2 of the original *Interpretive Notes* these forms of incrementation (and their consequences for API scores) are covered in detail, with the dual purposes of explaining the API scale and providing the groundwork for these accuracy calculations.

My best attempt at a short introduction to these calculations is through the results for two "typical" schools shown on the final page. The first school is an Elementary School of median size, middle state decile in API, with 3 significant subgroups (the modal number for decile 5 Elementary Schools). The second school is a High School also of median size, middle state decile in API, having 4 significant subgroups.

The calculation starts with the actual 1999 data for the school. The "Base" row indicates the probabilities that statistical variability alone (i.e., null improvement I0) will result in school eligibility for GPA (PrAPI&Subgr>Targ1) and AB1114 (PrAPI&Subgr>Targ2). These False Positive probabilities for GPA are .10 for the Elementary School (with s.e.API = 13.7) and .01 for the High School (with s.e.API = 7.8). The False Positive probabilities for the doubled growth targets in AB1114 eligibility are much smaller: 0.035 and .0002. (The PrAPI&Subgr>Targ2 values are of interest independent of the existence of AB1114 awards as an indication of the properties of an award program with a higher threshold than is used in GPA; i.e., would the GPA program have better properties with larger growth targets?). Turning back to the theme of "saved by the subgroups," if there were no subgroup requirements the False Positive probabilities would be .273 and .103 for the Elementary School (>Targ1 and >Targ2 respectively), and for the High School these False Positive probabilities would be 0.1255 and 0.0087 (>Targ1 and >Targ2).

False Negatives represent the chances of denying award status due to statistical variability in the scores, for a specified level of underlying improvement. The I5 row indicates an underlying level of improvement of 5 percentile points on each Stanford 9 test (representing 3-4 additional questions correct) before adding on the statistical variability in the scores. A pure I5 incrementation to this Elementary School data would result in an school API of 658 (shown in the second column); that "improvement" of 45 API points from the 613 Base score is just slightly greater

than the median improvement of 42 points seen for decile 5 Elementary Schools for 1999-2000 (see section 2 of the year 2000 *Interpretive Notes*). The False Negative probability for GPA award (Targ1) for an I5 incrementation is seen from column 3 to be 1 - .9299, about 1/14. Closer correspondence to that median improvement would be between the I5 and P5 rows, yielding a False Negative probability of about 1/10.

There is a long list of technical issues, necessary justifications, and alternative calculations arising in the estimation of these award probabilities to be discussed in the actual reports. The attempt here is to introduce some small portion of these rather complex calculations in the most straightforward manner I could muster.

The plots at the bottom of the next page simply reiterate the message that one can't jump from values of the standard error of the school API score to conclusions about properties of the award programs. These plots show False Positive probabilities for groups of Elementary (left frame) and High Schools (right frame); all schools have 3 significant subgroups (the modal value) and are in the middle deciles of the statewide API distribution. Roughly, one can calibrate that a API standard error approaching 15 corresponds to a False Positive probability approaching 1/10.

Probabilities of Award Eligibility: Two School Examples.

Elementary School CDS 19643376011951

n= 350, CA Rank = 5, Sim Rank = 6, API = 613, s.e.(API) = 13.7

Sig Subgroups: Socially Disadvantaged, Hispanic, White

Incrementation	API	PrAPI&Subgr>Targ1	PrAPI&Subgr>Targ2
P0	610	0.0655	0.0198
Base (I0)	613	0.1002	0.0354
P1	615	0.1275	0.0513
I1	621	0.2446	0.1125
P2	624	0.3111	0.1576
I2	630	0.4590	0.2787
P3	634	0.5321	0.3553
I3	640	0.6515	0.4832
P4	642	0.7136	0.5540
I4	647	0.7927	0.6609
P5	651	0.8639	0.7543
I5	658	0.9299	0.8657
P6	661	0.9564	0.9097
I6	668	0.9832	0.9625

High School CDS 15635291530708

n= 1115, CA Rank = 5, Sim Rank = 3, API = 609, s.e.(API) = 7.8

Sig Subgroups: Socially Disadvantaged, African-American, Hispanic, White

Incrementation	API	PrAPI&Subgr>Targ1	PrAPI&Subgr>Targ2
P0	605	0.0015	0.0000
Base (I0)	609	0.0097	0.0002
P1	613	0.0307	0.0015
I1	618	0.1457	0.0165
P2	622	0.2700	0.0450
I2	626	0.4480	0.1352
P3	629	0.5737	0.2122
I3	634	0.7207	0.3857
P4	638	0.8717	0.6002
I4	644	0.9555	0.8335
P5	648	0.9792	0.9180
I5	653	0.9935	0.9635
P6	655	0.9932	0.9750
I6	662	0.9982	0.9925

False Positives (FP) and Standard Error API