Statistical Misunderstandings of the Properties of School Scores and School Accountability

David Rogosa
Stanford University

The work reported herein 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 are those of the author and do not necessarily reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences (IES), or the U.S. Department of Education.

In public policy settings, statisticians constantly urge that the statistical uncertainty in relevant measures be reported and incorporated into decision-making. Current work in educational assessment and school accountability provides a vivid instance of "be careful what you wish for" as, unfortunately, educational researchers' attempts to consider statistical uncertainty have caused serious missteps in the design and presentation of school accountability systems, in particular, the federal No Child Left Behind Act of 2001 (NCLB) .

This chapter provides an overview of statistical issues arising in work on educational assessment and school accountability, organized around some of the most consequential misunderstandings. What are good descriptions of statistical properties (e.g. uncertainty) of school and subgroup scores? What are the consequences of the statistical uncertainty for school accountability systems? Solid answers to these questions would seem critical for sound and credible educational policy. Accountability is not a bad thing, but it can be done badly. A real threat to the viability of school accountability is the poor quality of technical work in educational research on the design and properties of accountability systems.

The chapter proceeds via a series of nine vignettes describing misunderstandings of statistical uncertainty in educational assessment and school accountability. The flow of the exposition is from assessments of the statistical uncertainty in a group summary (school-score) to the consequences of uncertainty for properties of accountability decisions to attempts to adjust accountability criteria for the effects of uncertainty. The first two vignettes describe misunderstandings of the accuracy of scores. The next three vignettes depict the consequences of statistical uncertainty for properties of accountability systems (including wayward claims of a substantial "diversity penalty" and "small school advantage"). The next set of three vignettes concentrates on the oft-cited "margin of error" and its misapplication to accountability, as in the NCLB confidence intervals. The final vignette describes the frequent misinterpretation of correlation coefficients to indicate the strength of the relation between student demographics and educational performance.

Vignette 1. Volatility Scam, Part 1: Precision versus Reliability of a Group Score

An obvious starting point for properties of a school accountability system is the properties of a single-year school score. The school summary score constructed from the test results from the school's students can have a variety of forms: a composite such as the California API (ranging from 200 to 1000), the proportion of students scoring proficient or above that is used in NCLB, or an average national percentile rank as often reported for standardized tests. Each of these school scores (as with any measurement) contain statistical uncertainty, and one general message is that the accuracy of a measure should be judged by reference to the purposes to which the data are put, a common mistake being to ask too much of the data.

Claims of "volatility" in the school-level scores from testing programs by Kane and Staiger (2002) (also Linn & Haug, 2002, among others) represent a serious threat to defensible policy uses of test scores in school accountability systems. That is, if the volatility claims were true, then accountability decisions would not be credible. However, the misunderstandings of reliability versus precision for school scores render Kane Staiger methodology and conclusions of no value (see Rogosa 2002b, 2003a). Kane and Staiger use North Carolina and California data to demonstrate their volatility claims, with representative assertions:

> We would infer that 14 to 15 percent of the variation in fourth-grade math and reading test scores was due to sampling variation.(p.241)
>
> We estimate that the confidence interval for the average fourth-grade reading or math score in a school with sixty-eight students per grade level would extend from roughly the 25th to the 75th percentile among schools of that size. Such volatility can wreak havoc in school accountability systems. (p.236)

In reality, Kane-Staiger volatility is a disguised form of the reliability coefficient:

Kane-Staiger proportion of variance in group summary due to error =
1 – Reliability Coefficient(of group summary).

Even using the Kane-Staiger criteria, the California data show little volatility. In terms of the California Academic Performance Index (API) using 1999 data, reliability coefficients for a population of schools of a specified size are given in Table 1 (c.f., calculations in Rogosa, 2002b, section 1, Rogosa, 2002d.) For median size elementary or high schools less than one percent of the between school variance in API scores is attributable to error. Moreover for the single-grade results (and California does not report API scores by grade for reasons such as insufficient precision) which have relatively poor accuracy, little Kane-Staiger volatility would be seen, as the single-grade reliability coefficient is very high.

INSERT TABLE 1

Further results reveal that Kane-Staiger methods and results err in both

Table 1
California API Reliability Coefficients

Elementary Schools		Grade 4		High Schools	
n	reliability	n	reliability	n	reliability
150	0.982	68	0.965	500	0.991
350	0.992	79	0.970	1000	0.996
500	0.994	103	0.976	1500	0.997

directions: Kane-Staiger methods find high volatility even when accuracy is very good, and Kane-Staiger methods determine the absence of volatility even when accuracy is moderate to poor. The primary misunderstanding is a confusion between the accuracy or precision of a school score (relevant to accountability) and relative standing measures such as reliability, which are irrelevant.

Technical caricature. An artificial, analytic, demonstration of the deficient properties of Kane-Staiger volatility methodology is taken from the caricatures in Rogosa (2002b, 2003a). Start with a single year cross-section of perfectly measured school scores; "perfectly measured" indicates school scores with no statistical uncertainty, e.g. schools composed of an infinite number of students with student test scores obtained from very long tests. For the population of schools, the distribution of true measurements for the collection of schools is specified to be Discrete Uniform [498, 502], i.e., mass 1/5 at 498,..., 502. Thus, under perfect measurement the school scores would have mean 500, variance 2. The error process obscuring the perfectly measured score is specified to be the same for each school (assumption for simplicity; think of all schools being the same finite size). The error process is Discrete Uniform [–2, 2]; that is, this error process has mass 1/5 at –2, –1, 0, 1, 2, and thus the mean is 0 and error variance is 2 points. Consequently, in this caricature the standard error for a school score is 1.41, which appears small compared to the magnitude of the school score of 500. Another way of expressing the accuracy of the school score is in terms of a hit-rate: P{observed school score is no more than 1 point different from the perfectly measured score} = .6. (Note: the discrete distribution of the observed school scores for this population of schools has support on [496,504] with distribution: Pr{school score = 500 + i} = (5 – |i|)/25 .) By these criteria a score for an individual school appears to be quite accurate.

However, the Kane-Staiger assignations of volatility use different criteria than accuracy of a school score. Over this collection of schools, the observed school scores have mean 500 and variance 4, and the reliability coefficient for the population of school scores is 2/4. Thus Kane-Staiger methodology would determine 50% of the observed variance to be due to error and would pronounce the school scores to be highly volatile. Furthermore, the Kane-Staiger explanatory vehicle of a confidence interval for the average school score would extend from the 15th to 85th percentiles of the score distribution (using interpolation, probability score 497 or less, probability score 503 or more both equal .12). (Note: Rogosa, 2002b, shows how this Kane-Staiger confidence interval is also expressed as a function of the score reliability coefficient.)

Vignette 2. Volatility Scam, Part 2: Consistency of Year-to-Year Improvement

Another venue for claims of volatility in school scores is year-to-year improvement and consistency of year-to-year improvement. As with the single-year school scores in the first vignette, the claims from the educational research literature of volatility in year-to-year improvement result from methodological misunderstandings: here, as simple as a failure to understand a basic property of the correlation coefficient, invariance under translation. The mis-attributions of volatility in year-to-year improvement are consequential for accountability because these misunderstandings have influenced NCLB discussions. In particular, the CCSSO (2002) AYP report dismisses year-to-year improvement as an accountability criteria (see their Figure 4) because of "wide variability and lack of reliability (Kane & Staiger, 2002; Linn & Haug, 2002)" (p.33) and "lack of technical merit" (p.34). And recently many states have made unsuccessful initiatives for year-to-year improvement criteria in NCLB (c.f., O'Connell, 2004).

Linn and Haug (2002) present a correlational methodology for determining the "stability" or "volatility" of year-to-year improvement, and apply that methodology to school-level scores (successive fourth graders) from four years (1997-2000) of Colorado assessment data (CSAP). Linn and Haug (2002) heavily cite and build upon similar methodology and assertions of year-to-year volatility contained in Kane and Staiger (2002). A sampling of the assertions in Linn and Haug (2002):

> Year-to-year changes in scores for successive groups of students have a great deal of volatility.(p.29),
>
> It was found that the year-to-year changes are quite unstable (p.29),
>
> The estimates of improvement, however, are quite volatile. This volatility results in some schools being recognized as outstanding and other schools identified as in need of improvement simply as the result of random fluctuations.(p.35)

Similarly, the assertions of volatility in Kane and Staiger (2002) would seem to indicate a ubiquitous lack of consistency in improvement. In California, almost all improvement is found to be transient or "fleeting":

> Although the California schools tend to be larger, the data reveal slightly more volatility in the California Academic Performance Index for any given school size. For the smallest fifth of schools, the correlation in the change in adjacent years was –.43, implying that 86 percent of the variance in the changes between any two years is fleeting. For the largest fifth of schools, the correlation was –.36, implying that 72 percent of the variance in the change was nonpersistent." (p.248-9)

Rogosa (2003a) provides detailed analytic results on the deficiencies of the Linn-Haug and Kane-Staiger methodologies and also presents useful data analysis methods for consistency in improvement.

The Linn-Haug stability measure, denoted as r_{LH}, uses four successive yearly measurements to obtain a correlation indicating "stability in the two-year change scores" (Linn & Haug, 2002, p.33). Similarly, the Kane-Staiger procedure uses three successive years of data to compute a correlational measure indicating "the proportion of the change in test scores that is attributable to nonpersistent factors" (Kane & Staiger, 2002, p.247). The detailed definitions, using school scores $Y_s(1)$, $Y_s(2)$, $Y_s(3)$, $Y_s(4)$, are r_{LH} = Correlation[Y(3) – Y(1), Y(4) – Y(2)], and for the Kane-Staiger proportion persistent (i.e., the proportion not nonpersistent for directional convenience) the measure $p_{KS} = 1 + 2r_{KS}$, where r_{KS} = Correlation[Y(2) – Y(1), Y(3) – Y(2)]. Technical results from Rogosa (2003a) reveal both stability measures to be functions of the reliability of the year-to-year difference score, and thus the volatility misunderstanding is another confusion between reliability and relevant properties.

Table 2 displays small examples (taken from Rogosa, 2003a) showing year-to-year improvement for a set of 5 schools each with four years of test data (Scores are in the California API scale, yearly scores on a 200 to 1000 scale, and year-to-year improvement of 50 points is very strong. Alternatively, the reader can divide these numbers by 10 and regard these as improvement in NCLB proportion proficient on a percent scale so that improvement of 5 percent in percent proficient is strong improvement). Two different configurations are shown (Examples I and II). In Example I each school has healthy improvement each year of approximately the same amount. School officials and parents in these schools would cheer. Yet, remarkably, Linn-Haug methods would determine 0% stability, 100% volatility for these school scores ($r_{LH} = 0$). Kane-Staiger methods, which use the first three years of data, would also determine extreme volatility as $p_{KS} = .062$, indicating 94% of change nonpersistent. An obvious question: If Example I in Table 2 represents "volatile", then can consistent, stable improvement ever be identified?

Example II in Table 2 shows that both Linn-Haug and Kane-Staiger methods can indicate strong stability, but not usefully so. In Example IIa of Table 2 the initial improvements from year 1 to year 2 are erased by the declines of year 2 to year 3, resulting in no overall improvement from year 1 to year 3. Yet Kane-Staiger methods would indicate 0% volatility ($r_{KS} = 0$ and thus $p_{KS} = 1$). In the Example IIb with 4 years of data, schools initially improve, then flatten out, and then decline, such that the overall improvement over the four years is exactly 0. Yet for these schools $r_{LH} = .977$, and thus Linn-Haug methods would determine great stability for these year-to-year improvements in school scores.

INSERT TABLE 2

Even though the Kane-Staiger and Linn-Haug indices are not useful, the empirical research question of consistency in year-to-year improvement is important. The common-sense data analysis is to ask whether schools (or

Table 2
Five-school examples, Misunderstandings of Consistency in Improvement

		School				
		A	B	C	D	E
I: Extreme Volatility ?						
Improvement	Yr1toYr2	40	40	50	59	45
	Yr2toYr3	50	40	40	41	45
	Yr3toYr4	50	50	40	49	45
IIa: Up-and-Down, KS persistance						
Improvement	Yr1toYr2	40	40	50	50	45
	Yr2toYr3	-40	-50	-50	-40	-45
IIb: Up-and-Down, LH stability						
Improvement	Yr1toYr2	39	39	49	49	44
	Yr2toYr3	7	-3	-3	-1	2
	Yr3toYr4	-47	-54	-41	-36	-44

subgroups) whose scores improve one year continue to improve (and how much). Demonstrations of consistency in improvement for California schools and disadvantaged subgroups are shown in Rogosa (2003a,b). Here, consider the elementary schools in Los Angeles Unified (LAUSD). For the 418 LAUSD elementary schools in the four year period 1999-2002, 406 improved their school API in 1999-2000 (median improvement 42 points), and the proportion of those 406 also improving in 2000-2001 is 0.938 (median improvement 37 points). For these 381 schools improving both in 1999-2000 and 2000-2001, the proportion of those also improving in 2001-2002 is 0.955 (median improvement 46 points). Furthermore, for the disadvantaged subgroups in concentrated poverty schools, 341 of the 353 subgroups improved their API in 1999-2000 (median improvement 43 points), and the proportion of those 341 also improving in 2000-2001 is 0.944 (median improvement 43 points). For the 322 disadvantaged subgroups improving both in 1999-2000 and 2000-2001 the proportion of those also improving in 2001-2002 is 0.972 (median improvement 50 points). The amount of year-to-year improvement is substantial, and slightly larger for the disadvantaged subgroups.
In California, one aspect of the ongoing debate on school accountability was the contention, put forth by the California Teachers Association and others, that year-to-year improvement in API scores was well-described by a "see-saw" metaphor, in that schools with scores that showed strong gains (and achieved awards) in one two-year cycle reversed those gains in the succeeding two-year cycle. The "see-saw" represents a legitimate empirical conjecture about (the lack of) consistency in improvement, similar to empirical conclusions voiced in Linn and Haug (2002), but that is a conjecture that is just not supported by these California data.

Vignette 3. Properties of Scores and Properties of Accountability Decisions

Understanding the accuracy of school or subgroups scores (or more to the point not misunderstanding the properties) is an important first step, but far from fully informative about the properties of an accountability system. The standard error of a school score has some usefulness as a global indicator of statistical uncertainty. But the properties of scores and properties of an accountability system are (almost) two separate, but not separable topics. Knowing standard errors does not directly reveal properties of an accountability system; it is better to have more precise scores, but properties can be good even in the face of some statistical uncertainty.

For descriptive purposes Table 3 provides some results for the California API in elementary schools (1999 scores). Results are listed by state decile with the s.e.(API) entry being the median standard error for the roughly 480 schools in each decile (more detail in Rogosa, 2002d), and with the bottom and top of each decile also shown. In addition to the obvious dependence of standard error on school size, plots in Rogosa (2002d) show substantial differences in s.e.(API) for schools of the same size, mainly a result of the additional dependence on the level of school's API score, seen in Table 3 as the largest values are in the middle of the score distribution. One lesson from Tables 1 and 3 is that scores with very high reliability (> .99, see Table 1) still may have non-trivial uncertainty. Another important lesson is that accuracy can only be evaluated in terms of the use of the scores–for the California examples statistical uncertainty in the school scores allows solid determination of the state-wide decile rank, but relatively poor determination of the similar schools rank. It's not the size of the standard error, it's how you use it.

INSERT TABLE 3

The statistical approach to the properties of accountability systems follows standard ideas from medical diagnostic and screening tests. The properties are expressed in terms of false positive and false negative events, which are depicted in the chart in Table 4. 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. Table 4 shows two layered labelings for the rows (observed outcome, success or not) and columns (true state, deserving or not); the top layer pertains to NCLB accountability in which proportion proficient for schools and subgroups are compared to a performance goal (to earn AYP), and the bottom layer pertains to past California accountability in which year-to-year improvement by schools and subgroups was compared to a growth target (to earn Awards). The entries are the joint probabilities a, b, c, d. In this education context, false positives describe events where statistical variability alone produces an (undeserved) successful outcome. False negatives describe events for which success is denied due to statistical variability in the scores, despite underlying ("real") performance or improvement. The tradeoff between false positives and false

Table 3
Standard Error of California Elementary School API by State Decile

Decile	1	2	3	4	5	6	7	8	9	10
s.e.(API)	10.24	11.99	12.74	13.24	13.55	13.67	13.15	12.40	11.35	8.76
Min(API)	302	449	497	543	587	629	670	715	763	818
Max(API)	448	496	542	586	628	669	714	762	817	958

negatives is the important policy decision in the formulation of an school accountability award or sanction system. The common derived quantities from Table 4 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.

INSERT TABLE 4

Inclusion of subgroup criteria in school accountability creates a disconnect between properties of a school score and the properties of decisions from the accountability system. Statistical variability in the school and subgroup scores serves to make the accountability criteria far more formidable than they might appear, because to have high probability that all subgroup scores (each of the subgroups has larger uncertainty than the school index) will meet the criteria requires underlying performance that exceeds (blows through) the often seemingly modest target (see Rogosa, 2003c for NCLB calculations). One explanation is the "herding cats metaphor": it is unlikely that a set of cats will all move in the same direction (past the performance goal or growth target) by chance, but a strong enough probe (performance or improvement) may persuade all the cats to move in unison.

An illustration of the effect of subgroups and the lack of a strong relation between the properties of the accountability system and the standard error of the school score is provided by Figure 1. The data are California elementary schools all having 3 significant subgroups (the state modal value) and with API scores located in the middle deciles of the statewide API distribution. The false positive (FP) probability is probability of award given no improvement. For schools with standard errors around the median value, the false positive probabilities have a wide range of .04 to .16, even for this homogeneous grouping. Another property of the accountability system to note is that, if subgroups were not included in the award criteria, the false positive probabilities would be increased by approximately a factor of three.

INSERT FIGURE 1

Table 4
2x2 Diagnostic Accuracy for School Accountability: NCLB (and California API)

	Good Real Performance (Good Real Improvement)	Poor Real Performance (NO Real Improvement)
Meet AYP (Award)	TRUE POSITIVE (a)	FALSE POSITIVE (b)
Fail AYP (NO Award)	FALSE NEGATIVE (c)	TRUE NEGATIVE (d)

Figure Caption

Figure 1. Illustration of less-than-perfect relation between false-positives (FP) in school accountability and standard error of school score (s.e.API) for California elementary Schools scoring in deciles 5 or 6 and having three significant subgroups

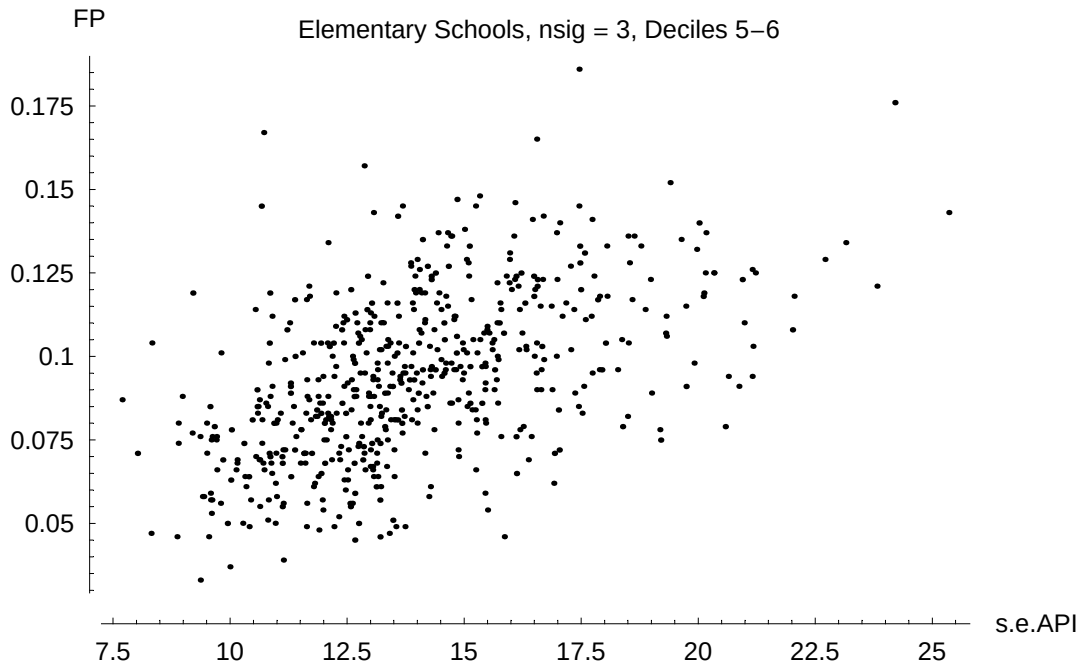

Figure 1. Illustration of less-than-perfect relation between false-positives (FP) for school accountability and standard error of school score (s.e.API)

Vignette 4. False Negatives and the Fallacy of Small School Advantage

Disparities in observed accountability results (i.e., small schools faring better) and flawed statistical analyses have led to many press reports of a "small school advantage" in accountability (e.g. Sharon Tulley, & Campbell, 2002). In their prescriptions for remediating claimed flaws in accountability systems, Kane and Staiger (2002) in the section "Implications for the Design of Incentive Systems" advise that the effects of school size ("Lesson 1") are so severe as to justify structural changes in accountability systems: "A remedy would be to establish different thresholds for different size schools, such that the marginal net payoff to improving is similar for small and large schools, or offer different payoffs to small and large schools.(p.257)"

The misunderstanding that generates concerns about a small school advantage is in neglecting or ignoring false negatives in considering the properties of the accountability system (e.g. Kane & Staiger, Figure 7). The one clear fact is that statistical variability (in a school or subgroup score) does decline as the number of students increases. A small school having made no real improvement (or poor real performance) 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 (or has strong real performance) has statistical uncertainty as its foe, in that a false negative result may occur more often than for a large school.

A useful calculation is to obtain probability of success (award) for different size schools under different levels of underlying real achievement. The numerical illustration here (adapted from Rogosa, 2002b) uses a set of four California elementary schools, where each of theawardse schools has a 1999 API score in state decile 5 (i.e., slightly below the state median) of about 610 with the same three numerically significant subgroups (Socioeconomically Disadvantaged, Hispanic, and White). The contrast in the four schools is the progression of school sizes (specifically number of students included in the school API score) n = 148, 244, 350, 491 which represent the 5th, 20th, 50th, and 80th percentiles of the elementary school size distribution. The size labels for the four schools in Table 5 are: smaller n (=148), small n (=244), medium n (=350), and large n (=486).

INSERT TABLE 5

The message of Table 5 is that a small school advantage is only seen if false positives (no real improvement) are solely considered; any combination of false positives and false negatives appears to erase (if not reverse) such an effect. The leftmost column in Table 5 gives, for reference, the standard error of the school API, which does decrease with school size. The next three columns, true improvement levels 0, 29, 41, reflect incrementation of all student scores to produce an increase

Table 5
Small School Advantage???

		Probability of Award			
		True Improvement			
	se(API)	0	29	41	Average
smaller n	20.2	0.131	0.447	0.596	0.3755
small n	14.26	0.094	0.620	0.797	0.5244
medium n	13.7	0.101	0.714	0.890	0.5523
large n	11.1	0.084	0.822	0.961	0.595

in the school API of 0, 29, 41 (see Rogosa 2002b for details of the calculations). The entries are the probability that school API and significant subgroups meet or exceed growth targets (the Governor's Performance Award, GPA, criterion, about 10 points year-to-year improvement for these schools and 8 points for the subgroups). The true improvement = 0 column gives the false positive probability, probability that statistical variability alone (i.e., null improvement) will result in school eligibility for GPA, and the false positive probability does decrease in Table 5 as school size triples. On the other hand, when substantial real improvement (set slightly below and above the observed state-wide median improvement) is imposed, the probability of award increases with school size. The far right column represents a combination of false positive and true positive probabilities: probability of award averaged over all incrementation, 0 to 50 API points. This probability increases with school size.

A separate group size misunderstanding–determinations of minimum subgroup size in NCLB accountability– provides a segue to the next vignette. Typical guidance (see CCSSO, 2002; Linn, Baker, & Herman, 2002) has unfortunately focused on the standard error of a single proportion, and such examination of just one component often provides little useful information on the properties of the accountability system. In contrast, Rogosa (2003c) shows "on the margin" calculations for comparing the properties of the accountability system, as constituted, with the properties resulting from setting a smaller (or larger) minimum n.

Vignette 5. How Large is the "Diversity penalty"?

NCLB accountability can be described as "no subgroup left behind," and there are strong policy and political motivations for including the subgroup criteria in AYP. Under NCLB, schools must meet the proportion proficient criteria not only for the school a whole, but also for numerically significant subgroups defined by race/ethnicity, socioeconomically disadvantaged students, English language learners, and students with disabilities. Legitimate policy research questions are, What is the effect of including subgroup criteria on the properties of the accountability system? What should be the rules for subgroup inclusion and criteria? As previously mentioned, subgroup criteria will reduce false positive probabilities (presumably a good thing). For diversity penalty claims, unlike the school size claims in the previous vignette where false negatives were ignored, for the diversity penalty claims, the focus of accountability critics is on the false negatives. Claims of a large diversity penalty rest on a misunderstanding of the basic distinction between equality of opportunity and equality of results.

A misleading "coin-flip" metaphor that has been picked up in many discussions of the effect of subgroups was introduced by Kane and Staiger (2002): "For a racially integrated school, winning an award is analogous to correctly calling three or four coin tosses in a row, instead of a single toss. As a result, at any given level of overall improvement, a racially integrated school is much less likely to win an award than a racially homogeneous school."(p. 258). The misleading aspect here is the reader's instinct to think in terms of a fair coin (and also to overlook that the number of independent subgroups is often far less than the reported number– e.g. a school composed primarily of Hispanics, all English learners, almost all disadvantaged are reported as 3 subgroups.) For this coin-flip metaphor to apply in California, NCLB would require non-overlapping homogeneous subgroups with performance all near the NCLB criteria which are set at less than .2 proportion proficient, representing an extreme case given that state-wide proportion proficient is around .4.

One of the first analyses of the effect of subgroups on the properties of a school accountability system to claim a diversity penalty is the presentation in Kane and Staiger (2002, esp. Table 4), using results from the past California API award programs. Kane and Staiger assert as their Lesson 2: "Incentive systems establishing separate thresholds for each racial or ethnic subgroup present a disadvantage to racially integrated schools"(p. 258), and further state in their conclusion: "Rules making any rewards contingent on improvement in each racial group present a great disadvantage to integrated schools and generate a number of perverse incentives that may harm rather than help minority students" (p.269). The empirical evidence does not support these claims of diversity penalty. Disparity in tabled outcomes (observed number of subgroups crossed with school size or mean achievement) is not credible evidence of "unfairness", as schools differ in many ways other than number of reported (overlapping) subgroups (c.f., Rogosa 2002b, sec. 4 on

Kane-Staiger).

More recently, Novak and Fuller (2003) garnered coast-to-coast press attention for their claims (using California NCLB results) that "schools serving diverse students in California are less likely to achieve their growth targets...." and "Schools serving middle-class children, for example, are 28 percent more likely to be labeled “needs improvement” by the feds when serving five student subgroups than schools serving one group." (p. 1) . Rogosa (2004) refutes the Novak and Fuller empirical claims and presents simple probability calculations to provide some quantification for questions about the effects of subgroups. These calculations use an "on the margin" logic--what is the effect on probability of award or probability of meeting the NCLB criteria if another subgroup is added? All indications are that such effects are far less than that claimed by the empirical tabulations of disparity in results. One (very basic) result shows that a school that has high probability of meeting the AYP criteria if no subgroups were included, has only slightly reduced probability with multiple (representative) subgroups. Table 6 calibrates the effect of additional subgroups in an artificial setting; entries in the table are the probability of meeting both math and English NCLB criteria for a school whose size is proportional to the number of subgroups (n indicates subgroup size). The results show a decrease in AYP probability of about .015 for each additional subgroup (i.e., .97 for two subgroups decreases to .94 for four non-overlapping subgroups).

INSERT TABLE 6

Another way to calibrate the effects of subgroups (see also Rogosa, 2003c) is through the increase in underlying educational attainment (true proportion proficient) required to maintain the same probability of meeting the NCLB criteria, as the number of subgroups increases (which could be characterized as a slight headwind). In addition, Rogosa (2002b) has some examples and calculations, using the California accountability data, suggesting that since larger schools tend to have more subgroups the effect of additional subgroups to diminish probability of award is approximately balanced by the decreased false negative probability for the larger schools (combining vignettes 4 and 5).

In sum, implementation of subgroup criteria is an important topic for further technical research on the properties of accountability systems, such as the current NCLB. And it may well possible to develop superior alternatives, especially alternatives more directly focused on the announced aim of NCLB, to reduce (eliminate) achievement gaps..

Table 6
Effects on probability of meeting AYP of additional subgroups

	Number of Non-overlapping Subgroups			
	0/1	2	3	4
n				
50		0.970	0.956	0.941
100	0.985	0.970	0.956	0.941
200	0.985	0.970	0.956	0.941

Vignette 6. The Margin of Error Folly: Blood Pressure Parable

The applications of the *margin of error* in assessment and school accountability can be shown to lead to preposterous results. The uses of a margin of error, typically defined as 1.96 times the standard error of the score, represent a misunderstanding of basic statistical principles, such as the meaning of a confidence interval. One way of expressing margin-of-error logic: If it could be, it is. The succeeding vignettes take up two educational accountability instances of margin of error folly; the blood-pressure parable below (adapted from Rogosa 2005) serves as a lead-in to the school accountability examples.

Blood pressure parable. Consider an artificial setting, using diastolic blood pressure (DBP) in hypertension diagnosis; the example uses part of the hypertension standard: DBP, 90 or above. Like educational standards, standards for blood pressure diagnosis are subject to revision (Brody, 2003). 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 interpretation a reading of 101 is not distinguishable 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 below, 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. Does the application of the margin of error appear to promote good health (or good educational policy)?

Technical Details for Probability Calculation. The basic statistical facts for what is termed the normal/normal model can be found in Carlin and Louis (2000, secs. 1.5.1 and 3.3.1) and Lehman and Casella, (1998, sec. 4.2). The likelihood specification is that $Y_i \mid \theta_i \sim N(\theta_i, V_i)$, a Gaussian distribution with mean θ_i and variance V for units i = 1, ... ,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:

$$\theta_i \mid y_i \sim N(B\mu + (1 - B)y_i , (1 - B)V), \quad \text{where } B = V /(V + A).$$

In the DPB example θ_i is the perfectly measured DBP for individual *i*, and for the DPB measurement $V = (6.12)^2$. 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{true DPB > 89.5 | DBP = 101 } = .9117.

Vignette 7. The Margin of Error Folly: California School Accountability

Using margin of error reasoning, the *Orange County Register* (Orange County, California), launched in August 2002 a weeklong series of attack pieces against the state of California school accountability system, in particular criticizing the rewards offered for improvements in the Academic Performance Index (API). The main assertions by the *Register* were: "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" (Sharon, Tully and Campbell, 2002) and "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" (Sharon and Tully, 2002). Furthermore, these claims have had carry-over to NCLB; Gladwell (2003) in a wide-ranging attack on NCLB and school accountability uses the *Register* claims: "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."

Again, if these claims were true, then school accountability, whether through rewards or sanctions, would not be defensible. Was the California school award program at the mercy of statistical uncertainty? Is this mistake probability, P{no real improvement | award}, really as large as 1/3? (Note in terms of Table 4 quantities the *Register* quantity of interest, P{no real improvement | award} = b/(a + b) , which in epidemiology would be termed 1 – predictive value positive). To the contrary, Rogosa (2002a, 2005) demonstates that a good estimate of these mistakes is 2% of schools, 1% of money– not 33%. The technical explanation in Rogosa (2005) of this margin of error folly is informative for understanding properties of school accountability decisions. How could the *Register* (and their cited experts, Richard Hill and Thomas Kane) get it so wrong? The prose explanation would be: by piling "could be's" upon "might have's" and counting those as certainties.

To understand the correct calculations, and to see how the *Register* analysis could be wrong by a factor of 20 or more, consider a an example of a specific elementary school in Los Angeles which the *Register* counted as "impossible to tell if the school really improved." This average-sized elementary school (CDS code 19647336018253) has 2 significant subgroups, Socioeconomically Disadvantaged (SD) and Hispanic, and this school received a GPA award of $40,262 for API improvement, 1999 to 2000. The year-to-year improvement of 31 points for this 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. Thus, according to the *Register* we are to have "no idea" whether this school actually improved, and this school is included in their mistake tally.

From Table 7 (adapted from Rogosa, 2005, Table 1), the probability is .992 that the true improvement is greater than 0. In effect, *Register* margin of error methods for determining whether the school "really improved" rounds .008 up to 1.0 for the tabulation of "impossible to tell" schools. Moreover, for the school's growth target criterion, P{true change < 9 | observed data} = .0386 and thus the odds are 25 to 1 that the true improvement exceeded the year 2000 growth target.

That school example is not atypical; for the 3585 elementary school GPA award winners in 1999-2000, the median value of P{true change > 0 | observed data} was .9988; over 75% of the schools had probabilities above .99, over 90% of these schools had probabilities above .97, and the expected number of GPA schools showing no real improvement was 35. Thus there is strong empirical evidence that schools receiving awards did have real improvement. Moreover, for these 3585 elementary school GPA winners in 1999-2000, the median value of P{true change > growth target | observed data} was .993; over 90% of these schools had probabilities above .91, and the expected number of GPA schools not meeting the growth target was 104.

Table 7
Probability Calculation for API Improvement, CDS code 19647336018253

	API	n	se(API)	margin of error
1999	616	349	14.2	27.8
2000	647	355	13.2	25.9
Improvement	31			34.9

P{true change ≤ 0|observed data} = .00798

Vignette 8. The Margin of Error Folly: NCLB Confidence Intervals

Legitimate concerns about the effects of statistical uncertainty on the properties of NCLB accountability and poor understanding of basic statistical principles combine to morph the margin of error folly into NCLB Confidence Intervals. The important statistical and policy question is, How much benefit of the doubt in AYP criteria should be accorded to schools to offset the effects of statistical uncertainty? And the basic misunderstanding lies in not appreciating the consequences of the confidence interval adjustment (another example of needing to consider the tradeoff between false positives and false negatives).

Under the NCLB "Confidence Intervals" approach adopted by a large majority of states, AYP requirements are adjusted downward, below the AMO (Annual Measurable Objective) established by the state. The amount of adjustment depends on the number of individual student scores contributing to the school or subgroup score so that the adjustment is larger for smaller groups of students. The genesis of these confidence interval procedures appears to be CCSSO (2002) with additional development by the Center for Assessment as seen, for example, in Marion and Gong (2003). Although the confidence interval adjustment is labeled as a "Statistically-Based Approach" (CCSSO, 2002 Ch.3, p.87 onward), these NCLB confidence intervals (usually implemented via hypothesis tests) should not be confused with good statistical practice.

Rogosa (2003d) quantifies the extreme consequences of these NCLB confidence interval implementations, using calculations based on some of the summary data from Utah testing and on the Utah NCLB procedures (Utah SOE, 2003). Utah employs the 99% confidence interval adjustment typical of many states. The NCLB AYP requires that the school-wide proportion proficient scores meet the AMO in both subjects (in Utah for grades 3-8: Language Arts, proportion proficient .65 and mathematics, proportion proficient .57), and also these subject proportion proficient scores meet the AMOs for each of the included subgroups. Table 8 displays the number of proficient students needed to actually meet the stated AMO (the n*AMO column) and the number of proficient students that the Utah procedure would deem as close enough via the 99% confidence interval adjustment (labeled minimum number proficient) for group sizes from 25 to 500 (see Rogosa 2003d for computational details). For example, with a group of 25 students, the mathematics AMO of .57 would seem to require 15 proficient students out of the 25. But the one-sided hypothesis test described in the Utah NCLB plan would deem only 8 proficient students as representing close enough.

INSERT TABLE 8

In terms of the accountability decisions display in Table 4, the intent of the NCLB confidence interval adjustment is to eliminate (or make very rare) false negatives, statistical uncertainty in the observed proportions proficient creating a

Table 8
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
25	11	17	8	15
75	39	49	33	43
100	54	65	45	57
125	69	82	58	72
300	176	195	151	171
500	300	325	259	285

failure to meet AYP. But good intentions, such as to ensure (as much as possible) that schools are not falsely labeled as "needs improvement," have consequences. These consequences should be understood (perhaps before these state NCLB plans were approved). Among the many kinds of calculations that can be done, the example here from Rogosa (2003d) serves to illustrate the maximum amount of benefit of doubt that the confidence interval adjustments bestow.

Consider two schools, sizes 125 and 500 students, each composed of 60% Caucasian, and 20% each Hispanic and African-American students. With no 99% confidence interval adjustment, the number of proficient students in the larger school required to meet the math AMO for the three subgroups is (171, 57, 57), and for school total, 285 proficient. Then the simplest probability reasoning would say that, given those observed numbers proficient, the probability that the true proportion proficient met the math AMO for all 3 subgroups (and therefore school also) is about 1/8. Table 9 displays probability calculations from Rogosa (2003d) which show the consequences of the 99% confidence interval adjustment. Given the data--observed minimum number of proficient students meeting the subgroup and school proficiency criteria--what is the probability that the true (measured without statistical uncertainty) proportion proficient would meet the AMO for all subgroups (and therefore school-wide also) ? First consider the calculations for math in the larger school. The number of proficient students in the subgroups that meets the AMO for subgroup and school under the 99% confidence interval adjustment is (169, 45, 45) yielding a posterior probability for the true proportions proficient meeting the AMO of about 1/100,000. That's one way of quantifying the benefit of the doubt. The equivalent calculation for English gives a posterior probability of about 1/50,000 for the true proportions proficient meeting the AMO. As AYP requires satisfying both English and math criteria, the probability for both sets of true proportions proficient meeting the AMO is considerably less than either of these very small probabilities.

INSERT TABLE 9

Do the NCLB confidence intervals represent reasonable benefit of doubt for the effects of statistical uncertainty? Does the confidence interval scheme really satisfy the NCLB legislation's requirement for "statistically valid and reliable" AYP (see, CCSSO, 2002, p.21)? (Note the state NCLB Accountability Workbooks require under System Validity and Reliability: "9.1 Accountability system produces reliable decisions. 9.2 Accountability system produces valid decisions" [Utah SOE, 2003].) The press caught on early to this NCLB margin of error folly (e.g., Deffendall, 2003; Lynn, 2003; Rado & Little, 2003); researchers and policy makers should also.

Table 9
Probability Calculation for NCLB 99% Confidence Interval: Schools with Three Non-overlapping subgroups

	School with 125 students; 75 Caucasian, 25 Hispanic, 25 African-American	School with 500 students; 300 Caucasian, 100 Hispanic, 100 African-American
English/Lang Arts:		
AMO = .65	.000037 (47,11,11)	.0000196 (192,54,54)
Mathematics		
AMO = .57	.0000068 (42,8,8}	.0000098 (169,45,45}

entries are:
Probability true proficiency meets performance goal (AMO)
(number Caucasian, Hispanic, African-American observed proficient)

Vignette 9. Demographics are far from Determinant

The final vignette takes up a controversial and misunderstood topic: the relation between demographic variables and test scores (for schools and individuals). The misunderstandings arise from most analysts and interested parties not having a good understanding of the correlation coefficient: in particular, not appreciating how far away from the upper bound +1 (perfect linear association) is consequential in the interpretation of a correlation coefficient. Certainly, it is very rare for a school drawing from a student population regarded as highly advantaged to score extremely poorly. Similarly, most often a school drawing from a student population regarded as highly disadvantaged does not obtain a very high score. That's the reality that drives the value of the correlation coefficient, but it is very far from the whole story. Examples below and more extensive descriptive analyses in the citations refute the slogan of the California Teachers Association that "It's all zip codes," their way of indicating that demographic characteristics predominately determine school performance in testing and accountability.

The school demographic measure used here is the "School Characteristics Index" (SCI), a composite of demographic measures computed by California Department of Education for each school. (Note: Similar results are obtained using a simpler school demographic measure, proportion of disadvantaged students, Rogosa; 2001.) To examine the relation between demographics (SCI) and student achievement (school API), one common first resort is the school-level correlation coefficients for each school type. Many would regard the correlations below as quite large, and educational researchers would typically conclude a very strong relation between school results and demographic characteristics.

	Elem	Middle	High
1999 Pearson correlation of SCI and API =	0.924	0.951	0.946
2000 Pearson correlation of SCI and API =	0.923	0.951	0.939
2001 Pearson correlation of SCI and API =	0.920	0.943	0.936

(Equivalent results are obtained from analyses that use multiple regression to predict school scores from an assortment of demographic measures.) A more informative view is provided by the corresponding scatterplots of API vs SCI (see Rogosa 2003b, Figures 7 through 12). The scatterplots reveal that even a correlation of .94 is rather far from 1.0, because even though API scores increase as the SCI index increases, the plots also show considerable range on API (perhaps 250-300 pts) for a chosen level of SCI.

Useful Data Analysis for the Relation between API scores and demographic characteristics: Range of Similar School API

In the California API reporting, the SCI is used to identify the "100 other schools with similar demographic characteristics" which are listed as *Similar Schools*. For elementary schools, this list, composed of the 50 schools with closest SCI scores above the school and the 50 with the closest SCI scores below the school,

comprises a (reasonably narrow) 2% slice out of the distribution of elementary schools. The data analysis opportunity is to use each school's list of 100 Similar Schools to assess whether the API scores for these similar schools lie in a narrow range. E.g., Do schools that are similar on measured demographic characteristics obtain similar API scores? A quantitative answer is to use the quantity *Range Similar School API*, the range of the corresponding 100 similar school API scores: max(API) – min(API).

Table 10 presents results for Range Similar School API in the years 1999, 2000, and 2001 for California elementary schools (further tables in Rogosa 2003b, Section B2). The 1999 statewide result 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 schools have a range of their Similar Schools API scores of at least 243 points. A good way to calibrate these numbers is to note that for elementary schools the statewide decile categories typically span 40-45 API points. Thus 243 points represents a span of 5 to 6 statewide deciles and the median range 277 represents a span of about 6 (or more) statewide deciles. Contrast those results with the typical interpretation for an API,SCI correlation of .92.

In addition, the tables in (Rogosa 2003c) show that indications from the entire state data also hold up when examined for each decile, and the Range Similar School API is even larger for schools in the lower state deciles (i.e. lower scoring). For 1999 there are 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 (a width of 6 deciles). Are these results consistent with the claim that demographics are determinant?

Table 10
Range Similar School API: California Elementary Schools 1999-2001

N	Mean	Median	Q1	Q3	Minimum	Maximum
4849	281.50	277.00	243.00	304.00	154.00	522.00
4775	268.88	257.00	222.00	308.00	125.00	435.00
4895	242.22	235.00	203.00	282.00	101.00	384.00

Discussion

Measuring and judging schools. School accountability can be thought of as a process with two main components, measuring schools and judging schools. The measuring question is, What type of number should represent the performance of students in the school? and the judging question is, Was that performance, however quantified, *good enough*? The measuring component presents multiple choices for summarizing the student test scores for a school or subgroup. Among the measuring options are: single cut-point measures (proportion at least Basic, proportion at least Proficient); weighted composite indices (differential weighting for multiple performance levels as in California API); with the availability of matched longitudinal data, the proportion of students improving year-to-year; or reflecting NCLB policy goals a measure of achievement gaps among subgroups in the school. In judging a school, the "good enough" criteria for making a decision about the school can be expressed in terms of yearly status (e.g., meeting a standard) or year-to-year improvement (e.g., meeting a growth target), with the judging mechanism possibly applied to subgroup performance as well as school scores.

In a school accountability system a form of "measuring" is joined with a form of "judging" for making decisions about schools. Designing a useful accountability system requires an understanding of which combinations of measuring and judging best serve the intended policy objectives, while also having defensible statistical properties. The statistical challenge is to make good determinations in the face of some uncertainty in the data (i.e., not lean too hard on the data leading to mistakes about schools). In addition, a main policy issue is that the 'good enough' criteria used in judging the schools are educationally meaningful and achievable. An analogy is to medical diagnosis where medical measurements are joined with criteria for a disease decision, and statistical properties such as false positives and false negatives are routinely investigated. In No Child Left Behind accountability, the particular choice for measuring schools is the proportion of students proficient (where the proficiency standard is determined by the state), and the choice for judging schools is that the yearly status measure for English and math meet a specified level (AMO) for school and included subgroups. That the particular implementation of measuring and judging does matter can be seen in widely publicized disparities between pre-existing state accountability systems (such as Texas and California) and NCLB AYP results.

Issues for NCLB. The vignettes identify educational research misunderstandings of statistical issues important not only to NCLB but also to just about all systems of school accountability. Those issues include: the properties of group summary scores; use and consequences of subgroup criteria; proper evaluation of accountability decisions; and adjustments in accountability criteria for statistical uncertainty in group scores. The summative message from these statistical misunderstandings is that instead of good statistical practice guiding NCLB accountability, these education reform efforts are flying blind. And because

state Accountability Workbooks are unfortunately not required to contain any useful information on the statistical properties of NCLB accountability (e.g., false positives, false negatives or the more advanced false discovery rate indices), researchers, education professionals and the interested public are left totally in the dark. Perhaps the most important statistical misunderstandings reside in the series of vignettes on the margin of error folly, which demonstrate why states should not be allowed to employ the popular NCLB confidence interval adjustments. By no stretch of probabilistic imagination do the "close enough is good enough" NCLB confidence interval adjustments satisfy the requirements of the NCLB statute for statistically valid and reliable decisions about schools.

Another important disconnect between policy intent and NCLB accountability arises in the oft-stated goal of NCLB to close achievement gaps (such as between high and low poverty students or between ethnic/racial groups). One of the largest misunderstandings about NCLB is that the AYP criteria have little or nothing to do with the goal of closing achievement gaps. NCLB criteria do place a lower bound on subgroup performance by raising the level of lowest-scoring subgroups. But large achievement gaps may persist even as the proportion proficient requirement in AYP approaches 1.0. In fact, success under NCLB may be as likely to increase gaps because higher performing students may benefit even more from improved teaching regimens and curricula aligned with the content of tests than will students who are considered to be at-risk of failure. Although the NCLB AYP criteria do not directly address or assess achievement gaps, one can construct accountability-useful measures (perhaps as an alternative to the current NCLB implementation) that credit schools for closing achievement gaps (either change in gaps or gaps in change) and judge schools on that basis, resulting in accountability procedures more closely aligned with the stated policy goal.

Implications of the Vignettes. NCLB represents a situation in which educational policy is far out front of the existing technical knowledge and statistical expertise, with unfortunate results. At a minimum, constructing good educational policy requires confronting and refuting bad educational research. The refutations of prior work described in the vignettes function in part to bring accountability to bear on educational researchers, which only seems fair given that high-stakes accountability currently impinges on students, teachers and principals. Educational research and policy would be well-served by following the example of medical research, where there exists a strong tradition of research rebuttal and correction in many venues, including academic journals and popular press.

References

Brody, J. E. (2003, August 12). 'Normal' blood pressure: health watchdogs are resetting the risk. *New York Times,* F7.

Campbell, R. (2002, August 11). API's error margin leaves a lot to chance: Mathematical imprecision could lead to inaccurate interpretations. *The Orange County Register*.
Retrieved January 5, 2005 from
http://www.ocregister.com/features/api/text_version/index.shtml

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.
Retrieved December 12, 2004 from http://www.ccsso.org/content/pdfs/AYPpaper.pdf

Deffendall, L. (2003, October 19). No statistics are being left behind. *Lexington Herald-Leader*.

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

Kane, T. J., & Staiger, D. O. (2002). Volatility in school test scores: Implications for test-based accountability systems. In D. Ravitch (Ed.), *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.

Linn, R. L., Baker, E. L., & Herman J. L. (2002, Fall). Minimum group size for measuring adequate yearly progress. *The CRESST Line*, 1, 4-5.

Linn, R. L., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis*, *24*(1), 29-36.

Lynn, R. (2003, December 18). Federal mandate: A complex statistical formula allowed them to meet their targets, *The Salt Lake Tribune*. Available: http://166.70.46.216/2003/dec/12182003/utah/120969.asp

Marion, S., & Gong, B. (2003). Evaluating the validity of state accountability sytems. Center for Assessment Presentation at CCSSO, St. Louis MO, September 11, 2003. Retrieved November 12, 2004 from http://www.ccsso.org/content/pdfs/CCSSO_Validity_BGSM03.pdf

Novak, J. R., & Fuller, B. (2003). Penalizing diverse schools? Similar test scores, but different students, bring federal sanctions. PACE Policy Brief 03-4, December 2003. Retrieved January 3, 2004 from http://pace.berkeley.edu/policy_brief_03-4_Pen.Div.pdf

O'Connell, J. (2004, March 24). Fight for changes to NCLB. California Department of Education. Retrieved December 3, 2004 from http://www.cde.ca.gov/eo/ce/sl/nclbfight.asp

Rado, D. & Little. D. (2003, September 28). Schools toying with test results: Some states meet standards with art of statistics. *Chicago Tribune*,1, 22.

Rogosa, D. R. (2001). Year 2000 update: Interpretive notes for the academic performance index. California Department of Education, Policy and Evaluation Division, October 2001. Retrieved April 1, 2004 from http://www.cde.ca.gov/ta/ac/ap/researchreports.asp

Rogosa, D.R. (2002a). Commentaries on the Orange County Register series: What's the magnitude of false positives in GPA award programs?; Application of OCR "margin of error" to API award programs. California Department of Education, Policy and Evaluation Division. September 2002. Retrieved April 1, 2004 from http://www.cde.ca.gov/ta/ac/ap/researchreports.asp

Rogosa, D. R. (2002b). Irrelevance of reliability coefficients to accountability systems: Statistical disconnect in Kane-Staiger "Volatility in School Test Scores" CRESST deliverable, October 2002. Retrieved December 5, 2004 from: http://www-stat.stanford.edu/~rag/api/kscresst.pdf

Rogosa, D. R. (2002c). Plan and preview for API accuracy reports. California Department of Education, Policy and Evaluation Division, July 2002. Retrieved April 1, 2004 from http://www.cde.ca.gov/ta/ac/ap/researchreports.asp

Rogosa, D. R. (2002d). Accuracy of API index and school base report elements. California Department of Education, Policy and Evaluation Division. December 2002. Retrieved April 1, 2004 from http://www.cde.ca.gov/ta/ac/ap/researchreports.asp

Rogosa, D.R. (2003a). Confusions about consistency in improvement. CRESST deliverable, June 2003. Retrieved December 5, 2004 from:

http://www-stat.stanford.edu/~rag/api/consist.pdf

Rogosa, D. R. (2003b). Four-peat: Data analysis results from uncharacteristic continuity in California student testing programs. California Department of Education, Policy and Evaluation Division. September 2003. Retrieved April 1, 2004 from http://www.cde.ca.gov/ta/ac/ap/researchreports.asp

Rogosa, D.R.. (2003c). California's AMOs are more formidable than they Appear. California Department of Education, Policy and Evaluation Division. October 2003. Retrieved April 1, 2004 from http://www.cde.ca.gov/ta/ac/ap/researchreports.asp

Rogosa, D.R. (2003d). The NCLB "99% confidence" scam: Utah-style calculations. CRESST deliverable, November 2003. Retrieved December 5, 2004 from: http://www-stat.stanford.edu/~rag/nclb/utahNCLB.pdf

Rogosa, D.R. (2004) Assessing the effects of multiple subgroups: A rebuttal to the PACE Policy Brief December 2003 "Penalizing diverse schools? Similar test scores, but different students, bring federal sanctions". California Department of Education, Policy and Evaluation Division, January 2004. Rejoinder: Being fair to NCLB, February 2004. Retrieved December 5, 2004 from http://www.cde.ca.gov/ta/ac/ap/researchreports.asp

Rogosa, D. R. (2005). A school accountability case study: California API awards and the Orange County Register margin of error folly. In R. P. Phelps, (Ed.), *Defending Standardized Testing* (pp. 205-226). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Sharon, K. & Tully, S. (2002, August 16). State testing expert says API margin of error is insignificant: Leader who helped design index calls it as accurate as possible. *The Orange County Register*. Retrieved January 5, 2005 from http://www.ocregister.com/features/api/text_version/index.shtml

Sharon, K., Tully, S., & Campbell, R. (2002, August 11). Test scores unreliable: Error margin means state can't precisely measure how schools are doing, but the cash still flows. *The Orange County Register*. Retrieved December 5, 2004 from http://www.ocregister.com/features/api/text_version/index.shtml

Utah State Office Of Education (2003). State of Utah Consolidated State Application Accountability Workbook. Plan Approved by U.S. Department of Education June 10, 2003. Retrieved from: http://www.ed.gov/admins/lead/account/stateplans03/index.html