# Education Policy Analysis Archives

**Volume 4 Number 4** **March 15, 1996** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal.

Editor: Gene V Glass,Glass@ASU.EDU. College of Education,
Arizona State University,Tempe AZ 85287-2411




## Standard Errors in Educational Assessment: A Policy Analysis Perspective


**Greg Camilli**
**Rutgers University**

*camilli@zodiac.rutgers.edu*



**Abstract:** In many educational settings, educational gains are measured and evaluated rather than absolute levels of achievement. Gains might be estimated for individual students,teachers, schools, districts, and so forth. In some educational programs, schools are required to make "statistically significant" progress over the course of one school year. This would typically require and estimate of the standard error (SE for short) of the gain, which is a number representing the precision of the gain similar to the "margin of error" in polls. Because SEs can be used to define educational targets, it is important to understand precisely what a standard error is -- and this requires going beyond the simple textbook definition. Statistical methods are tools for understanding social processes, but there is no necessary connection between a statistical method and an empirical outcome. A policy analyst must ask how closely features of the statistical theory correspond to aspects of the measured outcomes for a given purpose. For example, how much does it matter if the assumption of random sampling is violated in certain ways? Can one assume that the children or educators at a particular school during a given year constitute a random sample of some population that is perhaps spread across time, space, as well as cultural and institutional dimensions?


### INTRODUCTION

In many educational settings, educational attainment is measured and evaluated. The units on which these measurments are taken might be students, but also teachers or school districts. The TVAAS program (Tennessee Value-Added Assessment System) is a statewide assessment

program that incorporates features such as these. Specifically, it is a "gain-oriented" statistical tool for collecting and analyzing student achievement test score data; that is, gains are the focus rather than absolute levels of achievement. The information provided by the statistical model is used in the evaluation of teachers, schools, districts, and the like.

It is not the purpose of this document to evaluate the TVAAS program itself. However, during discussion of this program on EDPOLYAN (Educational Policy Analysis Forum -- an electronic forum in which discussion is conducted on the Internet), the issue of standard errors (or SEs) arose (SEs are computed with the TVAA multi-level model). In particular, some schools are required to make "statistically significant" progress which entails a 1.5 to 2.0 SE gain over the course of one school year. Because SEs are used to define educational targets, it is important to understand precisely what a standard error is -- and this requires going beyond the simple textbook definition.

In most introductory statistics courses, the terms "population," "random sample," and "sampling distribution" are taught. No two samples give the same result, e.g., the average height of a sample will always differ to a greater or lesser extent across random samples. This is why polls always append the "margin of error" to reported percentages. For example, the polls report that 51% of respondents would have voted for Bill Clinton with a margin of error of plus or minus 3%. (The margin of error is the product of two components: the standard error (SE), and the critical value (CV). The former represents how much a result may vary from sample to sample while the latter is used in conjunction with the former to place a band of confidence around the obtained result. For the given example, 3%=SE*CV; and the band of confidence extends from 51%- 3% (48%) to 51%+3% (54%). It is common to interpret the latter by saying that even though there might be variation from sample to sample, 95 out of 100 samples would give a result between 48% and 54%.]

Why should the standard error (SE) serve as a standard against which gains are evaluated? This question must be answered at both a technical and policy level:

- Technical. If there were 10 people in a room and you wanted to know their average income, you could ask all 10 people and calculate the mean. But now suppose that a town has 500,000 people in it. Obviously the same strategy is not going to work. Rather, you must economize by sampling a representative cross-section and calculate the mean of this group. This method doesn't guarantee an accurate result all the time, but it does well most of the time -- especially with larger samples. Thus, a sample result is not an exact answer to one's overriding question. From a statistical point of view, numbers are fuzzy rather than precise creatures; and a statistician's concern is to keep the amount of fuzziness to a minimum.

  By metaphor, an exact number is a pin prick (or puncture) whereas a fuzzy number is a bruise. Two pin pricks would be easy to discriminate visually, but if two bruises were large and overlapping, they might be difficult to distinguish. Now imagine the radius of a bruise as the standard error: as the radius decreases, it becomes clear as to whether there is one bruise or whether there are two. And as the radius decreases to near zero, the bruise becomes a pin prick. In short, the standard error is the statistician's criterion for the separability of two numbers, and two numbers are conventionally thought of as separable if they are at least 1.5 or 2 SEs apart. This is equivalent to requiring that the confidence bands around two numbers not overlap.

- Policy. Statistical methods are tools for understanding social processes. There is no necessary connection between a statistical process and an empirical outcome,so policy

analysts must ask how closely features of the statistical theory correspond to aspects of the measured outcomes of an educational program. An important part of this analysis concerns how well sentences typical of the statistical theory support actions based on the separability criterion of 1.5 or 2.0 or some others numbers of SEs.

For example, a typical sentence of classical statistical inference might read "A random samples is taken from a population." To which the analyst might respond "What is the population?" Furthermore, if the population can't be defined, one might conclude that it is not possible to determine whether the sample was indeed random. Thus, the language of the statistical theory might not satisfactorily explain the SE criterion, in which case more analysis is necessary to arive at a pragmatic understanding.

These issues are explored in the following discussion among members of EDPOLYAN. The discussants are in alphabetical order Greg Camilli, Sherman Dorn, Gene Glass, Harvey Goldstein, Bill Hunter and Leslie McLean. Passages have been edited to focus on the issue of standard errors. The original postings contained more ancillary issues as well as parenthetical comments. However, the participants have reviewed the following text for accuracy and completeness. In addition, further summary comments were provided by Harvey and these are given at the end of the discussion section. Original messages were posted in late December, 1994, through January, 1995.

**EDPOLYAN DISCUSSION**

*Goldstein*: I have come in on what I gather is the tail-end of a discussion of missing data in the analysis of TVAAS system data to produce estimates of school effects. Apologies therefore if the issue has been discussed already, and also because I'm from a different educational system but one where we have had quite a lot of debate about value added analysis using longitudinal data. ... in the UK the value added debate has been looking at problems with the sampling errors (standard errors) of value added gain scores...it turns out that these are typically so large that you cannot make any statistically significant comparisons between most of your schools...only those at opposite extremes of a ranking. Is this also the case in Tenessee? If so what do you do about it when reporting?

*Camilli*: I've been wondering what the standard errors mean. Usually, I have in mind that a sample is drawn from a population, and an effect (say gain score) is estimated from the sample data. The standard error then conveys how precise this estimate is (much like the "margin of error" that pollsters use). For TVAAS, what are the sample and population?

*McLean*: The purpose of this post is to focus on two aspects of the TVAAS that I feel have received too little attention: validity and standard errors. This is not to say that the political nature of any evaluation is not important or to take anything away from the discussion of formative vs. summative evaluation.

Harvey Goldstein stated on Dec. 19, 1994, "..it turns out that these [standard errors] are typically so large that you cannot make any statistically significant comparisons between most of your schools...only those at opposite extremes of a ranking. Is this also the case in Tennessee? If so, what do you do about it when reporting?"

Below are listed the mean gains for math with their standard errors for schools within one of the larger school systems in Tennessee. These means are three year averages and were calculated from the TVAAS mixed model process. This should give an idea of the sensitivity of the process.

| TYPE OF SCHOOL | GRADE | RANK | GAIN | MEAN STD.ERR. |
|---|---|---|---|---|
| INTERMEDIATE | | | | |
| | 3 | 1 | 71.6 | 4.9 |
| | 3 | 2 | 71.2 | 3.7 |
| | 3 | 3 | 67.0 | 2.6 |
| MIDDLE | | | | |
| | 6 | 1 | 22.6 | 1.0 |
| | 6 | 2 | 20.1 | 0.8 |
| | 6 | 3 | 15.6 | 1.0 |
| | 6 | 4 | 13.9 | 1.0 |
| | 6 | 5 | 13.1 | 1.2 |
| | 6 | 6 | 12.5 | 1.0 |
| | 6 | 7 | 11.1 | 0.9 |
| | 6 | 8 | 9.8 | 1.1 |
| | 6 | 9 | 9.3 | 1.3 |
| | 6 | 10 | 8.4 | 0.9 |
| | 8 | 12 | 13.5 | 6.6 |
| | 8 | 13 | 13.2 | 1.7 |
| | 8 | 14 | 11.0 | 0.9 |

The problem for those of us who have calculated, pondered and puzzled over such results as these, in national and international assessments, is that the reported standard errors are unbelievable (impossibly small). We can't say they are wrong, of course, because we lack the details of the calculations, but Harvey Goldstein has analyzed at least as much data and written several books and taken the lead in multilevel modeling (sometimes called, by others, hierarchical linear modeling), and his informed and experienced "opinion" is not to be taken lightly. The standard errors remind me of those Richard Wolfe found faulty in the first International Assessment of Educational Progress--the fault being that the estimates of error failed to include all the components reasonable people agree should be included. Moreover, the Std. Errors above are clearly proportionate to the mean scores, not a desirable outcome. There must be at least one error (three lines from the bottom of those displayed above). I, too, will leave to later a comment on the statement below from TVAAS, except to say that whatever it is they do is not "certainly sufficient":

*Camilli*: I thought that some of you might want to take a look at some statistics regarding the metric of the scores that TVAAS uses. Below, I've given the mean, median and standard deviation of the IRT metric for fall reading comprehension as reported in the CTBS/4 Technical Bulletin 1 (1989).(I hope this isn't too far out of date.)

| Grade | mean | median | STD |
|---|---|---|---|
| 1 | 473 | 481 | 84.3 |
| 2 | 593 | 606 | 81.1 |
| 3 | 652 | 657 | 59.6 |
| 4 | 685 | 694 | 53.6 |
| 5 | 707 | 714 | 48.6 |
| 6 | 725 | 730 | 43.8 |
| 7 | 733 | 738 | 43.6 |
| 8 | 745 | 750 | 43.1 |
| 9 | 760 | 764 | 38.6 |
| 10 | 770 | 774 | 39.6 |
| 11 | 776 | 780 | 38.2 |
| 12 | 780 | 782 | 38.0 |

If you plot these data by grade, some interesting possibilities emerge. For example, one wonders why students below average gain as much as students above average. The explanation I see is that there is much less room for growth at higher grade levels, but this is a function of the scoring metric. A transformation of scale might lead to different results.

*Goldstein*: I see that Les McLean and one or two others have taken up my query about how well schools can be separated taking the estimated standard errors into account. I don't yet know how the standard errors have been calculated, but based upon a table Sandra Horn xxx sent me, I would say that the results (e.g. for grade 3 based upon a 3 year average in one of the larger school systems, are in line with our own results. What you do (roughly) is multiply each standard error by about 1.5, use this to place and interval (i.e. +-1.5 s.e.'s) about each gain estimate and judge whether two schools are significantly different at the 5% level by whether or not the intervals overlap. Most intervals do! BUT if you average over 3 years then you get smaller standard errors so fewer do.

*McLean*: My observation that the gains were proportional to the standard errors does NOT seem to be true--within grades. If you lump all grades together, the correlation is over 0.5, but within grades (the correct plot, IMHO) it is essentially zero. Grade six shows a substantial NEGATIVE correlation, but there are only 12 observations.

What are these standard errors anyway? In a separate post to me, Greg Camilli points out that if all students are tested, then the "sampling error" has to be zero. What we need to make sense of this is, as I have said already, a technical report. How are they (TVAAS) modelling the error in their multilevel models? What explanatory variables do they use? Do they include covariance terms? Is the "standard error" an estimate of measurement error? Just how much data are missing?

*Goldstein*: Re Les Mclean's message about standard errors. He quotes Camilli as stating that the standard errors given are 'sampling errors' and that if all students are tested then these are zero. I am confused! The usual standard errors quoted in this context are those relating to the accuracy of the estimated school effects where there is a conceptually infinite population of students of whom those measured (whether they are all those in the school at a particular time or not) are a random sample. If they are not this, then what are they?

*Hunter*: [In response to Goldstein's last question] I cannot say what they are precisely, but I am quite confident that those in any particular school at any particular time are NOT a random sample.

*Camilli*: Harvey Goldstein is wondering what I mean about standard errors. TVAAS probably doesn't test random samples of students, or samples at all, given what I know about the program. Given that it's not a random sample, one could always imagine that this were the case anyway (a counterfactual): imagination is required by all theories of statistical inference. However, without some sensible restrictions any set of numbers whatsoever can become a "random sample from a population." And once the "population" is in place, it can be of any size at least as great as the "sample." Now if it is imagined as infinite, then we can go on to imagine our "sample" as one realization of an infinite possible samples. Thus, we arrive at an estimate of "sampling error" called the "standard error."

Here's a fictitious dialogue between an Educator and a Statistician regarding this point:

> E: What is this standard error associated with a gain score?

S: It's like this. Suppose you have a conceptually infinite population of 8th graders, and from this population you took an infinite number of random samples and computed the gain for each sample. You'd want these to all give about the same result; it's like witnesses at a trial corroborating a claim made by the defendant. Small standard errors are analogous to a high degree of corroboration; while high SE's indicate a lot of uncertainty.

E: What if the tested group isn't a sample?

S: Then you just imagine it's a sample.

E: OK, I'll imagine it's a sample, but what's the population?

S: Just imagine that the population is pretty much like the sample.

E: So if I get a small SE, then I can be confident in the gain score because most of the samples from the population will have similar gains, because the population is pretty much like my sample?

S: Basically.

E: What if the standard error is large? Does this mean that I shouldn't be confident because most of the sample in the population that is similar to my sample will give substantially different estimates of gain?

S: Well, you've got the basic idea.

E: Okay, but just one more question. If the SE is large, doesn't it mean that the population isn't similar to my sample? If so, how can I imagine an infinite number of samples from that population?

S: Look, SE's are really theoretical quantities. They're things that are defined by equations -- and the equations can be explained in different ways. Population/sample is the easiest way, but don't get bogged down. Most statisticians agree that they are useful and that small ones are better than large ones.

E: OK, but just one more question. How small do they have to be to be good?

S: That depends on your question. Suppose you want to test whether one teacher's gain is larger than another's. If the difference is one the order of 1.5 or 2.0 SEs, then you can have confidence in it.

E: Why should I have confidence?

S: Because the difference is large relative to the sampling error ... er, I mean, standard error.

E: I see. Well, I have to go now. By the way, could you write something up that I could give to parents that explains this? Thanks.

*Goldstein*: Well, I enjoyed Greg Camilli's imaginary conversation, but of course the reality is that standard errors are not things statisticians invented to make life difficult. Most non-statisticians have little difficulty in understanding that if you only have a measurement on 1 student there ain't

much to be said about the rest. The bigger the sample the more confident you become that what you have observed is a good guide to what you would get on repeated samples with also suitably large numbers...assuming of course that you adopted a sensible randomly based sampling strategy.

Now we come to the philosophical bit. Social statisticians are pretty much forced to adopt the notion of a 'superpopulation' when attempting to generalise the results of an analysis. If you want to be strict about things then the relationship you discovered between parental education and student achievement back in 1992 from a sample of 50 elementary schools in Florida can only give you information about the physically real population of Florida schools in 1992. Usually we are not interested just in such history, but in rather more general statements that pertain to schools now and in the future...we may be wrong of course and that is why we strive to replicate over time and place etc. BUT the point is that, getting back to value added estimates for a school, if we want to make a general statement about an institution we do have to make some kind of superpopulation assumption....what we happen to observe for the students we have studied is a reflection of what the school has done, and would have done, for a bunch of students, given their measured characteristics such as initial achievement. The more students we measure the more accurate we can be and that's why we need an estimate of uncertainty (standard error).

*Glass*: Greg answered with a hypothetical conversation between an educator and a statistician. I think Greg exposed some key problems with this notion of standard errors, and it is no more a problem with TVAAS than it is a problem with most applications of inferential statistics in education.

Harvey asks, in effect, what is wrong with regarding standard errors as being measures of the accuracy of samples as representations of "conceptually infinite populations" from which the samples might "conceivably have been drawn at random."

After more than thirty years of calculating, deriving, explaining and publishing "standard errors" and their ilk, I have come to the conclusion that I don't know what they mean and I doubt seriously that they mean anything like what they are protrayed as meaning.

Consider this: if the population to which inference is made is one that is conceptually like the sample, then the population is just the sample writ large and the "standard error" is much larger than it ought to be. If you show me 25 adolescent largely Anglo-Saxon boys who love sports and ask me the population from which they could conceivably have been sampled, I'll conceive of an "infinite" population of such boys. If no population has actually been sampled and all I know about the situation before me is the sample, then I will conceive of a population like the sample. This is surely the very opposite of inference and standard errors are surely beside the point.

Consider something even more troubling: I present you with a sample- - Florida, Alabama, Tennessee, South Carolina. N=4. I calculate the state high school graduation rates, average them and calculate a standard error. What is the population? States in the Southern U.S.? Fine; that's certainly conceivable, even if not "infinite." But suppose that someone else conceives of "States in the U.S." Well, that's conceivable too. But it is surely ridiculous to think that these four states can be used to infer to both of these conceivable populations with equal accuracy (standard errors). Or to make matters worse, suppose that I suddenly produce a fifth "state": Alberta. Now it raises the question whether the conceptual population is "geo-political units in North America"-- or the entire Western Hemisphere.

I can't imagine that there is much wisdom in attaching a number accurate to two decimal places

when we can't even be certain whether it is referring to an "inference" to the Southern U.S., North America or the Western Hemisphere. Now, if you think I am playing with your head and will suggest a way out of this dilemma that rescues the business of statistical inference for us, let me assure you that I have no solution. In spite of the fact that I have written stat texts and made money off of this stuff for some 25 years, I can't see any salvation for 90% of what we do in inferential stats. If there is no ACTUAL probabilistic sampling (or randomization) of units from a defined population, then I can't see that standard errors (or t-tests or F-tests or any of the rest) make any sense.

Does any of this apply to TVAAS? Just this. If one is worried about "stability" (in any of the many senses in which the word could be interpreted) then why not simply compare teachers' scores across all years for which data are available. That would answer in very straightforward ways whether the ranking of teachers jumps around wildly for whatever reasons or is relatively steady. (I hasten to add that I don't approve of such things as ranking teachers with respect to their students' test scores.)

*Goldstein*: Gene Glass also takes me to task on standard errors and raises the interesting question of when a sample should be considered as having a reference population and when not. There is no general answer...it depends on what you want to do. As I said in my response to Greg, I cannot easily see how you can have empirical social science without assuming that the units (people, schools etc) you happen to have measured are representative (in the usual statistical sense) of a (yes) hypothetical population whose members exhibit relationships you want to estimate. Such populations must (I think) be hypothetical because they have to embrace the present and future as well as the past when the data were collected. The issue is therefore the general philosophical issue and not a statistical one - statisticians simply try to provide tools for making inferences about such populations.

*Camilli*: Harvey replied to my previous post with "The bigger the sample the more confident you become that what you have observed is a good guide to what you would get on repeated samples with also suitably large numbers...assuming of course that you adopted a sensible randomly based sampling strategy."

Bigger is better, I agree. Another issue is whether it is the correct standard error, and still another is whether the SE has a meaningful referent. If the sample consists of all kids in the system, how can imagining a larger group possibly create more information. If I want to understand the behavior of my three cars (I wish), how would it benefit me to imagine I had a fourth? This is not a statistical issue at all. "Population" has always been a heuristic device.

Generalizing beyond known populations is risky business, and requires more than statistical knowledge. This was the focus of the long and interesting dialogue between Lee Cronbach and Don Campbell. Standard errors have something to do with the precision of estimates. Perhaps they convey something about how well a model fits certain data. You might want to argue, on this basis, that the model is likey (or not) to generalize; but model fit at one instant does not logically imply model fit one second later. This, I think, is the difference between induction and inference.

The standard errors will apparently be used to measure whether statistically significant progress is being made by schools that fail to meet the standard (whatever that turns out to be), so it is important to be clear about what SEs mean. I find it fascinating that they are being used as policy tools with legal implications. In this regard, it is important to understand what drives the SEs. I'm guessing that missing data will add to SEs (it really would be helpful if the TVAAS staff would

respond), and am sure that unit size will decrease SEs. Thus, standard errors for schools will typically be smaller for districts than for schools than for teachers than for students. As far as I can tell, only certain districts are required to make statistically significant progress; this may turn out to be a pretty easy criterion to satisfy.

*Goldstein*: When you try to enshrine complex technicalities in the law you certainly ask for trouble - especially, as would appear here, when those drafting the law have a rather meagre understanding of the technicalities. My interpretation of $49-1-601 is that it requires (say from one year to the next) that the differnce in value added scores for a school between two years is statistically significantly different from zero (at 5%?). If each years scores are on the same metric then this question can certainly be asked and one can even think of a suitable interpretation. The problem arises if we require this to be the case for all those schools below the mean (note that the legislation does not say STATISTICALLY SIGNIFICANTLY BELOW the mean.). If the schools are successful then the mean for all schools inevitably goes up!! and it isn't difficult to envisage a scenario where every school makes a real (even statistically significant) gain leaving the ranking of all schools the same! This raises the issue of the measurements used. Are these standardised each year on the Tennessee population? If so then not only is the ranking the same, so are the actual scores! All this needs some careful unpicking I would have thought and raises very serious issues for the interpretation of TVAA.

*McLean*: The discussion of standard errors has gotten so involved that a look at the Tennessee legislation should tell us where standard errors are needed and what interpretations reasonable people ought to be able to put on them. Below, the text from Sherman Dorn's post [who is quoting and paraphrasing from TVAAS statutes] and Les McLean's reponses are indicated by "-->".

> -->*Dorn*: The goal is for all school districts to have mean gain for each measurable academic subject within each grade greater than or equal to the gain of the national norms.
>
> -->*McLean*: How will anyone decide whether the mean gain is greater than or equal to the gain of the national norms? Publication of "standard errors" must mean that an error bound will be established around the national norms--perhaps 1.5 Times the median std. Error per grade--one "harvey", or 2.0 Std. Errors--one "dorn".
>
> -->*Dorn*: If school districts do not have mean rates of gain equal to or greater than the national norms based upon the TCAP tests (or tests which measure academic performance which are deemed appropriate), each school district is expected to make statistically significant progress toward that goal.
>
> -->*McLean*: ok, gang, the veil is lifted from our eyes--there is no such thing as "statistically significant progress" without standard errors and the assumption of samples from some population.
>
> -->*Dorn*: schools or school districts which do not achieve the required rate of progress may be placed on probation as provided in $49-1-602. If national norms are not available then the levels of expected gain will be setupon the recommendation of the commissioner with the approval of the state board.
>
> -->*McLean*: Yo, commish! I do not envy you your task.
>
> -->*Dorn*: value added assessment means: (1) a statistical system for educational

> outcome assessment which uses measures of student learning to enable the estimation of teacher, school, and school district statistical distributions; and (2) the statistical system will use available and appropriat data as input to account for differences in prior student attainment, such that the impact which the teacher, school and school district have on the educational progress of students may be estimated on a student attainment constant basis.
>
> -->*McLean*: I could write a rationale for a "statistical system" that did not need standard errors, given that they test all the students. It would contain careful, modern descriptive statistics that would gladden John Tukey's heart.
>
> -->*Dorn*: On or before July 1, 1995, and annually thereafter data from the TCAP tests, or their future replacements, will be used (notice the 'will'-- the language is not just permissive here) to provide an estimate of the statistical distribution of teacher effects on the educational progress of students within school districts for grades three (3) through eight (8).
>
> -->*McLean*: Here we are again--these gains are to be interpreted as "teacher effects". Peace, TVAAS, but I do not believe that anyone's models and techniques are yet good enough to isolate the teacher effect from all the other effects on standardized test scores in schools with all their complexity. Next to this concern--it is a concern about validity and is not vague or complex-- the definition and estimation of standard errors is too small a matter to take our time.

*Goldstein*: Les McLean's comments have inspired some more thoughts. In the simplest value added model, an outcome score is regressed on an input score so that generally each school will have a different regression line - perhaps with varying slopes but in the basic model with parallel slopes so that schools can then be ranked on the resulting regression intercepts. (The actual analysis is a bit more complex but this simple model captures the essence). We find, typically, that the variation among these intercepts is relatively small compared to the residual variation of student scores about the regression lines for each school (5% - 30% depending on which educational system you are studying). In addition, the regression itself will account for quite a lot of the variation in outcome...maybe as much as 50-60%.

This means that there is a substantial remaining variation (among students) unnacounted for and it is this residual variation which determines the standard error values. Thus, for example, if this residual variation was zero, we would exactly predict each schools (relative) mean and the standard error of that prediction would be zero. This would mean also that once we knew each student's input score (and anything else we were able to put into our regression model) and the school that student was in, we would have a perfect prediction of the student's outcome. Of course, we are nowhere near that situation and it is this uncertainty about the individual prediction that translates into uncertainty about the school mean (think of the mean roughly as the average of the student residuals about the regression line for each school). If you took another bunch of students with exactly the same set of intake scores you would NOT therefore expect to get the same set of outcome scores - this is what the uncertainty implies - nor the same mean for the school. In the absence of being able to predict with certainty we have to postulate some underlying value for each school's mean (otherwise we are pretty well lost) which we can think of as the limit of a series of conceptual allocations of students to the school. Thus an estimate of uncertainty, conventionally supplied by calculating the appropriate standard error,is important if you want to make any inference about whether the underlying means (that is, the population means) are different and, more importantly, to set limits (confidence intervals) around the

estimated difference for any two schools or around the difference between a school's estimate and some national norm. Hence my original remark some time ago that when you did just that you found that most institutions could not statistically be separated, and I suspect also for TVAAS that very many cannot statistically be separated from a National norm, whether they are actually above or below it. It would be good to hear from the TVAA people on this issue.

*Camilli*: Harvey continues the standard error saga, and I want to reiterate: if you had all the students in the school there wouldn't be any uncertainty at all; you'd know the mean. I think we need a "superpopulation" to get us out of this predicament. Harvey said "If you took another bunch of students with exactly the same set of intake scores you would NOT therefore expect to get the same set of outcome scores - this is what the uncertainty implies - nor the same mean for the school."

This bunch of students is from the superpopulation, no? They are students who might exist, but don't, who are substantially like the students in the sample. I'll say it again, Harvey, this is a heuristic. It simply doesn't convey any additional meaning regardless of how many times it is repeated. I think we're lost when we accept statistical inferences based on data that weren't observed, and moreover, do not exist conceptually. If "all the students in the school" doesn't really have that meaning, then we are playing a game with language.

If we can get away from the superpopulation for a moment, we can begin to analyze what drives the standard error. It certainly isn't sampling error; nonetheless, it is a quantity that exists in a real sense. As you've implied above, SEs have something to do with model fit. Thus, we should be interested in those things that cause models to fit more loosely to the data. District size is certainly one factor; but correlation of effects within the model will also inflate SEs. Effects like teachers within schools, teachers with school, schools with district might be some examples. As Gene implies, separating these effects may take some doing.

*McLean*: Harvey Goldstein's exposition on standard errors (17 Jan, "Standard Errors: yet again") may have been more than some wanted, but I found it instructive and thought-provoking. If you deleted without reading, reconsider--it gets at the heart of the matter of TVAAS. While still wanting to retain the concept of the sample from some (unspecified) population, Harvey's main lesson for us was to highlight the crucial role of the model adopted by the statistician in estimating scores--gain scores, in the case of TVAAS. A model is a formula that the statistician considers a reasonable try at relating the desired quantity, the 'gain' in achievement (not directly measurable because of nuisances such as social class and prior learning) to aspects of schooling, such as teacher competence.

Advised by statisticians with wide experience outside of education (and maybe in education--we have not been told), the policy-makers decide to give the statisticians their head and to accept their estimate of 'gain', knowing that the formula will be complex and the procedures well beyond the understanding of all but a very few. The statisticians make a persuasive case that their formula and their procedures will provide the policy-makers with an estimate of gain that will distinguish the bad teachers from the poor from the average from the good from the excellent. "National norms" are invoked, unspecified, but responsibility given to the Commissioner of Education to provide norms if the national government lets the side down.

All this tedious repetition is needed to give a context for Harvey Goldstein's description of standard errors. In essence (correction, Harvey, please, if needed) the errors are S&E, not SE--errors of Specification & Estimation, not of sampling. A 'specification' error is made when our model, our formula, does not accurately link the target (the gain) with the data (the item

responses or scale scores plus proxies for prior learning and social class and the like). We ALWAYS make a specification error--the only question is how large. If we limit ourselves, as in the TVAAS, to linear models, and we try to estimate gains across big, complex societies such as states, the error can be huge--and there is not consensus how to estimate the size of the error. Here is a source of error.

Even though they do not sample students and schools, sampling cannot be avoided--people are absent, times of testing vary, the tests cannot possibly cover all the content (hence content sampling), items are omitted, test booklets get lost, some teachers do not cover the material on the test, ..., and so on and so on and so on. This is why we do not use a very simple formula:

Gain = (Avg. score end - Avg. score beginning)

After all, when we test everyone, and when the goal is to measure gains by THESE students THIS year in THESE places with THESE teachers, who needs an error term? With well-constructed tests, the measurement errors will cancel out when we calculate school and class means. Oh--there is measurement error in individual pupil scores, but we can report that (from the test publisher's manual) and besides, these scores don't count in the student's grade--the teacher does not get them in time, and even if they do they do not use them.

Ok, so I seem to have lost the tenuous thread of the argument--NOT SO! We have learned over the years that the simple formula is more likely to mislead than to lead--to distort our view of gain rather than to clarify it. Raw score comparison tables (called 'League Tables' in the UK, after the rankings of sports teams), however compelling they seem, are statistically invalid, immoral, racist, sexist and stupid. Apart from those few flaws, they are fine. But would Tennessee put up with such poor procedures? Not on your life--scaling, imputation, hierarchical linear models and prayer are brought into play. Here is another source of error.

All this talk of standard errors and models and politics keeps coming back to one key aspect: VALIDITY. Do those numbers represent gains in achievement? The formulas and procedures are complex enough that evidence is needed. Even if they do, how accurate are they--and I mean how much do they tell us about better learning, class-by-class, teacher- by-teacher; or has the TVAAS traded in science for voodoo? Without a better explanation, the use of these scores to label teachers as competent or incompetent seems a lot like sticking pins in dolls.

It is possible to validate the numbers--but it would take a lot of thinking, a lot of hard work and maybe 0.01 of the budget of TVAAS.

*Glass*: Harvey, and are these future batches of students "random" or "probabilistic" samples from that "conceptual" superpopulation? It seems highly doubtful. So what sense can possibly be made out of probability statements that surely assume random sampling? None that I can imagine.

I think Les had it right last night. The "errors" in these teacher measurement schemes are model specification errors and not sampling errors. And the important questions to ask about them are not "will they be different in some conceivable 'population'?" but "what do they contain: ability differences in students, effects from previous teachers, etc.?"

*Camilli*: Les, I think your distinction between SE and S&E is a clear and elegant statement. It is a must-read for anyone interested in how statistical models are likely to behave in policy contexts. I'd like to throw in two additional cents:

1. I think TVAAS is certain to encounter a related problem with its "linear metric." How is it,

the press may ask, that gains are so much larger in the earlier than the later grades? Does this mean that students aren't learning very much in high school? Moreover, because the standard errors are likely to be different across districts, larger districts might have to achieve smaller gains to be consistent with the law. Does this imply different standards for different districts? (I recognize that larger districts have to pull up more kids to achieve a SE's worth of gain -- but I'm not sure this type of argument would wash since a SE may be only a baby step toward the national average.)

2. The "natural" sample that exists on any given day does, I suppose, give rise to a superpopulation of the sort that Harvey Goldstein writes of. However, this is not the population about which most people think of when evaluating gains since, as Bill Hunter points out, it is not a random sample from the school's student body.

*Hunter*: Per Camilli who wrote "The "natural" sample that exists on any given day does, I suppose, give rise to a superpopulation of the sort that Harvey Goldstein writes of. However, this is not the population about which most people think of when evaluating gains since, as Bill Hunter points out, it is not a random sample from the school's student body."

I need to clarify a bit. I think it is not the case that a sample of convenience "gives rise to" or "implies" a population of any sort (unless one chooses to regard the sample _as_ a population). As far as I can tell this thinking is exactly backwards--samples derive their meaning and existence from populations: I cannot see that the reverse order has any meaning at all. I also question the utility of Harvey G.'s conception of such samples as samples from a population in time. This _might_ make sense in a time/space of great stability, but I see little reason to believe that children four or five years from now will have experiences of the world (especially the world of information) that is comparable to children of today (or five years past). The kinds of changes that required revision and re-norming of intelligence tests every 15 or 20 years half a century ago now take place in five years or less--probably about the same time scale that would be required to conscientiously develop and renorm the test.

Moreover, I think it is not just that such a sample is not a random sample from some _specific_ population (as Greg suggests above), but that it is not a random sample of ANY population for two reasons: 1) the process of selection did not insure equal and independent likelihood of selection for all members of the population and, more importantly, 2) no population was specified (to which the above process was not applied).

*Goldstein*: Brief response to Greg. The point about imagining another bunch of students like the ones you used to compute the school mean is that this seems to me just what one always has to do. The information about the students whose data you analysed may be of historical interest, but for most people they really want to assume that, given no evidence to the contrary, if and when a fresh set of similar students passes through the school (as is happening by the time they get to read the report)they would expect a similar outcome. The superpopulation is not just a heuristic device it is a reality in th esense that further batches of students are samples from it. How else would you make sense of anything?

Now to Les' points: Specification error actually, I think, sits on top of what I mean by standard errors, the latter assume that the specified statistical model is a good description. This raises what I think is perhaps the more important issue. Are we using the right measures? have we adjusted for all th econfounding factors? Have we adjusted properly for measuring errors (unreliability). On this side of the Pond we have I believe won the intellectual (not the political - we are used to losing that one)argument against Les' RAW league tables and are beginning to make people aware of the limitations of value added ones. The standard error argument is only one point of

reference but it is quite important because it does, I believe, point out the inherent scientific limits to any kinds of institutional comparison in terms of how finely ranked you can get. There is a kind of uncertainty principle operating; you can establish that there are institutional variations without being able to determine exactly which institutions are actually different from each other. That's perhaps difficult to live with but does seem to be a fact of life.

*McLean*: On January 18, Bill Sanders wrote (via Rick Garlikov--and along with many other topics):

> --> *Sanders*: To Leslie McLean, your plots of standard errors as calculated make no sense. Middle schools in the example school system we provided have more students than intermediate schools in almost every case. Thus, their standard errors tend to be lower. Middle schools also have smaller expected nominal gains. Therefore, your attempt to show a relationship over grades is nonsense.
>
> --> *McLean*: It was indeed the point I was making--that the plot (or correlation) over grades made no sense. That is why I argued that the within- grade correlations were the ones to look at--and that they were around 0.0. BTW, if means in a table are based on widely different Ns, you would do your readers a good turn to say so, don't you think? Your remark that "middle schools also have smaller expected nominal gains" is ambiguous and interesting. In what sense "expected"; in what sense "nominal"?

*Camilli*: About superpopulations: these are entities that don't exist, except in the imagination. Yet it is contended that it is a "reality in the sense that further batches of students are samples from it. How else would you make sense of anything?" A lot of people have sought to answer this question, among them Alan Birnbaum who paraphrased the likelihood principle as the "irrelevance of outcomes not actually observed." He went on the write of the "immediate and radical consequences for the everyday practice as well as the theory of informative inference." As for the superpopulation, it exists in one's mind as a vehicle for generalization. But generalization itself requires more worldly knowledge. For example, consider the standard error of statistic calculated from a poll during an election. You might say a population exists, but only for a limited amount of time. Experience with the rate of change in public sentiment (and the way the question is asked) is required for a valid generalization. Happily, however, we are in full agreement on the role of specification error, as masterfully articulated by Les.

**SUMMARY COMMENTS BY PARTICIPANTS**

*Goldstein*: I am a bit confused by the TVAA requirement to make a gain of 1.5-2.0 STANDARD ERRORS. Shouldn't this refer to STANDARD DEVIATIONS? The standard error is a measure of the accuracy with which a statistic (e.g. mean gain score) is estimated. The standard deviation is a measure of population spread and is the appropriate unit to use.

*Camilli*: Sherman, a question has come my way from Harvey Goldstein. He asks whether STANDARD ERROR should be STANDARD DEVIATION?" It's my recollection that the law specifically states that SEs are to be used for assessing gain, not SDs. Could you send me the relevant section?

*Dorn*: Okay, here is the relevant section of the TN law, and the answer's "none" -- at least explicitly:

$49-1-601. (c) If school districts do not have mean rates of gain equal to or greater than the national norms based upon the TCAP tests (or tests which measure academic performance which are deemed appropriate), each school district is expected to make statistically significant progress toward that goal.

But statistically significant is a strange concept for TVAAS, since there is no random sampling -- it's supposedly everyone in the relevant universe. Does it mean statistically significant considering test-retest reliability? Does it mean statistically significant considering the norming population? Does it mean statistically significant considering a hypothetical "let's pretend this is a random sample" thought experiment? Yeesh. In point of fact, courts have not had a chance to even consider this, since probation is not a question until this fall, and the legislature has delayed individual teacher reports for an additional year, at least (Nashville TENNESSEAN, 31 May 1995). I find it amusing that a state court will decide what statistical significance is here.

*Goldstein*: The discussion has certainly been interesting and useful for me in forcing me to be explicit about a number of 'taken for granted' assumptions. There seems to me to be three separate issues being debated.

1. If we have a collection ( sample) of individuals on whom we make measurements, is there some sense in which we can and should regard these as members of a larger collection or population of individuals. Does this population have to exist in reality (i.e. it can be enumerated in principle) or can we think of a hypothetical 'superpopulation' and when might this be useful
2. If we accept that there is a population about which we may wish to say something (e.g. what is the mean gain score among ALL 6-7 year old boys), how can we obtain a RANDOM probability sample so that we can then apply the statistical techniques which require such samples in order to draw valid inferences?
3. Any member of a sample of human (social) beings simultaneously belongs to more than one recognisable population; thus a child belongs to a particular social background grouping and a neighbourhood of residence and a school etc. Which is the appropriate population for inference?

Let me tackle 1) first. There are clearly some real enumerable populations which we can sample and make statements about. Surveys of voting intentions are a case in point where we wish to say something about how the whole (voting) population thinks, based on a suitable (preferably random) sample. A great deal of statistical sampling theory exists to help us do just this. At the other extreme you have something like what has been called 'generalisability theory' in educational testing that chooses to regard a set of chosen test items as being 'sampled' from some (conceptually) infinite population of such items contained within a 'domain'. I personally have great difficulty with this concept since, as Gene Glass points out, what people seem to be doing here is to imagine the population as just a larger version of the items they happen to have (unless they really have sampled, for example words from a dictionary for a spelling test). This then tends to come down to a sleight of hand whereby you choose your own test items, declare that they allow you to make inferences about an undefined domain, and then use statistical procedures to describe how accurately you have been able to describe that domain. And don't believe them when they tell you they have rules for generating the items and the rules implicitly define the domain - it doesn't work!

There are, however, other cases where I simply don't see how you can make any substantial progress without the notion of a hypothetical population of which your sample is a realisation. In effect this is nothing more than saying that you want, on the basis of what you observe on one

group of individuals, to make some statements about other, unobserved individuals. If you are doing ethnographic, case study research, you are interested in what you find for what it may tell you about other (similar?) cases. Likewise, if you observe a relationship between race and school achievement you are concerned to make a more general statement and set of speculations about the relationship as it may apply to other children It seems to me that without this there is no empirical social science possible. This is a philosophical not a statistical issue. If you can't make inferences about future individuals then all social science is just descriptive history. The notion of a superpopulation is simply a formalistaion which allows us to use the tools of statistical inference. It is the ONLY formalisation I know of which allows a satisfactory method of generalising from the observed to the unobserved.

This leads to the next issue, which is how one can conceive of drawing a random sample from such a population. Gene's example of the four US states as a sample is instructive. Suppose, instead of merely calculating the mean graduation rate across States you compared the probability of graduating between Florida and Tennessee. In 1994 you found a moderate difference. You might be happy to stop there and leave it at that. On the other hand as a social scientist you might want to contextualise the difference, noting that Florida and Tennessee had different social compositions and you wondered whether these might 'explain' the observed differences. You might go on to look at other factors, and soon you would be constructing quite sophisticated statistical 'models'. The point about these models is, in general, that they don't explain all the differences - there is residual variation between children in whether or not they graduate.. We might, in principle, be able to explain everything but in practice this is extremely rare, and Les McClean's discussion of specification errors is relevant here. So the unexplained variation is assumed to be random - a reflection of our ignorance if you like. It is at this point when we invoke the statistical assumption of random variation that we are forced to assume some kind of sampling (or exchangeability if you insist on being a Bayesian) from a population. Whether you wish to confine inferences to Florida and Tennessee or wished to make some tentative inference about the factors which 'explained' graduation variations in general, across time and space, is a matter of debate and presumably disagreement, but generalise we surely wish to do?

This gets into my third issue about the appropriate population of reference. In brief then, I am not arguing that we always need a superpoluation notion, which then leads on to the statistical apparatus of standard errors, etc., but I am saying that to make sense of school comparisons (as with State comparisons), adjusting for those factors extrinsic to schools (gain scores and much more than these of course, such as race and class and gender) the notion of a superpopulation is really indispensable for us to make any progress. Let me ask the question again which I don't think anyone answered: If you have two schools, each with 2 students following a particular course, who would stake their academic reputation on reporting a moderate difference in average (over 2 cases) gain score as a judgement that the schools were REALLY differentially effective? Or suppose there was only 1 student in each school? Of course this is extreme - I merely wish to pursue the logic of refusing to recognise a superpopulation to an absurdity.

*Camilli*: Harvey frames the discussion with three questions:

1. When should we think of sample as members of a superpopulation. Clearly, there is a way to draw a sample in which it makes sense to think of a sampling distrubution. Frequency does have meaning in this situation. But Harvey thinks that you can "make progress" by imagining a sampling distribution when the population is poorly defined and sampling isn't random, and "In effect this is nothing more than saying that you want, on the basis of what you observe on one group of individuals, to make some statements about other, unobserved

individuals. If you are doing ethnographic, case study research, you are interested in what you find for what it may tell you about other (similar?) cases." But the logic here is circular: I will assume my sample is similar to other nonrandom samples, then I will assume that the results in these other samples (similar by assumption) will yield similar results. In short, I think generalization is possible, but classical frequency theory is a lazy metaphor. You can make inferences about the future, but don't think statistical theory provides a formal basis for this. Ian Hacking in The Emergence of Probability recounts how Hume demolished this notion in 1739 (see p. 181).

2. Models are proposed by scientists to account for variation, and few if any models fit perfectly. In this case, a measure of mistfit is a measure of ignorance, but whoa! How does one equate ignorance with random variation? It seems to me that this is an attempt to reify frequency theory. I agree that generalize is what we surely wish to do, but what is happening here is 1) a statistical theory is adopted which is a mathematical formalization, 2) a strict correspondence between the terms of the theory and real events is assumed, and 3) results and manipulations of the theory are presumed to have counterparts in the real world. That is, real world events are now assumed to follow statistical laws. (Marx is spinning like a top.) We will make progress when we can usefully distinguish descriptive theory from observed covariation.

Suppose one has two students from each of two schools with gain scores, yet knows nothing of how these students were encountered. Does one want to determine whether these students are representative of the schools to which they belong, or assume that they ARE representative of Population X? In the latter case, we are 100% certain that we have a valid sample; this is an easily recognized tautology. In the former case, we have more work to do. Generalization isn't impossible, but we must make an argument for doing so and defend its validity. The argument is based on evidence, completeness, and persuasiveness. None of these qualities is based in statistical theory.

## References

Hacking, Ian. (1975) *The Emergence of Probability.* Cambridge University Press.

---

To receive a publication guide for submitting articles, see the *EPAA* World Wide Web site or send an e-mail letter to LISTSERV@asu.edu and include the single line GET EPAA PUBGUIDE F=MAIL. It will be sent to you by return e-mail. General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, Glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692)

## Editorial Board