

**Assessment of the Statistical Methodology
Used in the Tennessee Value-Added Assessment System (TVAAS)**

*by Walter W. Stroup, Ph.D.
Professor of Biometry
University of Nebraska-Lincoln*

1. Introduction

I visited the University of Tennessee Value-Added Research and Assessment Center in Knoxville from March 29-30, 1995. The purpose of my visit was to review the statistical methodology used in the TVAAS. My specific charges were

- a. Review the statistical model and its assumptions and evaluate the evidence that data used in TVAAS are in fact consistent with these assumptions.
- b. Verify that the software used in TVAAS is accurate, i.e. it performs the required statistical calculations correctly.

I approached objective (a) by questioning Drs. William L. Sanders and Arnold Saxton, the UT statisticians responsible for the analysis of the TVAAS data, at length about the model. I asked about its assumptions, evidence in support of the assumptions, alternative models considered but not used, why these alternatives were discarded, difficulties in working with the models, and how these difficulties were handled.

For objective (b), I had Dr. Saxton, who wrote the computer code for the mixed model component of the TVAAS data analysis software, explain how the program was supposed to work. Then I created my own hypothetical data and analyzed it using SAS-PROC MIXED, a commercial software package designed to analyze models like those used in TVAAS. After doing the PROC MIXED analysis, I gave my data to Dr. Saxton. The results of his program and my analysis were then compared.

My main conclusions are

- a. The statistical model being used is reasonable and the data are consistent with the assumptions of the model.
- b. The software used in TVAAS computes the statistical analysis correctly.

The next two sections contain the details of my review.

2. Review of the statistical model and adequacy of its assumptions

The TVAAS data consist of test scores taken annually on each student between grades 2 and 8 at every school in Tennessee in each of five subjects. Each observation in the data set is identified by

system
school
grade
calendar year
subject
student

Statistical analysis of the TVAAS data uses two models:

The "school model"

For a given school, an observation for a given school-grade-year-subject-student is characterized by

$$\text{observation} = \text{school-grade-year-subject mean} + \text{error}$$

The school model is analyzed separately by system, except for counties with multiple systems. Analysis is done by county for multiple system counties.

The "system model"

For a given system, an observation for a given system-school-grade-year-subject-student is characterized by

$$\text{observation} = \text{system-school-grade-year-subject mean} + \text{error}$$

Like the school model, analyses of the system model are done separately by system for single system counties and separately by county for multiple system counties.

Both models share a common set of assumptions. Specifically,

1. within a given year-grade, the five observations for the different subjects for each student are correlated.
2. observations on the same student in different grade-year's are correlated
3. all other correlations are assumed to be zero
4. the variance-covariance structure is homoscedastic, i.e. the same structure is used for every student in a given model

5. the relationship between student scores for a given subject in adjacent grades is linear
6. characterizing the observation in terms of a cell mean implies that the main-effects and 2-way, 3-way, etc. interactions are assumed to be non-zero (at least potentially).

It is worth noting that this is basically a standard model used to analyze an incomplete-block, repeated measures design. In this case, the blocks are students. The blocks are potentially incomplete in that some students may stay in one system for grades 2 through 8 but others will enter or leave the system, or switch schools within a system, in any given year. The repeated measures aspect results from testing students annually. The one difference between this model and the standard set-up is the fact that these are multivariate data, i.e. each student has five observations - one per subject - each year. The model must therefore account for among-subject, within-student correlation (assumption (1) above).

Blocking by student is essential to account for variability among students. If blocking was not used, covariates corresponding to all the relevant items contributing to variation among students would have to be included in the model. Blocking implies that each student becomes his/her own control making these covariates unnecessary.

Blocking is thus highly convenient. The question is, does it actually "work" for these data? Blocking is modelled through the correlation among test scores for the same student in different years. If the data are consistent with this aspect of the assumed correlation structure, then blocking "works." The two relevant questions are: 1) is the within-student correlation significantly different from zero 2) is it positive, and 3) is it homoscedastic? Looking at the available evidence strongly suggests that the answer to each of these questions is "yes." The equality of variances was tested using Bartlett's procedure. The hypothesis of equal variance-covariance by student was rejected for some schools. However, Bartlett's is an extremely sensitive test, and with the numbers involved in TVAAS data, rejection based on trivial differences is common. In examining the variance-covariance matrices leading to rejection, this is exactly what I found - the actual differences were trivial. The bottom line is that the data strongly support the validity of the blocking assumption used in the analysis.

The second aspect of the correlation assumption is that the test scores for the different subjects for each student were correlated. An unstructured pattern of covariances were used. The data suggest that the covariances were significantly different from zero and that the assumption of homoscedasticity is valid. No further test is needed since no structure or pattern was assumed.

The only aspect of the covariance structure that could be justified from the nature of the data but was not used in the model would be a formal structure of covariance over time. For example, student scores seemed to be most strongly correlated with scores in adjacent years and decreasingly correlated as the time between scores increased. Formal correlation models, such as multivariate autoregression models, might be used to more efficiently model the data. Instead, an unstructured model was used. However, with the size of the data sets, the difference in efficiency is negligible. Moreover, the unstructured model results in a more conservative estimate of differences between schools or systems - making it more difficult to

identify a school as distinctly below- or above-average. Errors that are made using the unstructured models would thus be on the side of caution.

The assumption that all other correlations are zero seemed well justified by the evidence. Other potential correlations were estimated at the same time those used in the model were measured. None of the other correlations proved to be significant.

Thus, assumptions (1) through (4) are consistent with the data.

One issue with the model deserves mention. The cell means model implicitly includes all interactions. Many have noted that test results for given grades are inconsistent from year-to-year. This inconsistency would show up as a grade-by-year interaction. There is no question from the data that this interaction exists. The question is whether to regard it as a fixed effect - as it is in this model - or as a random effect. The statistical meaning of a FIXED grade-by-year effect is that inconsistencies result from causes that can be identified and addressed. Regarding the grade-by-year effect as RANDOM implies that the inconsistencies have no identifiable cause and therefore cannot be predicted or addressed.

Which is the "right way" to model grade-by-year interaction? This is a "subject matter" decision - educational subject matter knowledge and philosophy are crucial - not a "statistical" decision. Dr. Sanders did present compelling evidence that the more effective schools and systems tend to be much less affected by large grade-by-year interaction. This means that despite having variation in the ability of students entering a given grade comparable to other schools or systems, the better schools manage higher gain regardless of their students' starting point. This suggests that for whatever reason, the more effective schools are better able to "tailor" their program to fit the students currently in their classrooms. From a statistical viewpoint, data that behave this way are more consistent with a fixed grade-by-year effect; there DO seem to be aspects of the grade-by-year effect which can be predicted and addressed. To stress this point, regarding grade-by-year as random is the "I give up, there's nothing I can do" position. Any teaching-improvement program will invariably be more successful when such effects are regarded as fixed until all reasonable efforts to identify and address causes of variation in results have been exhausted.

Assumption (5) was that there is a linear relationship between student score in successive years. After removing "edge effects" - that is, students who performed well one year but were ill or otherwise had a "bad day" the previous or following year - the scatter plots of the remaining valid data were strongly linear. There was no evidence whatever that the degree of linearity or the spread in the data was affected by test score. Thus, assumption (5) seems justified.

The final assumption (6) is that all interactions contain in the cell mean are significant. For some interactions, especially the 3- and higher-way terms, this may be overkill. Dr. Sanders stated that these terms had not been broken out and tested specifically. There are plans to do so at some point in the future, but it is not a high priority. From a practical point of view, simplifying the model to remove negligible interactions would have little impact of the conclusions about school-wide or system-wide gain. The actual estimate of

gain would not change. Removing interactions would increase the degrees of freedom for error and thus make it easier to identify above- or below-average schools or systems. However, with the number of students in the TVAAS data sets, the practical difference would be trivial. Removing the negligible interactions has more aesthetic than practical consequence.

3. Review of the Statistical analysis

There were two aspects of the statistical analysis to review

1. How the covariance matrix was estimated, and
2. The accuracy of the computer programs Dr. Saxton wrote to analyze the data.

The estimation of the covariance was a relatively minor part of my review. The procedure Drs. Sanders and Saxton use is to estimate the school-grade-year-subject or system-school-grade-year-subject means for the "school" and "system" models, respectively, then obtain the residuals for each student and compute the covariance. SAS procedures whose accuracy has been well-established for over 20 years were used. This method of estimating covariance is a standard statistical procedure and is not the least bit controversial.

The most important part of my review concerned the software written by Dr. Saxton to analyze the data. Formally, the statistical procedure to analyze the model discussed in the previous section is called generalized least squares. It is a special case of "mixed model methodology" which is a generalization of analysis of variance and regression methods. Mixed model methodology is rapidly becoming standard practice in a wide variety of statistical applications.

Well-tested and accurate commercial software exists to do mixed model analysis. The most versatile is SAS's PROC MIXED. While PROC MIXED is well-established as an accurate and reliable program, its main limitation is that it only handles relatively small data sets. Dr. Saxton's program is designed to reproduce the analysis PROC MIXED would compute for the "school" and "system" models if PROC MIXED had the capacity to deal with size of data sets involved in TVAAS.

To test Dr. Saxton's software, I created hypothetical data sets with the characteristics that would typify TVAAS data. My hypothetical data sets were about as large as PROC MIXED can handle. I computed the covariance estimates and then used PROC MIXED to compute the school and system effects. Then I gave Dr. Saxton my data without showing him the results of my analysis. He ran the analysis using his software and we compared results. In each case, our results agreed for all the model effects at least to the 6th decimal place.

There is no question based on the testing we did that the software programs Dr. Saxton has written for TVAAS are computing the analysis correctly.

4. Summary

No statistical model is ever an EXACT description of the data being analyzed but a properly-chosen, excellent model does yield an accurate picture of the main processes driving the data. After examining the TVAAS model for statistical analysis and the software being used to compute the analysis, I am satisfied that legitimate and defensible methodology is being used. There are some modifications that could *modestly* improve the precision. However, I emphasize the word *modestly*; moreover, the impact of improving the precision would simply be to make it easier to classify schools or systems as above- or below-average. The result: to the extent that these models err, they err on the side of caution.

A Review of the Tennessee Value-Added Assessment System (TVAAS)

David A. Harville
Professor of Statistics
Iowa State University
June 6, 1995

General Comments

TVAAS is a system for estimating the influence of individual teachers, schools, and school systems on the educational progress of their students. It was developed by combining (in a rather ingenious way) a very powerful statistical methodology, known as mixed model methodology, with the concept that, during each year, the effect of each teacher (or school or school system) is to add incrementally to each student's ability in one or more subjects. TVAAS provides a sound and objective basis for comparing the additions made by each teacher (or school or school system) to their students' abilities with those made by their peers or with national norms. Its use serves to identify teachers, schools, and school systems that are seriously deficient.

For each teacher (or school or school system), TVAAS can provide an accurate and unbiased estimate of the average annual addition they have made to their students' abilities in each of their subjects. It can also provide a standard error (a measure of the precision of the estimate). The nature of the estimates, together with the availability of standard errors, insures that no teacher (or school or school system) will be "targeted" unless there is very substantial evidence of serious deficiencies. TVAAS has significant advantages over other statistical methodologies that have been proposed for assessing the influence of individual teachers, schools, or school systems. In particular, systems that (unlike TVAAS) attempt to account for student differences by using concomitant information tend to be plagued with problems occasioned by missing information. TVAAS accounts for such differences in a more direct and more effective way.

In the development of TVAAS, a number of choices had to be made with regard to the underlying model, model assumptions, etc. These choices appear to have been made in a very competent and appropriate way. The overriding considerations were accuracy and fairness.

TVAAS is a relatively sophisticated and complex system. It is computationally intensive and requires powerful computer facilities for its implementation. Great care has been exercised in developing the requisite software and procedures, and the software and procedures have undergone extensive testing.

In summary, TVAAS appears to be a statistically sound and appropriate system for estimating the influence of individual teachers, schools, and school systems. A great deal of

time, effort, and thought have gone into its development, and the assumptions underlying every "new wrinkle" have been validated for a variety of real data sets. While TVAAS itself is relatively new, the statistical methodology that underlies TVAAS has been successfully adapted for use in many areas of application — refer to Robinson's 1991 review (*Statistical Science*, vol. 6, pp. 15-51). I anticipate that this methodology will be just as successful in its application to the evaluation of individual teachers, schools, and school systems as it has been in its application to problems in other areas.

Some Specific Comments

1. *Quality control.* Adequate safeguards have been taken to insure that the assessments provided by TVAAS are essentially error-free. Much merging of test scores and much other data preparation must be carried out before the test scores can be processed by TVAAS. In developing procedures for accomplishing these preliminary tasks, great care has been taken to avoid misidentification of students or teachers.

2. *Estimation of variances and covariances.* It is assumed that a student's test scores (corresponding to as many as 5 years and 5 subjects) are correlated. The variances and covariances of these scores must be estimated. These estimates are required as input to TVAAS; TVAAS exploits the information in these estimates in estimating the influence of individual teachers, schools, and school systems. Separate estimates are obtained for each school system. For purposes of estimating the influence of individual teachers, the variance among teachers must also be estimated. Two approaches to the estimation of the variances and covariances have been devised: the more sophisticated (and more computationally intensive) of the two approaches is known as REML (for restricted or residual maximum likelihood); the second approach is simpler but possibly less accurate. REML is to be used when it is computationally feasible to do so. The methods that are being used for estimating variances and covariances are adaptations of widely used and widely accepted methodologies and can be expected to produce very satisfactory estimates.

3. *Teacher evaluation.* In the case of teacher evaluation, TVAAS is based on a linear statistical model that considers a student's test score to be the sum of the following three quantities: a parameter that is specific to year, grade, and subject, a sum of teacher effects (that are specific to year, grade, and subject), and a residual effect. The teacher effects are those for the teachers that the student had for that subject during the year of the test and during previous years. Appropriate modifications are introduced to account for team teaching and for any other deviations from standard practice. The teacher effects are modeled as realizations of random variables; and as a consequence the estimates of the teacher effects are what are known to statisticians as shrinkage estimates — this insures that

no teacher will be identified as significantly worse than other teachers in the same school system unless the evidence for such a difference is very substantial. In short, the model provides a very appropriate basis for teacher evaluation and has produced very sensible results when applied to a wide variety of data sets.

4. *Variability of results.* A report released recently (in April 1995) by the Comptroller of the Treasury of the State of Tennessee (entitled "The Measure of Education: A Review of the Tennessee Value Added Assessment System") was critical of TVAAS. In particular, the report raised questions about the relatively large yearly differences in the value-added scores for certain school systems and about the rather large differences in the value-added scores for two seemingly similar schools. The apparent implication was that these differences reflect poorly on TVAAS and raise questions about the assumptions that underlie it. However, in the example (cited in the reports) of yearly differences, it appears that TVAAS was accurately reflecting the information in the data. That is, the differences are "real," even though the reasons for the differences may not be obvious. Moreover, in the example (cited in the report) of school-to-school differences, the authors of the report made a basic error in determining the standard error of a difference of two estimates — the variance of the difference (which is the square of the standard error) should be the sum of the variances of the two estimates, not the larger of the two variances.