DOCUMENT RESUME

ED 476 134 TM 034 888

AUTHOR	Doran, Harold C.
TITLE	Value-Added Analysis: A Review of Related Issues.
SPONS AGENCY	Carnegie Corp. of New York, NY.
PUB DATE	2003-04-15
NOTE	44p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).
CONTRACT	B7361
PUB TYPE	Information Analyses (070) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Accountability; *Educational Assessment; Educational Improvement; *Measurement Techniques; Models; *School Effectiveness; Statistical Analysis
IDENTIFIERS	*Value Added Model

ABSTRACT

This paper makes the case that value added assessment (VAA) is one potential tool for increasing the accuracy of inferences regarding school effects, even though it is far from being a panacea. The promise of VAA rests in its focus on individual growth trajectories and its perceived ability to isolate the influence of a school from other factors believed to influence a student's learning. VAA may decrease the mistakes that have been made when measuring school effectiveness, but not eliminate them completely. VAA may be one method for reducing false positives, or the Type II error rate, by more appropriately estimating growth over time as compared to conventional cross-sectional techniques. This paper explores issues related to VAA with the aim of balancing statistical concerns with logistical issues. Value added analysis should be one component of accountability, and if the models are appropriately specified, then they will be one source of information to support appropriate classroom and public action. Appropriate specification is not easily accomplished. Value added models are complex, but can be an important tool for the accountability movement. (Contains 57 references.) (SLD)

Running Head: VALUE-ADDED ANALYSIS (AERA)

Value-Added Analysis: A Review of Related Issues

Harold C. Doran

New American Schools, Alexandria, Virginia

April 15, 2003

hdoran@nasdc.org

This research was supported by Grant B7361 from the Carnegie Corporation of New York.

The statements and views expressed within are solely those of the author.

Draft for comment and review

Introduction

The argument for increased accountability as measured by educational outputs has become the *sine qua non* of school reform. Test-based accountability assumes that assessment results can serve two masters, the internal school function and external accountability function. The internal function describes the extent to which school officials use assessment and accountability data to improve the school's instructional program. The external function describes the extent to which data from accountability models can provide valid representations regarding the adequacy of the school's instructional program and its staff to support public action. This has motivated researchers, practitioners, and policymakers to explore methods by which data can be analyzed to extract the impact of non-school related factors and attempt to estimate the contribution of a school to a student's increase in knowledge. Among these methods, value-added analysis (VAA) has emerged as one promising tool.

The promise of VAA rests in its focus on individual growth trajectories and its perceived ability to adequately isolate the influence of a school from other factors believed to influence a student's learning. Yet, this potential should not equate to wide-scale implementation before VAA has been submitted to empirical and practical scrutiny.

In this paper, I argue that VAA is one potential tool for increasing the accuracy of inferences regarding school effects, yet it is far from a panacea. It may considerably decrease the mistakes that have been made when measuring school effectiveness, but not completely eliminate them. In the parlance of methodologists, VAA may serve as one method for reducing false positives, or the Type II error rate, by more appropriately estimating growth over time as compared to conventional cross-sectional techniques.

This paper explores six issues related to VAA. My purpose is to balance statistical concerns with logistical issues, such that the ultimate merit of VAA is not based only on its robust statistical properties. Instead, I argue that these types of analyses should be judged by the extent to which they provide information that leads to more appropriate classroom and public action beyond that which is currently available. This implies positive social impact, otherwise referred to as the consequential aspect of validity (Messick, 1989).

What is Value-Added Analysis?

The educational vernacular is often overflowing with jumbled phrases that do more to confuse than clarify. A layperson walking into this arena often feels like a humanoid sitting at the Star Wars bar, without any capacity to understand surrounding conversations. However, the term "value-added" can be readily understood when it is reduced into a simple question—How much *value* has a school *added* to a student's learning? I refer to this as the fundamental question. A more appropriate methodological definition will be suggested later in this paper.

The question, as simple as it may sound, is filled with cavernous concerns, limitations, and problems. Yet it may be one source of nourishment needed to supply information relevant to school effects and to support appropriate accountability action and policy, such as the No Child Left Behind Act of 2001 (NCLB).

From the outset, the fundamental question requires that the implicitness hidden within it be made explicit before it can be answered. Specifically, an operational definition of learning is needed as well as realization that schools are only one source of learning. Last, one must consider the quality measures typically used to assess changes in student abilities.

First, *learning* implies that an *individual* has changed in some domain from one measured occasion to the next. Despite this relatively straightforward, yet crude, definition of learning, evaluations of schools have often been based upon a single year's test score, otherwise referred to as a cross-sectional method. Additionally, inferences regarding individuals have been plagued with analyses at the aggregate level, not the individual student. Measurement in this form is inconsistent with the VAA question at hand; change over time at the individual student level.

One possible explanation for the sparseness of longitudinal research methods in education may be rooted in the rather forceful and extensive debates regarding the measurement of change (Cronbach & Furby, 1970; Linn & Slinde, 1977; Rogosa, Brandt & Zimowski, 1982). Undoubtedly, many researchers recall days when the classical formula for the reliability of the difference score was presented and used to chastise the gain score. Unfortunately, many remain convinced that change cannot be reliably measured; therefore it should not—at least in education. However, unreliability does not imply that gains are imprecise measures of change (Collins, 1996). Rogosa et al summarize:

> The crucial message is that low reliability does not necessarily imply lack of precision. Although individual differences in growth are necessary for high reliability, the absence of such differences does not preclude meaningful assessment of individual change (p. 731).

Too often the gain score has been eschewed as an artifact that renders the measurement of change to be useless. This is as untrue today as it ever was; yet misunderstanding has impeded progress.

A second belief implied in the value-added question is related to *who* has added the value. Because many factors affect student learning, any attempt to assign causality must disentangle other non-school related factors from the distribution of scores. In fact, much of the popularity that value-added analysis has gained over the last few years is due to the supposition that these methods have the capacity to isolate a school (or teacher's) contribution to a student's learning. However, in the absence of random assignment to treatments, school evaluations are quasi-experiments. As such, the assignment of causality without strong methodological and/or statistical controls is often fallacious. Given the non-random nature of schooling assignment, sufficient care is warranted before assuming schools are the reason scores are measurably different.

Third, if change can be measured, then the quality measures must be considered. That is, a central concern surrounding the debate is the quality of instruments used to measure student knowledge. Using standardized tests instead of other more authentic forms of assessment concerns practitioners as a limited view of student knowledge. Nonetheless, standardized assessments often include metrics that are more desirable for measuring change than other forms of assessment. This issue is discussed in the last section of this paper.

A Brief History

Although a variety of analyses falling under the banner of value-added analysis have been in place, it is likely that the Tennessee Value-Added Accountability System (TVAAS) developed by Dr. William Sanders is the most well recognized and first system to have been employed for an entire state (Ceperley & Reel, 1997). TVAAS uses mixed-model equations as the statistical methodology to analyze student outcomes from standardized assessments collected from individual students over time.

The longitudinal nature of responses presents the opportunity to use each student as his or her own control, otherwise known as a blocking design. As such, the error variance may be further decomposed allowing for more precise estimates of the treatment to be obtained (Kirk, 1995). No other covariates are used to control for non-random assignment in the TVAAS model. The longitudinal nature of student responses also presents the threat of missing data. However, the methodology employed by Sanders does not require imputation nor are any fractured response vectors deleted listwise as is often found the case when traditional software packages are employed.

A second well known value-added example can be found in the Dallas Independent School District (DISD). The methodology developed by Webster and Mendro (1997) also uses longitudinal test score information to estimate the contributions of schools to each student's learning. However, the unique methodology employed makes use of two statistical approaches, Ordinary Least Squares (OLS) and Hierarchical Linear Models (HLM)[1].

[1] Because HLM includes a fixed and a random effect, the terms HLM and mixed-models are used interchangeably.

In the first stage, student responses from a standardized test are regressed on "fairness" variables. These fairness variables are covariates thought to influence a student's learning and include gender, ethnicity, language proficiency, and free and reduced lunch status. The residuals from the OLS model are then used as the outcome in a two-level HLM approach, where covariates are again entered as statistical controls.

Both the TVAAS and DISD approaches have been used to identify effective teachers and included as components in state and district accountability plans. Additionally, both approaches include the use of longitudinal standardized test scores and a complex statistical approach to estimate school effects.

However, the contrasts between the two approaches elevate uncertainties that warrant further exploration before large-scale implementation or selection of a model. These uncertainties set the stage for the remainder of this paper as described below.

First, both the TVAAS and DISD approaches employ complex statistical methods. Therefore, can any gain score analysis produce unbiased estimates of school effects as well as those that employ complex methodologies? Second, both make use of a test metric from a commercially available test. Therefore, what considerations should be given to the choice of a test score metric when designing a value-added approach? Third, both methodologies seek to ascribe effectiveness labels to those that have produced gains beyond a certain criterion. How then does one operationalize effectiveness when the only consideration is a test score? Fourth, should covariates be included to control for non-random assignment? Fifth, because the data requirements are extensive, what systems must be in place to facilitate the use of a value-added methodology? Last, what are the threats to validity?

What Can Legitimately Be Considered a Value-Added Methodology?

The range of potential methods for evaluating school effects leads one to question whether one methodology is better suited for the task than others. Even more to the point, can some approaches, such as a simple gain score analysis on NCE scores, be considered among the tools for performing VAA? To answer this question, it seems that formulating criteria, rather than a blanket decision, is more appropriate. In this regard, one may relate the VAA to the criteria and determine whether the model appropriately answers the fundamental VAA question. Therefore, I borrow from statistical theory and suggest that value-added models must be multivariate, account for correlated errors, be sufficiently flexible, and consider methods for increasing the precision of estimates.

The Multivariate Nature of Schools

Schooling is by definition multivariate. It is a process that occurs over time, affects students in more than one way, and more than one dependent variable (DV) is often reported when students are assessed. The multivariate nature of schooling suggests that VAA should consider the covariance structure between individual responses and among the DV set. Stevens (2001) presents three compelling reasons for considering more than one criterion for evaluating treatment effects:

1) Any worthwhile treatment will affect the subjects in more than one way;
2) The use of multiple DVs permits a more complete and detailed description of the phenomena under study; and

3) Treatments can be expensive to implement, while the cost of obtaining data on several DVs is relatively small, and maximizes the information gain.

By means of matrix algebra, it can be shown that the covariance between the DVs affects the determinant (generalized variance) of the variance-covariance matrix, a single number that describes the variability among the DV set. When the DVs are correlated, then the generalized variance summarizes the shared variability among the DV set. In other words, the variables have something in common and should be analyzed concomitantly. If one can assume that the DVs are orthogonal, then the multivariate considerations may be ignored as the generalized variance is not affected in this case. However, this is rarely, if ever, the case with educational data.

Performing multiple statistical tests further compounds the problem, and may lead to the acceptance of spurious results. For example, performing separate regressions for different grades and on the different outcome variables would likely result in school or teacher estimates that are not trustworthy[2]. Introducing a Bonferroni adjustment is often used as an alternative solution to control for experimentwise error rates. However, this often results in nominal alpha rates that are extremely low as well as decreasing statistical power. For this reason, the degree to which the multivariate nature of the data is respected is posed as one criterion for evaluating a value-added model.

Accounting for Correlated Errors

Second, dependence among the student responses violates the assumption of independent errors, thereby producing standard errors that are too small and inaccurate significance tests (Osgood, 2001). These correlated observations can lead to a dramatic

[2] Assuming the tests are independent, the overall alpha can be found by $[1-(1-\alpha)^k]$

effect on the Type I error rate (Scariano & Davenport, 1987). Ignoring the intra-class correlation and using statistical methods such as (M)ANOVA, (M)ANCOVA, or OLS regression techniques increases the likelihood that the model is misspecified. When model misspecification occurs, unbiased estimates of the treatment effect cannot be reliably obtained.

Not only is it imperative that the pure clustered nature of the data is accounted for (i.e., when all cases are nested within only one larger unit), but that consideration is given to the cross-classified nature of the data (Raudenbush & Bryk, 2002; Goldstein, 1995). The cross-classified model recognizes that students transfer teachers each year (in most cases) and permits for the variance components to be further decomposed into variance among schools, variance among teachers within school, among students within classrooms, and within students over time (Rowan, 2002). As such, the teacher variance component may be compared to other sources of variability to assess the teacher's contribution to student learning. For example, by comparing the ratio of teacher variance to the growth rate variance, one can assess the proportion of gain variance that can be attributed to teacher (Raudenbush & Bryk, 2002). When the model has been appropriately specified, the proportion of variance accounted for by teacher can be considered the value-added. This definition appropriately recognizes that multiple factors affect learning (gains) and identifies the fractional portion of change that may be attributed to the classroom teacher.

The nested structure of educational data almost guarantees that student responses are not independent. Consequently, accounting for the intra-class correlation among students is suggested as a second criterion.

Model Flexibility

Third, the assumption of learning requires the measurement of change over time. There are at least two methodological concerns to consider in this regard. First, the spacing of the measurements has historically presented researchers with limitations. The general assumption required for the measurement of change using structural equation models requires that all measurements be equally spaced over time and for each unit to have a balanced data set (Raudenbush & Bryk, 2002). Mixed statistical models relax this assumption and permit for the time-series data to be unequally spaced and for individuals to have incomplete data without being deleted from the analysis (Raudenbush, 2001). Maas & Snijders (2003) report:

> An important advantage of the multilevel approach is that incompleteness of the data on the dependent variable does not complicate the analysis, provided that missingness is at random. The missing observations simply can be omitted from the data set (p. 72).

Although the missing data permit for the analysis to proceed without losing cases, it is incumbent upon the researcher to examine the reason for the missing data. The flexibility of the mixed-model should not relax the attention one pays to the sparseness of the data set.

Longitudinal test score data from schools are generally administered in intervals that are not truly equal. Additionally, students often change schools or miss tests. Therefore, I suggest that a third criterion be the flexibility of the statistical model for dealing with unbalanced data and unequally spaced time-series.

Improving Precision

Educational data are inherently unreliable. However, a number of methods exist for improving the reliability of estimates, and should be used to assess true growth rates. First, shrinkage estimators, or Empirical Bayes estimates, may be employed to increase the precision with which we measure treatment effects. Using Kelley's Equation (Wainer et al, 2001), $\rho_{.j}x + (1 - \rho_{..})\mu$, observed score estimates are improved when the they are regressed proportional to the reliability of measurement. Specifically, when the reliability of the observed score (x) is high, more value is placed on the observed score. Otherwise, strength from the larger data set is "borrowed" such that more value is placed on the grand mean (μ).

One benefit of Item Response Theory (IRT) over Classical Test Theory (CTT) is that measurement error varies as a function of ability, θ, rather than remaining constant over the entire score distribution. Weighting observed scores by their measurement error may also provide more reliable estimates. This issue is explored in greater detail in the next section as it relates specifically to the choice of the test metric.

Another consideration surrounds the number of observations in the time-series before growth can reliably be assessed. Of course, only linear models can be formulated when two time points are available. This limits the analysis of change, especially when one suspects a trend exists in the data. The simple posttest on pretest regression introduces an artifact referred to as regression to the mean (RTM). RTM occurs by mathematical necessity whenever two variables are imperfectly correlated, which is almost always the case with educational data. When present, the degree of regression to the mean can be calculated as RTM = $(1 - \rho) * 100$, where ρ is the Pearson Product

Moment Correlation. RTM describes the amount of regression towards the population mean that would be expected, given the imperfect correlation. This psuedo-effect (Trochim, 2002) may suggest progress, or even decline, when in fact the change is solely due to regression, not teaching. In this scenario, the unsuspecting researcher may incorrectly claim that change is due to teacher or school effects.

Venter, Maxwell, and Bolig (2002) demonstrated the power of adding an intermediate time-point to a pre/post design. They conclude that it permits a researcher to assess whether non-linearity may exist in the data and adds to the statistical power to detect treatment effects. Three observations are better than two, but having only two observations shouldn't exclude the possibility of VAA. However, the two-wave study provides very limited information about true student change over time. For this reason, adding the entire vector of available student scores provides more reliable parameter estimates and permits for the true growth rate to be more reliably measured (Rogosa et al, 1982).

Improving the reliability of growth estimates can be improved via a number of empirical and design methods. Because the degree of imprecision is often estimated it should be considered as an element to improve statistical analyses. Therefore, the degree to which reliability is considered is posed as the last criterion.

The criteria posed are certainly not exhaustive. However, analyses falling under the banner of value-added should respect the multivariate nature of the data and the correlated observations, use statistical models that are sufficiently flexible, and implement design or statistical corrections to account for the unreliability of data. Otherwise, justification should be offered. If the justification cannot be substantiated in light of statistical or methodological theory, the results from the models cannot be appropriately used to produce credible estimates of school effects. As such, they are also not appropriate for answering the VAA question.

Does the Metric Really Matter

The array of metrics that accompany most commercially available standardized tests presents the researcher with yet another decision. Choosing between one of the norm-referenced scores (e.g., normal curve equivalent, grade equivalent) and one of the content-referenced scores (e.g., scale scores, performance levels) may also have consequences for the estimates of the treatment and interpretation of results. To increase interpretation, I argue in support of content-referenced scores as the appropriate metric for measuring changes in learning.

It was common for many years to see Normal Curve Equivalent (NCE) scores used in Title I research applications (Davis, 1991). However, Rogosa et al (1982) have demonstrated that working with standardized measures constrained to have equal variances has two potential negative impacts. First, it guarantees that the correlation between initial status and change will always be less than or equal to zero. In this scenario, a researcher would inappropriately infer that an inverse relationship exists between initial status and average rates of change over time. Second, standardization eliminates the potential to identify a trend in the mean or variance over time (Goldstein, 1995). Trends in the data often supply rich information. Moreover, when non-linear trends are present, basic techniques of differential calculus can be applied to examine the instantaneous rate of change with respect to time.

Thum (2002) contends that because the NCE score is norm-dependent, it is not well suited for measuring change over time. He maintains a more reasonable measure is the IRT scale score. Thum goes as far as to say, "the scale score is not just the preferred alternative for the study of change; of the various choices discussed above, it is the only option" (p. 7).

The benefits of the developmental scale scores from commercially available tests permit for student progress to be analyzed over an extended period of time, exactly the type of metric needed for growth curve models. Second, change in scale score units over time represent increases in knowledge and skills, not changes in a measurement scale. Peterson, Kolen, and Hoover (1989) report:

> Developmental score scales, in which tests designed for use at different grade levels are calibrated against one another to span several grades, facilitate the estimation of an individual's growth (p. 231).

Hoover (1984) cogently argues in favor of the grade-equivalent score (GEs) over other extended developmental scales for measuring the progress of individual students over time. Even more recently, Hoover (2003) presents his position for selecting norm-referenced scores as the metric of choice for measuring change. He challenges the conception that norms can only be used for comparison purposes.

However, GEs are still norm dependent and cannot be tied to content attainment. That is, GE scores are developed based on the distribution of raw scores (Nitko, 1986). As such, it is difficult to conceive of their relationship with the actual content of the test. Furthermore, no methodology has been produced for relating a GE score to the knowledge and skills a student possesses.

One convincing reason to avoid the GE score is the misrepresentation of student abilities often associated with any given GE. For example, a Grade 3 student with a GE score of 5.6 in math does not imply that this student has the knowledge and skills necessary to join other Grade 5 students in their curricular activities.

Ebel (1962) documented the fallacy that norm-referenced scores relate to the knowledge and skills covered in a test. He aimed to increase the interpretability of test scores via an approach referred to as item mapping. Ebel explains:

> It is not very useful to know that Johnny is superior to 84 per cent of his peers unless we know *what it is that he can do better than they* (emphasis added), and just how well he can do it! (p.18).

Unlike the GE score, Zwick, Senturk, Wang, and Loomis (2001) demonstrated the potential of item mapping techniques to characterize IRT scale scores as what a student knows and can do. Among their findings, it is reported that exemplar items were identified that provide illustrative support for the knowledge and skills a student must have in order to be within a cut score category (viz., below basic, basic, proficient, advanced). Techniques such as these are much more helpful to practitioners as they cement the abstract cut scores with illustrative items, thereby increasing the interpretability of a scale.

Burket (1984) presents another view in response to Hoover:

> From the standpoint of scaling, the most significant potential advantage of IRT stems from the fact that IRT models the probability of the correct response of an examinee to an item, and therefore permits the interpretation of a test score in terms of what that score implies about the examinee's ability to perform. Given an

> appropriately chosen set of calibrated benchmark items, this means that true criterion referencing of test scores is possible (p. 15).

IRT provides an additional benefit over CTT statistics. Specifically, IRT reports a unique standard error of measurement (SEM) for each maximum likelihood estimate and quantifies the degree of imprecision at θ. Formally, the SEM is reported as the squared inverse of the test information function, $I(\theta)^{-1/2}$, and varies as a function of θ. Because IRT more appropriately recognizes that error is not the same at each level of ability, it is possible to use the SEMs as weights in the VAA. Specifically, weighting each observed score by the inverse of its SEM, or SEM^{-1}, may provide an additional element of precision to the results of the VAA.

Seltzer, Frank, and Bryk (1994) compared the use of GEs and IRT-based scale scores in analyses of growth to determine whether they lead to congruent, or dissimilar inferences. Seltzer et al report that IRT scales and GEs do not produce different results when considering educational status. However, the different metrics do suggest differing rates of growth as well as different patterns of variability over time. They also corroborate the argument that item maps can be formulated from IRT-based scales to increase scale interpretability. They conclude that IRT-based metrics provide a more realistic portrait of change within schools and are better suited to support the formative role of assessment.

Although IRT scales may be more appropriate for measuring growth, it is assumed that they are interval in nature and can be used in parametric analyses. Ballou (2002), however, challenges this assumption of IRT score scales. He speculates that gains made by students at different ends of a score distribution may differ in the amount of gain, as defined by the scale, but they may not actually differ in the amount of knowledge gained. He states that:

> As prominent psychometricians have pointed out, many of the usual procedures for comparing achievement gains yield meaningless results if the ability scales lack this [equal interval] property (p. 15).

It is important to note that no citation of any prominent psychometrician is included around this statement. Reviewing a few classical textbooks, one can readily find statements arguing that the use of parametric tests is warranted when the quantitative variable approximates an interval scale (Jaccard & Becker, 1997). Nunnally and Bernstein (1994) state:

> We have already stressed the issue of whether scores on a conventionally scored test form an interval scale, and they [proponents of nonparametric methods] have often argued that they do not. We strongly suggest that this position can easily become too narrow and counterproductive (pp. 20-21).

No test metric provides a perfectly interval scale. Assuming, latent traits can be scaled in the same manner as time or Kelvin's is unreasonable. Rather, scores that are at least suitably equated and approximate an interval scale can justifiably be used in parametric analyses for measuring change over time. However, some consideration may be given to the nature of the growth required by students at the different ends of a score distribution.

Consider the following cumulative distribution functions (CDF) for the Stanford 9 in Grades 1 through 5. The y-axis is the percentile rank while the x-axis represents the scale score corresponding to each percentile rank at each grade level.

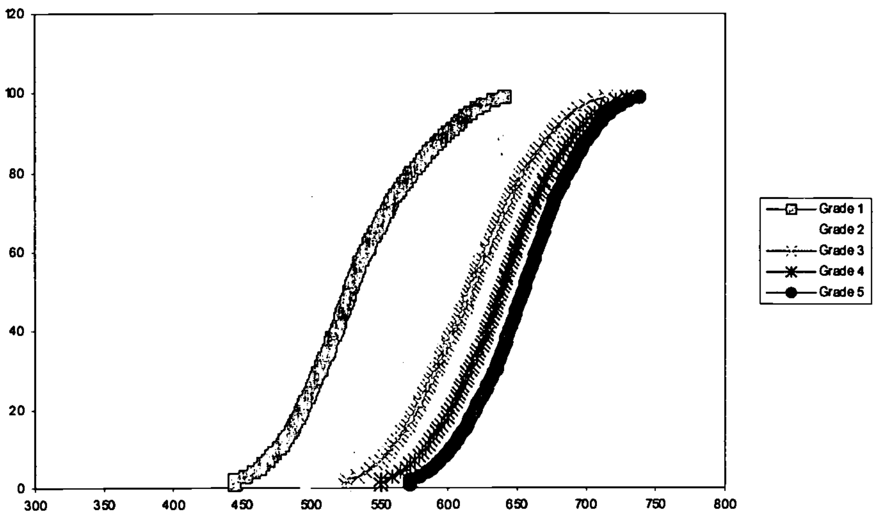

Figure 1 Stanford 9 Reading CDFs, Grades 1 through 5

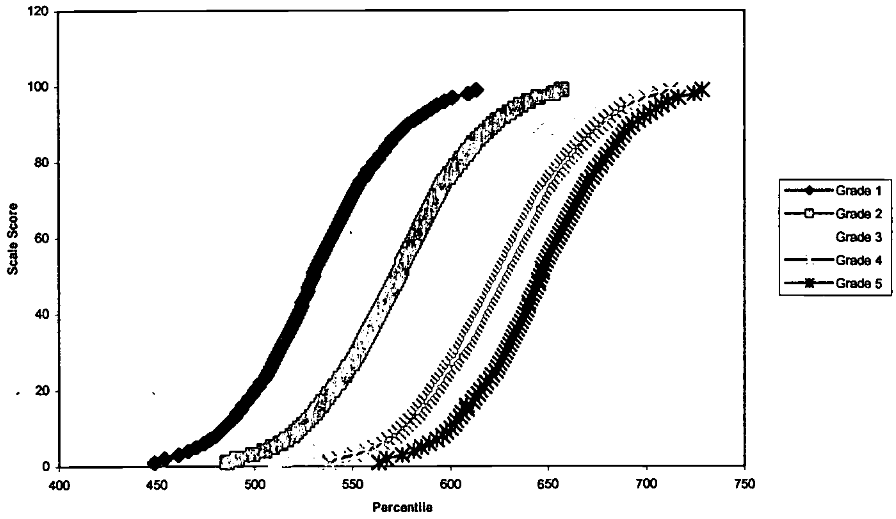

Figure 2 Stanford 9 Math CDFs, Grades 1 through 5

The scale score values were obtained from the Stanford 9 Table of Norms (1997) for each percentile rank. Holland (2002) demonstrated that at least two types of gaps could be measured from CDF distributions, vertical and horizontal. Holland defines

F(x) = *proportion* of the group with test scores less than or equal to x and assumes the gaps can be measured when both distributions are stochastically ordered. This permits for one to measure the vertical gap as D(x) = F(x) – G(x). This measure describes the percentile associated with the same scale score over different CDFs.

The horizontal gap is defined by Holland as p = F^{-1}(p), where *p* is a proportion between 0 and 1. This measures the horizontal difference between two CDFs at a percentile rank and describes the amount of change in scale score units a student must gain simply to maintain the same percentile rank.

Figures 3 and 4 present the horizontal gap measured at each fixed percentile rank across Grades 1 through 5 for reading and math Stanford 9 scores, respectively. These figures show the amount of gain a student would need to make, in terms of scale scores, to maintain their pretest percentile rank on a subsequent administration of the reading and math tests. Clearly, the amount of growth required to maintain the same percentile rank differs as a function of percentile.

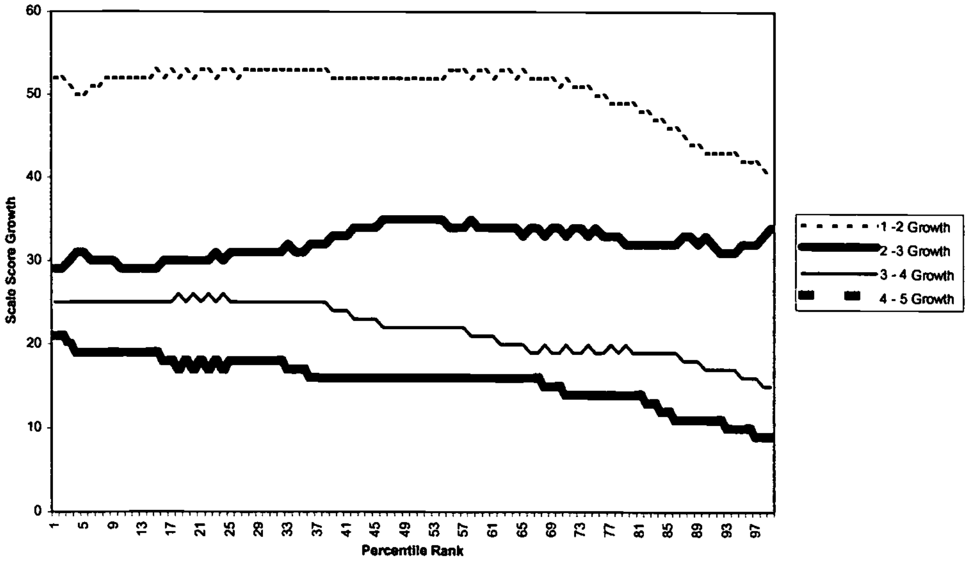

Figure 3 Horizontal Gaps for Stanford 9 Reading, Grades 1 through 5

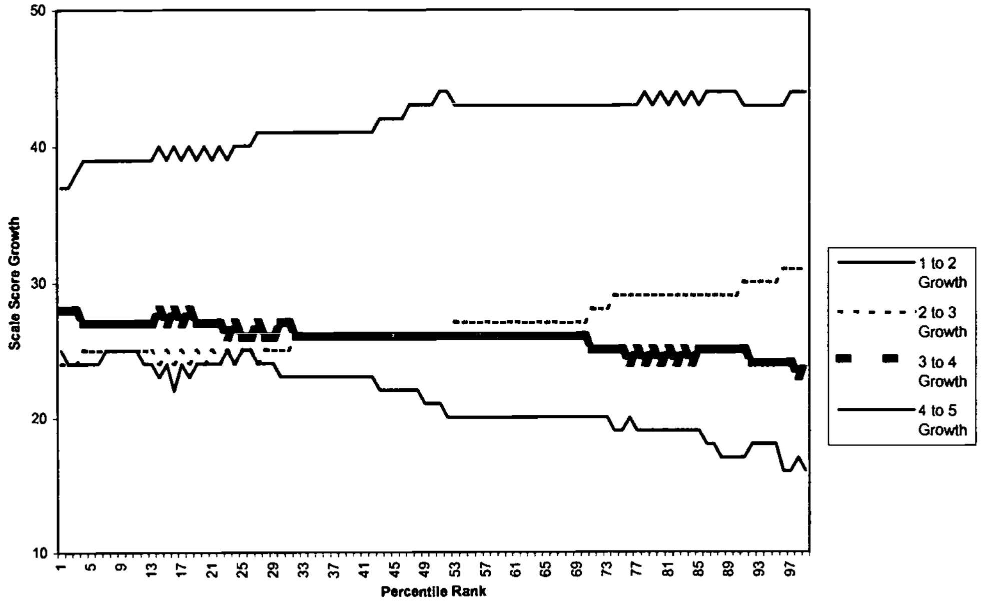

Figure 4 Horizontal gaps for Stanford 9 math, Grades 1 through 5

Because item parameters for the Stanford 9 are estimated using the Rasch model one may question the logic of Rasch for the multiple choice test. Divgi (1986) reports "As pointed out by Traub (1983, p. 63), the assumptions of equal discrimination and zero guessing parameters for multiple choice items 'fly in the face of common sense and a wealth of empirical evidence accumulated over the 80 years" (p. 292). Curiously, the Stanford 9 Technical Data Report (1997) reports both *p* values and point-biserial correlations, the classical cousins to difficulty and discrimination in IRT. It is worth noting that the guessing parameter becomes less of a concern in computer adaptive testing (CAT) situations, if the CAT is working properly (Wainer & Mislevy, 2000).

Furthermore, Slinde and Linn (1979) found that in certain conditions the Rasch model does not "result in an adequate vertical equating of existing tests" (p. 162). If the equating were working properly, one might expect a horizontal line across all percentile ranks, suggesting that the amount of growth required to maintain a pretest percentile rank is the same regardless of percentile rank. It may be of interest to superimpose obtained and predicted scores over Figures 3 and 4 and compare the results.

Although norm-referenced scores may provide a scale for measuring change, these measures are less desirable than content-referenced scores for two primary reasons. First, content-referenced scores can be related to student skills, where norm-referenced scores cannot. Second, changes in the norm-referenced developmental scale may be an artifact of the distribution, and not related to a student's increase in skills. In the spirit of providing useful information to support classroom action, IRT-based scale scores are better suited to analyze changes in learning.

Defining Effectiveness

Aside from the statistical and psychometric considerations, questions surround the means by which effective schools are classified. If value-added models purport to measure school effects, then an operational definition of effectiveness must relate any identified effects to the school. In other words, policies aiming to hold schools accountable are only palatable when measurement models can provide information relevant to schools effects, not school *and* contextual effects.

The central measure of success in Title I school evaluations was centered upon, and will continue to be, Adequate Yearly Progress (AYP). The fabric of AYP has often taken the form of test scores, where those wearing the robe in the winner's circle are deemed effective by demonstrating changes in test scores at or above a specified criterion.

Schwarz, Yen, and Schafer (2001) maintain that measures of AYP are a policy decision rather than a measurement decision, and must consider both the adequacy and attainability of AYP goals. Adequacy describes the extent to which the goals can be described as meaningful and rigorous. Attainability, on the other hand, balances the adequacy consideration by ensuring that the goals are realistically within the reach of the school population.

AYP is also a policy decision, not a measurement decision, because assessments are not constructed with any metric that provides a benchmark defining how much growth a student must make to be considered adequate. Although some may argue that a student should increase their GE score by 1.0, or maintain their pretest percentile rank, these definitions are unsupportable (Nitko, 1986).

If AYP is to be recognized as the benchmark of "good", then schools should be recognized as good for the work that *they* do, not the work they do in combination with other events outside their locus of control.

Historically, good schools have been perceived as those with high test scores, not test score gains (or even non-quantifiable characteristics). The legend of Lake Wobegon, where all of its children are above average, is still the aim for many practitioners and policymakers. However, few Title I evaluations have considered a school's test scores aside from non-school related factors.

Test scores are tainted by many factors exogenous to the school. For example, schools located within rich contextual environments may have higher test scores. However, this may be an artifact of location, and not due to instructional quality. Any failure to reconcile the influence of non-school related factors in quasi-experiments of school performance confounds the assessment of treatment and produces estimates of school performance plus that of the exogenous variables. Controlling for exogenous variables may take the form of research design or statistical controls, or both. This is explored in greater detail in the next section.

The optimal method for mitigating the influence of exogenous variables would be through randomization. In this regard, the correlation between these variables on student performance could be considered null. However, students are not randomly assigned to schools or even teachers. Instead, schools are highly homogenous in their populations given that many factors, including economic status and parent level of education, drive home purchase and school selection.

A review of the literature suggests a number of methods for defining school effects. Raudenbush and Willms (1995) differentiated between two types of school effects, Types A and B. They define these effects as:

> The Type A effect is the difference between a child's actual performance and the performance that would have been expected if that child had attended a "typical school".... The Type B effect, then, is the difference between a child's performance in a particular school and the performance that would have been expected if that child has attended a school with identical context but with practice of "average" effectiveness." (pp. 309-310)

The Type A effect includes the effects of the school, but is confounded by the exogenous variables likely to affect a student's test score such as the socio-economic status of the school. This is the most common effect reported, but is misleading as it does not serve the true accountability function. If the fundamental purpose of accountability is to determine whether the *school* is performing well, then labeling schools as effective or not under a model estimating Type A effects would clearly be inappropriate. In addition, Type A effects do not relate to the fundamental question of VAA, how much value has the school added.

The Type B effect, on the other hand, would more appropriately assess the value-added question given that its aim is to isolate the effects of the school from the impact of the exogenous variables. However, formulating a model to estimate Type B effects forces one to reconcile the relationship between exogenous variables and school performance. Any model that fails to consider the influence of known exogenous variables will inherently produce biased estimates of the treatment effect.

Meyer (1994) offers two methods for considering school effects, the *total* school performance and the *intrinsic* school performance. Although the methodology used by Meyer differs from that of Raudenbush and Willms to assess school effects, the *total* and *intrinsic* performance measures are consistent with the Type A and B effects as defined above, respectively.

Thum (2002) has formulated an operational definition of effectiveness that aligns with the NCLB expectation of AYP and is consistent with Type B effects. In fact, he refers to this measure as AYP-NCLB. It is described as the minimum amount of growth an individual student would need to make if he or she is to be at proficient by the end of the NCLB timeline, 2014. Using this measure of AYP, states and autonomous districts could report the proportion of students at proficient each year and modify the annual measurable objective based on yearly performance of the population. This would not have the effect of lowering expectations. In fact, it would have the opposite effect. The annual measurable objective would be modified to be more or less rigorous based on student performance.

Residuals from OLS regression-based models have also been discussed as one method for estimating school effects (Darlington, 1997). This method is justified on the basis that $E(\varepsilon_{i2}|X_i) - E(\varepsilon_{i1}|X_i) = 0$ for all *i* (Taris, 2000) and that ε_{i2} and ε_{i1} are independent (Nunnally & Bernstein, 1994). Lord (1963) stated:

> All this has led some people to assert that deviation from the regression line is the real measure of change and that the ordinary difference between initial and final measurements is not a measure of change. This can hardly be correct" (p. 23).

It is relatively easy to dismiss the credibility of OLS residuals as unbiased estimates of school or teacher effects for two reasons. First, OLS assumes that the intra-class correlation is zero, a highly untenable assumption when considering the nature and organization of schools. As such, the model is likely to have been incorrectly specified from the outset. Second, deviation from the regression line is confounded by the RTM artifact in the simple posttest on pretest analysis and may lead to incorrect inferences regarding teacher or school effects.

Defining effectiveness in the absence of random assignment forces a researcher to consider the relationship of all exogenous variables on the outcome variable of interest. In addition, empirical estimates of effectiveness should be carefully considered before they are used to classify schools.

The Inclusion of Covariates

Given the confounding of the exogenous variables, and the quasi-experimental nature of school evaluations, VAA must consider the impact of covariates as statistical controls. However, the role of covariates wears at least two masks, an empirical role and that of appropriateness. Including covariates into a model may play a role computationally, but may have the unintended effect of setting lower growth expectations for certain populations than for others. Meyer (1994) aptly notes that statistical methods do not preclude teachers, parents, and school officials from setting learning standards.

Sanders (1997) believes that the blocking design sufficiently accounts for the non-school related variables and renders other covariate adjustments unnecessary. Because pretest scores are highly correlated with posttest scores, they will account for an appreciable portion of error variance in a blocking design. He further posits that collecting all relevant covariates is a "hopeless impossibility" and that the TVAAS methodology alleviates these concerns.

However, Raudenbush and Bryk (2002) suggest that covariates play an important role when estimating treatments effects in quasi-experiments. They advocate:

> In general, statistical adjustments for individual background are important for two reasons. First, because persons are not usually assigned at random to organizations, failure to control for background may bias the estimates of organization effects. Second, if these level-1 predictors (or covariates) are strongly related to the outcome of interest, controlling for them will increase the precision of any estimates of organizational effects and the power of the hypothesis tests by reducing unexplained level-1 error variance, σ^2 (p. 111).

Previously, shrinkage estimators were posed as one criterion for increasing the precision of treatment effects. It is worth expanding on this concept and reviewing the benefits of conditional shrinkage estimators (Raudenbush and Bryk, 2002) to further enhance the precision of estimates. Simply stated, conditional shrinkage estimates modify Kelley's Equation (Wainer et al, 2001) to include the covariates used in the analytical model in place of the grand mean. These adjust the regressed values towards a predicted value, given the covariate, rather than towards the grand mean. If covariates are used as statistical controls, one has the added benefit of conditional shrinkage estimators to increase the precision of the estimates.

The absence of random assignment to schools guarantees that factors outside the control of a school affect its student composition. As such, students may be afforded extended learning opportunities beyond the school setting that are not completely accounted for via initial starting position. Therefore, in the spirit of decomposing variance components so that true teacher variance can be properly estimated, it seems that covariate adjustments are warranted when their relationship to the outcome variable is substantially related.

What Systems Must Be In Place

In order for value-added analyses to proceed, a number of preliminaries must have been carefully attended to. Specifically, six components must be in place including an annual testing system, data in electronic format, unique student and teacher identification numbers (ID) that remain constant regardless of school attended, sufficient samples sizes, and software programs capable of mixed-model analyses.

The requirement for annual testing will soon be met, to a limited extent, given the testing requirements found in NCLB. If schools are using a measurement system, such as a computer adaptive test or tailored test, that is aligned with their curriculum and produces desirable test metrics, then this system may be more appropriately used in the analysis.

Although many school systems receive individual student data, few have access to these data in electronic format. Typically, paper reports are returned to schools within a few months of testing. It is possible to request the student scores in electronic format from most testing companies. However, managing these data in an information system suggests a financial investment that is often ignored when other fundamental needs are of organizational precedence. Schools already using CAT as one method of assessment will already have access to data stored in electronic format.

When student names are misspelled or changed, or schools have high degrees of mobility, year-to-year merges of longitudinal data become difficult. Two consequences become readily apparent. First, only analyzing cases for which there is an exact match may reduce the size of the data set, reducing the power to detect effects. Second, taking the time to identify cases over time using a variety of matching criteria increases time and energy, ultimately affecting the cost. One way of resolving this issue is by assigning and maintaining unique student identification numbers that remain consistent over all years, regardless of which school the student attended. The potential for intra-state mobility suggests that the assignment and maintenance of IDs is a function of the State Education Agency (SEA) and not one of an individual school or district.

The development of the cross-classification structure is possible only when a researcher can identify which classroom teacher students have had. For the same reasons described above, teacher names are insufficient. Therefore, I advocate for numeric teacher IDs to provide support for cross-classified data structures.

The power to detect treatment effects is contingent upon the nominal alpha specified by the researcher and the sample size. Improved estimates are generally gained when sample sizes are large. Of course it is also possible that large samples simply produce significant results. As such, researchers should assess whether sufficient power exists to detect an effect as well as the practical significance of the identified effect (Kirk, 2001; Thompson, 2002). Furthermore, any estimates of school or teacher effects should be accompanied by confidence intervals.

Last, the computational power required to estimate the enormous number of fixed and random parameters could consume most desktop computer for a number of hours, if not days. Furthermore, many traditional software packages are incapable of allowing the user to specify the mixed methodologies employed by many analysts. Software programs such as HLM, MLWin, R (nlme), SAS (Proc-Mixed), and S-Plus (nlme3) all have the capacity to deal with the mixed-model methodologies employed by most analysts.

Validity Concerns

Complex statistical models alone cannot adequately document quality teaching within a school. Certainly, the sophistication afforded via growth models permits for fewer mistakes when drawing inferences about school performance over traditional cross-sectional approaches. However, considerations to research validity including statistical conclusion, internal, construct, and external (Cook & Campbell, 1979) as well as the construct (Messick, 1989) aspects of validity should weigh heavy before causality is presumed.

Establishing a Causal Relationship

Reviewing all potential threats to research validity would fall outside the scope of this paper and could serve as a thesis all on its own. So, choosing the more salient issue seems appropriate, the establishment of causality. To establish causality, three considerations are generally required: temporal precedence, covariation of the cause and effect, and consideration of other plausible alternatives (Trochim, 2002).

Establishing temporal precedence requires the researcher to demonstrate that the treatment occurred before the effect. Given the cumulative nature of test scores (Meyer, 1994) the legitimacy of higher gains may not be attributable to the school or teacher unless explicitly controlled for. Second, many social interventions have a lagged effect. That is, one may not see an immediate increase in test scores, but rather an increase at some later point, even when instructional quality is high.

Establishing covariation between cause and effect is usually not a problem in social research. Certainly, two variables are often highly correlated, but have no meaningful relationship. Such is the case when you correlate vocabulary size with speed with which one runs the 100-meter dash. Of course, children are slower than adults, and they have smaller vocabularies than adults. Correlating speed with vocabulary establishes a positive mathematical relationship, but of course not a causal one. In a similar vein, relating higher gains to a particular school presumes that the instructional program is the reason that the test scores increased. On the other hand, if one were to infer a causal relationship between gains and school, then one must also be able to demonstrate that these students would not have experienced this gain if they attended another school. This is, of course, relates to the Type B effect discussed by Raudenbush and Willms (1995).

The last consideration relates to the causality presumed in the covariation. That is, the researcher must be able to demonstrate the school—and no other credible plausible alternatives—can be offered to substantiate the observed gains. For example, some students in suburban neighborhoods are often afforded tutoring opportunities or out of school experiences which have a direct effect on their learning. However, the increased gains that may be observed on a given test would say nothing about the school's effectiveness. Instead, it might suggest that the student's parents are actively involved in their child's education and provide learning opportunities beyond the school setting.

Appropriate formulation of a statistical model is only part of a researcher's responsibility. Providing substantial evidence that schools are in fact the reason for the change in scores is paramount in quasi-experiments.

The Consequential Aspect of Validity

Just as the more salient issue related to research validity is explored, only one consideration is given within construct validity, its consequential aspect. Consequential validity is primarily concerned with an evaluation of the intended effects (Doran, 2001; Lane & Stone, 2002). Specifically, it considers the relationship that the actual social consequences have with the desired intended effects of the program. In general, test-based accountability programs have historically been used as a policy lever, rather than a lever for classroom change.

Linn (2000) purported that assessments had become a prominent tool in the reform movement for four reasons. First, assessments were relatively inexpensive when compared to real and meaningful change efforts such as reducing class size, providing instructional aides, or implementing program changes. Second, assessments could be externally mandated. Linn stated, "It is far easier to mandate testing and assessment requirements at the state or district level than it is to take actions that involve actual changes in what happens inside the classroom" (p. 4). Third, assessment instruments could be rapidly implemented. It had been demonstrated that policymakers could develop and implement an assessment within a four-year political term. Fourth, test results were visible. The results of the test provide information that could be used to demonstrate a significant need for change.

The four preceding points elegantly describe the history of test-based educational accountability. It has been an inherently political ideal, with little focus on instructional diagnosis and increasing the professional dialogue among teachers and learners. However, the historic context of accountability should not be equated with the potential future of accountability.

Instead, accountability can, and should, be directly tied to a framework for learning; one that connects the web of internal and external actors in a coherent fashion. When appropriately applied, data from accountability models could serve the internal function by challenging schools to reflect upon their own teaching, consider whether students are being afforded full opportunities to learn, modify their policies and practices, and reallocate resources to fully support the areas identified as in need. Externally, accountability data should inform appropriate and justifiable social consequences that support the improved quality of services provided to children. Although social consequences are likely to occur, negative consequences should not be the result of construct underrepresentation or construct irrelevance (Messick, 1994). O'Day (2002) states:

> In particular, I argue that accountability systems will foster improvement to the extent that they generate and focus attention on information relevant to teaching and learning, motivate individuals to use that information and expend effort to improve practice, build the knowledge base necessary for interpreting and applying the new information to improve practice, and allocate resources for all of the above (p. 294).

When considering O'Day's context for improvement, one wonders whether VAA can provide information that is relevant to teaching and learning. Teachers have often viewed standardized test information as less credible than other classroom based measures. Moreover, returning test scores and analytical results to teachers after those students have left their classrooms does little to support the formative role of assessment. This, of course, does not imply that test results cannot be viewed as credible or have an impact of tailored instruction. In fact, I argue that they should. Nitko (1989) states:

> The position taken in this chapter is that appropriately used educational tests, whether created by classroom teachers or other agents, are potent educational tools that enhance the instruction process (p. 447).

Yeh (2001) also articulates a similar view that assessments including a variety of item types may focus on higher cognitive skills, and therefore serve as better methods for driving instruction. Rather than dismissing instructional tests as invalid and less credible, the challenge is to develop "better" tests.

Defining better through item type is a step in the right direction, but still insufficient. Many new applications such as computer adaptive tests enhance the science and application of assessment programs. Tailoring items to the ability level of students has significant potential to provide more reliable estimates of student performance, return data in a more timely fashion, and report scores on the type of scale necessary for conducting value-added analysis.

Better may not even be a psychometrically defined, although I do not suggest that psychometric rigor be ignored. However, psychometric properties alone are unlikely to encourage increased and appropriate classroom or public action.

Better is likely not to be accomplished by developing state tests under the pretense that they are criterion-referenced tests, when in fact they are only criterion-*like* tests. For example, many of the tests that have been implemented in states are purportedly aligned with the state content standards. Yet many of the items making up the test were purchased from an item bank, also likely to be found on a typical norm-referenced test. These state tests also report criterion-referenced scores. Specifically, they make use of scale scores and performance levels. However, these tests are not really criterion-referenced, they are criterion-like.

My remarks are not a yearning for new psychometrics. On the contrary, they are a yearning for innovative methods of assessment and accountability, which in turn support appropriate classroom and public action. Hargreaves (2003) articulates his view for an innovative education system, much of which is exemplified when systems focus on assessment for learning (AFL). Hargreaves characterizes AFL as:

> It is the process by which the teacher provides feedback in such a way that *either* the teacher adjusts the teaching in order to help the students learn more effectively *or* the learner changes his/her approach to the learning task *or* both of these (p. 10).

The AFL ideal runs parallel to the notion of formative assessment. If formative assessment is still viewed as a method to support radical innovation in education, then one might question why this has failed to occur in the past. Quite possibly, the means by which feedback mechanisms have remained static via insufficient measurement methodologies focused on cross-sections of students is partly to blame.

Assuming we will be able to provide information that is relevant to teaching and learning, schools and teachers must be motivated to use VAA information to modify instructional practices, change existing policies, and reconsider the allocation of resources. Of course, providing better information does not imply that results will be appropriately interpreted leading to justifiable actions. For this reason, teachers must be supported to interpret and apply the results from complex analyses such that appropriate classroom modifications can begin. For this reason, I believe that information should be reported in ways that relate the VAA results to the state content standards and provide unique information about individual students and their academic progress.

For sake of argument, assume relevant information can be gleaned from assessments and that teachers would be motivated to use this information for instructional modifications. One method of exploring the consequential aspect of validity is to explore the extent to which VAA results influence schools and teachers to make to more valid instructional/programmatic modifications than do the results of conventional analyses using less sophisticated methods. Certainly, I recognize the danger in selecting the term "appropriate". However, I submit that although an assortment of "appropriate actions" could be rationalized, this is best left to the judgment of those in their local context, and not to an author distanced from the goals of a school and its community.

An *a priori* expectation would be that more appropriate classroom and public actions are taken given the VAA results than would have occurred from a conventional analysis. Empirical methods such as predicted pattern testing (Levin & Neumann, 1999) could be used to study the extent to which schools and teachers modify instructional practice (Doran, 2001). If VAA is more likely to produce results that lead to more appropriate action, then validity evidence must be catalogued to support these methodologies over more conventional methods.

Conclusion

This paper has explored the breadth of issues as they relate to value-added analysis. Although the issues discussed are not exhaustive, they provide a methodological outline to define and consider value-added analysis. In addition, I argue that value-added analysis should be one component of accountability. If the models are appropriately specified, then they will be one source of information to support appropriate classroom and public action. Otherwise, accountability models will continue to provide stale and irrelevant information to teachers and policymakers.

However, appropriate specification is not easily accomplished. The complexity required for an analysis to be considered value-added extends beyond the reach of simple gain score models and OLS regression techniques. Ignoring the intra-class correlation and non-random assignment of students to schools will not produce results that can be appropriately used to support high-stakes educational decisions.

This is not to say that value-added models are too complex and alternatives should be sought in lieu of VAA. On the contrary, value-added analysis is one potent tool for the accountability movement. As has been previously argued by Drury and Doran (2003), "Trading rigor and accuracy for simplicity is an indefensible strategy—the stakes are simply too high" (p. 2).

Future research should consider the consequential aspect of validity. The ultimate merit by which value-added analyses should be judged rests upon the ability of the models to provide relevant information that can be used to support justifiable and appropriate classroom and public action. If results from value-added can fill this void, which I believe they will, then their role as one component of accountability plans is justified.

References

Ballou, D. (2002). Sizing up test scores. Education Next, 2, 10-15.

Burket, G.R. (1984). Response to Hoover. Educational Measurement: Issues and Practice, 3,15-18.

Raudenbush, S.W. & Bryk, A.S. (2002). Hierarchical linear models, applications and data analysis methods. Thousand Oaks, CA: Sage.

Ceperley, P.E. & Reel, K. (1997). The impetus for the Tennessee value-added accountability system. In J. Millman (Ed.), Grading teachers, grading schools: Is student achievement a valid evaluation measure? Thousand Oaks, CA: Corwin Press.

Collins, L.M. (1996). Is reliability obsolete? A commentary on "Are simple gains scores obsolete?" Applied Psychological Measurement, 20(3), 289-292.

Cook, T.D. & Campbell, D.T. (1979). Quasi-experimentation- design & analysis issues for field settings. Boston: Houghton Mifflin.

Cronbach, L.J. & Furby, L. (1970) How should we measure "change"—or should we? Psychological Bulletin, 74, 68-80.

Darlington, R.B. (1997). The Tennessee value-added assessment system: A challenge to familiar assessment methods. In J. Millman (Ed.), Grading teachers, grading schools: Is student achievement a valid evaluation measure? Thousand Oaks, CA: Corwin Press.

Davis, A. (1991). Upping the stakes: Using gain scores to judge local program effectiveness in chapter 1. Educational Evaluation and Policy Analysis, 13, 380-388.

Divgi, D.R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. Journal of Educational Measurement, 23, 283-298.

Doran, H.C. (2001). Evaluating the consequential aspect of validity on the Arizona instrument to measure standards. Unpublished doctoral dissertation, University of Arizona.

Drury, D. & Doran, H. (2003, January). The value of value-added analysis. National School Boards Association Policy Research Brief, (3), 1

Ebel, R.L. (1962). Content standard test scores. Educational and Psychological Measurement, 22, 15-25.

Goldstein, H. (1995). Multilevel statistical models. London: Arnold.

Harcourt Brace. (1997) Stanford 9 technical data report. San Antonio, TX.

Hargreaves, D.H. (2003, January). From improvement to transformation. Paper presented at the meeting of the International Congress for School Effectiveness and Improvement 'Schooling to Knowledge Society,' Sydney, Australia.

Holland. P.W. (2002). Two measures of change in the gaps between the CDFs of test-score distributions. Journal of Educational and Behavioral Statistics, 27, 3-17.

Hoover, H.D. (1984). The most appropriate scores for measuring educational development in elementary schools: GE's. Educational Measurement: Issues and Practice, 3, 8-14.

Hoover, H.D. (2003). Some common misconceptions about tests and testing. Educational Measurement: Issues and Practice, 22, 5-14.

Jaccard, J. & Becker, M.A. (1997). Statistics for the behavioral sciences, third edition. Pacific Grove, CA: Brooks-Cole.

Kirk, R.E. (1995). Experimental design: Procedures for the behavioral sciences. Pacific Grove, CA: Brooks-Cole.

Kirk, R.E. (2001). Promoting good statistical practices: Some suggestions. Educational and Psychological Measurement, 61(2), 213-218.

Kreft, I. & De Leeuw, J. (1998). Introducing multilevel modeling. London: Sage.

Lane, S. & Stone, C.A. (2002). Strategies for examining the consequences of assessment and accountability programs. Educational Measurement: Issues and Practice, 21, 23-30.

Levin, J. R., & Neumann, E. (1999). Testing for predicted patterns: When interest in the whole is greater than in some of its parts. Psychological Methods, 4(1), 44-57.

Linn, R.L. & Slinde, J.A. (1977). The determination of significance of change between pre- and posttesting periods. Review of Educational Research, 47, 121-150.

Linn, R.L. & Slinde, J.A. (1979). A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. Journal of Educational Measurement,16, 159-165.

Linn, R. L. (2000). Assessment and accountability. Educational Researcher, 29(2), 4-16.

Lord, F.M. (1963). Elementary models for measuring change. In C.W. Harris (Ed.) Problems in measuring change. Madison, WI: University of Wisconsin Press.

Maas, C.J. & Snijders, T.A. (2003). The multilevel approach to repeated measures for complete and incomplete data. Quality & Quantity, 37, 71-89.

Messick, S. (1989). Validity. In R.L. Linn (Ed.) Educational measurement (3^{rd} ed. pp. 13-103). Englewood Cliffs, NJ: Prentice Hall.
Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher, 23(2), 13-23.

Meyer, R.H. (1997). Value-added indicators of school performance: A primer. Economics of Education Review, 16, 283-301.

Nitko, A.J. (1989). Designing tests that are integrated with instruction. In R.L. Linn (Ed.) Educational Measurement (3^{rd} ed. pp. 447-475). Englewood Cliffs, NJ: Prentice Hall.

Nitko, A.J. (1996). Educational Assessment of Students. Englewood Cliffs, NJ: Merrill.

Nunnally, J.C. & Bernstein, I.H. (1994). Psychometric Theory. NY:McGraw-Hill.

O'Day, J.A. (2002). Complexity, accountability and school improvement. Harvard Educational Review, 72, 293-321.

Osgood, D.W.(2001). Advances in the application of multilevel models to the analysis of change. In L.M. Collins & A.G. Sayer (Eds.) New methods for analysis of change. Washington, DC: American Psychological Association.

Peterson, N.S., Kolen, M.J. & Hoover, H.D. (1989). Scaling, norming and equating. In R.L. Linn (Ed.) Educational measurement (3^{rd} ed. pp. 221-262). Englewood Cliffs, NJ: Prentice Hall.

Raudenbush, S.W. (2001) Toward a coherent framework for comparing trajectories of individual change. In L.M. Collins & A.G. Sayer (Eds.) New methods for analysis of change. Washington, DC: American Psychological Association.

Raudenbush, S.W. & Willms, J.D. (1995). The estimation of school effects. Journal of Educational and Behavioral Statistics, 20, 307-335.

Rogosa, D., Brandt, D. & Zimowski, M. (1982). A growth curve approach to the measurement of change. Psychological Bulletin, 92, 726-748.

Rowan, B. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the prospects study of elementary schools. [On-line]. Available: http://www.sii.soe.umich.edu/documents/LargeScaleSurveyResearch.pdf

Sanders, W.L., Saxton, A.M., & Horn, S.P. (1997). The Tennessee value-added assessment system: A quantitative, outcomes-based approach to educational assessment. In J. Millman (Ed.), Grading teachers, grading schools: Is student achievement a valid evaluation measure? Thousand Oaks, CA: Corwin Press.

Scariano, S., & Davenport, J. (1987). The effects of violations of the independence assumption in the one way ANOVA. The American Statistician, 41, 123-129.

Schwarz, R.D.,Yen, W.M., & Schafer, W.D. (2001). The challenge and attainability of goals for adequate yearly progress. Educational Measurement: Issues and Practice, 20, 26-33.

Seltzer, M.H., Frank, K.A. & Bryk, A.S. (1994). The metric matters: The sensitivity of conclusions about growth in student achievement to choice of metric. Educational Evaluation and Policy Analysis, 16, 41-49.

Stevens, J.P. (2002). Applied multivariate statistics for the social sciences. Mahwah, NJ: Lawrence Erlbaum Associates.

Taris, T.W. (2000). A primer in longitudinal data analysis. London: Sage.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. Educational Researcher. 25-32.

Thum, Y. M. (forthcoming). Measuring Progress towards a Goal: Estimating Teacher Productivity using a Multivariate Multilevel Model for Value-Added Analysis. Sociological Methods & Research.

Trochim, W.M. (2002). Establishing a cause-effect relationship [On-line]. Available: http://trochim.cornell.edu/kb/causeeff.htm

Venter, A., Maxwell, S.E. & Bolig, E. (2002). Power in randomized group comparisons: The value of adding a single intermediate time point to a traditional pretest-posttest design. Psychological Methods, 7, 194-209.

Wainer, D. & Mislevy, R.J. (2000). Item response theory, item calibration, and proficient estimation. In H. Wainer, et als. (Eds.) Computerized adaptive testing: A primer. (2nd Ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Wainer, H., Vevea, J.L., Camacho, F., Reeve, B.B., Rosa, K., Nelson, L., Swygert, K.A., & Thissen, D. (2001) Augmented scores—"borrowing strength" to compute scores based on small number of items. In D. Thissen & H. Wainer (Eds.), Test scoring. Mahwah, NJ: Lawrence Erlbaum Associates.

Webster, W.J. & Mendro, R.L. (1997). The Dallas value-added accountability system. In J. Millman (Ed.), Grading teachers, grading schools: Is student achievement a valid evaluation measure? Thousand Oaks, CA: Corwin Press.

Yeh. S.S. (2001). Tests worth teaching to: Constructing state-mandated tests that emphasize critical thinking. Educational Researcher, 30, 12-17.

Zwick, R., Senturk, D., Wang, J. & Loomis, S.C. (2001). An investigation of alternative methods for item mapping in the national assessment of educational progress. Educational Measurement: Issues and Practice, 20, 15-25.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

TM034888

I. DOCUMENT IDENTIFICATION:

Title: Value-Added Analysis: A Review of Related Issues	
Author(s): HAROLD C. DORAN	
Corporate Source:	Publication Date: APRIL 15, 2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ☑	Level 2A □	Level 2B □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please →

Signature:	Printed Name/Position/Title: HAROLD C. DORAN	
Organization/Address: 675 N. Washington, Suite 220 Alexandria, VA 22314	Telephone: 703-765-0301	FAX:
	E-Mail Address: hdoran@nasdc.org	Date: 4/03/03

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse: **University of Maryland**
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Lab, Bldg 075
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Lab, Bldg 075
College Park, MD 20742
Attn: Acquisitions