A Potential Outcomes View of Value-Added Assessment in Education

Donald B. Rubin, Elizabeth A. Stuart, and Elaine L. Zanutto

November 12, 2003

1 Introduction

1.1 Assessment and Accountability in Education

There has been substantial interest in recent years in the performance and accountability of teachers and schools, partially due to the No Child Left Behind legislation, which requires states to develop a system of sanctions and rewards to hold districts and schools accountable for academic achievement. This focus has lead to an increase in "high-stakes" testing with publicized school rankings and test results. The papers by Ballou et al. (2004), McCaffrey et al. (2004) and Tekwe et al. (2004) approach the estimation of school and teacher effects through a variety of statistical models, known as "value-added" models in the education literature. There are many complex issues involved, and we applaud the authors for addressing this challenging topic.

In this discussion we approach value-added assessment from a "potential outcomes" (Rubin Causal Model, RCM) point of view (Rubin 1974, 1978, 2003; Holland 1986; Little and Rubin 2001), with the goal of clarifying the estimation goals and understanding the limitations of data for the types of comparisons being sought. We discuss the challenges in conceptualizing and obtaining reliable estimates of the causal effects of teachers or schools. We also present an idea for future research that focuses on assessing the

effect of implementing reward structures based on value-added models, rather than on assessing the effect of teachers and schools themselves, which we feel is a more relevant policy question, and also one that is more easily addressed.

1.2 Value-added Models–Causal or Descriptive?

The value-added models used to estimate the effectiveness of teachers and schools in Ballou et al. (2004), McCaffrey et al. (2004), and Tekwe et al. (2004) range from a relatively straightforward fixed effects model (Tekwe et al. 2004) to a relatively complex and general multivariate, longitudinal mixed-model (McCaffrey et al. 2004) with either test scores or test score gains as outcomes. These models incorporate parameters for school and teacher effects (including lagged teacher effects), parameters for student-level, classroom level, and school-level covariate effects, and parameters to allow for residual intra-class correlation among outcomes for students in the same class. These models attempt to address problems such as apportioning school effects to more than one school (for students who attended more than one school in the year prior to the test) and the persistence of teacher effects into the future. But none of these articles attempts to define precisely the quantity that is the target of estimation, except in the rather oblique sense of seeing what the models estimate as samples get larger and larger. Thus, there is a focus on the estimation techniques rather than the definition of the estimand.

The goal of the value-added literature seems to be to estimate the "causal effects" of teachers or schools; that is, to determine how much a particular teacher (or school) has "added value" to their students' test scores. It is implied that the effects being estimated are causal effects: the effect on students of being in school A (or with teacher T) on their test scores, where schools and teachers are to be rewarded or punished "because" of their estimated effects on students. But is it possible to get reliable causal estimates

in this setting? The potential outcomes perspective (RCM) provides a framework to clarify this issue concerning whether we are seeking causal or descriptive answers. Before delving into this perspective, it may be helpful to connect this "value-added" problem to one in "hospital profiling", where a relatively substantial literature already exists on a similar problem.

1.3 A Related Problem–Hospital Profiling

The problem of comparing the performance of schools, accounting for the backgrounds of the students they serve, is similar to that addressed in the literature on hospital profiling. In hospital profiling, the aim is to assess the performance of particular hospitals in treating diseases, after accounting for the varying patient populations served by each hospital, so called "case-mix adjustment" (e.g., Goldstein and Spiegelhalter 1996; Christiansen and Morris 1997; Burgess et al. 2001). Additionally, hospital profiling is often done only within subgroups of hospitals, such as by type (teaching, general, psychiatric). Making comparisons within types may also be useful in the school setting, for example comparing public inner-city schools only with other public inner-city schools. This issue will be discussed further in Section 2, which stresses the importance of comparing comparable units.

The school setting appears to be even more complicated than the hospital profiling one, for a variety of reasons. First, there is interest in longitudinal effects with a desire to separate out the effects of last year's and this year's teachers. Thus, we seem to need to have "vertically-linked" test scores that can be compared over time. Second, longitudinal data are not strictly hierarchically nested since students do not remain together as a class over time; not only are students' teachers changing each year, but their classmates are also changing. Finally, there is substantial missing test score data in the school setting and obviously relevant unobserved covariates, such as the motivational levels of the students. These are all issues that

appear to be more complex in the school setting than in the hospital profiling setting; nevertheless, workers in the school assessment setting could possibly find relevant ideas in the hospital profiling literature.

The hospital profiling literature also points out other possible problems when such methods for assessing "successful" teachers or schools are implemented. For example, there is the possibility that once any system of assessment is implemented, schools will "game the system" to obtain results that unduly benefit them. In the hospital profiling setting, Green and Wintfeld (1995) describe an increase in reported incidence of risk factors that would increase expected mortality, such as congestive heart failure, after implementation of a system to generate case-mix adjusted physician-specific mortality rates. Presumably, doctors hoped to improve their performance ratings by inflating the entry-level risk statuses of their patients. In the school setting, schools may place more students in special-education or English-as-a-second-language courses so that their student body appears to be more disadvantaged or so that some groups of students are excluded from the overall analysis of test scores.

2 Defining and Estimating Causal Effects in Value-Added Assessment

2.1 Causal Inference and the RCM

Causal effects are inherently comparisons of potential outcomes, measured at the same point in time (e.g., test scores at the end of fifth grade, Y) on a common set of units (e.g. a specific classroom of students); Rubin (1974, 1978), Holland (1986), Little and Rubin (2001). To estimate the effect of being in school A versus school B for a particular student, say Q, Q's test score at the end of fifth grade if Q had been in school A, $Y_Q(A)$, is compared with Q's test score at the end of fifth grade if Q had been in school B, $Y_Q(B)$. The "fundamental problem of causal inference" (Holland 1986) is that only one of these potential

outcomes, $Y_Q(A)$ or $Y_Q(B)$, can be observed for student Q: Q is in either school A or school B. Thus, causal inference can be thought of as a missing data problem, with at least half of the potential outcomes missing. Inference proceeds by estimating the unobserved potential outcomes, either implicitly or explicitly.

The first task in causal inference is to identify the "units", "treatments", and "potential outcomes". A challenge in the value-added assessment setting is that it is difficult to define even these fundamental concepts. The units are the objects to which treatments are applied. Should we think of the schools as the units? Or are the units the individual students? Or the classrooms? The potential outcomes are likely to be test scores at the end of the year at the unit level (for example, an individual student's test score if individual students are the units, or an average [or median] test score for an entire school, if the school is the unit). Districts or states may also be interested in other measures of improvement.

The treatments are the interventions of interest, for example school A versus school B. However, even that can be difficult to define; are we interested in only the "administrative" effect of being in school A for each student, which would be an effect due to institutional changes such as a different set of teachers, different curriculum, better facilities, etc., or is it a more general effect on an individual student of being placed in a fully different environment, with both a different institutional set-up as well as different classmates? Assigning student Q into school A rather than school B is very different from assigning student Q and all of student Q's current classmates into school A rather than school B.

Suppose the treatment action is defined to be school A; are the potential outcomes under that treatment to be compared to the potential outcomes in school B, as just assumed, or to the average potential outcome at a collection of "average" schools, as seems to be done in the papers by Ballou et al. (2004), Tekwe et al. (2004) and McCaffrey et al. (2004)? Or should the potential outcomes in school A be compared to the potential outcomes when being in no school? Students and parents choosing between schools will

presumably want to know what their test scores would be in a different "possible" school. School boards comparing teachers may want to compare teachers' individual performance to some overall average teacher's performance, or to a set of other teachers the students could have had.

These various questions seem like they would have different answers, and few of them seem to be like the questions addressed by the current articles under discussion. Precisely, what are the causal effects being estimated by the methods in the papers? Or are they instead simply estimating descriptive quantities? The meaning of this last question is well illustrated by "Lord's Paradox."

2.2 A Classic Example of Poorly Formulated Causal Assessment–Lord's Paradox

Lord's Paradox is a classic example to illustrate the importance of defining appropriate comparisons and stating clearly any assumptions underlying estimates implied to be "causal". This example originally arose in a similar educational setting, in discussion of gain scores vs. covariance (regression) adjustment (Lord 1967). Lord described the following "paradox:" A university is interested in estimating the effect of the university diet on student's weight, and is particularly interested in any differential effect on males and females. Simple descriptions are given of the data at the beginning and end of the year. For both males and females, the distribution of weights is the same at the beginning and end of the year (the average female weight is the same, the female variance is the same, the average male weight is the same, the male variance is the same, the correlation between September and June weight is 0.8 for both males and females, etc.). Lord then posits two statisticians. Statistician 1 uses gain scores (comparing the change in weight from September to June between males and females) and claims that since on average neither males nor females gained or lost weight during the year, then there is no differential effect of the diet on males or females. Statistician 2 computes a covariance adjusted difference of the two group means and sees that, for males

and females of the same initial weight, the males weigh more at the end of the year. He thus concludes that there is a differential effect of the diet for males and females, with males gaining more weight on average. For a graphical representation of the analyses of statisticians 1 and 2, see Bock (1975). Lord's primary question concerned which of these statisticians was correct.

2.3 Lord's Paradox Resolved

Holland and Rubin (1983) explain the apparent paradox by noting that either statistician can be correct, depending on the assumptions made. In the hypothetical scenario, all students receive the new diet; no students receive the undefined "control" diet, whatever it is (no diet? the "old" university diet? the diet the students ate before attending university?). Thus, the only potential outcome that is even observed is that under the treatment (university diet). The potential outcomes under the control diet are completely missing.

If it is assumed that under the control diet each student's weight in June would be the same as their weight in September, then statistician 1 is correct. Statistician 2 is correct under the assumption that weight gain under the control diet is a linear function of the student's weight in September with a common slope but varying intercept for males and females. This simple example is very instructive regarding the importance of thinking carefully about what is being estimated and what is the quantity of interest. It is easy to focus on estimation methods without thinking about the underlying problem–what the technical methods are trying to estimate. Many statisticians and educational researchers were perplexed by Lord's paradox–valid causal inference does not come easily or naturally, except in randomized experiments, and even there only with no complications such as those discussed in Sections 2.6 and 2.7.

2.4 Post-test Scores versus Gain Scores

Despite the debate about whether post-test scores or gain scores should be used as outcomes, the RCM perspective makes it clear that the same causal effect is the estimand whether using either post-test scores or gain scores because test score before treatment assignment is a covariate, unaffected by treatment assignment. Consider $Y_Q(B)$ to be the student Q's test score at the end of the year when in school B and $Y_Q(A)$ to be the student's test score at the end of the year when in school A, and X_Q to be student Q's baseline test score. The causal effect of being in school A versus being in school B for student Q is $Y_Q(A) - Y_Q(B)$. Using gain scores instead we have $(Y_Q(A) - X_Q) - (Y_Q(B) - X_Q) = Y_Q(A) - Y_Q(B)$. Although gain scores will often be used to be more precise, post-test scores can be used to estimate the same causal effect as gain scores. However, it is important to remember that, as in Lord's paradox, the gain score itself is not a causal effect, except under the strong assumption that the potential outcome under control equals the baseline observation ($Y_Q(B) = X_Q$, where school B is considered the "control" treatment). It is unlikely that this will be true in this setting. Even without formal instruction, students still learn (and forget) over time and it is likely that their test scores would change as a result.

2.5 Estimating Causal Effects of Schools in an "Ideal" Setting with Randomization

Randomized experiments are the "gold standard" for estimating causal effects. To think clearly about what is being estimated, it is thus useful to think about what we would do in an ideal world where random assignment of students to schools is feasible.

At first glance, to estimate the causal effect of being in school A versus school B (where individual students are the units), students could be randomly assigned to the two schools, thus ensuring that the schools would have similar mixes of students in their classes. Under randomization, the difference in observed outcomes

between the students in school A and the students in school B is an unbiased estimate of the true effect of being in school A versus being in school B. Randomization could also be extended to facilitate comparisons across more than two schools.

Examining the scenario of a randomized experiment helps to conceptualize the problem and think about how to estimate causal effects, however, even in this idealized setting there are a number of complications.

2.6 Complications: Interference Between Units and Versions of Treatments

An assumption commonly invoked (either implicitly or explicitly) in causal inference is the Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1980, 1986). Without this assumption, inference becomes substantially more complex. There are two components to SUTVA. The first is that there is only one version of a specific treatment: all individuals assigned to "treatment" receive the same active treatment, and all individuals assigned to "control" receive the same control treatment. The second component of SUTVA is that there is no interference between units: the value of each unit's potential outcome does not depend on the treatment assignments of other units. In other words, student Q's test score in school A is not affected by whether her best friend, student R, is in school A or school B. Anyone familiar with education will realize that this is probably a fairly unrealistic assumption; students in schools talk to and interact with one another, both inside and outside the classroom. In other words, the treatments the other students receive, and not just the school itself, are likely to affect each student's test scores. This interaction is also an issue in the methods used by Ballou et al. (2004), McCaffrey et al. (2004) and Tekwe et al. (2004), in which the students are often implicitly considered to be non-interfering.

It is sometimes possible to alleviate deviations from SUTVA through design; for example, by considering schools to be the unit of analysis with randomization done at the school level rather than at the individual

student level. However, considering schools to be the unit of analysis creates its own complications, and may or may not be addressing the question of interest. What are the treatments to be randomized to the schools? Are collections of teachers to be randomly assigned in groups to schools? If a randomized experiment cannot even be conceptualized, it is difficult to conceptualize the causal question being addressed.

2.7 Another Complication: Missing Data

Another issue is missing data. Most of the methods in the papers use complete cases only, which is only appropriate if the missing data are missing completely at random (MCAR, Little and Rubin 2002) and is not appropriate in settings such as this with longitudinal data and missing outcomes (see, e.g. Barnard et al. 1998, 2003), even when the missing data are ignorable (Rubin 1976c; Little and Rubin 2002). Furthermore, because the missingness likely depends on the missing values themselves (for example, students who sense that they will likely not do well on the test may be more likely to miss school the day of the test), it is possibly non-ignorable, and thus methods for non-ignorable data may be relevant.

Evidence of non-MCAR can be seen in data presented in some of the value-added papers. For example, Ballou et al. (2004) state that "In all subjects and years, mean scores are higher among claimed students"– relative to unclaimed students, who are not attributed to any particular teacher, and thus are not used in the estimation of teacher effects (in essence, they are considered to be missing values in the model estimation). The unclaimed students may move classes throughout the year, and are likely to be students whose performance is worse than average. Table 1 of Tekwe et al. (2004) also suggests non-MCAR missingness: the difference between overall average scores for 1999 and overall average scores for 1998 is always larger than the average change score (which is calculated using only those students with both sets of scores), possibly due to dropping low-scoring students who were not promoted to the next higher grade

out of the change score analysis (Tekwe et al. 2004, p. 8). Non-MCAR and non-ignorable missing data can be an especially large problem when using longitudinal data, as students who move may be more likely to be students who perform at lower levels, and their scores will be more likely missing for at least one year.

2.8 Observational data

Thus we see that many complications exist when thinking about an ideal randomized experiment, and even more complications arise when thinking about using observational data, which, of course, is the more realistic scenario. With observational data, one key goal is to find treated and control units that look as similar as possible on background covariates. If the groups look very different on background covariates, the results are likely to be based on untestable modeling assumptions and extrapolation.

Implicit extrapolation in models of outcome data (e.g., test scores) is common, and is particularly hard to diagnose with complex models such as those in Ballou et al. (2004), McCaffrey et al. (2004) and Tekwe et al. (2004), because common model diagnostics do not assess the overlap in covariate distributions. Because the values of "percent minority" and "percent in poverty" differ widely in different schools, as illustrated in Table 2 in Tekwe et al. (2004), it is likely that the estimates adjusting for such covariates using models rely heavily on extrapolation, even if students were randomly assigned to those schools after being subclassified into blocks (with dramatically different probabilities of treatment assignment between blocks but similar probabilities within blocks). This situation implies extreme sensitivity to these models' assumptions. If school A has no students who "look like" students in the other schools, it is impossible to estimate the effect of school A relative to the comparison schools without making heroic assumptions.

2.9 Replicating a Randomized Experiment

With observational data, the goal is to replicate a randomized experiment as closely as possible. Matching methods, such as ones that use a multivariate distance measure (Rubin 1976a, b) and propensity scores (Rosenbaum and Rubin 1983a, 1984, 1985; Rubin and Thomas 1992a,b, 1996), enable observational data to replicate two key features of randomized experiments. First, the comparison is done on groups of units who are similar with respect to the observed covariates. Second, the study is "designed" in that the treated and control units are matched without using the observed outcome variable, thus preventing bias due to manipulating the samples to get a desired result. Before any analysis of the outcome data, the matched samples can be assessed for covariate overlap ("balance") to make sure that, within each matched group, treatment assignment looks as if it could have arisen from a randomized experiment where treatment assignment probability is a function of the observed covariates.

2.10 Model-based Analysis versus Propensity-based design in an Observational Study

Reliable estimates can only be made where covariate distributions overlap (Rubin 1977), as illustrated by Lalonde (1986) and follow-up work by Dehejia and Wahba (1999). More specifically, using large data bases as a control group to try to replicate estimates from a randomized experiment on the effect of a job-training program, Lalonde found that the estimates were highly sensitive to the choice of model (for example, linear versus quadratic, choice of covariates, ignorable model versus non-ignorable model), with answers ranging wildly, yet each with very narrow and non-overlapping associated "confidence" intervals. Dehejia and Wahba instead used propensity score methods to choose a set of well-matched comparison individuals and were able to replicate closely the experimental results. Of course, propensity score methods will not always work this well. Having an adequate set of observed covariates is critical.

The ability of propensity score analyses to reveal the extent to which two groups serve similar types of students and have similar educational environments is an important diagnostic tool to identify whether the data can support causal comparisons between these two groups. Comparing treated and control groups with very different distributions of background covariates will lead to extreme extrapolation in models relating outcome variables to covariates, thus making any estimates highly sensitive to untestable modeling assumptions.

Propensity score matching also avoids any specification of regression models for the relationship between the outcome and the covariates. Although propensity score models must be fit to estimate the probability of receiving treatment, estimates of treatment effects are generally less sensitive to misspecification of the propensity score model than regression models are to misspecification of the regression model (Drake 1993; Rubin 1997).

The fact that propensity score methods match on background covariates, without any use of the outcome at the matching stage, has several desirable properties, as indicated earlier. Comparable units can be found before the outcomes are even observed. This is especially helpful in high stakes situations such as school assessment, to prevent researchers from intentionally or unintentionally manipulating matched samples to generate desired results and also protects them from such claims by others (for more on this perspective see Rubin, 2001).

Furthermore, because the outcome variable is not used in the matching process, the same matched samples can be used to study multiple outcomes, as with randomized samples; when model-based analyses are used, separate regression models are needed for each outcome variable. Once the matched samples are chosen, inference can proceed using modeling methods, however the results will be relatively less sensitive to the model assumptions because there will be less extrapolation (Rubin 1973, 1979; Rubin and Thomas

2000). If it is impossible to obtain overlapping covariate distributions using matching or subclassification, the conclusion should be that reliable causal inferences cannot be drawn from the existing data, without relying on explicitly stated, and typically heroic, assumptions.

2.11 The Critical Advantage of Randomized Experiments

One key remaining difference between observational studies and randomized experiments is that randomization assures balance on all covariates, observed and unobserved, between the treated and control groups. In contrast, with observational studies, we can only balance the observed covariates. It is thus very important to try to measure all of the relevant covariates such that treatment assignment will not depend on the potential outcomes, given the covariates (termed "strongly ignorable" (or "unconfounded") treatment assignment) (Rosenbaum and Rubin 1983a).

3 Another approach

We have seen that even in an ideal setting with randomization, estimating relevant causal effects of teachers and schools is extremely difficult to conceptualize. If causal inference here is so difficult, how are we to guide policy and think about the benefits of value-added assessment? Whatever policies are being compared, probably none will ever involve moving large numbers of individual students from one school to another school. Rather, in terms of policy questions, we should be more interested in comparing interventions that are realistic, such as implementing various reward structures based on value-added assessment models, and seeing which structures are most effective at improving performance.

We advocate a position of taking the current value-added models at face-value and considering their pa-

rameter estimates as descriptive measures that we hope are of some relevance to the question of educational assessment. The real question then is, do these descriptive measures, and proposed reward systems based on them, improve education?

3.1 Estimating the Causal Effect of Value-added Assessment

To think about how to answer this question, we propose the design of a study to assess the effects of the reward systems implied by the different assessment models on educational improvement, thereby shifting the focus away from estimating the effect of teachers or schools to estimating the effect of implementing a reward system based on one of these models.

We again must first consider the fundamental concepts: units, treatments, and potential outcomes. The units could be states, school districts, or perhaps even individual schools. The active treatment is the implementation of a reward structure based on the results from a value-added assessment model. The control treatment would be no implementation of such a reward structure–life "as it was". The potential outcomes may be quantities such as average test scores at the end of the school year, or equivalently gain scores, or perhaps average test scores in sub-groups defined by covariates such as gender or baseline test score, or measures of parent satisfaction, or graduation rate. Each unit has a potential outcome under each treatment (reward structure in place, no reward structure in place), and again, the main task is to estimate the unobserved potential outcomes because each unit can be assigned to only one treatment.

Ideally, units (states, school districts, or schools) would be randomized to either receive this new reward structure or not receive it. For example, half of the states would be assigned to the new treatment, whereas the other half would not receive the new treatment. Matched pairs or blocking could also be used in the design. After a specified period of time (perhaps one or two years), the outcomes in the treated and control

groups would be compared. Due to the randomization, any observed differences in these observed outcomes could not be due to differences in baseline covariates, and the average difference in outcomes would be a valid estimate of the average causal effect.

3.2 A possible design using observational data

Unfortunately, randomization is probably infeasible, and thus observational data will need to be used to estimate the causal effect of implementing a reward structure associated with a value-added assessment. Using the ideas of replicating a randomized experiment, we can think about how to approach the design of such an observational study. For example, we might consider the treated units to be states that have implemented some form of value-added assessment (VAA) into their state accountability systems, such as Tennessee and North Carolina.[7] We thus are thinking of a hypothetical randomized experiment, where some states were randomized to treatment, and all of the others were randomized to control.

To replicate this hypothetical randomized experiment with observational data, we would like to find a well matched control state for each of the treated states. This matching could be done using propensity score matching, estimating the propensity scores, for example, using a logistic regression model where the response is whether a state has a reward system based on VAA, and the predictors are state-level covariates such as poverty level, high school dropout rate, population, etc. Alternatively, matching could be done using a multivariate matching method, or exact matching on a few key variables, or some combination of these matching methods. A benefit of matching in this situation is that the number of states that have implemented VAA is relatively small, thereby ensuring a large pool of potential control states. Matching is generally more successful in settings such as this, where there is a large set of potential controls.

However, in some cases it may be difficult to find a comparison state for each treated state; there may be

no other states that "look like" Tennessee on the background covariates. In this case, it may be useful to create "pseudo-states," as in Irie (2001) or Abadie and Gardeazabal (2003). Assuming school district level data was available, the propensity score would be estimated on a school district level, and individual school districts within Tennessee could be matched to individual school districts from multiple other states (perhaps restricted to the South). The control "pseudo-state" that is matched to Tennessee would consist of this set of matched districts from multiple control states, for example a mix of districts from Alabama and Arkansas. Once the pairs of matched treated and control states (or pseudo-states) are chosen (in essence, the observational study "designed"), analysis would proceed by analyzing the state-level outcomes in the matched pairs.

The same framework would apply if interest instead focused on a hypothetical randomized experiment at the district level, for example in Pennsylvania or Ohio, where pilot programs have implemented VAA in only a subset of the school districts within the state. In that case, ideally, VAA districts in Ohio would be matched to non-VAA districts within Ohio (and similarly for Pennsylvania). If it is difficult to find well matched districts, pseudo-districts could be formed by matching individual schools within the districts.

With observational data, fully unobserved covariates that may affect both the decision to implement VAA (treatment assignment) and the outcome are a concern. It may be that states in which there is high value placed on education and measurement and thus implement VAA also have higher values of the outcome. Because assignment to implement VAA is not randomized, two states may look identical on observed covariates, but have different political environments or educational values that will affect both whether they implement VAA and their outcomes. In addition to simulations such as in McCaffrey et al. (2004) that assess the effect of omitted variables, sensitivity analyses such as those described in Rosenbaum and Rubin (1983b) could be used to explore the sensitivity of results to an unobserved variable that affects

both treatment assignment and outcome, i.e., non-ignorable treatment assignment.

4 Conclusion

Value-added assessment is a complex issue, and we appreciate the efforts of Ballou et al. (2004), McCaffrey et al. (2004) and Tekwe et al. (2004). However, we do not think that their analyses are estimating causal quantities, except under extreme and unrealistic assumptions. We argue that models such as these should not be seen as estimating causal effects of teachers or schools, but rather as providing descriptive measures. It is the reward structures based on such value-added models that should be the objects of assessment, since they can actually be (and are being) implemented. Of course, this focus requires a dramatic shift from current thinking, but a shift towards studying interventions that can be implemented and toward evaluations of them that can be conducted. We look forward to discussion of this approach.

5 References

Abadie, A. and Gardeazabal, J. (2003). "The economic costs of conflict: A case study of the Basque country." *American Economic Review* 93 (1): 113-132.

Ballou, D., Sanders, W., and Wright, P. (2004). "Controlling for student background in value-added assessment of teachers." *Journal of Educational and Behavioral Statistics.*

Barnard, J., Du, J., Hill, J., and Rubin, D.B. (1998) "A Broader Template for Analyzing Broken Randomized Experiments", *Sociological Methods and Research*, 27: 285–317.

Barnard, J., Frangakis, C., Hill, J, and Rubin, D.B. (2003) "A Principal Stratification Approach to

Broken Randomized Experiments: A Case Study of Vouchers in New York City" (with discussion and rejoinder) *Journal of the American Statistical Association* 98: 299-323.

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research.* New York, NY: McGraw Hill.

Burgess et al. (2001). "Medical Profiling: Improving Standards and Risk Adjustments Using Hierarchical Models" *Journal of Health Economics* 19: 291-309.

Christiansen, C.L. and Morris, C.M. (1997). "Improving the Statistical Approach to Health Care Provider Profiling" *Annals of Internal Medicine* 127: 764-768.

D'Agostino, R.B. Jr. and Rubin, D.B. (2000). "Estimating and using propensity scores with partially missing data" *Journal of the American Statistical Association* 95: 749-759.

Dehejia, R. and Wahba, S. (1999). "Causal Effects in Non–experimental Studies: Re-Evaluating the Evaluation of Training Programs" *Journal of the American Statistical Association* 94: 1053–1062.

Drake, C. (1993). "Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect" *Biometrics* 49: 1231–1236.

Goldstein, H. and Spiegelhalter, D.J. (1996). "League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance" *Journal of the Royal Statistical Society, Series A* 159:3:385-443.

Green. J. and Wintfeld, N. (1995). "Report Cards on Cardiac Surgeons: Assessing New York State's Approach" *New England Journal of Medicine* 332:1229-1232.

Holland, P.W. (1986). "Statistics and Causal Inference," (with discussion and rejoinder) *Journal of the American Statistical Association* 81: 945-970.

Holland, P.W. and Rubin, D.B. (1983). "On Lord's Paradox". Chapter 1 (pages 3–25) in *Principals of Modern Psychological Measurement*, ed. Wainer, H. and Messick, S. Hillsdale, NJ: Lawrence Erlbaum Associates.

Irie, M. (2001). *Handgun Waiting periods and Their Effectiveness on Crime Rates and Homicide: a Statistical Analysis on Gun Control.* AB/AM Thesis, Department of Economics and Department of Statistics, Harvard University.

Lalonde, R. (1986). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data" *American Economic Review* 76: 604–620.

Little, R.J. and Rubin, D.B. (2001). "Causal Effects in Clinical and Epidemiological Studies via Potential Outcomes: Concepts and Analytical Approaches" *Annual Review of Public Health* 21: 121–145.

Little, R.J. and Rubin, D.B. (2003). *Statistical analysis with missing data, 2nd Edition.* Wiley Series in Probability and Statistics. New Jersey: Wiley Interscience.

Lord, F.M. (1967). "A paradox in the interpretation of group comparisons" *Psychological Bulletin* 68:5: 304-305.

McCaffrey, D.F., Lockwood, J.R., Koretz, D., Louis, T.A., and Hamilton, L. (2004). "Models for value-added modeling of teacher effects." *Journal of Educational and Behavioral Statistics.*

Rosenbaum, P. and Rubin, D.B. (1983a). "The Central Role of the Propensity Score in Observational Studies for Causal Effects" *Biometrika* 70: 1: 41–55.

Rosenbaum, P. and Rubin, D.B. (1983b). "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome" *Journal of the Royal Statistical Society, Series B* 45: 212–218.

Rosenbaum, P. and Rubin, D.B. (1984). "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score" *Journal of the American Statistical Association* 79: 516–524.

Rosenbaum, P. and Rubin, D.B. (1985). "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score" *American Statistician* 39: 33–38.

Rubin, D.B. (1973). "The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies" *Biometrics* 29: 185–203.

Rubin, D.B. (1974). "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies" *Journal of Educational Psychology* 66: 688-701.

Rubin, D.B. (1976a). "Multivariate matching methods that are equal percent bias reducing, I: Some examples" *Biometrics* 32: 109-120.

Rubin, D.B. (1976b). "Multivariate matching methods that are equal percent bias reducing: II: Maximums on bias reduction" *Biometrics* 32: 121-132.

Rubin, D.B. (1976c). "Inference and missing data (with discussion)" *Biometrika* 63: 581-592.

Rubin, D.B. (1977). "Assignment to treatment group on the basis of a covariate" *Journal of Educational Statistics* 2: 1-26.

Rubin, D.B. (1978). "Bayesian Inference for Causal Effects: The Role of Randomization" *Annals of Statistics* 6: 34–58.

Rubin, D.B. (1979). "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies" *Journal of the American Statistical Association* 74: 318–328.

Rubin, D.B. (1980). Discussion of "Randomization Analysis of Experimental Data: The Fisher Randomization Test," by D.Basu, *Journal of the American Statistical Association* 75: 591-593.

Rubin, D.B. (1986) Which ifs have causal answers? Discussion of Holland's "Statistics and causal inference." *Journal of the American Statistical Association* 81, 961-962.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons.

Rubin, D.B. (1997). "Estimating causal effects from large data sets using propensity scores" *Annals of Internal Medicine* 127: 757-763.

Rubin, D.B. (2001). "Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology* 2: 169-188.

Rubin, D.B. (2003). "Teaching statistical inference for causal effects in experiments and observational studies." To appear in *Journal of Educational and Behavioral Statistics.*

Rubin, D.B. and Thomas, N. (1992a). "Affinely Invariant Matching Methods With Ellipsoidal Distributions" *Annals of Statistics* 20: 1079–1093.

Rubin, D.B. and Thomas, N. (1992b). "Characterizing the Effect of Matching Using Linear Propensity Score Methods With Normal Distributions" *Biometrika* 79: 797–809.

Rubin, D.B. and Thomas, N. (1996). "Matching Using Estimated Propensity Scores, Relating Theory to Practice" *Biometrics* 52: 249–264.

Rubin, D.B. and Thomas, N. (2000). "Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates" *Journal of the American Statistical Association* 95: 573–585.

Tekwe, C.D., Carter, R.L., Ma, C., Algina, J., Lucas, M., Roth, J., Ariet, M., Fisher, T., and Resnick, M.B. (2004). "An empirical comparison of statistical models for value-added assessment of school performance." *Journal of Educational and Behavioral Statistics.*

Barbara Lea-Kruger
March 18, 2004

JEVS – Evaluation

Background

For most of its recent history, program evaluation at JEVS consisted of simply satisfying the requirements outlined by their primary funders - the various government agencies that provided the bulk of the organization's resources. For the most part, these evaluation measures were fairly non-informative and short term. For instance, a job-training program would be required to place a certain amount of its clients in jobs, but would not be asked to track how long they stayed in those jobs past, at the most, 90 days. Most of the measures also tended to be generic. They were the same for any program in any locality regardless of other factors.

As a result, a culture arose that revolved around simply meeting these goals. If the program was required to place 10 percent of its participants, JEVS would consider the program a success if it met this goal, even if it could have conceivably placed more of the clients. The structure of the agency revolved around getting the grant, fulfilling at least the minimum requirements of the grant and not worrying about other possible outcomes.

In the last six years, JEVS experienced a tremendous growth spurt due to adding a welfare-to-work effort that almost doubled its size and budget. The rapid expansion made it more difficult for executive staff to effectively monitor all the agency's program activities and keep the diverse programs and their staff all focused on JEVS' overall mission and goals. As a result, the executive staff began a process of establishing an agency wide system to measure the outcome and impacts of its services.[1] While some of the impetus to undertake this effort did come from external pressure – government regulations were beginning to demand more stringent evaluation measures – the biggest push for instituting more appropriate evaluation was internal, driven by a strong commitment at the executive level to move beyond the funders' expectation and look at ways to add value to the client's experience.

The evaluation process

According to the readings, one of the first things that needs to be decided when attempting to institute an evaluation plan is who is going to do it. JEVS' approach was to incorporate both an internal and external evaluator. The Executive Management Team hired a consulting firm to help them with this process. The consultants presented a variety of Organizational Effectiveness Models for consideration including ISO9000, the Baldridge Criteria and Total Quality Management (TQM). The Executive Team decided to implement TQM.

The first step was for the Executive team to understand and internalize TQM principles. As part of this process the team came up with a set of core principles. For example, the team identified *quality customer service* and a *commitment to excellence* as two core principles. The next step was to spend a half a day developing and agreeing upon benchmarks of success. Once these two steps were completed the team felt ready to consider how best to implement TQM at the program level.

---

[1] The growth spurt created an opportunity for JEVS to reexamine not only its evaluation measures, but also its fundraising and strategic planning. Rather than simply reacting to the changing culture that resulted from the expansion, the executive team decided to manage the change by instituting neworganizational frameworks.