Journal of Public Administration Research and Theory Advance Access published December 16, 2004

"Best Practices" Research: A Methodological Guide for the Perplexed

Stuart Bretschneider
Frederick J. Marc-Aurele Jr.
Jiannan Wu
Syracuse University

ABSTRACT

Like many applied fields, public administration has a long-running love affair with the idea of "best practices" research. Although occasional reviews and critical examinations of approaches to best practices research have appeared in the literature (Overman and Boyd 1994), very little critical examination and reflection have been devoted to core methodological issues surrounding such work. The purpose of this article is twofold. First, we critically examine the underlying assumptions associated with "best practices research" in order to distill an appropriate set of rules to frame research designs for best practice studies. Second, we review several statistical approaches that provide a rigorous empirical basis for identification of "best practices" in public organizations—methods for modeling extreme behavior (i.e., iteratively weighted least squares and quantile regression) and measuring relative technical efficiency (data envelopment analysis [DEA]).

INTRODUCTION

Over forty years ago, Nobel Laureate Herbert Simon noted in his classic work, *Sciences of the Artificial* (1963), that many applied fields shared a common concern over design and action. It is therefore not surprising to see that many of these same fields (e.g., medicine, management, computer science, and law) are concerned with how to improve actual performance through identification and codification of something typically referred to as a "best practice." Though overly simplistic, the logical appeal of a "best practice" is profound, particularly when individuals and organizations operate in competitive and hostile environments (Scott 1998). While most of the so-called best practices literature comes from management consultants and practicing managers, it has also found its way into a variety of more academic settings (Hatry 1999; Holzer and Callahan 1998; Keehley 1997).

The first objective of this article is to examine the concept of a "best practice" and to attempt to derive what would be the necessary and sufficient conditions to the identification of a true "best practice." Second, this article will critically review various methods and

The authors would like to thank Lynne D'Amico for her help with some of the computations and Ken Meier for his helpful comments. Address correspondence to Stuart Bretschneider at sibretsc@maxwell.syr.edu.

doi:10.1093/jopart/mui017
Journal of Public Administration Research and Theory

techniques that are currently in use to identify "best practices." The article begins with a short section that argues why concern over "best practices" is warranted within the field of public management and why a formal and careful research agenda is needed. Next, we will review the two general approaches used to identify a "best practice," in order to establish what might be the necessary and sufficient conditions to perform such a task. This is then followed by a lengthy review of various analytic approaches that are currently in use to empirically identify "best practices." We conclude by summarizing what we feel are the hallmarks of a good "best practices" research design for public administration scholars, as well as a list of key critiques that continue to be problematic, even within this design.

WHY SHOULD WE CARE ABOUT "BEST PRACTICES" RESEARCH?

Dwight Waldo envisioned public administration as an applied field in his classic work, *The Enterprise of Public Administration* (1980), and emphasized the concept of administrative technologies. While Simon and Waldo disagreed on issues of emphasis, there is clear agreement that the relevant domain of public administration was deliberative actions taken to achieve some ends. From this perspective, theory should inform action. While a theory might provide an explanation of how changes in communications technology affect organizational structure, for example, the action issue derives from an additional notion of values about what organizational outcome is preferred. Even if one can affect the degree of hierarchy through technological action, preferences of the outcome must be clearly stipulated before deciding whether or not it is better to flatten out the hierarchy. Using this example, we take the view that the role of the researcher is to increase our understanding of cause/effect linkages and how they can be manipulated. Thus issues of controllable cause are important in the formulation of explanatory theory for public administration. The values that affect action, though, derive from political and cultural processes that include many other individuals and groups. Researchers have a role to play here, but they are not alone. Political and cultural processes, while affecting perceptions, are less capable of defining underlying cause/effect relationships.

This article focuses on the initial, and probably the more relevant, role for research in public administration: the development of causal theory with emphasis on the identification of controllable causes. The demand for such theory, though, far exceeds the capacity of our existing research infrastructure to respond. This discrepancy between the institutional capacity for scientifically based research and practical demands for useful knowledge has created tension between practitioner communities and academics, not only in public administration but also in all the design sciences. This disparity leads to the growth of nonacademic sources of causal and prescriptive knowledge, including consultants, professional associations of users (International City Managers Association, Government Finance Officers Association, American Society for Public Administration, etc.), and not-for-profit groups interested in social change (Brookings Institution, Hoover Institution, etc.). Most of these sources typically develop a common sense form of "best practices" approach to vet their information. These generally take the form of "best practices" competitions or awards, where a group of experts use "human judgment" to identify the "best practice" from a set of cases submitted (Altshuler and Behn 1997).

Unfortunately, the lack of a more rigorous approach to "best practices" results in several problems. First, a large body of knowledge in decision psychology identifies numerous problems of bias associated with the application of human judgment in forming

decisions (Kahneman, Slovic, and Tversky 1982). Second, there is a problem of comparability in that the cases that are typically considered in a "best practice" competition suffer from severe selection problems and do not form reasonably comparable groups. For example, a small city licensing program may be compared to a state-level welfare-to-work program. Third, since the focus of such activities is on action, the clear and careful identification of the causal chain is usually vague and ambiguous. Typically, case-based analysis suffers from weak internal validity; therefore, many of these studies do not specifically deal with alternative explanation for success. Generalizability is therefore also a problem in these competitions. Fourth, all of these problems end up requiring the reader to review the "best practices" and to evaluate how they might (or might not) be used within their specific context. That is, application to another setting is not straightforward and is often extremely problematic. A final problem arises from the context that created the demand for such "practical" knowledge. There exists a real market for "practical" knowledge. Thus, various informational anomalies are created, including incentives not to test or verify and the creation of differentiation even when it is not present to enhance revenues.

While academic research in public administration can produce the type of causal theories that are useful to practice, the existing institutional infrastructure cannot keep up with the huge demand. This knowledge gap creates a market for developing multiple alternative sources of "useful" knowledge, but these "best practices" may possess substantially problematic characteristics. To illustrate this point, consider the almost universal situation of taking your car to the repair service to go through several fixes before the actual solution is found (if ever). This illustrates that in a complex physical system it is often impossible to identify the causal chain in order to prescribe an action that rectifies the situation. The weakness is not in the action; it is in our poor understanding of the cause/effect relationships at work. Thus, it is useful and important to understand how to conduct "best practices" research in order to not only identify actions but also to make sure that the action is the appropriate cause to some desired effect.

DEFINING A "BEST PRACTICE": THE NECESSARY AND SUFFICIENT CONDITIONS

The term "best practice" implies that it is best when compared to any alternative course of action and that it is a practice designed to achieve some deliberative end. Hence, there are three important characteristics that are associated with a "best practice":

1. a comparative process,
2. an action, and
3. a linkage between the action and some outcome or goal.

For example, we might compare how several organizations conduct strategic planning and relate that to how successful each organization is at garnering resources from its environment.

While this definition seems simplistic, there are several important issues that arise from it. *Comparability* across actions and outcome is important to the identification process, as well as to context. Does it make sense to include hospitals, schools, manufacturing firms, and government agencies together as a single group of organizations while searching for a best practice; or, should they be segmented, and if so, what is the

basis for forming the different groups? A second important issue that arises from our definition is how to characterize the *linkage between the action and the outcome.* This is essentially the problem of specifying a cause/effect relationship between the practice and the outcome. From a social science perspective, we know that understanding cause/effect relationships is complex and difficult. Findings are especially subject to errors in specification. One approach to dealing with the specification issue comes from economics, which provides one of the powerful empirical approaches to developing these types of relationships. Production theory in economics is used to relate inputs (including resources and activities) with outputs.[1]

These two points are also related to each other. The *definition of comparability* in cases for comparative purposes will in part derive from our theories of cause and effect. For example, if we focus on schools as an organizational unit, our theory for linking actions with outcomes might be different if we separate elementary schools from secondary schools, thus affecting how we form comparable groups.

While comparability is a necessary condition for identification of a "best practice" it is *not* sufficient. To be sufficient the cases selected for comparison must include all comparable cases for the relevant domain. The reason for this is that any successful comparative approach can only find the "best case" within the sample. If a better case occurs outside the sample, then the idea of a "best practice" is inappropriate. Although it might be argued that the results from such a limited approach could yield a "good" practice, even that notion can be misleading, depending on the mix of performances inside versus outside the comparison sample. Hence, to identify a "best practice" in elementary schools in Texas at a particular point in time would require looking at all elementary schools in Texas. Any results from such a study would have difficulty generalizing to other states and, possibly, to future years within Texas, depending on the magnitude of change in the state, national, or international environments that impact Texas elementary schools.

Given this problem, one might think that the solution is to pull a representative sample of cases, but unfortunately, that research strategy does not solve the problem. There is no guarantee that such a sample will include the ultimate "best case." The best possible statements that can be made from a random sample will be probabilistic statements about how far the sample "best case" is from the population "best case," which will also require assumptions about the probability distribution of cases. This is essentially a form of sample inference on the maximum. For an example of this type of logic, see "MULTILOC: A Multiple Store Location Problem" (Achabal, Gorr, and Mahajan 1982).

As noted above, it is also necessary to have a clearly articulated causal structure to relate inputs to outputs. Some studies have attempted to focus on variation in outputs only to rank cases accordingly. Unfortunately, this approach does not adequately ensure comparability across cases. For example, given two elementary schools, one with 60 percent of sixth graders passing a standardized math test and another with 85 percent, simplistic ranking does not take into account that the school with the lower pass rate is larger, includes a more racially diverse population of students, and has a lower teacher-student ratio than the other school. Such input differences are used to help form more appropriate comparisons with controls for theoretically and statistically important

1 While there is a literature that differentiates between the use of "output" and "outcome," for this article, we use the two terms interchangeably and mean the broader results from organizational activity that are typically called "outcomes."

differences. To do so, it is necessary to articulate a cause/effect structure that identifies both key causal elements and how inputs are related to outputs. The various frameworks provided within the field of production theory in economics are a valuable approach to organizing such a causal model. This process must accept the existence of variance in inputs and outputs and uses the causal framework to develop appropriate comparisons.

The final piece of our definition turns on the mechanism for conducting the comparison across cases. Currently, there are two approaches: one is to apply human judgment, and the other is to apply statistical models. Both approaches have strengths and weaknesses. The problem with most of the current applications of human judgment is that they are not nested within an appropriate research design framework. For example, the typical application of judgment (Behn 1993) presents the results, and not the process. Little information, if any, is given about the comparison groups or the criteria for sorting. In essence, this approach is based on the authority of the judge, and it is inherently antiscientific and, therefore, unreproducible. This is not, by necessity, the only way human judgment can be used. In more statistically oriented approaches, human judgment comes into play in the various modeling and method decisions. However, unlike the authority model, such judgment decisions are carefully articulated and open to scrutiny and criticism. In principle, one can envision a situation where a small sample of comparable cases are enumerated, but the nature of the cause/effect relationships are so highly complex that human judgment techniques are a reasonable approach to develop rankings and identify a "best practice." Even in this context, great care must be exerted to apply scientific practices, including full disclosure and scrutiny by peers. The problems associated with statistical models have already been noted. They include problems of identification from relevant samples to including the true best case, appropriate comparability of cases, and sufficient definitions of cause/effect elements and relationships.

The basic results from this analysis have helped us to define the joint necessary and sufficient conditions for finding a "best practice." There are essentially two conditions that must be met:

1. completeness of cases; and
2. comparability of cases.

To obtain the second condition, it is necessary to have a complete and accurate statement of the causal relationships linking inputs to outputs. While admittedly these conditions are rarely attainable, they at least establish a foundation for judging any research design that is aimed at the identification of a "best practice." The remainder of the article will focus on the actual techniques that are currently available for researchers to sort through the issues of completeness and comparability of cases in the context of causal models.

ISSUES OF COMPLETENESS: SAMPLING AND MEASUREMENT

We have already established that the condition of completeness requires the design to obtain empirical information on all the relevant cases. It has also been noted that while completeness is the goal, it will rarely be obtained. Some studies have delimited the domain in such a way to closely approximate the completeness condition. Geographic and temporal limits are typically used, such as looking at a high school in one state or city for one year. This highlights another major characteristic of typical "best practice" studies—the limited

generalizability of results or problematic external validity. This situation is generally true in most applied and action-oriented research settings, including evaluation studies.

The approach of limiting the domain of a study is preferable to the typical social science approach of random sampling. The logic of random sampling is to use statistical inference to enhance generalizability about typical situations. But as noted earlier, the appropriate use of statistical inference in the context of "best practices" requires one to focus on extremes, not mean or typical values. While some inference theory to do this is available, it is highly complex and limited.

Completeness depends on comparability. That is, a theory-based definition of comparability is a necessary antecedent for delimiting the set of cases to be compared. The schools example illustrates that segmenting by elementary and secondary status not only helps to deal with the comparability concern, but also reduces the number of study units that need to be observed. The next section will take up the issue of comparability more completely.

Once a set of cases is defined, there are some important measurement issues that need to be addressed. Again, measurement is related to issues of comparability and what constitutes relevant inputs and outputs. One important early consideration in the research process is that of establishing reasonable approaches to measuring desired outcomes from the system. If one is concerned with "best practices" from management, it is critical to be able to accurately and reliably measure results or outputs. This requirement is even more restrictive than completeness and comparability concerns, particularly in public administration, where the major activities are managerial in nature. What are the results of, for example, planning as an activity? While we are able to measure different aspects of a planning activity (resources uses, level of involvement), it is much more difficult to identify the relevant outcomes. This can be referred to as an observability condition.

In most situations, the outcomes from action are complex and usually multidimensional. Consider the case of primary education, where we generally tend to think about outcomes in two broad categories: knowledge and skills. Then, within each of these categories, the lists of specific knowledge areas and skills are quite large. The design issue here is what are the implications of narrowing in on a single or small subset of outcomes. To do this, we need either to assume or demonstrate that those excluded outcomes are independent from those included in the analysis. This again is a rather stringent condition, since the most typical process for generating an outcome requires the use of resources (inputs) that, once they have been used for that purpose, cannot be used for the generation of alternative outputs. Thus, within an organization that produces multiple outcomes, resource utilization implies trade-off decisions and interdependence across outcomes. The exceptions would be if two outcomes are co-produced automatically or if one is a complete by-product from the production of the other. Trying to examine how elementary schools generate reading skills without considering math skills requires the assumption that there are not internal trade-offs in resource utilization in the process.

This analysis has identified that any "best practice" design will be, by its very nature, less generalizable than standard social science research design. While not surprising, given the applied nature of "best practices" work, it does identify some clear suggestions. Delimiting the domain of cases in space and time to define a complete and exhaustive set best approximates completeness. This approach is preferable to the traditional social science approach of random sampling. We have also identified an observability requirement with regard to outcomes and determined that delimiting a design by focusing on selected outcomes is also problematic since most are interdependent.

Table 1
Modeling Approach for "Best Practices" Research Designs

Number of Outputs	Model Form	Estimation
One Output	$y = f(x_1, \ldots, x_n)$ (Linear and Nonlinear Models)	Average Relationship (Ordinary Least Squares Estimation) Extreme Behavior (SWAT, Quantile Regression)
Multiple Outputs	$g(y_1, \ldots, y_m) = f(x_1, \ldots, x_n)$ (Parametric and Nonparametric Models)	Average Relationship (Canonical Correlation) Extreme Behavior (Econometric Methods and Nonparametric Data Envelopment Analysis)

ISSUES OF COMPARABILITY: CAUSAL MODELING

A General Framework

The general framework we propose in considering a formal model to link inputs and output(s) is the theory of production from economics. Production theory models have several important strengths. First, they recognize that some formal relationship between inputs and output(s) exists and that a ''best practice'' is essentially about comparing how these units transform inputs (resources) into output(s). These models are conditioned on a comparability criterion; thus, the theory requires that units all use generally the same production or transformational process. Also implicit in these models is the existence of an optimal situation for converting inputs to output(s) where all firms or units are assessed relative to that optimum. Clearly, the idea of a ''best practice'' is an integral part of production theory. And, finally, the comparisons are conducted in a way to compare across similar cases within the larger group of cases.

Table 1 organizes various forms of production models by how they deal with outcomes and how the technical process of estimation is conducted.

The estimation of relationships from data, typically, focuses on the expected or average case. Regression analysis generates a model of mean or average behavior. While this is useful in statistical modeling from random samples, especially when the researchers want to generalize to a large group, it is not appropriate when searching for a ''best practice.'' Here we are interested not only in the behavior of the ''best'' units in the sample, but in how their behavior is different from the typical and poor-performing unit. The remainder of this section will discuss the technical options available to researchers for estimating relationships that focus on ''extreme'' behavior.

One Output Models

Assuming that one can obtain measurements on a reasonably complete and comparable set of units, and that those measurements include one output and a set of inputs, there are three alternative approaches available for investigating extreme behavior. All three of these approaches assume a basic understanding of multiple regression analysis for estimating a conditional mean relationship between inputs and the output. While is it not necessary to

the following discussion, we will assume that the relationship between inputs (x_j's) and the output (y_i) is linear. The form of the model is

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i, \text{where} \quad [1]$$

the β_j's are parameters that describe the specific relationships between the input variable x_j and the output y, and ε_i is a random disturbance associated with the ith unit. In the context of standard estimation, the results describe the expected value or conditional average unit's relationship:

$$E(y_i) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}. \quad [2]$$

This approach does identify extreme cases, but only by focusing on the residuals from the estimation process.

The difference between the conditional average for a particular case and the actual measurement of output is called the residual. That is,

$$e_i = y_i - E(y_i). \quad [3]$$

This quantity measures how far each unit is from what the model estimates to be the typical case for the specific combination of values of the inputs. For example, if two cases had identical values for the inputs but different output values, the residuals would also be different; but since the input combination was the same, we could directly compare the two. The larger residual value implies that the associated unit had somehow done a better job at using these inputs to produce the output. By this logic, the maximum-value case is one way to think about identifying the ''best practice.'' For an example of how to make use of these residuals to sort high- from low-performing organizations, see Pandey and Bretschneider (1997). One of the drawbacks of this approach is that it does not provide any information concerning how these high- (or low-) performing units converted their inputs to outputs. The only information about that process applies to the average or mean case. Nevertheless, this type of residual analysis is a powerful heuristic approach that can be used to identify atypical cases and explore for ''best practices.''

The other two approaches for identifying and analyzing the extreme organizations derive directly from the residual concept noted above. They both reason that the concept of deviation from the mean (or median) should provide useful information as to how the units out in the ''tails'' of the distribution behave. Using that idea, these approaches attempt to differentially weight the information by essentially providing more weight (i.e., importance) to the cases that are away from the center of a dataset. The notion of a weighted estimate is not new, but it is typically used to correct unequal variances across the data in order to allow for standard statistical inference (Western 1995). The shaded area in figure 1 reflects how data points are distributed around a regression line. This is a situation where weighted estimation might be used in traditional regression. The problem is that when the input values for x are small, the variation around the conditional average is small, but when the values of the input are large, there is more variation around the regression line. It is instructive to note that the usual weighting strategy is to weight larger residuals less and smaller ones more in order to produce a more uniform pattern, so that inference from the sample mean can be used to generalize to a larger population. But if one has a complete and comparable set of units, inference is not an issue. An alternative way to view figure 1 is to see it as if there were three distinctive groups of units as depicted in figure 2. While the overall display is the same, the interpretation is different. Considering

Figure 1
Unequal Variance in a Simple Regression

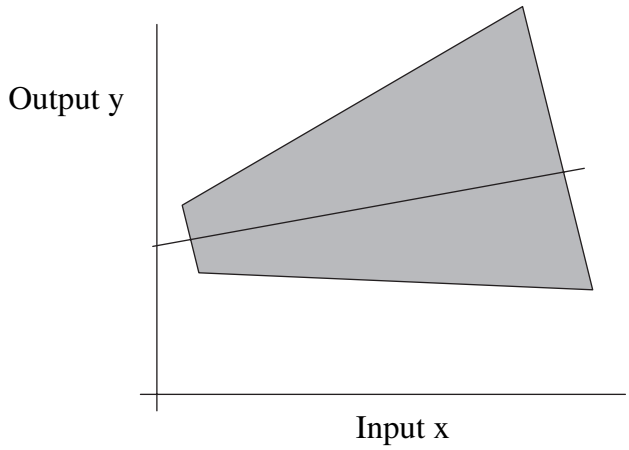
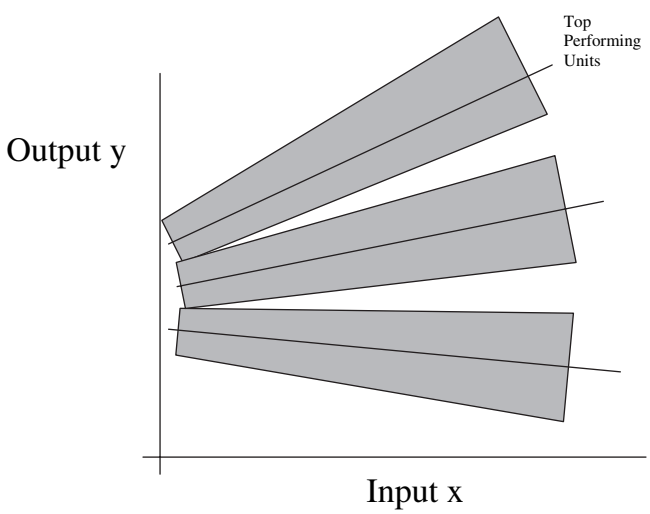


figure 2, the goal is to identify unique and distinctive conditional means for each group, with an emphasis on the upper group of high-performing organizations. Marc-Aurele, D'Amico, and Bretschneider (2000) simulated data of the form depicted in figure 2 in order to develop comparative information on the two alternative weighting strategies aim to estimate models of the extremes. Although figure 2 depicts the situation with only one input variable, these concepts generalize to situations with multiple inputs and nonlinear relationships.

While there are important differences in the two approaches, it is important to remember they both attempt to do the same thing by using a form of weighted estimations. The first approach is a form of iterated weighted least squares popularized by Meier and

Figure 2
Three Subgroups of Cases



Gill (2000) as the substantively weighted analytical technique (SWAT). Their book, *What Works: A New Approach to Program and Policy Analysis*, contains an excellent description of the approach, with detailed instructions for how to actually conduct the analysis. In essence, through a process of repeated least squares estimations followed by differentially weighting the observations with positive (above the regression plane) residuals, this technique generates an estimate of the conditional relationship out in the tail of the distribution for the top-performing cases (Gill and Meier 1998; Meier, Gill, and Waller 1997; Meier and Keiser 1996; Meier, Wrinkle, and Poliand 1999). Note that in figure 2 the summary line has a different intercept and slope, thus identifying how these high performing units convert inputs to output(s) differently in comparison with either the typical or low-performing units.

The second weighted approach is quantile regression. This technique is less widely diffused, especially among applied social science and public administration scholars. Historically, it comes from the original work in nonparametric and order statistics (Bassett and Koenker 1978; Koenker and Bassett 1978). Like SWAT, this approach generates a conditional relationship between inputs and output(s) for the tails of the cases. Specifically, it generates the conditional q quantile, where q can be any percentage cut point in the distribution of cases. The resulting relationship for, say, the 95 percent quantile generates a model where 95 percent of the cases are on the plane formed by the model or below it and 5 percent are above it. Dielman and Pfaffenberger (1982) provide a clear explanation of how to estimate the base case of the median regression or 50 percent quantile. Quantile regression uses the following linear programming (LP) model to generate the estimates:

$$\text{Minimize } z = \sum_i (we_i^- + (1-w)e_i^+), \text{ subject to} \qquad [4a]$$

$$(b_0 + b_1x_{1i} + \ldots + b_px_{pi}) + e_i^- - e_i^+) = y_i \quad i = 1, \ldots, n, \text{ where} \qquad [4b]$$

z is the value of the objective function to be minimized, e_i^- and e_i^+ are the negative and positive residual from estimating parameters b_j, and w is a weight.

When $w = .5$, this LP model generates estimates for the parameters such that the sum of the absolute value of the residuals is minimized, and 50 percent of the observations are above the plane and 50 percent are on or below the plane. This is known as the median regression, or 50 percent quantile regression. By changing the value of w, the researcher can estimate any conditional quantile. For example, if $w = .95$, then $(1 - w) = .05$, and the LP model generates the 95 percent conditional quantile. This formulation is presented to demonstrate that both quantile regression and SWAT form weighted estimations, and that they both result in the generation of models that are focused on explaining the behavior of units in the tails (i.e., ''extremes'') of the distribution, in contrast to those in the middle (mean or median).

Table 2 summarizes the three approaches for identifying and analyzing ''best practices'' in situations with one output and multiple inputs.

The use of residuals from a standard regression analysis may help in identifying high-performing units, but it cannot account for specific and unique relationships in the tails of the distribution, as do SWAT and quantile regression. Note that when using SWAT and quantile regression techniques the user should consider the inputs relevant to the high-performing units, which may not be the same as those used by average-performing units.

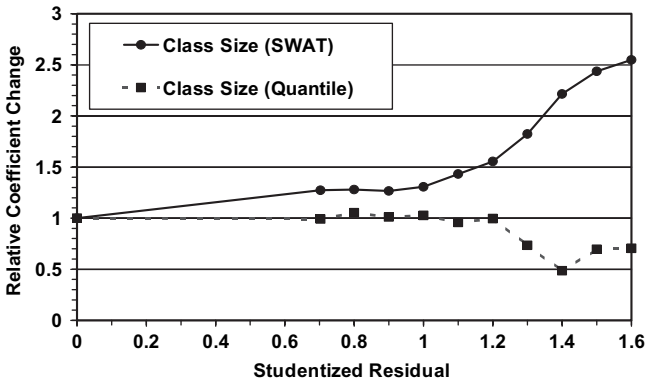
Table 2
Comparing Approaches to One Output Model Estimation

Criteria	Regression Residuals	SWAT	Quantile Regression
Specific Model	No	Yes	Yes
Estimation	One step	Multiple steps	One step
Complexity	Simple	Moderate	Moderate to complex
Focus	Deviations from conditional mean	Heuristic and approximate	Exact conditional quantile

Complexity is difficult to assess since it depends on the background of the user. In general, though, quantile regression is more complex to use, for a variety of reasons. Most social science and public administration scholars are not familiar with median regression or linear programming. Also, the theory for conducting inference from samples associated with the estimated quantile and the parameters in the conditional relationship is complex, involving asymptotic theory that applies only in large samples (Judge et. al 1985). Quantile regression's advantage is that it generates an exact estimate of a desired quantile, while SWAT procedures are heuristic and approximate as to how far out in a tail the estimates actually are.

Before leaving this section, it is useful to reconsider our central issue: how well can each approach identify and analyze a ''best practice''? Clearly, both SWAT and quantile regression are superior to just examining regression residuals by themselves, though the simpler approach offers a lot to researchers who are interested in exploratory analysis. In the field of statistics, a great deal of work has been done both analytically and through the use of simulation to investigate the strengths and weaknesses of traditional regression techniques based on means, and squared error loss functions with nonparametric and order-based techniques such as median regression and quantile regression. Most of this work focuses on sampling error and inference, but it tends to find quantile regression more robust and more statistically efficient, that is, smaller, tighter, confidence intervals on estimates for comparably sized samples (Bassett and Koenker 1978; Buchinsky 1994, 1995, 1998a, 1998b; Koenker and Bassett 1978; Koenker 2000). While most of this is only tangentially relevant, one simulation study finds that when comparing SWAT and quantile regression head to head, both approaches become more and more inaccurate as they attempt to estimate relationships in the ''extreme'' tails of a distribution. In general, these results suggest that anything beyond the 85 percent quantile may be problematic (Marc-Aurele, D'Amico, and Bretschneider 2000). This problem was even more pronounced when both techniques were applied to the same empirical dataset, resulting in different effects for each of the inputs on the output. Using the data and model described by Meier and Gill (2000, table 4.1), figure 3 graphs how the regression coefficient changes relative to its initial value as the estimation process moves further into the upper tail (high performers) of the distribution. These results control for all the variables in the model and the various points in the distribution as estimation moves out into the tails. While no real differences occur in these estimates until a weight of more than 1.2, from that point forward SWAT suggests an increasing effect of reducing class size while quantile regression suggests a decreasing effect! Which one is correct? What would be the policy implications of selecting the wrong one? Our suspicion is that neither is correct, and that the lack of data in the extreme tails causes increasing uncertainty in the estimation process for both techniques. This clearly suggests that researchers must be careful about pushing either of

Figure 3
Comparison of SWAT and Quantile Regression Estimates for Class Size in General Production Function Model of Schools in Texas



these techniques too far in the direction of finding the ''extreme'' (i.e., ''best'') cases in a sample, leaving us with ''good'' practices instead of ''best'' practices.

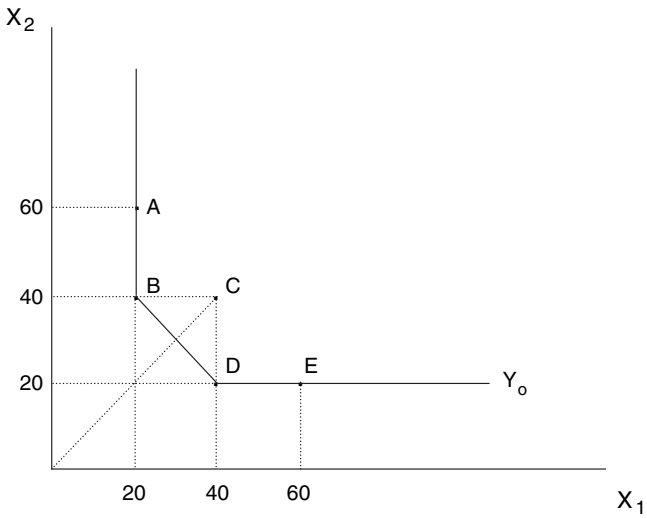
This last point also highlights that an important focus of research in this area must include issues of method and technique. Analytic approaches have provided some basic information about how these methods work, their underlying assumptions, and their limitations. But they do not fill in all the detail associated with implementation and use. In general, studies that simulate specific situations and then investigate the strengths and limitations of these techniques under known conditions are the only way to develop evidence about the techniques. Examples of this type of work include Marc-Aurele, D'Amico, and Bretschneider (2000) and Wu and Bretschneider (2001).

Multiple Output Models

At this point, it is useful to review the basics of economic production theory. The theory assumes that a firm has multiple outputs and inputs, but that for a given technology of production, there exists a relationship known as the production possibility curve or frontier. Figure 4 provides an example for a firm that produces two outputs. The curve shows how there are various mixes of the two outputs possible and that as more of one is produced, less of the other is produced. This is because the theory assumes that the inputs are fixed at some level. The theory goes on to assume that all firms in an industry (e.g., automobile manufacturers, high schools, etc.) have a single, common core technology and that, for a given set of inputs and a fixed technology, the curve represents the ''best'' or optimal level of outputs. Figure 4 indicates that firms A, B, D, and E are all on the frontier and are therefore efficient, but firm C is not on the curve and is therefore inefficient. All the applications of production theory attempt to estimate from data about a firm's inputs and output(s) the production frontier and identify which firms are on the frontier and which ones are not. Also, these applications are able to generate measures of relative technical productive efficiency by measuring how far inefficient firms are from comparably (in terms

2 Two possible exceptions to this are the concepts of MANOVA and canonical correlations.

Figure 4
Production Possibility Curve and Technical Productive Efficiency



of similar inputs) efficient firms. It is important to remember that these techniques only generate *relative* measures, which are relative to the cases in the sample, again emphasizing the need for completeness of the data.

Traditional statistical models do not generally deal with functional forms where there are multiple outputs.[2] One statistical technique that is worth considering is that of canonical correlation. This approach considers a set of inputs and a set of outputs. The procedure attempts to find a set of weights to form a linear composite of the inputs and a set of weights to form a linear composite of outputs such that the correlation between the two resulting variables is maximized (Gyimah-Brempong and Gyapong 1991; Ruggerio 1996c, 1998; Vinod 1968).

A second, much more commonly applied approach is known as data envelopment analysis (DEA) (Charnes, Cooper, and Rhodes 1978). This approach was developed in operation research and requires that a series of linear programming problems be run for each of the firms or organizations in the sample. Banker, Charnes, and Cooper (1984) have shown how the estimated production frontier can be used to determine the relative technical efficiency of producing units. Allowing the production possibilities frontier to exhibit variable returns to scale, the DEA program can be operationalized with the following linear programming model to evaluate the technical efficiency, η_0, of each firm or organization in the sample:

Minimize η_0, such that [5a]

2 Two possible exceptions to this are the concepts of MANOVA and canonical correlations.

3 Some examples from the area of education include Bifulco and Duncombe (1998), Ruggiero, Duncombe, and Miner (1995), Bessent and Bessent (1980), Bessent et al. (1982), and Bessent et al. (1983).

$$\sum_{i=1}^{n} \theta_i y_{ij} \geq y_{0j} \quad \forall\, j = 1, \ldots, S \qquad [5b]$$

$$\sum_{i=1}^{n} \theta_i x_{ik} \leq \eta_0 x_{0k} \quad \forall\, k = 1, \ldots, M \qquad [5c]$$

$$\sum_{i=1}^{n} \theta_i = 1 \qquad [5d]$$

$$\theta_i \geq 0 \quad \forall\, i = 1, \ldots, n, \text{ where} \qquad [5e]$$

y_{ij} is the ith organization and the jth output; x_{ik} is the ith organization and the kth input; η_0 is the efficient for firm 0, the reference firm; and θ_i are estimated weights.

This linear programming problem is repeatedly solved n times for each of the n organizations. The maximum possible value for η_0 is 1, and the solution also provides information on what additional output would be necessary for each inefficient firm to move to the frontier, given its current set of inputs.

Two points are important when thinking about DEA and other methods for analyzing multiple input and multiple output situations as a method for studying and identifying ''best practices.'' First, while there have been many applications of DEA in public administration,[3] like the use of SWAT and quantile regression, it is difficult to assess in a comparative sense which technique is best given a particular empirical condition. This, once again, forces us to include methods studies and simulation designs to better understand each technique's operating characteristics (Bifulco and Bretschneider 2001; Banker, Conrad, and Strauss 1985; Banker, Gadh, and Gorr 1993). Second, while DEA does permit the analyst to consider more complex and realistic situations with multiple outputs, it does reduce down to a single measure of performance for determining ''best,'' which theoretically is tied to the concept of productive efficiency.

CONCLUSIONS

While this article has consistently used the idea of "best practice," it has really focused on the methodological requirements associated with the evaluation and ranking of organizational performance. The idea that a "best" practice exists, and can be identified analytically, is essentially a "pure type." Any empirical attempt will fall short, but the intent of this article is to provide guidance to public administration scholars on how to find "good" practices and understand how organizations can be differentiated.

The appendix contains a first attempt to develop a checklist of issues that must be addressed before any empirical measurement or analysis is undertaken. These questions provide the framework for any study that is interested in sorting "good" from "bad" practices. As with any design, the answers that are given will force the analyst to make some simplifying assumptions and trade off some characteristics in the design against others. The heavy emphasis on both methodology and technical methods reflect our

3 Some examples from the area of education include Bifulco and Duncombe (1998), Ruggiero, Duncombe, and Miner (1995), Bessent and Bessent (1980), Bessent et al. (1982), and Bessent et al. (1983).

support of Meier and Gill's (2000) pronouncement that "individuals with normative concerns, those who seek to give policy advice, need to be better methodologists than those interested solely in building empirical theory."

APPENDIX

Checklist for "Best Practices" Research Design Issues

1. How complete (or representative) is the sample?
   a. Comparisons are only as good as the cases in the group.
   b. Random representative samples cannot guarantee a "best" outcome.
   c. Random samples can provide probability limits on how far the population "best" is from the sample "best."
   d. Limiting geographic and temporal space to achieve completeness is a better strategy than random sampling.
2. How comparable (and on what basis are you determining comparability) are the units in the sample?
3. Have you identified all the major inputs and outputs for the system?
4. Have you used an appropriate structure (i.e., linear or nonlinear) for relating inputs and outputs?
5. Does the nature of the cause and effect relationship (i.e., linear or nonlinear) for the typical case remain structurally the same for the "extreme" cases?
6. Does the set of input and output variables for the typical case remain the same for the "extreme" cases, or are different models necessary?

REFERENCES

Achabal, D. D., W. L. Gorr, and V. Mahajan. 1982. MULTILOC: A multiple store location problem. *Journal of Retailing* 58 (2): 5–25.

Altshuler, A. A., and R. D. Behn. 1997. *Innovation in American government: Challenges, opportunities, and dilemmas.* Washington, DC: Brookings Institution Press.

Banker, R. D., A. Charnes, and W. W. Cooper. 1984. Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science* 30:1078–92.

Banker, R. D., R. F. Conrad, and R. P. Strauss. 1985. A comparative application of DEA and translog methods: An illustrative study of hospital production. *Management Science* 32:30–44.

Banker, R. D., V. M. Gadh, and W. L. Gorr. 1993. A Monte-Carlo comparison of two production frontier estimation methods: Corrected ordinary least squares and data envelopment analysis. *European Journal of Operational Research* 67:332–43.

Bassett, G., and R. Koenker. 1978. Asymptotic theory of least absolute error regressions. *Journal of the American Statistical Association* 73:618–22.

Behn, R. D. 1993. Case-analysis research and managerial effectiveness: Learning how to lead organizations up sand dunes. In *Public management: The state of the art*, ed. B. Bozeman, 40–54. San Francisco: Jossey-Bass.

Bessent, A., and E. Bessent. 1980. Determining the comparative efficiency of schools through data envelopment analysis. *Educational Administration Quarterly* 16:57–75.

Bessent, A., E. Bessent, A. Charnes, W. Cooper, and N. C. Thorogood. 1983. Evaluation of educational program proposals by means of DEA. *Education Administration Quarterly* 19:82–107.

Bessent, A., E. Bessent, J. Kennington, and B. Reagan. 1982. An application of mathematical programming to assess productivity in the Houston Independent School District. *Management Science* 28:1335–66.

Bifulco, R., and S. I. Bretschneider. 2001. Estimating school efficiency: A comparison of methods using simulated data. *Economics of Education Review* 20:417–29.

Bifulco, R., and W. Duncombe. 1998. The identification and evaluation of low-performance schools: The case of New York City. Paper presented at the annual meeting of the American Education Finance Association, Mobile, AL.

Buchinsky, M. 1994. Changes in the U.S. wage structure, 1963–1987: Application of quantile regression. *Econometrica* 62 (2): 405–58.

———. 1995. Estimating the asymptotic covariance matrix for quantile regression models: A Monte Carlo study. *Journal of Econometrics* 68 (2): 303–38.

———. 1998a. The dynamics of changes in the female wage distribution in the USA: A quantile regression approach. *Journal of Applied Econometrics* 13:1–30.

———. 1998b. Recent advances in quantile regression models. *Journal of Human Resources* 33 (1): 88–126.

Charnes, A., W. W. Cooper, and E. Rhodes. 1978. Measuring the efficiency of decision making units. *European Journal of Operational Research* 2:429–44.

Dielman, T., and R. Pfaffenberger. 1982. LAV (least absolute value) estimation in linear regression: A review. *Journal of Studies in the Management Sciences* 19:31–52.

Gill, J., and K. J. Meier. 1998. Ralph's Pretty Good Grocery versus Ralph's Super Market: Separating excellent agencies from the good ones. Paper presented at the annual meeting of the Midwest Political Science Association, Chicago.

Gyimah-Brempong, K., and A. Gyapong. 1991. Characteristics of education production functions: An application of canonical regression analysis. *Economics of Education Review* 10:7–17.

Hatry, H. P. 1999. *Performance measurement: Getting results.* Washington, DC: Urban Institute Press.

Holzer, M., and C. Callahan. 1998. *Government at work: Best practices and model programs.* Thousand Oaks, CA: Sage.

Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lutkepohl, and T. Lee. 1985. *The theory and practice of econometrics.* 2d ed. New York: John Wiley and Sons.

Kahneman, D., P. Slovic, and A. Tversky. 1982. *Judgment under uncertainty: Heuristic and biases.* Cambridge, UK: Cambridge University Press.

Keehley, P. 1997. *Benchmarking for best practices in the public sector: Achieving performance breakthroughs in federal, state, and local agencies.* San Francisco: Jossey-Bass.

Koenker, R. 2000. Galton, Edgeworth, Frisch, and prospects for quantile regression in econometrics. *Journal of Econometrics* 95:347–74.

Koenker, R., and G. Bassett. 1978. Regression quantiles. *Econometrica* 46:33–50.

Marc-Aurele, F. J., L. C. D'Amico, and S. I. Bretschneider. 2000. Examining the 'extremes' of bureaucratic performance: A comparative examination of quantile regression and substantially weighted analytic techniques. Working Paper 00-01, Center for Technology and Information Policy, Maxwell School of Citizenship and Public Affairs, Syracuse University, Syracuse, NY.

Meier, K. J., and J. Gill. 2000. *What works: A new approach to program and policy analysis.* Boulder, CO: Westview Press.

Meier, K. J., J. Gill, and G. Waller. 1997. Optimal performance vs. risk aversion: An application of substantively weighted least squares. Paper presented at the Fourth National Public Management Conference, Atlanta, GA.

Meier, K. J., and L. R. Keiser. 1996. Public administration as a science of the artificial: A methodology for prescription. *Public Administration Review* 56 (5): 459–66.

Meier, K. J., R. D. Wrinkle, and J. L. Poliand. 1999. Equity and excellence in education: A substantively reweighted least squares analysis. *American Review of Public Administration* 29 (March): 5–18.

Overman, E. S., and K. J. Boyd. 1994. Best practice research and postbureaucratic reform. *Journal of Public Administration Research and Theory* 4 (1): 67–83.

Pandey, S., and S. I. Bretschneider. 1997. The impact of red tape's administrative delay on public organizations' interest in new information technology. *Journal of Public Administration Research and Theory* 7 (1): 113–30.

Ruggiero, J. 1996. Measuring technical efficiency in the public sector: An analysis of educational production. *Review of Economics and Statistics* 78 (3): 499–509.

———. 1998. A new approach for technical efficiency estimation in multiple output production. *European Journal of Operational Research* 111:369–80.

Ruggiero, J., W. Duncombe, and J. Miner. 1995. On the measurement and causes of technical inefficiency in local public services: With an application to public education. *Journal of Public Administration Research and Theory* 5:403–28.

Scott, W. R. 1998. *Organizations: Rational, natural, and open systems.* 4th ed. Upper Saddle River, NJ: Prentice Hall.

Simon, H. A. 1963. *The sciences of the artificial.* Cambridge, MA: MIT Press.

Vinod, H. D. 1968. Econometrics of joint production. *Econometrica* 36:322–36.

Waldo, D. 1980. *The enterprise of public administration.* Novato, CA: Chandler and Sharp.

Western, B. 1995. Concepts and suggestions for robust regression analysis. *American Journal of Political Science* 39:786–817.

Wu, Jiannan, and S. I. Bretschneider. 2001. Identifying the best practice: A Monte Carlo evaluation of quantile regression. Working Paper 01-1, Center for Technology and Information Policy, Maxwell School of Citizenship and Public Affairs, Syracuse University, Syracuse, NY.