Journal of Public Administration Research and Theory Advance Access published December 8, 2012

Subjective Organizational Performance and Measurement Error: Common Source Bias and Spurious Relationships

Kenneth J. Meier*[*,†], Laurence J. O'Toole Jr.[‡]
**Texas A&M University; †Cardiff University; ‡The University of Georgia*

ABSTRACT

In any design science such as public management, the importance of measurement is central to both scholarship and practice. Research built on measures that are not valid or reliable can generate misleading results and produces little value in the world of practice. This article applies measurement theory to administrators' self-perceptions of organizational performance, measures commonly used in the literature. Such measures can be prone to common source bias whereby spurious results are highly likely. This article uses measurement theory to show why common source bias can be a problem and then introduces an empirical test for bias based on a data set that includes both perceptual measures of performance and archival measures. The empirical test shows that spurious results are not only common but that they might include as many as 50% of all statistical tests. The article further examines the specific types of questions and measures most likely to generate measurement bias and provides guidelines for scholars conducting research on public organization performance.

The questions, are public programs effective? and what is the role of public management in this regard? have occupied scholars in both the United States (Moynihan 2008) and numerous other countries (Pollitt and Bouckaert 2000). Missing in the rush to performance appraisal and performance management is any effort to tie empirical efforts to the extensive literature on measurement theory (Ghiselli, Campbell, and Zedeck 1981; Hand 2004; Lance et al. 2010; Shultz 2005). This article uses measurement theory to assess one potential problem in measuring organizational performance. It considers both subjective and data-based measures; these can be measures

Preliminary versions of this article were presented at the Annual Meeting of the American Political Science Association, September 1–5, 2010, Washington, DC, Copyright American Political Science Association, where it received the Herbert Kaufman Award as the best article presented on public administration; and the Fall 2011 meeting of the Association of Public Policy Analysis and Management, November 4, Washington, DC. We would like to thank the anonymous reviewers and Ling Zhu for perceptive comments on an earlier draft of this article. Address correspondence to the author at kenneth-j-meier@pols.tamu.edu.

doi:10.1093/jopart/mus057

internal to the organization and those imposed by external stakeholders. Using several hundred public organizations based on an original survey conducted in 2009, the empirical illustration shows that perceptual measures of organizational performance by organization members can and frequently do lead to spurious results in scholarly research. To the extent that public management scholarship produces flawed research, it provides little guidance to practitioners seeking to improve performance.

The basic problem this article addresses is measurement bias with special attention to what is alternatively termed "common source bias" or "common method variance," which is in turn a special case of correlated measurement error.[1] First, the article defines common source bias and illustrates how some subjective measures of organizational performance can be contaminated by common source bias. Second, we move to the theoretical level with a discussion of measurement theory and the problem that common source bias creates for finding reliable and valid results. Third, we undertake an empirical analysis of some 84 public management survey items to determine what types of survey questions from a common source are more susceptible to bias; the results show that many but not all survey items have such biases. Although this study only addresses empirically problems with administrators' subjective measures of performance, the more general objective of this work is to illustrate how to assess the overall dangers of measurement bias and, if possible, to provide guidelines for scholars in terms of how and under what conditions the challenges of common source bias might be problematic.

ORGANIZATIONAL PERFORMANCE AND COMMON SOURCE BIAS

The growing field of public management has focused on the question of when and under what conditions management affects organizational performance. Progress has been slow because what "performance" means in the public sector is open to considerable debate (see Boyne 2003; Moynihan 2008). An entire subfield of performance measurement discusses this question, both at the organizational level and at the individual level. For many public organizations, established performance indicators do not exist—despite efforts by the federal government through such vehicles as the Government Performance and Results Act and the Program Assessment Rating Tool (PART scores). For other organizations such as schools, performance indicators are subject to substantial controversy about the reliability and validity of the measures. Those scholars working with an array of organizations that perform different functions face dauntingly complex problems in the measurement of performance because one somehow has to compare, say, the processing of social security checks with the military's ability to fight a war successfully.

One inviting apparent solution to this problem is to rely on perceptual measures of performance by citizens, by knowledgeable experts, or by the managers involved. Perceptual measures, particularly by the managers associated with the organization or program, are an integral part of the Federal Human Capital Survey, the National

1 The problem is also called "monomethod bias." We shall use common source bias as the term since it is most common in the public management literature; the generic management literature generally uses "common method variance."

Administrative Studies Project, the Merit Principles Survey, and the American State Administrators' Project, all of which have spawned a substantial literature linking public management to performance.[2] A general question like "how would you rate the performance of your workgroup compared to others in ——?" seeks to create a performance measure that can be used across a variety of different organizations with different purposes.[3] The payoff to such an effort means that theories can be tested on a wider range of organizations and thus are likely to be more generalizable.

The benefits of using administrators' self-assessments of performance, however, need to be weighed against the costs. If an analysis uses a survey of managers both to ascertain the level of organizational performance and to collect information about management practices, common source bias needs to be considered (see Campbell and Fiske 1959). Common source bias exists when some of the common variation between two concepts is a function of the common measurement and/or source used to gather the data. It can be defined as measurement error that is not random or "systematic error variance shared among variables measured with and introduced as a function of the same method and/or source" (Richardson, Simmering, and Sturman 2009, 763).

Why might we think that common source bias is a problem with subjective measures of organizational performance? Numerous studies show that individuals' self-assessments of themselves or their organizations—including those comparing public managers' perceptions of organizational performance to external objective measures of performance—find that individuals consistently overestimate the level of performance in the organization (see Bazerman 2005; Hastie and Dawes 2001; Kahneman and Tversky 2000; Meier and O'Toole forthcoming) and that this overestimation is not related to more difficult tasks or the availability of resources. More problematic than the overestimation is that the error contained in the assessments of organizational performance can be correlated with measures of management, even if these management measures are uncorrelated with the objective performance measure. Correlated errors—that is, situations in which the common measurement contains a source of error that shows up in both the independent variable (management) and the dependent variable (performance)—can contribute to spurious results (Doty and Glick 1998; Podsakoff et al. 2003, 879). Managers may respond to surveys in ways that reflect favorably on themselves in terms of both organizational performance and the adoption of the most current managerial practices.[4] Although common source

2 Some of these data sets also include bureaucratic respondents beyond management as well as agency managers. Nothing we have to say here applies to either citizen surveys of government performance or the use of outside experts to assess performance. Citizen surveys in particular are a valuable tool for performance assessment and at times contain valuable information that can be gathered in no other way. For an excellent assessment of citizen surveys versus putatively objective measures of performance, see Schachter (2010).

3 The number of studies using subjective measures from the four mentioned surveys is substantial. A Google Scholar search (July 2, 2012) using both the database title and the word performance generated the following number of studies: Merit Principles Survey—158, Federal Human Capital Survey—209, National Administrative Studies Project—130, and American State Administrators' Project—71. Not all of these studies actually use the subjective performance measures, but these numbers do indicate extensive use.

4 Correlated errors are positively correlated or negatively correlated. This means that the impact on the significance of the relationship (see the discussion that follows) is different from the impact on the size of the regression coefficient. The direction of bias on the regression slope will depend on the direction of the true relationship and the direction of the correlated error.

bias often generates false positives—the conclusion that a relationship exists when one does not—it can also generate false negatives—insignificant relationships when an actual relationship exists.

MEASUREMENT THEORY

Although the notion of common source bias has intuitive appeal, linking the problem to measurement theory will indicate both the general nature of the problem and also why it is difficult in advance to determine if common source bias will be a problem. Any theoretical concept (T) used in research should be distinguished from the indicator (X) that is used to operationalize it. Classical test theory (Lord and Novick 1968; as applied to management, see Conway and Lance 2010 or Lance et al. 2010)[5] expresses this with a simple equation:

$$X_i = \lambda_{T_i} T_i + e_i \tag{1}$$

where X_i is the realization of the ith person's characteristic or trait (T_i) and e_i represents nonsystematic measurement error.

No concept of any sophistication can be measured without error. The error is generated as the result of one or more of: poor conceptualization, an insensitive measurement instrument, respondent fatigue or inattention, or a variety of other factors. The degree of error, or actually the lack thereof, can be considered as directly linked to the validity of the measurement (Zeller and Carmines 1980).

We can extend equation (1) by adding measurement error that is generated by the measurement process itself where M_j represents the jth method.

$$X_{ij} = \lambda_{T_{ij}} T_i + \lambda_{M_{ij}} M_j + e_{T_{ij}} \tag{2}$$

Lance et al. (2010, 437–8) show that under traditional measurement assumptions—that T_i, M_j, and e_{ij} are uncorrelated with each other and X_{ij}, T_i, and M_j have unit variances (i.e., the λs are factor loadings or standardized regression coefficients)—the variance in observed scores ($\sigma^2_{X_{ij}}$) is composed of three independent parts—the individual's true score variability ($\lambda^2_{T_{ij}}$), the variance attributed to the method of measurement ($\lambda^2_{M_{ij}}$), and any nonsystematic error score variance ($\sigma^2_{e_{ij}}$) or

$$\sigma^2_{X_{ij}} = \lambda^2_{T_{ij}} + \lambda^2_{M_{ij}} + \sigma^2_{e_{ij}} \tag{3}$$

Equation (3) clearly shows that both method variance and nonsystematic error score variance contribute to the variance in observed scores and in the process affect the construct validity of the measure. To the extent that the two sources of error variance increase relative to the true variance, the validity of the measure declines. If one can make a simple assumption, measurement error by itself does not create an intractable problem. If one can assume that the measurement

5 This discussion follows closely what is a parsimonious and elegant presentation by Lance et al. (2010) and similar presentations in Conway and Lance (2010) and Lance et al. (2009).

error in two concepts *Tx* (independent variable) and *Ty* (dependent variable) are uncorrelated with each other, then the measurement error will simply attenuate the relationship between two variables, and the analyst will be more conservative in rejecting a null hypothesis.[6] With common source bias, however, one cannot assume that the errors are uncorrelated, simply because the data-generating process is producing the same errors in both measures. Both of these situations and the related problems can be illustrated by examining the correlation coefficient between two concepts *Tx* and *Ty* that are measured with the same method. The observed correlation can be represented by

$$r_{xy} = \lambda_{XTx}\lambda_{YTy}\rho_{TxTy} + \lambda_{XM}\lambda_{YM} \quad (4)$$

where λ_{XTx} is the reliability index for X and λ_{YTy} the index for Y, ρ_{TxTy} is the correlation between the true scores of X and Y, and λ_{XM} and λ_{YM} are the effects of the common method M on X and Y.

When two methods are uncorrelated (i.e., no common source variation), the second term on the right hand side becomes zero and the equation represents the common demonstration that random measurement error will result in an observed correlation that is smaller than the true correlation. When X and Y have correlated errors (including from common measurement), the second term of the model becomes positive and, thus, inflates the observed correlation between X and Y relative to the true correlation.[7]

The most important aspect of equation (4) is that it shows common source bias can have impacts that either inflate relationships (i.e., make false positives more likely) or deflate relationships (make false negatives more likely). Common source bias is a form of measurement error; measurement error generally affects the reliability of the two measures and contributes to the first term in the model by reducing it in magnitude and thus reducing the size of the observed correlation. At the same time, the common source bias contributes to the second term in the model and biases the correlation upward. Whether common method variance inflates correlations or deflates correlations, in practice, depends on the how much common source bias contributes to the second term in the model (inflation) versus measurement error in the first term (deflation). So unlike random measurement error that attenuates relationships and thus makes hypothesis tests more conservative, common source bias in theory can generate false positives, produce false negatives, or have no impact whatsoever (for a formal demonstration in regard to regression see Siemsen,

6 Following the literature, we have assumed that T_i, and M_j are independent. It might be the case that M_j is dependent on T_i if managers who do very poorly overestimate more than managers who only slightly underperform. In that case, equation (3) would need a nonzero covariance term, which would complicate the breakdown in equation (4).

7 Even with different measurement methods, there is still a possibility of common source bias if there is covariance among the method bias across measurement techniques (see Brannick et al. 2010, 411). We have also simplified this discussion by focusing on relationships in the same direction (e.g., all positive). Depending on whether the bias affects both questions the same way or pushes in opposite directions (social desirability versus undesirability), spurious relationships can be positive or negative.

Roth, and Oliveira 2010, 461).[8] This ambiguity in terms of result is the pernicious aspect of common source measurement error; it can create positive results where none exists and it can generate null results when in fact a relationship does exist.[9]

This ambiguity is complicated because the degree of common source bias in a set of measures can vary from item to item (see Richardson, Simmering, and Sturman 2009, 764 on what is called the congeneric perspective of common method variance; see also empirical analyses by Cote and Buckley [1987], Rafferty and Griffin [2004], and Williams and Anderson [1994]). In a survey, for example, not all questions answered by a single respondent will necessarily contain the same amount of common source bias; if that were the case, one could easily create a marker variable that measured the bias and remove the bias from the other measures. For example, assume a respondent who is concerned about the social desirability of his or her responses and who will color the responses to questions accordingly. Because not all questions have the same element of social desirability or because this element of social desirability varies from question to question, there will be some relationships that can be assessed that will not be threatened by the bias (if the questions lack social desirability) and some questions where spurious results are likely. Spector (2006, 224) notes that in addition to social desirability, negative affectivity (a predisposition to experience negative emotions that lead to a negative view of the world; see also Williams and Anderson 1994), and acquiescence (the willingness of respondents to agree with questions on a survey) can generate common source bias. Podsakoff et al. (2003) identify common rater effects, item characteristic effects, item context effects, and measurement context effects as potentially generating common source bias. Each of these general categories has a series of subcategories that also generate bias. In principle, any concept that contributes measurement error to both the independent and dependent variable can generate common source bias.

Even the name "common source bias" is somewhat misleading, since the problem of correlated errors can be generated in a variety of ways (see Weisberg 2005).[10] One source of the bias might be the questions themselves, because questions might have more than one stimulus; that is, they tap performance but also the respondent's self-interest in looking competent. In this case, there might be more than one source of contamination, since some questions might contain both social desirability and also a concern with consistency, and some questions might have one but not the

8 Lance et al. (2010) find that on average the inflation impact of common method bias is countered by the deflation impact of unreliable measures. One should not interpret this finding as a reason to ignore common measurement bias for three reasons. First, common method bias varies a great deal by question and by study and there still will be many cases where it either inflates or deflates relationships. Second, the extent of common source bias appears to vary substantially across substantive area/discipline (e.g., management, information technology, etc., see Cote and Buckley 1987; Malhotra, Kim, and Patil 2006). Cote and Buckley (1987) find approximately one-fourth of the variance in the measures they studied to be the result of correlated measurement error. Third, arguing that one's findings are valid or more valid because they contain two sources of error rather than one puts a scholar in an unenviable position.

9 This article only deals with common source bias that affects the dependent variable and an independent variable. It does not deal with the problems created when the bias affects only the independent variables.

10 Similarly, common method variation is also somewhat misleading, since the source may not be the method—that is, a survey—but rather any common invalidity source such as faulty measures and the timing of the measures (e.g., asking questions about national security after 9/11).

other. If the source of the bias is the questions, then using different individuals as respondents for the independent and the dependent variables might not resolve the problem. The second source of bias is the respondent who will interpret the question in terms of his or her own position, set of cognitive biases, and other individual factors. For respondents the biases might also be multidimensional—that is, a desire to demonstrate consistency with widely accepted management practices and an interest in avoiding negative stigma. The actual common source bias can be a combination of both question impacts and respondent impacts (with both affected by context), thus, making the determination of its cause extremely complex.

Because measurement theory tells us that common source bias can be a serious problem but cannot provide an unambiguous answer for the type of problem (direction and significance), scholars need to rely on empirical tests involving the types of questions that are used in the specific substantive area (see Cote and Buckley 1987; Malhotra, Kim, and Patil 2006, 1866). The next section of the article takes on this task by introducing a data set in public management designed in part to assess common source bias for one specific measure of performance, providing a method for determining the degree of bias, and then empirically examining 84 survey items about public management to determine the types of questions that might be problematic.

TESTING FOR COMMON SOURCE BIAS

The generic management literature and the psychological literature have long been concerned with common source bias and possible solutions (Avolio, Yammarino, and Bass 1991; Brannick et al. 2010; Cote and Buckley 1987; Doty and Glick 1998; Kemery and Dunlap 1986; Lance et al. 2010; Malhotra, Kim, and Patil 2006; Podsakoff and Organ 1986; Podsakoff et al. 2003). As a result, these fields have often designed studies to take common source bias into account and in the process have developed some sophisticated detection strategies and some suggested solutions (although consensus on solutions is lacking; see below). In public management, data sets with both subjective and objective measures of performance are rare, and an ideal data set to compare various detection techniques and solutions is nonexistent. This study takes a first step in that direction to illustrate the potential problems in the field of public management.

The Data Set

Data for this analysis come from two sources, the Academic Excellence Indicator System of the Texas Education Agency and an original survey of Texas school superintendents. The state of Texas operates an elaborate accountability system for Texas schools that collects information on a variety of performance indicators as well as student and financial data. All data other than administrators' perceptions of performance and management style are taken from this source for the academic year 2008–09, the year immediately preceding the gathering of the perceptual data. For the perceptual data, school superintendents were surveyed via a four-wave mail survey between July 2009 and November 2009. The response rate for public school superintendents was 58%; the survey also included some charter school superintendents for

a total of 642 respondents.[11] The sample is representative of the entire population with no apparent biases. There were no statistically significant differences between the respondents and the nonrespondents in terms of standardized test scores, college-bound scores, the racial and income distribution of students, and instructional expenditures. Respondents did receive $480 less in per pupil revenue than nonrespondents even though educational expenditures were similar.

The Texas education system uses a standardized test called the Texas Assessment of Knowledge and Skills (TAKS), a criterion-based test that is given in elementary grades and as an exit exam at the high school level. TAKS is a high stakes test, and students must pass this test to receive a regular diploma from a Texas high school. The specific performance indicator is the percentage of students who pass all TAKS tests (math, reading, writing, social sciences, etc.) using what is termed the accountability sample.[12] This is the official criterion used to evaluate schools and school districts in the state and forms a major part of the annual grades that the state assigns to schools. The state of Texas also collects other outcome indicators such as attendance and performance on college aptitude tests.

For perceptual measures of performance, superintendents were asked "compared to similar districts, my assessment of our — performance is" on a 5-point scale using the categories "excellent," "above average," "average," "below average," and "inadequate." Three different stimulus items were used: "TAKS performance," "college-bound performance," and "overall quality of education in the district." These three items were included on the survey to obtain questions that reflect the range of performance indicators that might be tapped via perceptual data in public management surveys. TAKS scores are highly specific and widely disseminated; managerial perceptions of them should generate the maximum level of congruence with the objective performance data (see Doty and Glick 1998). The TAKS scores are measured with little or no error as long as one interprets them as measuring test performance rather than a some general indicator of educational attainment. For college-bound performance, there is no single indicator, and thus the reference point will be more vague and more likely to be subject to some perceptual biases and more measurement error. Finally, the overall quality of education is very vague and could refer to performance on many dimensions. This final indicator is perhaps closest to those general indicators contained in other data sets where respondents are asked to assess how well their organization or their work group is performing. We might expect this measure to have the greatest measurement error.

11 The survey contains responses from 595 public superintendents and 47 charter school superintendents. The response rate for charter schools is difficult to determine since each charter school is treated by the state as a separate district; but if two or more schools are operated by the same organization, they would have the same superintendent. Inclusion of the charter school superintendents had no impact on any of the results presented in this article.

12 Excluded from the accountability sample are those students exempted as the result of some special education classifications, as the result of limited English abilities, as the result of mobility (recent arrival to the school) and absences. The rules for exemptions have been tightened in recent years, and the state collects test data on all students, whether in the accountability sample or not (e.g., several elementary exams are given in Spanish). The results throughout the entire state were posted on the TEA Web site in December of 2009, with results made available to district superintendents considerably earlier.

To compare the college-bound indicator of performance, we use the Texas Education Agency's definition of a "college ready" student: one who scores above 1,110 on the SAT or its ACT equivalent. This score is equivalent to the top 20% of scores nationwide. Although this is an official government-sanctioned definition, it is not a particularly robust measure of college-level performance, since college board scores do not correlate that strongly with student success in college. As a result, we will use the college assessments as a check on the TAKS analysis and as a guide to whether or not the problem becomes more severe as the dependent variable becomes less precise (and thus is more like the measures used in the literature). We will not propose to create a measure of the "overall quality of education," a daunting task that would require a great many subjective and controversial determinations. We note, however, that this measure is less precise than the other two (in terms of specific referents) and, thus, is closest to the measures actually used in the public management literature. To the degree that relatively precise measures of perceived performance have problems, this less precise measure should suffer even greater ill effects. We will illustrate these more severe effects with an analysis that compares the perceptual overall quality measure to both TAKS scores and college board scores.

Method of Analysis

Before proceeding to the assessment of common source bias, a brief look at the responses to the subjective performance item in table 1 will illustrate why an analyst should be concerned with such measures. Fully two-thirds of the respondents rate their organizations above average on the TAKS; the ratio of above average to below average responses is 9.9 to 1. The college preparation indicator has slightly less overestimation with 46.4% above average and a ratio of 3.4 to 1 above to below average. The overall quality of education question, the most vague of the items, has the greatest bias with fully three-fourths of respondents rating their organization above average and the ratio of above to below average is a stunning 25 to 1. Clearly, top managers

Table 1
Upward Bias in Self-Assessments of Performance

Compared to similar districts, my assessment of our ____ performance is

	TAKS	College Bound	Overall Quality
Excellent	21.4%	12.7%	25.2%
Above average	46.0%	33.7%	50.3%
Average	25.5%	40.0%	21.5%
Below average	6.6%	13.2%	2.8%
Inadequate	0.2%	0.5%	0.2%
N	635	615	636
Mean	3.83	3.45	3.97
SD	0.85	0.89	0.77

Intercorrelations	TAKS	Quality
College bound	.50	.63
Overall quality	.70	

are overestimating the performance of their own organization (recall TAKS scores for respondents were no different from those of nonrespondents). If these misperceptions are correlated with any response biases (i.e., errors) in other variables, then spurious correlations are quite likely.

The questions ask respondents to compare their organizations to similar organizations because that is the convention for subjective measures of performance in the literature. This raises the question as to whether the differences between the subjective measures and the archival measures are actually the result of respondents adjusting their own assessments based on the resources that they have and the difficulty of the tasks that they face. If managers made these adjustments, then the difference between the subjective and the archival measure would be correlated with measures of resources or task difficulty. In fact, these differences are rarely correlated with any of several measures of resources or task difficulty (see Meier and O'Toole forthcoming), and at times these correlations are in the wrong direction (i.e., respondents give themselves more credit when they have more resources or face an easier task). We also determine whether the comparison of similar districts to all districts is appropriate in two other ways. First, we treated this as a selection bias problem and ran a series of propensity score matching efforts (see Appendix A). In all cases the results showed no adjustment to the characteristics of the organization. Second, we ran a randomized experiment where school principals were asked to rate the performance of their school to either "similar schools" or "other schools." The questions produced identical results. (Details are reported in Appendix A.) These three tests provide strong evidence that the results in table 1 reflect a valid comparison.

One way to illustrate common source bias is to use an independent variable measured by the same process and respondent as the perceptual performance measure (i.e., from the same survey) and run regressions with perceptual performance indicator and the archival performance indicator—to determine if any differences result. We do this by creating what is called an educational production function whereby educational outputs are a function of resources and constraints. Assessing common source bias in multiple regressions is important because the impact on simple correlations might be countered in practice by the control variables, especially in complex specifications (see Siemsen, Roth, and Oliveira 2010, 471). To illustrate this process in table 2, we use multiple regression to estimate performance—both subjective and archival—with a production function specified with set of controls along with a survey item as the independent variable. We take the superintendent's response to the statement "Our district is always among the first to adopt new ideas and practices" on a 4-point agree-disagree scale as the illustrative independent variable. This item might be considered a measure of the degree to which a district operates with a "prospecting" management strategy (Miles and Snow 1978). The first set of control variables in the equation are the type of students—in this case the percentage of black, Latino, and low-income students. These students generally do less well on standardized tests and should generate negative relationships. The second set of control variables attempt to tap the resources applied to education and include average teacher salaries, class size, number of years of teacher experience, and teacher turnover. Class size and teacher turnover should be negatively related to performance and teacher salaries should be positively related. Teacher experience could generate either sign in that teachers need some experience to become good

Table 2
Common Source Bias: Management Style and Subjective and Actual TAKS Performance

	Actual TAKS		Subjective Assessment	
	Slope	*t*-Score	Slope	*t*-Score
Prospector style	0.0505	0.94	0.1748	3.69**
% Black students	−0.0577	1.66*	−0.0008	0.25
% Latino students	−0.0796	3.51**	−0.0020	0.99
% Low income	−0.2495	8.14**	−0.0117	4.35**
Teacher salary K	0.3726	3.11**	0.0160	1.53
Class size	−0.0299	0.20	0.0177	1.36
Teacher experience	−0.2168	1.40	−0.0193	1.42
Teacher turnover	−0.2951	7.06**	−0.0091	2.49**
SE	8.90		0.78	
F	69.31		14.83	
R^2	.47		.16	
N	629		626	

*$p < .10$ two-tailed test; **$p < .05$ two-tailed test.

at the job but also that younger teachers are likely to have had more rigorous training as the result of the increase in standards for teacher education. Such controls are commonly used in examining education outcomes. The results for both subjective and archival dependent variables (TAKS performance) are displayed in table 2.

The test of potential common source bias will be a comparison of the *t*-scores for the appropriate coefficients. Table 2 indicates a pattern that would result from common source bias. Here, we leave aside the point that the standard educational production function explains much more of the variance in the archival measure of TAKS performance than in the perceptual measure.[13] We focus for present purposes only on the issue of common source bias. The management measure of prospecting is positively and strongly correlated with the perceptual measure of performance but unrelated to the archival measure of TAKS scores. In short, *the survey measure generates a false positive result*. It is important to stress that the subjective performance question asks specifically about performance on the TAKS and the archival measure is the actual TAKS score. Although there might be some random measurement error, the official TAKS score is much closer to the "true" TAKS score than most other measures of performance used in the literature. The relationships in the archival regression, as a result, are unlikely to be attenuated owing to measurement error in the dependent variable.

To determine how extensive the problem of spurious results might be, a similar set of regressions were run for all 84 survey items in the 2009 management survey. These items cover a wide range of concepts that are frequently used in public management research. Rather than producing 84 tables with full results for all measures in the survey similar to table 2, we extract two bits of information from each regression: (1) the

13 We replicated these results as well as those in the remaining tables using archival performance measures for the year after the survey to determine if that might affect the results. The results were nearly identical to those reported.

Table 3
Measuring Potential Spurious Results for Perceived Organizational Performance: Summary Results from Regressions of Management Items

	Dependent Variables	
	TAKS tests	College-bound performance
Average *t*-score bias	1.42	2.09
SD of bias	1.05	1.52
Minimum bias	0.00	0.03
Maximum bias	4.11	7.20
False positives	20	31
False negatives	3	10
% Spurious results	27.4	50.0[a]

[a]One case has significant relationships in different directions.

absolute value of the difference between the *t*-score for the perceptual measure (3.69 in table 2) and the *t*-score for the archival measure (0.94 in table 2) as the measure of the degree of the potential common source bias (2.75 in this case; see Appendix B for all questions and bias estimates) and (2) using the .05 level of confidence whether or not the subjective equation generates a false positive or a false negative relative to the objective performance equation.[14] We count it as a false positive anytime the relationship for the subjective performance is statistically significant and the one for the archival measure is not. A false negative has an insignificant relationship for the subjective performance indicator and a significant one for the archival measure. For the *t*-score bias measure, given that a *t*-score of approximately 2.0 is associated with the .05 level of statistical significance, we take differences larger than 2.0 as clearly problematic.[15] We have run these two sets of regressions for each of the 84 variables in the survey and then grouped the survey questions by characteristics (see below) to determine if certain types of questions are more susceptible to common source bias than others.[16]

Table 3 shows the overall results for both the more specific dependent variable (performance on the TAKS) and the more general one (college bound). Even with the highly specific TAKS measure, the average *t*-score difference to indicate possible bias is 1.42, with a range of 0–4.11. Using our rough indicator for false positives, we found 20 cases of false positives among the 84 regressions. False negatives

14 Each of these items is only illustrative, since we do not have a good measure of how much random measurement error is affecting either of the performance indicators.

15 In essence this means if the actual relationship were exactly zero, common source bias would still produce a significant finding.

16 An alternative way to estimate the potential common source bias is to simply put the objective measure of performance into the equation in table 2 that predicts the subjective performance measure. Because the actual performance will control for how well the organization is doing, any remaining unique relationship between the survey item and perceptual performance will be generated by the correlated errors. The bias estimates using this approach were very similar to those reported in this article, with means of 1.87 for TAKS, 1.85 for college bound, and 2.83 for the general measure. The limitation of this approach is that it does not provide a specific count of false negatives and false positives. We would like to thank Ling Zhu for suggesting this alternative estimator.

were more rare; in only three cases were archival measures significant but perceptual measures not.[17]

The results for the college-bound measure were even more troubling. Because the college-bound measure was less specific, we expected that perceptions would be less constrained and thus be able to generate more common source bias problems. The average *t*-score difference for common source bias is 2.09, with a range of 0.03–7.20. Because the college-bound variable should have substantially more random error than the TAKS measure, we might expect fewer false positives (see equation 4), but that is not the case. The analysis found 31 false positives and 10 false negatives. In addition, one case (the survey item on environmental stability) generated a significant positive relationship for the perceptual measure and a significant negative relationship for the archival measure. This totals 42 (50% of the total cases) spurious results.[18] This should be considered a low estimate of potential spurious results, in that the number will increase with sample size, and in some cases the relationship for the perceptual measure carried a sign different from that in the comparison equation. The finding that should be stressed is that *using the .05 level of confidence for this question, the probability of a correct decision about a relationship is no better than 50-50 when one employs this perceptual measure of performance*. Given that even the college board question is far more specific than the very general types of questions used in most analysis, the problems in other data sets could well be far greater. It is important to reiterate that the contribution of common source bias to false negatives or false positives is theoretically ambiguous (see equation 4). The results in practice here—that show a large number of false positives and some false negatives—means that the random error that would generate false negatives does not necessarily compensate for the correlated error generating the false positives in this case.

Given these problematic results, one could ask if the correct method would be simply to ban perceptual measures by administrators as dependent variables tapping performance. Although these results suggest that scholars should always be skeptical of results that use administrative survey responses on both sides of the equation, there might be cases for which survey questions tap behaviors or attitudes that do not share common source bias or correlated error with the administrators' perception of organizational performance. As Podsakoff et al. (2003) contend in their extensive discussion, common source bias can result from problems linked to respondents, survey items, context, and source. This point suggests an examination of the results for the individual questions to determine if some generalizations can be made about when and what types of questions will be less likely to share common source bias.

Rather than discussing 84 different questions, it would be more helpful to group the questions by characteristics and determine if some such characteristics generate more problems of bias. This process, however, will only reveal the average potential bias in a set of questions; any individual question might have more or less bias, depending

17 Another modeling strategy might be to adjust the metric of the archival measure to match the mean and standard deviation of the subjective measure and then directly compare the regression coefficient. This strategy is problematic in that common source variation is biasing one of the sets of coefficients making the comparisons problematic.

18 Although measurement error in the SAT/ACT measure can attenuate the relationships in the archival set of equations and generate more false positives, it should not generate more false negatives.

on the error-generating process. Eighteen aspects of the individual questions were identified inductively from examination of the survey items and the existing literature on common source bias from other disciplines. Any question can have more than one characteristic (e.g., it can ask about management innovation and it can ask about the environment), so the coding is not mutually exclusive. They were coded as follows:

1. *Time*—does the question ask the manager about how much time the manager does *X*?
   For example, how frequently do you meet with local business leaders?
2. *Quality assessment*—does the question require an assessment of quality, of others etc.?
   For example, how would you rate the quality of teachers in your district?
3. *Environment*—the question asks about some aspect of the environment of the organization.
   For example, my district's environment—the political, social, and economic factors—is relatively stable.
4. *Environment general*—the question is about the environment in general.
   For example, see question immediately above.
5. *Environmental support*—the question asks for an assessment of environmental support.
   For example, how would you rate community support in your district?
6. *Environment the people*—the question calls for an assessment of clientele characteristics or behavior.
   For example, in general citizens and other people in the communities served by my school district are active in civic and community affairs.
7. *Internal management*—the question asks about internal management of the organization
   For example, I give my principals a great deal of discretion in making decisions.
8. *Networking*—the question relates to the managerial networking in the environment.
   For example, see item 1 above on time.
9. *Exploiting*—the question deals with efforts to exploit environmental opportunities.
   For example, we continually search for new opportunities to provide services to our community.
10. *Buffering*—the question deals with buffering the environment.
    For example, I strive to control those factors outside the school district that could have an effect on my organization.
11. *Performance appraisal*—the question deals with the use of performance management.
    For example, I use performance data to make personnel decisions.
12. *Strategy*—the question asks about general management strategy.
    For example, our district is always among the first to adopt new ideas and practices.
13. *Prospecting strategy*—the question relates to a strategy of prospecting (innovation).
    For example, see example for category 12.
14. *Defending strategy*—the question relates to a strategy of defending.
    For example, our district concentrates on what we already know how to do.
15. *Reacting strategy*—the question relates to a reactor management strategy.
    For example, what we do is greatly influenced by the rules and regulations of the Texas Education Agency.

16. *Diversity*—the questions ask about diversity or diversity management.
For example, in my district, employees generally value ethnic and cultural differences.

17. *Goals*—the question asks about organizational goals.
For example, what is the most important problem facing your district? Please rank order your choices.

18. *Observable*—the question asks about observable behavior.
For example, how frequently do you meet with City/County government officials?

Because the characteristics are not mutually exclusive, we opted to assess all 18 characteristics simultaneously by using these characteristics as independent variables in a regression with the dependent variable our *t*-score measure of bias. In this analysis, each of the 84 observations is a survey item (see the Appendix B for the complete list) that is scored with dummy variables for all 18 characteristics as independent variables (1 if the item has that characteristic, 0 if not). Our hypothesis for the analysis is that measures that ask for more specificity and more observable activities will generate less bias than those that are more vague or more likely to have an element of social desirability. The results of this regression appear in table 4 for the more specific standardized test indicator of performance, TAKS.

One key to interpreting table 4 is the intercept. This can be interpreted as the average bias of a question with none of the characteristics coded (i.e., zero for all characteristics). The significant coefficient of 1.57 in table 4 means that the average question with none of these characteristics is likely to inflate the *t*-scores of this variable by 1.57, by itself almost enough to generate a false positive. So to interpret the slope for the "quality assessment" characteristic of 1.49 means that a question asking for a quality assessment will generate a *t*-score bias of 3.06 (or 1.57 + 1.49) *t*-score units, and this amount of bias is significantly more than a question without any of the measured characteristics. The table indicates that the most problematic questions for common source bias are those where the manager is asked to make a quality assessment, questions that ask about environmental buffering, questions that ask about diversity, and questions that ask about exploiting the environment. Negative signed coefficients indicate less of a common source bias threat (recall these need to be interpreted in light of the intercept). Questions less likely to contain common source bias include strategy questions about reacting, questions about environmental support, questions about managerial networking, and questions about observable behavior. In all cases, these are the average expectations of a type of question; *t*-scores for these coefficients indicate the impact will vary across individual questions (see Appendix B). That is, survey items having (a) characteristic(s) in common nevertheless might vary considerably in the amount of bias elicited in response. For example, questions dealing with time allocation have a coefficient of −0.46 indicating little bias, but the question asking about whether the superintendent initiated the most recent contact with the school board carries a strong warning about bias (coefficient = 2.25; see Appendix B).

Another qualification is in order. This indicator of bias means that spurious results are more likely; it does not guarantee that a given set of results is, in fact, spurious. As an illustration, the question asking the superintendent to evaluate the quality of teachers in the district has a large bias coefficient (by our measure of the difference in *t*-scores, +4.11; see Appendix B), but the question still produces strong and positive

Table 4
Spurious Regression Results by Type of Survey Question: TAKS Performance

Type of Question	Slope	*t*-Score
Intercept	**1.57**	**3.74**
Time allocation	−0.46	0.12
Quality assessment	1.49	3.11
Environment	0.22	0.21
Environment general	0.58	0.06
Environmental support	−0.90	1.65
Environment people	0.27	0.27
Internal management	−0.28	0.73
Managerial networking	−0.69	1.42
Exploiting	0.96	1.85
Buffering	1.26	1.41
Performance appraisal	0.40	1.02
Strategy	0.07	0.11
Prospecting strategy	0.08	0.10
Defending strategy	−0.31	0.51
Reacting strategy	−1.66	1.47
Diversity	1.15	2.63
Goals	−0.20	0.43
Observable behavior	−0.63	2.05
SE	0.87	
R^2	.46	
N	84	

Note: Predicted bias is the sum of the slope plus the intercept.

results in a production function using the archival measure (*t*-score = 5.57). In this case, the common source bias overestimates the strength of this relationship in the perceptual equation, but the relationship is not spurious. This illustration shows how difficult it is to deal with common source bias when the dependent variable is a perceptual measure. False positives are more likely, false negatives are less likely but still a possibility, and in some cases the bias works similarly to random measurement error.

Table 5 provides parallel results but for the college-bound performance measures, a more general question that is more likely to be subject to bias problems. If the general pattern of relationships looks similar to those in table 4, it is because the two measures of bias are correlated at .59, an expected finding given that one would assume this bias to be correlated across dependent variables measured the same way. The large intercept (2.95) indicates a great deal of bias in questions with none of the listed characteristics. This essentially means that a type of question needs at least a coefficient of nearly −1.00 to get the bias below the basic standard of 2.0. Questions that ask for a quality assessment, those that ask about the environment in general, and those that ask about diversity contain the largest estimated bias. The least bias is found in questions of reacting as a strategy, defending as a strategy, and the goals of the organization. Even in these questions, however, the threat of bias is real and needs to be considered.

The significantly worse results in terms of potential common source bias for the college-bound measure versus the TAKS measure indicates that *as assessments*

of performance get more general, the threat of common source bias can become more severe. Although we cannot be sure that this is always the case, it is the case in the best existing public organization data set for assessing common source bias. This finding, furthermore, comports with expectations. Analysts using perceptual measures, even perceptual measures as grounded in specificity as these measures, need to be concerned with common source bias and should present evidence regarding why common source bias does not generate spurious results in their findings. Those analysts using very general measures of performance face an even more difficult if not Herculean task.

To illustrate this potential problem with exceedingly general subjective assessments, we provide two estimates approximating the analyses displayed in tables 4 and 5. The optimal illustration would involve an objective measure of the overall quality of education and then a replication of the analysis for the subjective assessment measure. Lacking such an archival measure of overall quality, we compare the regressions on the subjective overall quality of education measure to the objective equations for TAKS and the college-bound performance discussed above, as the best archival approximations available that tap aspects of educational quality. In support of this test, we note that the strongest correlate of the subjective measure of overall education quality is actually the district's TAKS pass rate. Table 6 provides both the assessment of false positive and false negatives and also the regression results for the questions for both comparisons.

Table 5
Spurious Regression Results by Type of Survey Question: College-Bound Performance

Type of Question	Slope	*t*-Score
Intercept	**2.95**	**5.03**
Time allocation	−0.26	0.50
Quality assessment	2.30	3.43
Environment	−0.50	0.34
Environment general	1.77	1.39
Environmental support	−0.95	1.25
Environment people	0.26	0.19
Internal management	−0.78	1.44
Managerial networking	−0.08	0.12
Exploiting	−0.10	0.14
Buffering	−0.57	0.46
Performance appraisal	−0.13	0.23
Strategy	0.22	0.25
Prospecting strategy	−0.87	0.82
Defending strategy	−1.07	1.24
Reacting strategy	−1.75	1.75
Diversity	1.02	1.67
Goals	−1.10	1.73
Observable behavior	−0.84	1.95
SE	1.21	
R^2	.50	
N	84	

Note: Predicted bias is the sum of the slope plus the intercept.

The table clearly shows that the potential problems of bias are more severe with this more general measure than with the more specific measures such as the subjective assessment of the TAKS. Compared to the TAKS equation, the overall subjective assessment dependent variable has a mean *t*-score bias of 2.51 and generates 32 false positives and 2 false negatives (or 40.5% of the cases). Comparing to the college-bound equation generates even worse results, with a mean *t*-score bias of 3.04, 34 false positives, 8 false negatives, and 2 cases of relationships that are significant in the wrong direction (for 52.3% of the cases). For individual types of questions, the results indicate that bias is virtually guaranteed. By adding the intercept to the slope coefficient, one gets common source bias *t*-scores of greater than 5 for quality assessments, exploiting the environment, buffering the environment, and diversity. Although some of the negative scores that indicate less bias are large, they need to offset the huge intercept; only for questions involving managerial networking, reacting strategy, and perhaps goals is this likely. The individual slope coefficients for the college-bound comparison tell an even more frightening story. The intercept bias alone is a massive 5.64, when this is added to the additional large biases in quality assessments (+5.98), the general environment (+3.95), and people in the environment (+2.91), the overall potential for bias is almost assured. In fact, the college board regressions suggest that all types of questions are problematic with the sole exception of questions that tap the reactor strategy for management.

Although this test on the overall quality measure is less than ideal, given that we do not have an objective measure of overall quality, it is fair to point out that this question is the one that most closely resembles those included in data sets typically used in the public management literature. Examining this question from two different perspectives shows that bias is not just frequently present but is almost always present and consistently leads to spurious conclusions.

A PRACTICAL NOTE TO SCHOLARS

Given the problems of common source bias, what are the guidelines for researchers using survey assessments? We offer three proposals in order of how well they are likely to deal with the problem. *First and most obvious, avoid the use of administrators' self-perceptions of performance as a dependent variable.* This is the case *especially when the independent variables are also gathered by survey* (see Podsakoff et al. 2003, 887), but the perceptual biases in self-assessments of performance can be so large that correlated measurement error can be a problem no matter how the independent variables are measured. This advice is merely a specific application of the admonition to use multiple sources of data (Brannick et al. 2010, 414). The use of different respondents for the dependent and the independent variables only solves the problem if the source of the bias is the respondents, not the questions (see Podsakoff et al. 2003); and since there is no literature on how to determine this post hoc for a survey, this strategy likely only increases the costs (more respondents) for no definite gain. Avoiding administrative perceptions of performance is the best practice for avoiding spurious correlations as the result of common source bias.

Second, if the researcher decides to use administrative perceptions of performance and independent variables gathered via the same survey, the focus should be

Table 6
Spurious Regression Results by Type of Survey Question: Overall Quality of Education

	Compared to TAKS		Compared to College	
Type of Question	Slope	*t*-Score	Slope	*t*-Score
Intercept	**3.25**	**5.29**	**5.64**	**6.12**
Time allocation	0.60	1.11	−0.07	0.09
Quality assessment	3.96	5.64	5.98	4.83
Environment	1.04	0.68	−3.03	1.33
Environment general	0.10	0.07	3.95	1.97
Environmental support	−0.46	0.58	−0.51	0.43
Environment people	0.60	0.41	2.91	1.34
Internal management	−1.19	2.09	−2.85	3.35
Managerial networking	−3.16	4.47	−2.48	2.35
Exploiting	1.85	2.43	−0.40	0.33
Buffering	2.45	1.88	0.27	0.14
Performance appraisal	−0.42	0.73	−0.39	0.46
Strategy	0.28	0.29	0.94	0.67
Prospecting strategy	0.32	0.28	−0.82	0.49
Defending strategy	−0.71	0.79	−1.15	0.85
Reacting strategy	−4.54	2.73	−6.42	2.58
Diversity	2.12	3.30	1.74	1.81
Goals	−2.25	3.38	−2.97	2.98
Observable behavior	−1.12	2.50	−1.67	2.49
SE	1.27		1.90	
R^2	.68		.63	
N	84		84	
Mean	2.51		3.04	
False positives	32		34	
False negatives	2		8	
Wrong direction	0		2	
% Spurious results	40.5		52.3	

Note: Predicted bias is the sum of the slope plus the intercept.

on the dependent variable. This article has demonstrated that the potential for common source bias is less when the performance question is tightly focused on a specific indicator of performances (TAKS scores versus general quality of education). With the general and vague measures of performance, the level of bias is so large that the researcher should have no confidence whatsoever in the findings.

Third, even after creating as specific a measure of performance as possible, the researcher needs to focus the analysis on independent variables that are also more specific and less likely to generate spurious results (see Brannick et al. 2010, 414). This article has found that questions that ask about how managers spend their time, questions dealing with observable behavior, questions about environmental support, questions about a reactive strategy, and questions about managing in the network seem to be less affected by common source bias than other questions. Even within these categories, however, individual questions can be significantly biased; hence we have included all our estimates of bias in Appendix B.

Although there are statistical techniques to rid a data set of common source bias, they all rely on the shared variance of the survey items (see Podsakoff et al. 2003, 893 for a full discussion of various methods). This means that the statistical solutions can remove both the correlated error and also the actual relationship between two variables in the process. This is the methodological equivalent of dealing with lung cancer by simply removing the entire lung via surgery. Null results are virtually guaranteed, even in the presence of strong conceptual relationships. Richardson, Simmering, and Sturman (2009, 796) conclude their assessment of these techniques that "based on our results, we cannot recommend any post hoc CMV [Common Method Variance] technique as a means for correcting CMV's potential effects in a given data base" (see also Conway and Lance 2010, 331 who term these techniques "ineffective" and "untested" and Lance et al. 2010). Spector (2006, 229) concisely expressed what needs to be done: "The trick is to minimize possible biases through the design of measures or to link self-reports to other measures that would provide confirmation about an observed relationship between variables." Conway and Lance (2010, 329) expand on this idea by focusing on the need to establish construct validity and list several sources of evidence that should be used in this regard (see also Brannick et al. 2010, 414). A promising alternative is proposed by Oberfield (2012) who advocates measuring the dependent variable at two different time periods and examining the change in the dependent variable.[19]

CONCLUSION

Public management deals with important and contentious issues of theory as well as practice. Given the high stakes involved, researchers need to attend to measurement theory as they design and execute systematic empirical investigations. Relatively little consideration has been given thus far to that subject among public management scholars; we hope that this study might stimulate further work that draws from and perhaps contributes to measurement theory as well as management theory.

Drawing from measurement theory, we have used a large data set to explore the key issue of common source bias, for the situation in which managers provide data both on their own actions and setting and also on their organization's performance. (It would also be useful to analyze this issue when the common respondents are other employees of public organizations, when they are citizens, and when they are knowledgeable experts.) Common source bias is also possible with archival measures, and probing the issue for these measures would also be beneficial.

The findings for the several hundred organizations included in our study clearly show that common source bias is a serious problem when researchers rely on the responses of managers. One general admonition, therefore, is for researchers to be aware of the issue and the definite tendency toward bias in the responses elicited in this fashion. Unfortunately for those desirous of unambiguous guidelines, however, the bias is not always in the same (positive) direction, nor is it even consistently present across survey items.

19 The logic is that the common source bias appears in the measure for both time periods and differencing the variable will subtract out all of the common source bias.

The preceding section has offered guidelines for researchers who are inclined to use survey assessments. Unfortunately, there are no short-term or easy fixes when managers' perceptual measures of performance are the dependent variables. A priority for research on public management and performance, therefore, should be to develop sound archival measures across a range of types of public organizations. With such measures available, researchers may be able to use perceptual measures of management, environmental characteristics, and the like—especially those that are quite specific—in estimating impacts on such archival measures—with correlated error unlikely, relationships may be attenuated but false positives are not as likely.

Meanwhile, the exceedingly general survey questions of performance that often appear in data sets currently under use by public management researchers clearly raise especially large red flags. The work reported in the present article certainly calls into question the findings of previously published research seeking to understand the determinants of public organizational performance as seen by managers. That sort of extant research in the literature cannot be salvaged, furthermore, by an argument that managerial perceptions of performance somehow incorporate valuable judgments made by the respondents. Meier and O'Toole (forthcoming) have shown strong evidence that managerial perceptions of performance are naive rather than sophisticated: such judgments do not take into account such matters as clientele characteristics, resources available, or task difficulty when judgments are rendered about how well an organization is doing.

This study demonstrates that for the types of items used in our survey, false positives are considerably more likely than false negatives. Researchers, therefore, need to focus especially strongly on this possibility, and positive findings from such data should be reexamined and, ideally, verified by additional analyses and other types of data.

Yet not all positives are false, when one uses perceptual performance measures. There are therefore few broad-brush guidelines aside from those sketched here. The proverbial devil is in the details. Some types of questions are much more likely to exhibit bias and others much less, but there is sometimes considerable variation in bias across questions that are broadly exploring fairly similar subjects. More work on this issue should be undertaken—especially in determining those kinds of survey items that researchers should avoid. But for now, we can say that considerable caution is clearly advised—recall the intercepts in the estimations reported earlier: 1.57, 2.95, 3.25, and 5.64. These are sobering findings.

This article has focused on common source bias, but measurement error is a more general problem in public management and any other field of inquiry. Common source bias generates correlated measurement error, and any form of correlated measurement error is likely to raise similar issues of bias, whether that error is from a common method or from some other process.

Our criticisms here are focused solely on administrative self-assessments of performance. The findings should be treated conservatively. In particular, we should not be interpreted as saying that all subjective assessments are necessarily flawed. None of the discussion necessarily applies to citizen assessments of agency performance or outside experts' assessments of performance. Similarly, the criticism should not

be taken as anti-survey. Surveys are a valid method of collecting information, and in many cases one is hard pressed to avoid using survey measures as both dependent and independent variables. An assessment of employee job satisfaction, for example, is difficult to measure in another way.

Developing valid knowledge about the determinants of performance is an important objective, and especially crucial is knowing how and how much management shapes performance. This topic has justifiably attracted the attention of a number of researchers. Unfortunately, much of the work developed from available data sets is likely to suffer from common source bias and, therefore, contain erroneous and misleading findings. To develop valid findings, other data sets and other approaches will need to be employed.

APPENDIX A: WHAT DOES "SIMILAR" MEAN?

An objection might be raised that the phrase "similar districts" creates some bias because the perceptual measure is then a comparison to similar districts rather than all districts. We address this problem in two ways. First, we treat it as a selection bias problem; and second, we run a randomized survey experiment to determine if the wording matters.

SELECTION BIAS

We treated the problem as the equivalent of selection bias and opted to create a set of statistically similar districts via a propensity score matching technique (see Caliendo and Kopeinig 2008; Dehejia 2005; Dehejia and Wahba 2002; Rosenbaum and Rubin 1985; Smith and Todd 2001). Propensity score matching essentially locates the data point in a multidimensional space and then selects the nearest similar case using the Mahalanobis distance or a similar criterion. To determine the dimensions of the space, we asked four school administrators what they thought about when the phrase "similar districts" was asked. The responses listed fell into four categories—size, resources, student characteristics, and, in one case, proximity. To be comprehensive, we included multiple indicators of size (enrollment, number of schools), resources (revenue per pupil, teachers' salaries, class size, teacher experience), and student characteristics (race, ethnicity, poverty). In a separate test, we also included a proximity measure to these variables (i.e., the geographic proximity to the district). To probe the robustness of the dimensions, we ran several iterations with one or more variables omitted or other variables included. The results were virtually identical to these presented here.

To test the impact of propensity score matching, we ran a regression with the perceptual measure as the dependent variable and the archival measure as the independent variable. We then created the residual from this equation. If there is selection bias relative to similar districts, one would expect the TAKS score of the matched district to be positively correlated with this residual (i.e., with the difference between the original district's subjective and archival scores). For the three dependent variables, the correlations were −.07 for TAKS, .02 for college bound, and −.02 for overall quality

of education. Alternative models for the propensity score matching generated similar correlations, that is, very small and rarely statistically significant.[20]

THE EXPERIMENTAL TEST OF SIMILAR VERSUS OTHER DISTRICTS AS COMPARISON CRITERION

We further probed whether the phrase "similar districts" created a bias by running a field experiment using school principals. That experiment using the same three questions employed in the present survey but referencing schools rather than districts showed that responses for "similar districts" were no different than responses for "other districts." The clear conclusion from the propensity score analysis and the field experiment is that the qualification "similar districts" does not create a selection bias problem, and one can directly compare the managerial perceptions with the archival measures.

Principals of 1,450 Texas school participating in an online survey in March and April 2011 were randomly assigned to two groups. One group was asked to rank their school on the three criteria used in this study against "similar schools." The other group was asked to rank their school on the same three criteria against "other schools." Although the random assignment statistically assures the two groups were comparable, we also checked the two groups in terms of type of school (elementary, middle, high school), student race, ethnicity, income, and school resources. In no case were the groups statistically different from each other. We then did both an analysis of variance on the means and a chi-square test treating the variables as categorical. The corresponding *t*-tests and chi-square statistics show that there was no difference on any of the dependent variables based on whether the comparison was similar schools or other schools.

	Mean	*t*-Score	*p*	Chi-square	*p*	*N*
Performance on the TAKS						
Condition						
Other schools	3.83	0.86	.39	0.87	.93	1,420
Similar schools	3.87					
Performance for college students						
Condition						
Other schools	3.34	1.76	.08	6.95	.14	1,363
Similar schools	3.42					
Overall quality of education						
Condition						
Other schools	3.95	0.76	.45	3.41	.49	1,427
Similar schools	3.99					

20 We ran 18 different propensity score models; the range of correlations across the models was very small. For TAKS, the range was −.051 to −.097, for college-bound performance, the range was −.011 to +.035, and for overall quality, −.025 to +.003. The most negative correlations occurred with the propensity scores that did not use the low-income student variable.

APPENDIX B: INDIVIDUAL QUESTION ESTIMATES OF COMMON SOURCE BIAS

The first value is the estimated *t*-score for TAKS, the second for college bound. The *t*-scores are the absolute value of difference of the *t*-scores in the perceptual and archival equations.

I. Time allocation

Indicate how frequently you interact with individuals in the following groups

School board members	0.29	0.46
Teachers' associations	1.30	0.80
Parent groups (e.g., PTA)	1.54	0.39
Local business leaders	0.11	3.09
Other superintendents	0.35	2.23
Federal education officials	0.89	2.83
State legislators	0.14	1.58
Texas Education Agency	1.33	0.99
City/County Government	0.66	1.80
Local Police/Fire Departments	0.79	0.03
Nonprofit organizations	0.99	2.84

Who initiated the last contact?

School board members	2.25	1.43
Teachers' associations	0.21	0.43
Parent groups (e.g., PTA)	0.46	1.96
Local business leaders	0.17	0.16
Other superintendents	0.25	0.78
Federal education officials	0.80	1.09
State legislators	1.16	0.18
Texas Education Agency	0.55	2.32
City/County Government	0.22	0.14
Local Police/Fire Departments	0.80	1.33
Nonprofit organizations	0.97	0.95

II. Performance appraisal

Superintendents are provided with substantial detail on the performance of students and employees. To what extent do you use this type of performance data to:

Make personnel decisions	1.50	1.16
Make strategic decisions	1.13	1.02
Make day-to-day management decisions	1.15	2.38
Advocate for my district to stakeholders	1.01	0.29
Allocate resources	0.58	0.11
Learn how to make services more efficient	0.85	0.22

III. District resources

How would you rate the following in your district?

Quality of teachers	4.11	7.20
Parental involvement	3.07	5.32
Professional development	3.62	3.30
Community support	0.89	4.19
Principals' management skills	2.63	2.87
School board support	0.66	1.78

IV. Leadership/management practices

I give my principals a great deal of discretion in making decisions.	0.81	2.12
I always try to limit the influence of external events on my principals and teachers.	2.01	2.03
Our district continually adjusts our internal activities and structures in response to stakeholder initiatives and activities.	1.93	1.16
Our district is always among the first to adopt new ideas and practices.	2.75	3.23
Our district frequently undergoes change.	1.11	1.26
There is a lot of conflict over educational issues in our community.	1.49	3.66
We continually search for new opportunities to provide services to our community.	2.41	1.63
I like to implement consistent policies and procedures in all my schools.	1.93	0.84
Our district emphasizes the importance of learning from the experience of others.	0.54	0.48
School districts are asked to do too many things; we should focus more on education.	0.55	0.98
What we do is greatly influenced by the rules and regulations of the Texas Education Agency.	0.47	1.19
I strive to control those factors outside the school district that could have an effect on my organization.	2.24	2.29
With the people I have in this district, we can make virtually any program work.	3.72	5.03
I am quite likely to recommend a subordinate for a superintendent position in another district.	1.71	2.56
I rely on advice from a senior management team to help make important decisions.	0.31	1.18
Our district resolves conflicts by taking all interests into account.	2.01	3.63
Our district works to build a common identity and culture among district employees.	3.11	2.89
Our district concentrates on making use of what we already know how to do.	1.57	1.96

V. Goals

What is the most important problem facing your district?
Please rank order your choices: 1 as most important to 8 as least important.

_____Bilingual education	1.85	1.30
_____College preparation	1.53	0.08
_____Compliance with No Child Left Behind	0.42	0.58
_____Student performance on the TAKS	1.42	1.85
_____Vocational education	1.70	0.89
_____Physical education	1.71	2.53
_____Nutrition issues	0.09	4.53
_____Discipline issues	1.96	3.32

VI. Diversity programs

There are special programs in place in my district to manage diversity among principals, teachers, and staff.	1.79	1.80
I have difficulty recruiting and retaining people of color.	3.23	4.53
Hiring and promoting employees from underrepresented groups is a priority in my district.	3.18	2.36
My district conducts special training and programs on cultural differences and values.	0.58	2.42
In my district, employees generally value ethnic and cultural differences.	3.01	4.15
I would characterize relations between diverse groups in my district as harmonious.	2.22	3.67

VII. The environment

My district's environment the political, social, and economic factors is relatively stable.	3.94	6.43
I would characterize my district's environment as relatively complex.	0.21	4.56
There is a great deal of uncertainty in the environment in which my district operates.	2.42	3.97
My district relies upon partnerships with others in order to accomplish policy goals.	1.15	1.27
State and federal laws put such limits on my discretion that it is difficult to run my district effectively.	0.66	3.45

VIII. Discipline issues

As a superintendent, how often do you spend time on discipline issues pertaining to:

Principals	0.00	1.17
Teachers	0.80	1.19
Students	0.20	0.75

IX. Social capital and trust

In general, citizens and other people in the communities served by my school district:

Exhibit a very high level of social trust towards others.	2.10	4.05
Make charitable contributions, give blood, do volunteer work, etc.	1.87	2.17
Are very active in civic and community affairs.	0.63	0.85
Participate in a wide range of community organizations (e.g., film societies, sports clubs, etc.).	0.10	2.18
The involved groups in this school district fulfill in general their agreements with one another.	2.01	1.93
The stakeholders in this school district fulfill in general their agreements with one another.	0.81	1.80
The stakeholders in this district give the other stakeholders the benefit of the doubt.	2.61	3.29
The stakeholders in this district keep in mind the intentions of other groups.	2.78	3.20
The stakeholders of this district generally do not use the contributions of other actors for their own advantage.	1.94	0.36
The stakeholders in this district can assume that the intentions of others in the district are good in principle.	3.10	3.72

REFERENCES

Avolio, Bruce J., Francis J. Yammarino, and Bernard M. Bass. 1991. Identifying common methods variance with data collected from a single source. *Journal of Management* 17 (3):571–87.

Bazerman, Max H. 2005. *Judgment in managerial decision making*, 6th ed. New York, NY: John Wiley and Sons.

Boyne, George A. 2003. Sources of public service improvement: A critical review and research agenda. *Journal of Public Administration Research and Theory* 13 (3):367–94.

Brannick, Michael T., David Chan, James M. Conway, Charles E. Lance, and Paul E. Spector. 2010. What is method variance and how can we cope with it? *Organizational Research Methods* 13 (3):407–20.

Caliendo, Marco, and Sabine Kopeinig. 2008. Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys* 22 (1):31–72.

Campbell, Donald T., and Donald W. Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56 (2):81–105.

Conway, James M., and Charles E. Lance. 2010. What reviewers should expect from authors regarding common method bias in organizational research. *Journal of Business Psychology* 25:325–34.

Cote, Joseph A., and M. Ronald Buckley. 1987. Estimating trait, method, and error variance. *Journal of Marketing Research* 24 (3):315–88.

Dehejia, Rajeev. 2005. Practical propensity score matching: A reply to Smith and Todd. *Journal of Econometrics* 125:355–64.

Dehejia, Rajeev, and Sadek Wahba. 2002. Propensity score matching methods for nonexperimental causal studies. *The Review of Economics and Statistics* 84 (1):151–61.

Doty, D. Harold, and William H. Glick. 1998. Common methods bias: Does common methods variance really bias results? *Organizational Research Methods* 1 (4):374–406.

Ghiselli, Edwin E., John P. Campbell, and Sheldon Zedeck. 1981. *Measurement theory for the behavioral sciences*. San Francisco, CA: W.H. Freeman.

Hand, D.J. 2004. *Measurement theory and practice*. London: Arnold Press.

Hastie, Reid, and Robyn M. Dawes. 2001. *Rational choice in an uncertain world*. Los Angeles, CA: Sage Publications.

Kahneman, Daniel, and Amos Tversky. 2000. *Choices, values, and frames*. New York, NY: Cambridge Univ. Press.

Kemery, Edward R., and William P. Dunlap. 1986. Partialling factor scores does control method variance. *Journal of Management* 12 (4):525–44.

Lance, Charles E., Lisa E. Baranik, Abby R. Lau, and Elizabeth A. Scharlau. 2009. If it ain't trait it must be method. In *Statistical and methodological myths and urban legends*, ed. Charles E. Lance and Robert J. Vandenberg, 337–60. New York, NY: Routledge.

Lance, Charles E., Bryan Dawson, David Birkelbach, and Brian J. Hoffman. 2010. Method effects, measurement error, and substantive conclusions. *Organizational Research Methods* 13 (3):435–55.

Lord, Frederic M., and Melvin R. Novick. 1968. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Malhotra, Naresh K., Sung S. Kim, and Ashutosh Patil. 2006. Common method variance in IS research. *Management Science* 52 (12):1865–83.

Meier, Kenneth J., and Laurence J. O'Toole, Jr. Forthcoming. I think (I am doing well), therefore I am: Assessing the validity of administrators' self-assessments of performance. *International Public Management Journal*.

Miles, Raymond E., and Charles C. Snow. 1978. *Organizational strategy, structure, and process*. New York, NY: McGraw-Hill.

Moynihan, Donald. 2008. *The dynamics of performance management*. Washington, DC: Georgetown Univ. Press.

Oberfield, Zachary W. 2012. Making change: How management overcomes organizational inertia. Paper presented at the annual meetings of the Midwest Political Science Association, Chicago, IL.

Podsakoff, Philip M., and Dennis W. Organ. 1986. Self reports in organizational research: Problems and prospects. *Journal of Management* 12 (4):531–44.

Podsakoff, Philip M., Scott B. MacKenzie, Jeong-Yeon Lee, and Nathan P. Podsakoff. 2003. Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology* 88 (5):879–903.

Pollitt, Christopher, and Geert Bouckaert. 2000. *Public management reform: A comparative analysis*. Oxford, UK: Oxford Univ. Press.

Rafferty, Alannah E., and Mark A. Griffin. 2004. Dimensions of transformational leadership: Conceptual and empirical extensions. *The Leadership Quarterly* 15:329–54.

Richardson, Hettie A., Marcia J. Simmering, and Michael C. Sturman. 2009. A tale of three perspectives: Examining post hoc statistical techniques for detection and correction of common method variance. *Organizational Research Methods* 12 (4):762–800.

Rosenbaum, Paul R., and Donald B. Rubin. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39 (1):33–8.

Schachter, Hindy Lauer. 2010. Objective and subjective performance measures: A note on terminology. *Administration & Society* 42 (5):550–67.

Shultz, Kenneth S. 2005. *Measurement theory in action*. Thousand Oaks, CA: Sage Publications.

Siemsen, Enno, Aleda Roth, and Pedro Oliveira. 2010. Common method bias in regression models with linear, quadratic and interaction effects. *Organizational Research Methods* 13 (3):456–76.

Smith, Jeffrey A., and Petra E. Todd. 2001. Reconciling conflicting evidence on the performance of propensity-score matching methods. *American Economic Review* 91 (May):112–8.

Spector, Paul E. 2006. Method variance in organizational research: Truth or urban legend? *Organizational Research Methods* 9 (2):221–32.

Weisberg, Herbert F. 2005. *The total survey error approach: A guide to the new science of survey research*. Chicago, IL: Univ. Chicago Press.

Williams, Larry J., and Stella E. Anderson. 1994. An alternative approach to method effects by using latent-variable models. *Journal of Applied Psychology* 79 (3):323–31.

Zeller, Richard A., and Edward G. Carmines. 1980. *Measurement in the social sciences*. New York, NY: Cambridge Univ. Press.