Organizational Performance: Measurement Theory and an Application

Or, Common Source Bias, the Achilles Heel of Public Management Research

Kenneth J. Meier
Department of Political Science
Texas A&M University
College Station, TX 77845 USA
kmeier@politics.tamu.edu
and
Cardiff School of Business
Cardiff University (UK)

and

Laurence J. O'Toole, Jr.
Department of Public Administration and Policy
School of Public and International Affairs
Baldwin Hall
The University of Georgia
Athens, GA 30602 USA
cmsotool@uga.edu

Paper presented at the Annual Meeting of the American Political Science Association, September 1-5, 2010, Washington DC, Copyright American Political Science Association. Would like to thank Ling Zhu for her perceptive comments on an earlier draft of this manuscript.

Organizational Performance: Measurement Theory and an Application

Or, Common Source Bias, the Achilles Heel of Public Management Research

The questions, Are public programs effective? and What is the role of public management in this regard? have occupied scholars in both the U.S. (Moynihan 2008) and numerous other countries (Pollitt and Bourckaert 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 et al. 1981; Shultz 2005; Hand 2004). This paper uses measurement theory to assess one potential problem in measuring organizational performance. It considers both subjective and data-based measures, as well as measures internal to the organization and those imposed by external stakeholders. The paper provides an illustration of the insights of measurement theory by an analysis of performance indicators for 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.

First, the paper illustrates how some subjective measures of organizational performance can be contaminated by common source bias. Second, we move this illustration 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 survey items to determine what types of questions are more susceptible to common source bias; the results show that many but not all survey items have such biases. The objective of this work is to assess the overall dangers of common source bias and, if possible, to provide guidelines for scholars in terms of how and under what conditions the problems 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 Administrative Studies Project II, the Merit Principles Survey, and the American State Administrators' Project, all of which have spawned a substantial literature linking public management to performance.[1] 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.[2] 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 Fisk 1959). A systematic study comparing managers' perceptions of organizational performance to external objective measures of performance found that managers consistently overestimated the level of performance in the organization (see Meier and O'Toole 2010) and that this overestimation was not related to more difficult tasks or the availability of resources. More problematic than the overestimation was that the error contained in the assessments of organizational performance was correlated with measures of management, even though these management measures were uncorrelated with the

[1] 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).

[2] The number of studies using subjective measures from these four surveys is substantial. A Google Scholar search using both the data base title and the word performance generated the following number of studies: Merit Principles Study – 101, Federal Human Capital Survey – 95, National Administrative Studies Project – 87, and American State Administrators Project – 41. Not all of these studies actually use the subjective performance measures, but these numbers do indicate extensive use.

objective performance measure. In measurement theory, this problem is called "common source bias," that is, the common measurement contains a source of error that shows up in both measures and thus can contribute to spurious results (Doty and Glick 1998). Managers 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. Although common source bias normally generates false positives – the conclusion that a relationship exists when one does not – under certain conditions 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 concept [C] used in research should be distinguished from the indicator [I] that is used to operationalize it. Theoretically, this can be repressed with a simple equation:

$$C = I + e$$

where C is the concept,

I is the empirical measure of the concept,

and e is an error term.

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). Most common analytical techniques assume that the error has a mean of zero (that is, the measure neither over- nor underestimates the concept) and that it is randomly distributed (that is, not correlated with the value of the concept, see Berry 1993). The assumption of random error is especially important in assessing the relationships among variables because it essentially allows the analyst to ignore the error in the estimation of results. To illustrate, let us assume two variables X and Y with the following measurement characteristics:

$$C_y = I_y + e_y$$

$$C_x = I_x + e_x$$

The research is interested in the theoretical relationship between X and Y, let us say the correlation between them:

$$r\ [C_x\ C_y]$$

but must estimate the relationship with the empirical indicators at hand or:

$$r\ [I_x\ I_y]$$

How might the fact that both indicators are measured with error affect this result? In general, the impact of measurement error on the relationship between two variables is ambiguous – that is, there is no way to tell. The assumption that the errors are random (or the regression assumption that the errors are uncorrelated) is the solution to this ambiguity. If the errors are random, they increase the variance of both I_x and I_y; but since they are random, there is no correlation of the errors – that is, r $[e_x\ e_y]$ is zero. Error, by adding additional variation to each term but by being uncorrelated with either term of the other, then merely attenuates the relationship between X and Y; in other words, measurement error of this type reduces the size of the correlation. A reduced correlation means that random measurement error can generate false negatives, the conclusion that a relationship does not exist when in fact it does. Statisticians generally prefer false negatives to false positives because they make the analysis more conservative – that is, less likely to conclude that a relationship exists when in fact one does not.

To understand the unique problems of common source bias, it is necessary to divide the error term into two parts – the part that is generated by the measurement technique itself (e_b) and the part that essentially results from a random process (e_r). Returning to our example of linking X and Y, this now yields:

$$C_y = I_y + e_{yr} + e_{yb}$$

$$C_x = I_x + e_{xr} + e_{xb}$$

The problem is that e_{yb} is no longer random; it is correlated with e_{xb} if the same measurement technique is used for both variables. If the common source error is positively correlated, then this correlated error will inflate the relationship between X and Y; in other words, it will overestimate relationships and perhaps generate positive findings when none exist. If the measurement error is negatively correlated (e.g., one question has an element of social desirability and another question has an element of social undesirability), then the size of the relationship is reduced and false negatives can be a problem. This ambiguity in terms of result is the pernicious aspect of common source measurement error; it can create positive results where none exist and it can generate null results when in fact a relationship does exist.[3]

This ambiguity is complicated because the degree of common source bias in a set of

[3] This paper 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.

measures can vary from item to item. In a survey, for example, not all questions will necessarily contain the same amount of common source bias. 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. Because not all questions have an 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 question lack social desirability) and some questions where spurious results are likely. In short, common source bias can generate false positives, false negatives, or no such effect whatsoever.

Even the name, common source bias is somewhat misleading since the problem of correlated errors can be generated by in a variety of ways (see Weisberg 2005). One source of the bias might be the questions themselves, because the question might have more than one stimulus, that is, they tap performance but also the respondent's self interest in looking competent. In this cases 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 or the 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, 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. The next section of the paper will do that by introducing a data set designed in part to assess common source bias, provide a method for determining the degree of bias, and then empirically examine 84 questions to determine the types of questions that might be problematic.

Testing for Common Source Bias

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 administrator perceptions of performance and management style are taken from this source for the academic year 2008-2009, 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.[4]

[4] 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 paper. The public school

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.[5] 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.

For perceptual measures of performance, superintendents were asked "compared to similar districts, my assessment of our ____ performance is" on a five 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 get questions that reflect the range of performance indicators that might be tapped via perceptual data. 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

sample is not statistically different from the non-respondents in terms of TAKS scores, college bound scores, the racial and income distribution of students, and instructional expenditures. Respondents did receive $480 less in per pupil revenue than non-respondents.

[5] 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 (for example, several elementary exams are given in Spanish). The results throughout the entire state were posted on the TEA website in December of 2009, with results made available to district superintendents considerably earlier.

and Glick 1998). 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. 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.

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 1110 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 closer to the measures actually used in the 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

One way to get leverage on common source bias is to use the responses to a given question tapping an independent variable and do parallel analysis with two dependent variables – one perceptual and internal and the other externally generated and objective – 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. To illustrate this process, in Table 1, we take the superintendent's response to the statement "Our district is always among the first to adopt new ideas and practices" on a four point agree-disagree scale. 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 education production function includes this independent variable along with several controls. 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 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.

[Table 1 about here]

The test of common source bias will be a comparison of the t-scores for the appropriate coefficients. Table 1 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. 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. Rather than producing a similar table with full results for all 84 measures in the survey, we will use the relevant information from the tables – that is, the difference between the t-score for the perceptual measure (3.69) and the t-score for the archival measure (0.94) as the measure of the degree of common source bias (2.75 in this case; see appendix for all questions and bias estimates). 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.[6] We have run two sets of regressions for each of the variables in the survey and then grouped the survey questions by characteristics to determine if certain types of questions are more susceptible to common source bias than others.[7]

Table 2 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

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

[7] An alternative way to estimate the 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 with means of 1.87 for TAKS, 1.85 for college bound, and 2.83 for the general measure. We would like to thank Ling Zhu for suggesting this alternative estimator.

specific TAKS measure, the average t-score difference to indicate common source bias is 1.42, with a range of 0 to 4.11. As a rough indicator of false positives, we counted any relationship where the relationship was statistically significant at the .05 level for perceptual performance and not for archival performance. There were 20 cases of false positives among the 84 variables. False negative were more rare; in only three cases were archival measures were significant but perceptual were not.

[Table 2 about here]

The results for the college bound measure were even more troubling. Because the college bound measure was less specific, we expected more common source bias problems. The average t-score difference for common source bias is 2.09, with a range of .03 to 7.20. The analysis found 31 false positives and 10 false negatives. In addition, one case (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 cases or 50% of the cases with spurious results. 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 random when one employs the perceptual measure of performance*. Given that even the college board question is far more specific than the very general types of question used in most analysis, the problems in other data sets are likely to be far greater.

Given these problematic results, one could ask if the correct method would be simply to ban perceptual measures by administrators as dependent variables of 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 administrator's perception of organizational performance. This suggest 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 discuss 84 different questions, it would be more helpful to group the questions by characteristics and determine if such characteristics generate more or less common source bias. This process, however, will only tell one the average common source bias in a set of questions; any individual question might have more or less bias, depending on the error generating process. Eighteen aspects of the individual questions were coded as follows:

1. Time – does the question ask the manager about how much time the manager does X?

E.g, How frequently do you meet with local business leaders?

2. Quality Assessment – does the question require an assessment of quality, of others etc.?

E.g, 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.

E.g, My district's environment – the political, social and economic factors – is relatively stable.

4. Environment general – the question is about the environment in general.

E.g, see question immediately above

5. Environmental Support – the question asks for an assessment of environmental support.

E.g, How would you rate community support in your district?

6. Environment the People – an assessment of clientele characteristics or behavior.

E.g, In general citizens and other people in the communities served by my school district are active in civic and community affairs.

7. M1 – the question asks about internal management of the organization

E.g., I give my principals a great deal of discretion in making decisions.

8. M2 – the question relates to the management of the environment.

E.g., see item 1 above on time.

9. M3 – the question deals with efforts to exploit environmental opportunities.

E.g, we continually search for new opportunities to provide services to our community.

10. M4 – the question deals with buffering the environment.

E.g, 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.

E.g, I use performance data to make personnel decisions.

12. Strategy – the question asks about general management strategy.

E.g., 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).

E.g, see example for variable 12.

14. Defending Strategy – the question relates to a strategy of defending.

E.g., our district concentrates on what we already know how to do.

15. Reacting Strategy – the question relates to a reactor management strategy.

E.g., 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.

E.g., In my district, employees generally value ethnic and cultural differences.

17. Goals – the question asks about organizational goals.

E.g., What is the most important problem facing your district? Please rank order your choices.

18. Observable – the question asks about observable behavior.

E.g, How frequently do you meet with City/County government officials?

The characteristics are not mutually exclusive; that is, a question could ask about managing the environment and it could also involve observable behavior on the part of the manager. As a result, we opted to assess all 18 characteristics simultaneously by using the characteristics as independent variables in a regression with the dependent variable our measure of common source bias. The results of this regression appear in Table 3 for the more specific standardized test indicator of performance, TAKS.

One key to interpreting table 3 is the intercept. This can be interpreted as the average common source bias of a question with none of the characteristics coded (that is zero for all characteristics). A significant coefficient of 1.57 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 "quality assessment" characteristic of 1.49

means that a question asking for a quality assessment will generate a common source bias of 3.06 (or 1.57 + 1.49) t-score units, and this amount of bias is significantly more than the 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 network management, 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. For example, questions dealing with time allocation have a coefficient of -.46 indicating little common source bias, but the question asking about whether the superintendent initiated the most recent contact with the school board carries a strong warning about common source bias (coefficient = 2.25; see the appendix).

[Table 3 about here]

Another qualification is in order. An indication of common source 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 common source bias coefficient (by our measure of the difference in t-scores, +4.11; see the appendix), but the question still produces strong and positive 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 4 provides parallel results but for the college-bound performance measures, a more general question that is more likely to be subject to common source bias. If the general pattern of relationships looks similar to those in Table 3, it is because the two measures of common source biased 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 common source 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 common source bias. The least common source 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 common source bias is real and needs to be considered.

[Table 4 about here]

The significantly worse results in terms of 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 becomes more severe. While we cannot be sure that this is always the case, it is the case in the best existing 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 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, we provide two estimates. The optimal illustration would involve an objective measure of the overall quality of education and then replicate the analysis for the subjective assessment measure. Lacking such an archival measure of overall quality, we compare the regressions on the subjective quality of education measure to the objective equations for TAKS and the college bound equations discussed above. 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 (see Meier and O'Toole 2010). Table 5 provides both the assessment of false positive and false negatives and the regression results for the questions for both comparisons.

[Table 5 about here]

The table clearly shows that the problems of common source 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 subject assessment dependent variable has a mean t-score bias of 2.51 and generates 32 false positives and two 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 common source 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. While some of the negative scores that indicate less bias are large, they need to offset the huge intercept; only for questions involving network management, 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 environment people (+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 questions used in the literature. Examining this question from two different perspectives shows that common source 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 when the independent variables are also gathered by survey.* 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; 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 gain. Avoiding administrative perceptions of performance is the only guaranteed method to avoid 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 on the dependent variable. This paper has demonstrated that the degree of common source bias is reduced 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 will 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. This paper has demonstrated that for 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 include all our estimates of bias in the appendix.

While there are statistical techniques to rid a data set of common source bias, they all rely on the shared variance of questions. This means that the statistical solutions will 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.

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.)

It is clear from the findings for the several hundred organizations included in our study that common source bias is indeed 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.

What then can be said by way of generalization that could be useful for research? Several points seem relevant. First, perceptual measures contain error, but they can be used without danger of common source bias provided that perceptual measures do not appear on both sides of the equation. In particular, the problem arises especially clearly in efforts to model and estimate performance when perceptual measures of performance are the dependent variables. The practical points here are that archival measures of performance are clearly to be preferred to managerial perceptions, that efforts should be made to develop sound archival measures across a range of types of public organizations, and that perceptual measures of management, environmental characteristics and the like are certainly useful in estimating such archival measures – with correlated error unlikely, relationships may be attenuated but false positives are not a particular issue.

Second, the more vague the items eliciting perceptual performance data, the more likely for common source bias to be a serious matter. Even specific perceptual performance measures, however, suffer from common source bias. The more general the measure, the worse the problem. The most general measures can be expected to be riddled with bias. So analyses based on the exceedingly general survey questions that often appear in data sets currently under use by researchers raise especially large red flags. The work reported in the present paper 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. In our earlier work (Meier and O'Toole 2010), we 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.

Third, 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 (those tapping socially desirable aspects like prospecting strategy, quality assessments, and managing diversity) and others much less (time allocation, observable behavior), 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 steer clear of. But for now, we can say that considerable caution is clearly advised – recall the intercepts in the estimations reported earlier: 1.57, 2.95, 5.29 and 6.12. These are sobering findings.

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. In particular creating management surveys for organizations that have existing performance measurement systems should be a priority. It is past time to begin that effort. General perceptual measures of performance are no substitute for the hard work of designing and implementing surveys for specific sets of organizations.

References

Berry, William D. 1993. *Understanding Regression Assumptions*. Newbury Park, CA: Sage Publications.

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.

Campbell, Donald T. and Donald W. Fiske. 1959. "Convergent and Discriminant Validation by the Multitrait-multimethod Matrix." *Psychological Bulletin* 56 (Number 2), 81-105.

Doty, D. Harold and William H. Glick. 1998. "Common Methods Bias: Does Common Methods Variance Really Bias Results?" *Organizational Research Methods* 1 (Number 4), 374-406.

Ghiselli, Edwin E., John P. Campbell, and Sheldon Zedeck. 1981. *Measurement Theory for the Behavioral Sciences*. San Francisco: W.H. Freeman.

Hand, D.J. 2004. *Measurement Theory and Practice*. London: Arnold Press.

Meier, Kenneth J., and Laurence J. O'Toole, Jr. 2010. "I Think (I am doing well), Therefore I Am: Assessing the Validity of Administrators' Self-Assessments of Performance." Paper presented at the annual meetings of the Midwest Political Science Association, Chicago, April.

Miles, Raymond E., and Charles C. Snow. 1978. *Organizational Strategy, Structure, and Process*. New York: McGraw-Hill.

Moynihan, Donald. 2008. *The Dynamics of Performance Management*. Washington: Georgetown University Press.

Pollitt, Christopher, and Geert Bouckaert. 2000. *Public Management Reform: A Comparative Analysis*. Oxford: Oxford University Press.

Schachter, Hindy Lauer. 2010. “Objective and Subjective Performance Measures: A Note on Terminology.” *Administration & Society* (forthcoming).

Shultz, Kenneth S. 2005. *Measurement Theory in Action*. Thousand Oaks, CA: Sage Publications.

Weisberg, Herbert F. 2005. *The Total Survey Error Approach: A Guide to the New Science of Survey Research*. University of Chicago Press.

Zeller, Richard A. and Edward G. Carmines. 1980. *Measurement in the Social Sciences*. New York: Cambridge University Press.

Table 1. Common Source Bias: Management Style and Subjective and Actual TAKS Performance

	Actual TAKS		Subjective Assessment	
	Slope	t-score	Slope	t-score
Prospector Style	.0505	0.94	.1748	3.69*
% Black Students	-.0577	1.66#	-.0008	0.25
% Latino Students	-.0796	3.51*	-.0020	0.99
% Low Income	-.2495	8.14*	-.0117	4.35*
Teacher Salary K	.3726	3.11*	.0160	1.53
Class Size	-.0299	0.20	.0177	1.36
Teacher Experience	-.2168	1.40	-.0193	1.42
Teacher Turnover	-.2951	7.06*	-.0091	2.49*
Standard Error	8.90		.78	
F	69.31		14.83	
R-Square	.47		.16	
N	629		626	

#p < .10 two tailed test

*p < .05 two tailed test

Table 2. Measuring Common Source Bias on Perceived Organizational Performance

	Dependent Variables = TAKS Tests	College Bound Performance
Average T-score Bias	1.42	2.09
Standard Deviation 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
Percent Spurious Results	27.4	50.0*

*One case has significant relationships in different directions.

Table 3. Common Source Bias by Type of Question: TAKS Performance

Type of Question	Slope	t-score
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
Network Management	-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
Strategy–Prospecting	0.08	0.10
Strategy–Defending	-0.31	0.51
Strategy–Reacting	-1.66	1.47
Diversity	1.15	2.63
Goals	-0.20	0.43
Observable Behavior	-0.63	2.05
Intercept	1.57	3.74
Standard Error	.87	
R-Square	.46	
N	84	

Table 4. Common Source Bias by Type of Question: College Bound Performance

Type of Question	Slope	t-score
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
Network Management	-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
Strategy–Prospecting	-0.87	0.82
Strategy–Defending	-1.07	1.24
Strategy–Reacting	-1.75	1.75
Diversity	1.02	1.67
Goals	-1.10	1.73
Observable Behavior	-0.84	1.95
Intercept	2.95	5.03
Standard Error	1.21	
R-Square	.50	
N	84	

Table 5. Common Source Bias by Type of Question: Overall Quality of Education

	Compared to TAKS		Compared to College	
Type of Question	Slope	t-score	Slope	t-score
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
Network Management	-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
Strategy–Prospecting	0.32	0.28	-0.82	0.49
Strategy–Defending	-0.71	0.79	-1.15	0.85
Strategy–Reacting	-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
Intercept	3.25	5.29	5.64	6.12
Standard Error	1.27		1.90	
R-Square	.68		.63	
N	84		84	
Mean	2.51		3.04	
False Positives	32		34	
False Negatives	2		8	
Wrong Direction	0		2	
Spurious Results Percentage	40.5		52.3	

Appendix
Individual Question Estimates of Common Source Bias

The First Value is the Estimated T-score for TAKS, the Second for College Bound

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 Depts.	0.79	0.03
Non-profit 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 Depts.	0.80	1.33
Non-profit 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