Evidence-based policy and performance management: Complementary or colliding movements?

Carolyn J. Heinrich
La Follette School of Public Affair, at the University of Wisconsin-Madison
cheinrich@lafollette.wisc.edu
Evidence-based policy and performance management: Complementary or colliding movements?

Carolyn J. Heinrich

La Follette School of Public Affairs
University of Wisconsin-Madison

February 2007
Two movements aimed at promoting government use of evidence on program performance and effectiveness to inform policymaking and program management have been gathering substantial momentum worldwide. One of these—evidence-based policy making, in which policies and practices are based on or determined by scientifically rigorous evidence—has long been espoused in the medical field. For example, there is a Center for Evidence-Based Policy (at Oregon Health and Science University), the Cochrane Collaboration (established in the United Kingdom), evidence-based policy networks, evidence-based journals and journal clubs, and evidence-based policy making newsletters and bulletins that review and disseminate current research findings on the effectiveness of healthcare interventions. Yet it is only more recently that evidence-based policy making has been embraced in social policy, asserts the Coalition for Evidence-Based Policy, with opportunities for significant advances still ahead:

“…in most areas of social policy—such as education, poverty reduction, labor and employment, crime and justice, and health care financing and delivery—government programs often are implemented with little regard to evidence, costing billions of dollars yet failing to address critical needs of our society. However, rigorous studies have identified a few highly-effective social interventions, suggesting that a concerted government strategy to build the knowledge base of these proven interventions, and spur their widespread use, could bring rapid progress to social policy similar to that which transformed medicine.”

(http://coexgov.securesites.net/index.php)

In fact, the UK’s social policy counterpart to the Cochrane Collaboration, the Evidence for Policy and Practice Information and Coordinating Centre (EPPI-Centre), just introduced in 2005 the first issue of a new journal, Evidence & Policy, intended to promote broader use of systematic approaches to reviewing and using scientific evidence in social policymaking.

The performance management movement—which endeavors to use information on agency or program outcomes to regularly assess government performance and hold managers accountable for program results—was growing and attracting global attention about the same
time that evidence-based policy making was expanding. The rise of the development of
performance management systems and practices has been nothing short of meteoric; both
nationally and locally, performance management is now a goal or function of most governmental
and non-governmental organizations, and in many countries, legislation and cabinet-level entities
have been created to support it. And although performance measurement as a management tool
in organizations has a long history dating back to the 19th century, it is primarily in the last two
decades that government performance management has adopted the explicit aim to regularly and
more rigorously measure outcomes and report results to the public (Radin, 2000, 2006). In
addition, governments worldwide are now attaching higher stakes to the achievement of
performance outcomes through the growing use of performance-contingent pay, organization-
wide performance bonuses or sanctions, and competitive performance-based contracting
(Heinrich, 2007).

As generally described above, evidence-based policy and performance management
appear to share an elemental goal: to improve government effectiveness by developing and
utilizing a more rigorous base of information and scientific evidence to guide decisions about
program design, funding, implementation and management. In fact, in ideal circumstances,
governments would use a full range of information in decision/policy-making in a logical flow—
from data on inputs and processes (e.g., staffing, resources, core technologies/procedures), to
outputs (e.g., services provided) and performance outcomes (intermediate results), to impacts
(i.e., value-added estimated through comparison with a counterfactual state, or what would have
happened had the policy or intervention not been implemented)—linking performance
monitoring of ongoing processes and results to the scientific evaluation of impacts and cost-
effectiveness (or net impacts). In practice, however, there are some important differences and
tensions between the approaches of these two movements to achieving this common objective, such as their methods and standards for assembling and analyzing data, and the strategic timing and use of this information to influence policy and/or hold public managers accountable for performance. The implications are that these two movements have not only intersected but have also, at times, collided—particularly in areas such as education, health care services delivery, welfare, and employment and training—where the federal government and other stakeholders are developing more specific requirements for the use of both scientific evidence and performance information in policy making and management and linking resources or rewards/sanctions to the outcomes.

The objective of this essay is to elucidate some of the critical differences and tensions between evidence-based policy and performance management and to address the thorny issues that could thwart their success if left unattended by researchers and policymakers. The paper begins with an overview of these two movements, including brief historical information and more recent developments in diverse policy areas. It then draws from a variety of cases and empirical studies to consider the following key issues/questions:

1. In building bases of information on government effectiveness, are the rigorous standards and principles of the scientific method for producing knowledge incompatible with the demands of performance management to produce timely information for decision-making?

2. What should count as evidence? Are there “gold standard” methods for producing evidence that should count more than others?

3. In communicating information and evidence, does the use of more technically sophisticated methods to produce evidence put it out of reach of policymakers and the public and jeopardize its value to policymaking?

4. Who should judge the quality and reliability of evidence and performance information used, and how can high standards for its production and use be upheld under accountability pressures?
Is it possible to achieve a balance between processes that produce rigorous information to be applied methodically in decision making, and those that facilitate public input and stakeholder representation with the goal of fostering democratic governance and accountability?

In considering these issues, this essay raises as many questions as it answers. And in addressing these questions, it points to more examples of challenges and ongoing debate than to concrete recommendations or resolutions. At the same time, this study suggests some promising directions for efforts to improve government effectiveness through the use of more rigorous information and evidence in decision making, and it also demonstrates the limitations and risks associated with such efforts, both in guiding policy making and advancing democratic values.

Background and Overview

Ian Sanderson (2003) traces the roots of evidence-based policy making to the 17th century era of “Enlightenment,” a time in which the focus of science shifted toward the application of knowledge for human betterment. Considerable time would pass, however, before confidence in a similar role for social science knowledge to shape political and social processes and improve policy interventions would take hold. In both Britain and the United States, the 1960s proved to be a watershed time for the influence of social science in public policymaking. Motivated by the rapid growth in public expenditures on social welfare and increasing demand for knowledge about how to effectively address social problems, the national governments in these countries took steps to expand the role of social science research in advising policy and executive agency decisions (Sanderson, 2002). For example, Britain established the Social Science Research Council to expand social science research, and in 1965, the U.S. issued an executive order requiring all federal agencies to develop measures of cost effectiveness and to integrate “sound logic, firm data, and systematic thinking” into their decision making (Institute for Research on Poverty, http://www.irp.wisc.edu/aboutirp/history.htm#intro).
As Sanderson (2003) recounts, however, optimism about the potential for social science to “change the world” was followed by skepticism about its accuracy and ability to have an immediate impact on policy decisions. Similar issues were emerging in the literature on the use of evidence-based practices in health care, where debate about the nature of evidence or “what counts as evidence” became more intense beginning in the 1970s, as the research community was coming to recognize that the production and use of evidence were social as much as scientific processes (Rycroft-Malone et al., 2004). For example, in their review of “research-into-practice” studies and meta-analyses with well-defined clinical foci, Rycroft-Malone et al. and Dopson et al. (2002) identified multiple, and at times, competing bases of evidence, with differing interpretations of the same evidence by diverse stakeholders and an important role for context in the study findings. And increasingly, significant variation and unexplained deviations in clinical practice from evidence-based practice were being attributed to managerial decisions or organizational characteristics, prompting some researchers to call for more attention to evidence-based management issues (Walshe and Rundall, 2001).

As these concerns about the role of management and implementation in achieving evidence-based policies and practices were growing, the performance management movement was taking off, with intensified public demands for better measurement and management of the explicit outcomes of government action. In the U.S., President Clinton signed the Government Performance and Results Act (GPRA) of 1993 into law with a promise to measure progress and hold federal agencies accountable for their results. Agencies are required under GPRA to establish performance goals, measures and plans; to provide evidence of their performance relative to targets; and to report their results annually to the public. Increasing demands for performance management and analogous reforms were also transforming governments in
Canada, Western Europe, New Zealand, Australia (and later Asia, Africa, and Latin America). Interestingly, some of the aphorisms employed in promoting performance management closely paralleled those being advanced to build support for evidence-based policymaking. For example, the new “ministerial mantra” used by the UK’s Labour government to argue for evidence-based policymaking—“What counts is what works” (UK Cabinet Office, 1999)—was akin to the Clinton administration’s call to determine through performance measurement “what works and what doesn’t” and to invest the public’s money accordingly.

These new demands for accountability and evidence of performance have raised the stakes for those involved in producing research evidence and program evaluations. For example, in early 2006, the American Medical Association signed an agreement with the U.S. Congress to develop more than 100 performance measures and standards based on clinical research evidence linking treatments and practices to better patient outcomes (Pear, 2006). One of the objectives of this effort to develop uniform measures of quality of care is to move toward a system that rewards quality and pays for performance. Not surprisingly, prominent medical specialists have voiced concern about who will be involved in defining, testing and validating quality of care measures and about the aggressive timetable set for their implementation (the end of 2007).

In effect, as the evidence-based policy and performance management movements have become more intertwined, some of their incongruities have become more apparent. An early observer of the implementation of performance management activities, Beryl Radin (1998:309) criticized the naïveté of the reforms’ architects for assuming that “performance information is objective and can serve multiple and conflicting masters… and the public in general.” And despite advances in our analytical tools and capacity for assembling performance information and scientific evidence, it has become increasingly clear that we are still far from a consensus—
intellectually or politically—regarding what should count as evidence, how it should be produced and validated, and how it should be used to influence policy making. These broader concerns are now considered below in a more detailed examination and discussion of the research/policy questions outlined in the introduction.

**Are Rigorous Standards for Knowledge Creation Incompatible with the Demands of Performance Management?**

Underlying the revered scientific method for producing *knowledge*—that is, an accurate, consistent and logical representation of the world and relationships operating within it—are simple and essential principles for minimizing the possibility of bias and error in our understanding and interpretation of the world. Conducted properly, the scientific method takes time—observation and description of phenomena over time—and is a collective endeavor, with scientists or scholars observing from different angles or lenses and working toward the formulation of a hypothesis to explain the phenomena. The testing of hypotheses follows standard (and typically) quantitative/experimental procedures, and it is advantageous to conduct multiple, independent tests of the hypotheses under carefully controlled conditions. Of course, it has long been acknowledged that it is more challenging to achieve highly-controlled, laboratory-like conditions in the generation of social science knowledge, and some efforts to do so have been criticized on ethical grounds and/or because of other concerns about the extent to which efforts to control conditions might interfere with the phenomena under observation.

In the current push for evidence-based policy making, two camps with differing views about the state of social science knowledge and its prospective role in policymaking appear to have emerged. One side argues that there is already a considerable, solid base of knowledge that has accumulated for a wide range of social problems, and that it is simply ignored by policy
makers who are not interested in trying to “digest” and use this evidence (Young et al., 2002; Shulock, 1999). A recommended response to this problem, consistent with the goals of the performance management movement, is to “democratize” the accumulation, presentation and debate of the evidence. Shulock (1999: 240) suggests that the traditional scientific approach to evaluating policies has been perceived by some as too positivist and elitist, and contrary to the stated goals of a few of the coalitions or collaborations for evidence-based policy, she proposes that analysts should not claim to be “seeking the truth,” but rather to be offering “compelling, evidence-based presentations of their perspectives.” In this regard, Shulock (and other like-minded critics) appear to discourage the idea of a hierarchy of evidence, in which researchers and analysts present their findings as superior due to their derivation from scientific methods. This perspective also argues for a more participatory discourse, a broader conceptualization of what should count as evidence, and a greater role for the public in evaluating the evidence, which should ultimately lead to its wider embrace.

Although Cook and Ludwig (2006) alternatively argue that there is little definitive evidence on what works among social policies, they likewise criticize the push for evidence-based policy that adheres to strict scientific standards for what constitutes acceptable evidence (i.e., the allowable probability of a type I error) and leads to the oft repeated conclusion that “more research is needed.” In particular, they suggest that the undue focus on identifying statistically significant impacts has led to the exclusion of much policy relevant evidence that is “concerned with basic beliefs about human nature and interactions” and is important to appropriately conceptualizing policy issues and problems (696).

The competing view argues that there is considerable room for strengthening the bases of scientific knowledge and evidence and their role in informing policymaking, and that generating
reliable, causal evidence through rigorous methods such as randomized controlled trials should be a priority of both researchers and policymakers (Maynard, 2006). For example, a primary goal of the U.S.-based Coalition for Evidence-Based Policy is to promote federal and state funding of randomized controlled trials that will strengthen the knowledge base of effective social interventions and facilitate their replication. And in its use of a new “evidence-based” Program Assessment Rating Tool (PART) to evaluate the effectiveness of federal programs and guide funding decisions, the U.S. Office of Management and Budget (OMB) purportedly accords higher ratings to programs that provide documentation of their effectiveness through experimental evaluations. In addition, the UK Cabinet Office has taken these ideas a step further, calling not only for more longitudinal and experimental research designs, but also for training for policymakers in economics, statistics and other relevant scientific methods to help them become “intelligent customers” of sophisticated policy evidence (Sanderson, 2002, p. 6).

In giving further consideration to these alternative perspectives on the production and use of evidence for social policy making, it is useful to draw examples from social policies such as employment and training in which there has been a comparatively large number of experimental and nonexperimental evaluations and a longer history with performance management activities. Large-scale experiments such as the National Supported Work Demonstration, the National Job Training Partnership Act (JTPA) Study, and the National Job Corps Experiment have been undertaken to produce evidence on program impacts and have drawn considerable attention from researchers, policymakers and other stakeholders, and performance measurement has been required in public employment and training programs in the U.S. since the early 1980s. Thus, with the accumulation of data and research over time, we have come to better understand what we can learn from different types of information and alternative approaches to gathering and
analyzing it and to more fully grasp the challenges and tensions embodied in these different approaches.

**Incompatibilities between knowledge creation and performance management: Examples from employment and training programs**

As discussed above, a judicious application of scientific research methods requires observation and analysis over time. In addition, the calculation of a program’s *impact* or value-added—based on the careful construction of the counterfactual state through random assignment or a comparably rigorous method—is typically a long and involved process that is undertaken with knowledge creation as a primary goal. However, most social welfare programs rely on short-term measures to regularly monitor program outcomes (without reference to a counterfactual) and produce performance reports.

For example, a common performance measure used in employment and training and welfare programs is the job entry (or entered employment) rate. The U.S. Department of Labor currently measures entered employment by dividing the number of adult participants who are employed in the first quarter after the program exit quarter by the number of adult participants who exit during the quarter. In addition, monetary bonuses are awarded to states depending in part on their performance as measured by the entered employment rate and other employment and earnings outcomes. Yet analyses of the relationship between measures of employment at program exit (and 13 weeks after program exit) and earnings *impacts* measured 18 and 30 months after random assignment in the National JTPA Study experimental evaluation have shown that these short-run outcome measures do not correlate with impacts (Heckman, Heinrich and Smith, 2002). In fact, some negative relationships have been identified, suggesting that policymakers’ efforts to regularly monitor and reward performance outcomes may come at the expense of promoting longer-term program impacts and efficiency.
Findings from the U.S. National Job Corps evaluation, an experimental study of youth who applied to the Job Corps program in 48 states in the mid-1990s, showed that incongruities may also exist between shorter-term vs. longer-term measures of program impacts (Schochet, 2006). Analyses of the program’s impact on earnings (at 12, 30 and 48 months after random assignment) showed that Job Corps participants’ earnings initially lagged behind those of the control group (of whom more than 70 percent also received similar services elsewhere). Job Corps participants’ earnings subsequently overtook those of control group members (after two and a half years), yet these impacts faded after four years. As a result, an earlier summary of findings that reported a positive net benefit of Job Corps to the federal government and public (Burghart et al., 2001) was contradicted in a later follow-up analysis showing that its costs to society exceeded the benefits by about $9,000 per participant (Schochet, 2006). Possibly most disconcerting for the Job Corps program managers, however, the experimental study also found that the performance rankings produced annually by the National Job Corps Office and used to identify high and low performing Job Corps Centers bore no relationship to impacts. That is, the program’s own performance calculations using administrative data on outcomes (e.g., completed GED, hourly wage, hours per week worked, and weeks employed in last 13 weeks) did not distinguish the effective from the ineffective programs.

The implications of these contradictory findings of alternative measures of program performance and impacts are not inconsequential for policymakers’ and the public’s understanding of program effectiveness. In the mid- to late 1990s, the National Job Corps office was under considerable pressure to produce evidence of the program’s superior effectiveness, given that the average per participant cost of $16,500 was about five times the cost of some of its youth training program competitors, such as Strive and America Works. Many of the Job Corps
programs’ competitors reported performance outcomes such as their job placement rate and average wage at placement that looked comparable to or better than those reported by the Job Corps training centers. Yet Job Corps Center staff and national program officers were convinced (based their experiences working with youth at the centers) that their residential intervention with intensive counseling and vocational training components had other important effects on youths’ lives that might not be reflected in their short-term labor market outcomes. The experimental study of Job Corps (Burghart et al., 2001) in fact confirmed that short-term performance measures (including those that might be reported annually as required by GPRA) would be unlikely to accurately reflect all of the program’s benefits. At the same time, Schochet’s (2006) longer-term follow-up study of Job Corps negated the interim report findings (which had showed positive net program benefits), leaving one to wonder if policymakers who placed confidence in the experimental evidence were misled in voting to continue funding the program (which had been scheduled to sunset in 2000).

**What Should Count as Evidence?**

In attempting to advance both performance management and evidence-based policy agendas, governments have sent mixed signals regarding their expectations for the nature and quality of evidence, which has also fueled the debate over the fundamental questions: what should count as evidence, and should some types of evidence count more? For example, while the U.S. OMB’s PART guidelines strongly encourage the use of randomized controlled trials or quasi-experimental methods and a longer-term evaluation focus, OMB also accepts historical performance data, GPRA strategic plans, annual performance plans and reports, financial statements, and inspector’s general reports as evidence in rating these programs’ effectiveness. And despite these fairly broad definitions of allowable performance information and evidence,
the U.S. Government Accountability Office (GAO) recently reported (2005: 7) growing friction between OMB and federal agencies due to the different purposes and timeframes of PART and GPRA and their “conflicting ideas about what to measure, how to measure it, and how to report program results.” For example, OMB and agencies could not even agree in some cases on an appropriate unit of analysis (or how to define a “program”) for use in both budget analysis and program management. In response (contrary to the intent of these initiatives), some federal programs appear in practice to largely ignore the requests for more scientifically rigorous evidence and quantitative information on performance outcomes, with little consequence for their program ratings and subsequent budget allocations (see also Gilmour and Lewis, 2006).¹

Although random assignment experiments (and double-blind clinical trials) are widely viewed as the “gold standard” of methods for producing rigorous scientific evidence, these methods are not without controversy. Consider a study (released in July 2006) by the Cochrane Collaboration, which compiled the results of eight randomized controlled trials involving 3417 patients, some of whom participated in Alcoholics Anonymous (AA). AA consists of self-help groups that offer a 12-step approach and emotional support to help recovering alcoholics achieve abstinence. The study (Ferri, Amato and Davoli, 2006: 2) concluded that the available experimental evidence failed to “demonstrate the effectiveness of AA or other 12-step

¹ For example, in my recent examination of the 2006 PART assessments for 42 programs administered by the U.S. Department of Health and Human Services, I (with a research assistant) reviewed and coded for analysis the evidence provided by programs and used by PART analysts to determine “whether a program is meeting its long-term and annual performance goals.” Two of these PART questions specifically require evidence of program effectiveness: (Q. 4.4) Does the performance of this program compare favorably to other programs, including government, private, etc., with similar goals? and (Q. 4.5) Do independent evaluations of sufficient scope and quality indicate that the program is effective and achieving results? For question 4.4, only 12 of the 42 programs provided any evidence comparing the program to public and private alternatives; and of these, 10 of the programs provided only qualitative, descriptive information (no data or empirical analysis). Most (36 of 42) programs offered some information on their effectiveness and results, yet only 10 of these provided anything more than qualitative, descriptive information. In fact, only one of these 42 programs provided some empirical evidence of program effectiveness for both questions; still, only 15 were classified as “results not demonstrated” and only one program as “ineffective.” Just seven of the programs had been externally evaluated. According to Gilmour and Lewis (2006), these ratings also bear little relationship to agency budgets, once the role of political influences is accounted for in budget allocations.
approaches in reducing alcohol use and achieving abstinence,” compared with other available treatments. However, the study authors also found the randomized controlled trials lacking in a number of respects. The design of these studies did not allow for the identification of specific factors that influenced treatment success (the “black box” critique), and the authors also recommended that longer follow-up periods be used in future studies. In addition, they criticized the studies for their minimal attention to quality of life outcomes for patients and their families and recommended new qualitative research to further explore these outcomes.

The reactions of other researchers and AA organizations to the study findings were likewise critical (Bakalar, 2006: A7). A professor of psychiatry at McGill University, acknowledging the randomized controlled trial as “the gold standard methodology,” argued that this method “washes out a factor that may be important in potentiating AA’s benefits, namely patient choice and preference.” Elaborating on this comment, a researcher at the National Alcohol Research Center commented that randomized controlled trials cannot address the question of whether individuals who voluntarily seek treatment benefit, “but observational data can.” Alternatively, a clinical psychologist at Harvard University simply dismissed the evidence and suggested that although AA and other step programs “are not cure-alls, at a minimum, they help.” A general theme among these reactions is that the observations and practical experiences of many appeared to contradict the experimental findings, leading them to discount the evidence.

More than a decade ago, Heckman and Smith (1995) challenged the common wisdom that randomized experiments should always dominate as the “method of choice.” Akin to researchers’ objections to the AA study findings, Heckman and Smith pointed out that an important assumption of experimental methods is that random assignment does not alter the typical process of selection into the program or intervention, so that those who participate do not
differ from those who would have participated in the absence of the experiment. Alternatively, nonexperimental methods aim to model the process by which program outcomes are produced, including the relationship between the process of selection into the program and the production of outcomes. In effect, both approaches require some assumptions, although those favoring experimental methods argue that randomized experiments involve more plausible assumptions and produce more reliable results. Heckman and Smith, in turn, counter that nonexperimental methods have the potential to provide much richer information about why a program works or does not work (not just whether it works or not).

A recent study by Greenberg, Michalopoulos and Robins (2006), acknowledging the consensus in favor of experimental methods, posed a fundamental question: do experimental and nonexperimental methods yield similar conclusions about program effectiveness? As they explain, if the two types of methods produce similar results, this would increase our confidence in using nonexperimental methods to broaden our understanding of the effects of programs; and if not, nonexperimental evaluation findings should be accorded less weight. Using a “between-study comparison” method and drawing from numerous (31) experimental and nonexperimental evaluations that have been conducted of public training programs, they pooled the estimates of program effects from these evaluations and assessed how study design, population and site-specific factors affected impact estimates. Although their results differed somewhat for the different subgroups included in these evaluations (men, women and youth), Greenberg, Michalopoulos and Robins found (particularly for women) that after adjusting for differences in program, participant and site characteristics, differences between experimental and
nonexperimental methods were minimal, and there was little to gain in terms of the precision of impact estimates through the use of experimental methods.²

The above discussions of contradictory study findings and competing methodological camps raise more questions than they satisfactorily answer. For example, is it imperative to establish a hierarchy and/or guidelines for the sifting and weighing of social policy evidence? Should the findings of experimental impact evaluations, which compare a program’s effects to a counterfactual state constructed through random assignment, always be accorded greater weight than observational data, nonexperimental analyses, or routinely collected performance outcome information? Or is information on impacts of limited use if it does not enable policymakers to attribute observed impacts to particular components of the intervention (and to understand the processes that produced them)? In addition, if experimental estimates of program impacts at different time points change, is it always preferable to give greater weight to the later estimates or to wait until such information is available to make policy decisions? And even then, can we be confident that we have measured all relevant impacts associated with a particular intervention to make fair judgments about program effectiveness?

Frank Furstenberg’s (2003: 29; 2006) reflections on his 35 years of research on teenage child-bearing lend support to a cautious approach to interpreting and using research evidence in policy making. He acknowledged that many earlier studies in this area suffered from selection bias and thereby erroneously “reinforced the stereotype of the teenage parent as a perpetrator of poverty.” Emphasizing the importance of long-term, longitudinal studies, including his own in which young women were interviewed at length in multiple follow-up surveys, he commented: “Had I stopped at a five-year follow-up, I would have concluded something very different than I

² Of course, as they note, there are limitations to the generalizability of their research, and comparable studies in other policy areas would help to reach a more definitive conclusion.
did 12 years later.” At the same time, he lamented that the “politics of sex” and public perceptions of teenage childbearing, which have unduly influenced public policy in this area, have not kept pace with (and have largely ignored) the research on this topic.

Communicating the Evidence: Do More Technically Sophisticated Methods Put It Out of Reach of Policymakers and the Public?

Another concern embedded in the debate about what should count as evidence and how to weigh it is that the use of more technically sophisticated methods for producing information on policy or program effectiveness might put it out of reach of policymakers and the public (or lead it to be discounted). In this respect, the production of readily interpretable average program impact estimates through simple random assignment experiments is touted as another advantage of experimental designs. Heckman and Smith (1995: 93) reject this argument, however, stating that it “mistakes apparent for real simplicity” and that evaluation results can never be judiciously communicated in “sound bites.” For example, as we have come to know over time, the knowledge produced from these experiments—the average impact of the “offer to treat” or treatment on the treated—is sometimes limited, particularly if the treatment is complex and varied in design, administration and take-up, and the population eligible to receive it is likewise heterogeneous. In these cases, designing an experiment that would produce average impact estimates for relevant participant and treatment subgroups, including cross-site estimates that would take into account variation in implementation, becomes considerably more complex and costly.

If there are already a large number of studies or a base of evidence that has been collected across sites but neglected by those with little time or resources to consume it, one strategy for compiling, synthesizing and disseminating findings to policymakers and the public is through the
use of meta-analyses (such as the Cochrane Collaboration analysis of randomized controlled trials of AA programs discussed above). Meta-analyses offer the opportunity to “pass on collective wisdom” or a cumulative understanding drawn from the successes and failures of past interventions or initiatives, explains Pawson (2001: 5). However, even meta-analyses may involve some degree of selection (regarding the studies included) and important assumptions about the comparability of interventions, as Greenberg, Michalopoulos and Robins (2006) recognized in their analysis. In this regard, it is not surprising that meta-analyses may also fail to provide clear guidance to policymakers. For example, Dyke et al. (2006) describe the conflicting findings of multiple meta-analyses of welfare-to-work programs. One review (Barnow and Gubits, 2002) reported that longer-term, intensive training strategies were more effective than short-term welfare-to-work strategies; another meta-analysis of 27 experimental evaluations of 116 interventions reached the opposite conclusion (Greenberg, Cebulla and Bouchet, 2005); and a third review of experimental evaluations of 20 programs concluded that the most effective programs were those that combined more intensive training options with employment-focused strategies and tailored approaches to individuals’ needs (Bloom and Michalopoulos, 2001).

Similarly, Cook and Ludwig (2006) discuss a recent National Research Council expert panel report on gun violence that determined (for one important topic after another) that existing research provides little credible (and mostly inconclusive) evidence.

Another tool that has been advocated for more effectively communicating performance information to the broader public is the use of organizational report cards. As Gormley (2003) elaborates, organizational report cards differ widely in their technical sophistication, from the “scientific” report cards that apply advanced techniques (e.g., risk adjustment using statistical controls) to produce “grades” or performance ratings to the more “popular” report cards that
stress ease of comprehension over validity in condensing and presenting the information. Unfortunately, as Gormley’s research demonstrates, the public does not always discriminate between “good” report cards (based on valid, accurate and comprehensive measures of performance) and “bad” report cards (based on biased, inaccurate and/or incomplete information). And I am less optimistic than my colleague that over time, the public will acquire the specific knowledge and capabilities required to discern the good from the bad. Does this imply that there is a role or need for a “rater” of the report cards and other information or evidence used to inform decision making by policy makers and the public?

Who Should Judge the Evidence, and How Can High Standards for its Production and Use Be Upheld Under Accountability Pressures?

At the urging of the evidence-based policy movement and in the face of accountability demands, government agencies are increasingly taking on the role or function of translator or “sifter” of the evidence to facilitate its effective communication and use. For example, in the area of education policy, the U.S. Department of Education has acted vigorously in this regard, partly in response to criticisms that “change in educational practice more resembles change in fashion” and that “lip service” has been given to research in education (Coalition for Evidence Based Policy Report, 2002: 2). The No Child Left Behind (NCLB) Act of 2001 has also put considerable pressure on education leaders to develop and use experimental evidence through requirements that funds spent on Elementary and Secondary Education Act (ESEA) programs be expended on activities supported by “scientifically-based research” (defined by ESEA as including experimental and quasi-experimental studies, with a preference for the former). In 2002, the U.S. Department of Education established the What Works Clearinghouse (WWC) “to provide educators, policymakers, researchers, and the public with a central and trusted source of
scientific evidence of what works in education” (http://www.whatworks.ed.gov/). The far-reaching communication and dissemination of this information—i.e., systematic reviews of existing research that describe the effectiveness of replicable educational interventions—in a form that is user-friendly and allows for access to data is a primary goal of WWC.

In many respects, the WWC is a model for the approach to policymaking advocated by the evidence-based policy movement. In regard to the issue of who should judge the quality and reliability of evidence and information used and how it should be accomplished, the WWC has established a thorough and systematic review process that determines whether a particular piece of evidence is allowed and assigns it a quality rating. This is a very intensive process, involving the organization of a review team (including a principal investigator, project coordinator, and research analysts); the development of a protocol for the review that incorporates tailored evidence standards; an extensive search of the literature (including submissions from intervention developers and the public) that screens studies for their relevance and validity; an assessment of the causal validity of the studies and the assignment of ratings ("Meets Evidence Standards", "Meets Evidence Standards with Reservations" or "Does Not Meet Evidence Screens"); and a more in-depth review and production of a report on those studies that meet evidence standards (with or without reservations). Furthermore, all reports produced are submitted to extensive review, including the opportunity for researchers, intervention developers, and a WWC Steering Committee to review and comment on the reports.

Although the WWC invites the involvement of intervention developers and aims to reach a broad cross-section of the public with its evidence-based reports, the WWC’s strict standards for study design—requiring a randomized trial, regression discontinuity methods, or a quasi-

---

3 This is not coincidental, given that the U.S. Coalition for Evidence Based Policy helped the Department of Education to develop this strategy and the infrastructure for its implementation.
experiment with equating of pretest differences—will clearly leave out of consideration a
preponderance of studies and reports that draw on administrative data, qualitative investigations,
and much routinely collected performance information. In this regard, the WWC may be of
limited use to some educators in coping with the immediate accountability demands of NCLB,
including requirements for annual report cards (tests) of students’ academic proficiency,
measurement of teacher qualifications, and documentation of the yearly progress of specific
student groups (by race, poverty status, disability, etc.). In fact, these requirements have
contributed importantly to the increasing focus on student annual achievement test scores and
engendered an active debate over their appropriate use and potential for improving educational
performance and accountability.

Compromised standards? Examples from high-stakes performance management in
education and employment and training

Standardized tests have long been used in the U.S. (since the 19th century) to monitor the
performance of educators and individual student achievement. For the purposes of performance
management, test score data are relatively inexpensive to collect, particularly since schools
commonly rely on commercially published multiple-choice tests to evaluate student
achievement. The results of these tests are also easily interpretable by the many stakeholders in
public education. Accordingly, in the face of growing (high) stakes associated with testing—
minimum test score requirements for grade advancement or graduation, monetary rewards or
sanctions for educators and schools, and budgetary re-allocations based on performance—test
scores have become increasingly important for performance accountability and are drawing
greater attention from the broader public, too.

Koretz (2002: 755), however, marshals a strong case against placing substantial emphasis
on test scores, arguing that “scores on most achievement tests are not inherently meaningful or
“useful” and that the domains typically selected for testing leave many important goals of education unmeasured.\(^4\) In other words, they produce limited knowledge about the impact of schools on student outcomes. He points to two other important problems in using test scores to measure student proficiency: they are fallible due to measurement error, and perhaps more importantly, easily subject to manipulation (inflation and distortion). Koretz’s own study of a “moderate-stakes” third-grade mathematics test, with “coached” and uncoached versions administered to the same students in a single year (1990), showed clear evidence of sizable test score inflation due to “teaching to the test.”

Jacob (2005) adds that although the intent of these test-based accountability systems is to help identify schools and students that are performing poorly and to provide incentives and extra support for them to improve, both economic theory and practice have shown that individuals will respond by focusing on the most readily observable aspects of the multidimensional goals or tasks at hand. In their study of performance targets in the English public health care system, Bevan and Hood (2006) describe performance targets as relating to a domain of total performance, for which there are some good measures, some imperfect measures, and some unmeasured (residual) components. They demonstrate that when the imperfect measures (like test scores in education) are the ones more easily observed and gamed, the opportunity for manipulation of measures can become a grave problem. For example, Bevan and Hood found that in response to standards set to reduce hospital accident and emergency waiting times, hospitals cancelled operations and required patients to wait in ambulance queues outside the hospital until they were confident the patients could be seen within the targeted (four hour) time.

---

\(^4\) Koretz elaborates that tests are small samples of domains of achievement (some of which are easier than others to assess), and that improvements in test scores reflect mastery of the sample included in the test; these gains are only meaningful if they generalize from the specific test to (indicators of) mastery of the domain of interest.
In his study of Chicago Public Schools’ high-stakes exam system, Jacob (2005) finds sharp gains in students’ reading and math test scores that are not predicted by prior trajectories of achievement but are consistent with the incentives established by the system (e.g., lower-performing schools and eighth graders experiencing the largest gains). In a careful empirical investigation that includes analysis of particular test question items, he also shows that the test score gains did not reflect broader skills improvement (particularly among younger students) but rather increases in test-specific skills (related to curriculum alignment and test preparation), primarily in math. Jacob also explored whether teachers or other school officials acted strategically to increase their students’ performance on tests, such as shifting effort and resources toward subjects included on the tests (math and reading) or changing low-achieving students’ status (e.g., to special education or through retention) to exempt them from testing. Jacob uncovers some evidence of shifting resources across subjects and increases in student special education and retention rates, particularly in low-achieving schools, which leads him to conclude that high-stakes performance accountability demands encourage strategic responses from those under pressure to meet performance standards. His findings of gaming behavior are consistent with those from other education studies (see summaries by Jacob, 2005 and Radin, 2006) and research on strategic responses to performance requirements in other public sector programs, such as welfare and employment and training (Courty and Marschke, 2004; Heinrich, 2007).

In their study of gaming responses in the Job Training Partnership Act (JTPA) program, Courty and Marschke (2004: 24) went a step further to address the question of whether such strategic behavior contributes to welfare losses, that is, “a costly misallocation of resources.” Although Jacob (2005) was not able to determine if educators’ strategic responses adversely affected student learning, Courty and Marschke demonstrated empirically that earnings impacts
were lower for individuals who received training from agencies that engaged in strategic manipulation of trainee termination (release) dates. Equally disconcerting are the implications of their findings that JTPA program administrators manipulated administrative data—immediately entering positive training outcomes and postponing entry (in effect, altering termination dates) of individuals with poor outcomes—for the credibility and validity of these data that are used in measuring performance and program impacts.

Unfortunately, stories of data manipulation have proliferated almost as rapidly as the use of administrative and other performance data. Radin (2006), for example, discussed a report by The Education Trust and other case examples of states’ intentional under-reporting of high school dropout numbers. Jacob and Levitt (2003) produced empirical evidence of outright cheating on the part of teachers, who systematically altered students’ test forms to increase classroom test performance. Wiseman (2004) described cases of highly inflated performance in state Temporary Assistance for Needy Families (TANF) performance reports and obvious discrepancies in the unaudited reports that were used to allocate performance bonuses. And Bevan and Hood (2006) reported that the British National Health Service found evidence of deliberate manipulation or misstatement of figures in approximately 6,000 patient records. Bevan and Hood also uncovered findings in a parliamentary select committee report that showed that as a consequence of strategic cancellation and delay of ophthalmology follow-up appointments in order to reach waiting time targets, at least 25 patients had lost their vision over a two-year period.

Do the extremely serious consequences of these types of strategic responses and data manipulation activities for the impacts and efficiency of the policies and programs being evaluated call for fundamental changes in our performance management and evaluation systems?
By design, many performance management systems—e.g., GPRA, PART, and those of state and local government agencies in the U.S.—engage stakeholders in the tasks of producing evidence on outcomes/performance. For example, welfare case workers enter information on individuals’ compliance with work requirements and employment outcomes into administrative data systems; state agencies assemble and submit data on workforce investment training activities and performance outcomes to the U.S. Department of Labor, and educators administer standardized tests and collect other data (e.g., attendance) on student outcomes. Indeed, one of the advantages of administrative data are their coverage of entire populations of individuals accessing government services. Should policy makers and other evidence users such as these, who have a stake in the measured performance outcomes, have a role in the production of such information? Does this diminish its objectivity and credibility (particularly in light of reported data manipulations and misuses)? Or if policy makers are more likely to understand, value and use information that they have had a hand in producing, do the potential benefits outweigh the risks of misuse?

**Promoting and protecting the objectivity and credibility of evidence**

One approach to addressing the problems associated with misuse of data or mendacious reporting of outcomes and evidence is to involve independent third parties in oversight and monitoring of performance data collection activities. Accounting/auditing firms and government agencies such as the U.S. Government Accountability Office may effectively perform these functions if they are given full access to administrative and other data used in performance management and evaluation. Likewise, independent organizations may be contracted to verify data quality prior to its use in evaluation activities, such as the U.S. Department of Labor’s
contractual arrangement with Mathematica Policy Research to perform external validations of states’ Workforce Investment Act Standardized Record Data (WIASRD) files.

The usefulness of such oversight and monitoring activities, however, depends considerably on how open or overt are the processes for addressing problems or irregularities that might be uncovered in audits and other assessments. For example, despite persistent concerns about inaccuracies and possible deliberate mishandling of performance (WIASRD) data by states, the results of Mathematica Policy Research’s data verification activities have been kept confidential and restricted not only from public access, but also from senior managers in federal and state programs (Heinrich, 2007). In Wisconsin, one of the private firms hired by the state to conduct audits of case information and performance data used by agencies providing public welfare services under contract with the state was also found to be providing auditing and accounting services to one of the largest (and subsequently failed) agencies it was supposed to be monitoring, leading to accusations of cover-ups and data misuse and misreporting (Heinrich and Choi, forthcoming). And in her study of New York public welfare agencies, Sanger (2003: 39) described the state’s monitoring and oversight activities as “notoriously poor,” with backlogs in accounting and inadequate assessments of expenditures and performance that made it difficult for external parties or the public to evaluate the integrity of the data and performance reports.

Marston and Watts (2003: 158) remind us that policy making is a political process that injects values into decision making, and that stakeholders will “assert various forms of knowledge in the context of material and hierarchical power relations.” In this regard, they caution that in-house performance data collection and other evidence gathering activities “will become a means for policy elites to increase their control over what constitutes knowledge about social problems,” and will lead to a devaluation of other forms and sources of knowledge,
including the input of ordinary citizens. In the context of science policy, Rosenstock (2002) adds that as government agencies get increasingly involved in the production of information and evidence (the “science base”), they are more vulnerable to forces that aim to politicize knowledge or to use it to justify policy choices, even when the data may be inadequate.

In light of the limitations of oversight and monitoring activities that are controlled by the public agencies being evaluated, one might ask if a more “democratic” approach to the assembly, use and verification of performance information and evidence might provide a check against political or other strategic uses of this information. The alternative advocated by the WWC is to impose very high standards for both the nature of the evidence and the process of review and corroboration (led by “technical elites”), with goal of producing more accurate and objective knowledge about policy impacts that can then be shared with policymakers and the public. These are two very different approaches with important, value-laden implications for the future direction of performance management and evidence-based policymaking movements, and this brings me to the final (broad) question addressed in this paper: Is it possible (or desirable) to achieve a balance between processes or procedures that produce high quality, accurate performance information to be weighed and applied methodically in decision making, and those that facilitate public input and stakeholder representation with the goal of fostering democratic governance and accountability?
Is It Possible to Balance the Scrupulous Application of Rigorous Evidence with the Obligation of Democratic Governance to Facilitate Public Input and Weigh the Public Interest in Decision Making?

James Madison (the fourth U.S. president) was one of the earliest advocates of the promulgation and use of information by citizens to hold government accountable and advance democratic values:

"Nothing could be more irrational than to give the people power and to withhold from them information, without which power is abused. A people who mean to be their own governors must arm themselves with the power which knowledge gives. A popular government without popular information or the means of acquiring it is but a prologue to a farce or a tragedy, or perhaps both."

Access to information empowers the public to observe, challenge and evaluate government decisions and actions and also deters official misconduct, expalciates Vaughn (2005). In the U.S., in addition to constitutional guarantees that compel government accountability, open government laws including the Freedom of Information Act, Administrative Procedures Act, Sunshine in Government Act, Federal Advisory Committee Act, Ethics in Government Act and other state statutes and local ordinances require government entities to share or publish information and allow free debate, discussion and association. The Government Performance and Results Act of 1993 goes a step further, however, in requiring government agencies to publish information on their performance or the results of their activities. This involves much more than just opening up records to the public; it has frequently imposed the need for new data collection and more sophisticated analysis of data, in addition to the equally challenging precursor task of defining empirical measures of performance goals. In effect, in the context of these new requirements for performance measurement, promoting open and accountable government requires more than disseminating information on results; it is also essential to make transparent the processes and methods by which information on results is produced.
At the same time, if the public is going to usefully “consume” this information and contribute to improved decision-making quality, as contemporary public administration theory suggests (Callahan and Gilbert, 2005), does it need to be better equipped to understand the more advanced or complex methods that are frequently required to produce accurate performance information or evidence on impacts? Are there important tradeoffs in the use of such methods and public participation and democratic accountability?

Consider, for example, the use of multivariate statistical methods in performance analysis to adjust performance standards for factors that are outside the control of public managers (e.g., external influences such as economic conditions) and to deter gaming behavior that is more likely to arise in the face of uncontrollable risks (Courty, Marschke and Heinrich, 2005). In the U.S. JTPA program, performance standards adjustments were implemented using regression-based models (with the base model specified at the federal level) to lower performance requirements for agencies that delivered services to hard-to-serve groups and/or in depressed labor market areas. The Workforce Investment Act (WIA, the successor program to JTPA) discarded this more technical regression-based approach to performance standards adjustment, however, in favor of a purportedly more democratic process in which state and local officials could “negotiate” standards using information and approaches of their choosing, with the idea that they would be more responsive to citizen input and local needs. The consequences of these changes were undoubtedly unintended; the processes by which performance standards were adjusted under WIA were poorly documented and thus became more opaque to the broader public (Heinrich, 2004). Furthermore, empirical research (Heinrich, 2004; Trutko et al., 2005) showed that the negotiated standards failed to adequately adjust for economic risks, with the
result that a majority of states failed to meet their performance targets and were unduly at risk for sanctions.

The introduction of customer satisfaction measures to promote “customer responsiveness” was also a new feature of WIA and of other government reforms advocated under the “new public management paradigm” that aimed to adapt private sector management foci and techniques to the public sector with the goal of increasing performance and democratic accountability. The National Academy of Public Administration has explicitly called for a greater focus on customer or citizen surveys as a gauge of government performance (Van Ryzin, 2004), and the use of customer satisfaction measures as leading indicators of public sector performance is also proliferating in Europe (Workshop on Statistical Methods for Performance Analysis, Cassino, Italy, 2006). It is implicitly assumed in the use of these measures, explain Callahan and Gilbert (2005) and Kelly (2005), that citizens will recognize improved services and government performance and that this will be reflected in higher measured levels of customer (or citizen) satisfaction. Empirical research has confirmed, however, that performance and satisfaction may be conceptually distinct constructs in some contexts (Churchill and Surprenant, 1982). As Kelly elaborates, unlike private sector transactions, public sector service provision involves some degree of legal coercion (i.e., taxation) and may not always be equitable. Thus, regardless of service quality, citizens who are required to pay for services that they do not consume or do not want may report low levels of satisfaction.

And even among those who voluntarily seek government services, their own definitions or premises of value may differ from those of public managers who rely on internal, objective (or professional) measures of quality or value. Smith, Whalley and Wilcox (2006: 3) refer to these customers of government services as “lay empiricists” and comment that “they are not compelled
to follow canons of formal inference” in making assessments of their satisfaction and may instead rely on “folk theories” or may be hindered by poor or biased memories and data limitations. Smith and colleagues undertake a unique study in which they evaluate the relationship between self-reported outcomes among JTPA participants and experimental impact estimates of their labor market outcomes. They find that self-reported outcomes and satisfaction levels bear little, if any, relationship to the estimated impacts of employment and training services. Kelly (2005: 80) reviews the literature on the relationship between citizen satisfaction and objective measures of service outputs and finds mixed results, with many studies relying on “vague satisfaction or evaluation questions” that at best reflect “some unknown mixture of different aspects of service provision.” In this regard, it may not be problematic that research suggests that public managers tend to discount citizen evaluations as unreliable or erroneous. Alternatively, does our increasing reliance on performance outcome measures obligate us to search for a shared definition (between public managers and citizens) of a successful program outcome that can be used in performance evaluations to also promote democratic accountability?

**Hope for compatibility in evidence-based policymaking and performance management? A final example**

I conclude this discussion with a hopeful example in the area of education, where it appears that policy makers and citizens might indeed come to a shared definition of value and where the prospect of compatibility between the goals of evidence-based policymaking and performance management looks promising. Recognizing that government entities are often ill-equipped to effectively use the information they are required to collect for accountability and evaluation purposes, the Wisconsin Center for Education Research (WCER) has been working with government (public school) officials for more than seven years to help school districts develop the capacity to more effectively analyze and use data on students and schools. As
discussed earlier, the No Child Left Behind Act requires school districts to produce “report cards” of their performance (the percentage of students achieving proficiency) based on test scores, but NCLB does not track individual student performance. Educators have criticized the NCLB approach to performance analysis as inaccurate and unfair—as failing to identify the schools where students are learning more (or less)—and in the face of federal sanctions, they have become motivated to work with researchers to improve these evaluation systems (Dillon, 2006).

WCER is working with school officials in Wisconsin, Michigan and Minnesota to develop a multi-state, longitudinal data system that “will enable education stakeholders to conduct value-added and other diagnostic and policy-relevant evaluation research and engage in data-informed decision making,” with ultimate goal to improve student achievement (Thorn and Meyer, 2006). The basic concept of a “value-added” approach to performance analysis is simple and appreciated by educators: it measures the productivity or contributions of teachers and schools to student achievement. The analytical methods for producing evidence on value-added, however, are not uncomplicated; they adjust for student selection bias, track student achievement over time and allow for decay effects of interventions, and they explicitly address the issue of measurement error in student achievement. In this regard, a value-added approach to measuring school effectiveness satisfies the high scientific standards set by the evidence-based policymaking movement for producing evidence.

Yet WCER’s partnership with public schools is exceptional in that it has worked closely with school district officials (who are investing the districts’ own resources) to build internal capacity for data assembly and understanding of the value-added modeling system. In other words, although the generalized value-added model is a sophisticated statistical tool used in
school performance analysis, the collaborative approach to its development and application has made it accessible to the public officials who have the most to gain from its use. As Robert Meyer of WCER explains, district staff, with the support of researchers, are using the value-added system to measure teacher and classroom performance, to diagnose problems and develop concrete steps to improve performance, and to remove incentives to teach narrowly to test outcomes or focus on short-term gains. Furthermore, WCER and public schools are taking the next steps to make this information available to the broader public and key stakeholders such as parents (with essential protections for privacy/protection of personal data). Thus, it is an uncommon but encouraging example of a case in which the objectives of evidence-based policymaking—to base policies and practices on scientifically rigorous evidence—and performance management—to regularly use performance information to assess and improve government effectiveness and hold managers accountable for program results—appear to be compatible and complementary in practice.

Conclusion

Since the 1960s, the quality of scientific evidence and performance information produced to guide the improvement of social policy and program effectiveness has come a long way. The greater sophistication of analytical methods—experimental, nonexperimental and others—reflects a superior understanding of the significance of selection (into interventions) and the implications of both within-program and cross-site variations in implementation and management for accurate estimation of outcomes and impacts. Researchers and policymakers have also come to appreciate the importance of time in estimating and understanding policy effects. Evaluations conducted six months, two years, five years or ten or more years following an intervention may lead to very different findings, which implies that the choice of time point(s)
for measuring performance or impacts may be critical to the conclusions reached. And experimental studies do not elude this limitation. As the discussion above suggested, although they are widely viewed as the gold standard of scientific methods, not all random experiments are “golden” in terms of the value or usefulness of the information they produce for policymakers.

In order to satisfy performance measurement requirements in current legislation, policymakers frequently have little choice but to consider and use a mix of different types of information (from various sources) in producing performance reports. Although government guidelines vigorously call for more experimental evaluations and other rigorous methods of performance analysis, research has shown that in the absence of high-quality, readily available information on outcomes and under high-stakes pressures to demonstrate performance improvements, data collected for these purposes are more likely to be misused or manipulated in performance analysis. In light of these problems, some of the steps recommended by those advocating evidence-based policies—such as systematic and intensive reviews of information and evidence that would reject information that failed to meet some minimum standards of quality—merit more serious consideration and broader application in the public sector. Correspondingly, the methods and processes through which performance data are analyzed and the findings of monitoring and oversight activities should also be open to closer scrutiny by the public.

The “What Works Clearinghouse” and Wisconsin Center for Education Research offer two different models for raising the standards for evidence production and review: the former emphasizes the role of technical elites in evaluating the evidence and broadly disseminating the findings to the public, and the latter suggests a more intensive, participatory approach to
engaging policymakers in the production of evidence and closer collaboration between researchers and policymakers in promoting effective use of the knowledge produced. These alternatives clearly have different implications for the scope and nature of public involvement in knowledge production and performance management, and accordingly, for the means by which democratic accountability is advanced. It is also the case that both of these examples come from education, and the most appropriate role for policymakers and the broader public in these types of activities (data assembly, analysis, monitoring and oversight, dissemination, etc.) will likely vary from one policy area to another. The work of WCER makes clear, however, that achieving success in these efforts requires long-term investments in developing not only the technical and methodological capacity for such an undertaking, but also in building strong relationships based on mutual appreciation and understanding between researchers and policymakers. Ultimately, such longer-term, more intensive investments might also yield greater net benefits in terms of the effective use of social policy evidence and performance information.

Finally, it is important to acknowledge what may be one of the most disconcerting (but salient) discoveries of our far-reaching efforts to promote evidence-based policy making and improve government performance: that is, the more we have come to know, the more aware we are of how tentative, limited and sometimes erroneous the bases of our information and evidence are. Recognizing this, do we want to impose strong penalties for failure to achieve performance goals, such as the dismissal of educators under NCLB, the reorganization of workforce investment areas under WIA, or budget reallocations under PART? If their use continues, “high stakes” performance management systems need to incorporate ample buffers for errors and imprecision in data collection and analysis and allow time for careful case reviews before sanctions are applied or bonuses awarded. In general, we would be better served if our demands
for performance accountability were less focused on identifying which entities achieve “bottom-line” results and more on producing information that enlightens public managers as to how they can improve performance, as in the case of WCER.
References


Kelly, J. M. 2005. The dilemma of the unsatisfied customer in a market model of public
administration. Public Administration Review 65(1): 76-84.


