SCHOOL of
PUBLIC POLICY

The Intersection of Performance Measurement and Program Evaluation: Searching for the Counterfactual

Douglas J. Besharov
School of Public Policy
University of Maryland

This paper argues that the key to using performance management to improve public and private programs is to manage toward a program's "outcomes," not just its "outputs" or "impacts." This requires the use of performance measures that can accurately gauge the outcomes attributable to program participation, which, in turn, requires a comparison with nonparticipants, that is, a "counterfactual."

These are all concepts from the field of evaluation. Hence, the title of this paper.

I conclude with examples of how scientifically valid counterfactuals can be identified.

Ineffective programs

Many government programs do not seem to achieve their goals, most objectively evidenced by the results of high-quality impact evaluations which often find that they have little or no effects.

For example, from a careful, national evaluation, we know that the U.S. Job Corps program, which provides job training services in residential settings for youth ages sixteen to twenty-four, fails any reasonable cost-benefit test. Although the program increased earnings and reduced the criminal behavior of participants, nine years after initial random assignment, the benefits of the program to society were less than about $4,000 per participant purchased at a cost

of about $17,000 per participant.[1] (All dollars are in 2010 dollars unless otherwise indicated.)

Impact evaluations, of course, have many limitations, a point I return to below. No single impact evaluation is likely to provide a definitive answer about program effectiveness, and most are subject to intense and continuing controversy among partisans on both sides. Nevertheless, they are often our best evidence of program effectiveness and, as long as their uncertainty is factored into the policy analysis, can help guide decision making.

For, appearances can be deceiving. Take early childhood education programs. Since 1965, the U.S. Head Start program has served about twenty-five million children, at a total cost of about $145 billion.

It matters how children are raised, of course. Romulus and Remus were suckled by a wolf, and they founded a city that became a great empire. The rest of us, though, need much more care and nurturing to reach our full potential.

Figure 1

So, it makes admirable sense that a program that provides compensatory education and socialization for disadvantaged children should give them a boost in their later school years. Yet, in the U.S. at least, our flagship early childhood education programs seem unable to help disadvantaged children develop cognitively or socially. Repeated studies have shown that they fail to achieve the vitally important goals assigned to them. In other words, regardless of the efficacy of the idea, the program, as implemented under real world conditions, does not seem effective.

- *The Infant Development and Health Program* (operating in eight medical centers between 1985 and 1988) provided home visits, parental education, and early childhood education services to low-birth-weight, pre-term infants and their parents from the birth of the child until age three. The cost of the program was about $20,400 per child per year. The program was evaluated using a randomized experiment. Although there were initial gains in the children's IQ, the gains faded by age five and there were no other significant differences in the children's school performance, health, or behavior through age

[1]Peter Z. Schochet, John Burghardt, and Sheena McConnell, "Does Job Corps Work? Impact Findings from the National Job Corps Study," *American Economic Review* 98, no. 5 (2008): 1864–1886, http://pubs.aeaweb.org/doi/pdfplus/10.1257/aer.98.5.1864 (accessed June 20, 2011).

eighteen.[2]

- *The Comprehensive Child Development Program* (operating between 1990 and 1995) provided case management, parenting education, early childhood education, and referrals to community-based services to poor children under the age of one and their parents for five years. The cost of the program was about $19,000 per family per year (about $60 million annually). The program was evaluated using a randomized experiment. After five years, there were no statistically significant differences on the children's cognitive development, socioemotional development, and health, and no statistically significant differences in the parents' parenting behavior, parenting attitude, employment, educational attainment, and welfare receipt.[3]

- *The Early Head Start program* (operating from 1995 until present) provides child development, parenting education, child care, and family support services to low-income children under the age of two and their parents. The cost of the program is about $18,500 per child per year (about $700 million in direct appropriations annually).[4] The program was evaluated using a randomized experiment, following children who had enrolled in the program between 1996 and 1998 through fifth grade. Although there were initial statistically significant gains in the children's IQ and vocabulary, the gains were very small, were largely reported by the parents, and disappeared by the fifth-grade follow-up. There were no other statistically significant differences on a number of cognitive,

---

[2]The Infant Health and Development Program, "Enhancing the Outcomes of Low-Birth-Weight, Premature Infants: A Multisite Randomized Trial," *Journal of the American Medical Association* 263, no. 22 (June 13, 1990): 3035–3042; Jeanne Brooks-Gunn, Cecelia M. McCarton, Patrick H. Casey, Marie C. McCormick, Charles R. Bauer, Judy C. Bernbaum, Jon Tyson, Mark Swanson, Forrest C. Bennett, David T. Scott, James Tonascia, and Curtis L. Meinert, "Early Intervention in Low-Birth-Weight Premature Infants: Results Through Age 5 Years From the Infant Health and Development Program," *Journal of the American Medical Association* 272, no. 16 (October 26, 1994): 1257-1262; Cecelia M. McCarton, Jeanne Brooks-Gunn, Ina F. Wallace, Charles R. Bauer, Forrest C. Bennett, Judy C. Bernbaum, Sue Broyles, Patrick H. Casey, Marie C. McCormick, David T. Scott, Jon Tyson, James Tonascia, and Curtis L. Meinert, "Results at Age 8 Years of Early Intervention for Low-Birth-Weight Premature Infants," *Journal of the American Medical Association* 277, no. 2 (January 8, 1997): 126-132; and Marie C. McCormick, Jeanne Brooks-Gunn, Stephen L. Buka, Julia Goldman, Jennifer Yu, Mikhail Salganik, David T. Scott, Forrest C. Bennett, Libby L. Kay, Judy C. Bernbaum, Charles R. Bauer, Camilia Martin, Elizabeth R. Woods, Anne Martin, and Patrick H. Casey, "Early Intervention in Low Birth Weight Premature Infants: Results at 18 years of Age for the Infant Health and Development Program," *Pediatrics* 117, no. 3 (March 2006): 771-780.

[3]Robert G. St.Pierre, Jean I. Layzer, Barbara D. Goodson, and Lawrence S. Bernstein, *National Impact Evaluation of the Comprehensive Child Development Program: Final Report* (Cambridge, MA.: Abt Associates Inc., June 1997).

[4]This figure is higher than that given by the program, because it includes expenditures and contributions not included in the program's official reports. See Douglas J. Besharov, Justus A. Meyers, and Jeffrey S. Morrow, *Costs Per Child for Early Childhood Education and Care Comparing Head Start, CCDF Child Care, and Prekindergarten/Preschool Programs (2003/2004)* (College Park, MD: Welfare Reform Academy, August 2007), http://www.welfareacademy.org/pubs/childcare_edu/costperchild.pdf (accessed July 13, 2011).

behavioral, and health measures.[5]

- *The Head Start program* (operating from 1965 until present) provides education, health services, social services, and parenting education to low-income children ages three to five and their parents. The cost of the program is about $10,800 per child per year (about $7.1 billion annually in direct appropriations annually).[6] The program was evaluated using a randomized experiment, following children who entered the program in Fall 2002 through first grade. After the first year, children in the program group had experienced small statistically significant gains on some cognitive measures. By Kindergarten and first grade, however, those gains had disappeared and there were no statistically significant differences on a number of cognitive, behavioral, and health measures.[7]

[5]John Love, Ellen Eliason Kisker, Christine M. Ross, Peter Z. Schochet, Jeanne Brooks-Gunn, Diane Paulsell, Kimberly Boller, Jill Constantine, Cheri Vogel, Allison Sidle Fuligni, and Christy Brady-Smith, *Making a Difference in the Lives of Infants and Toddlers and Their Families: The Impacts of Early Head Start. Volume 1: Final Technical Report* (Washington, DC: U.S. Department of Health and Human Services, Administration on Children, Youth and Families, Commissioner's Office of Research and Evaluation and Head Start Bureau, June 2002), http://www.mathematica-mpr.com/PDFs/ehsfinalvol1.pdf (accessed December 30, 2002); and Cheri A. Vogel, Yange Xue, and Emily M. Moiduddin, Barbara Lepidus Carlson, and Ellen Eliason Kisker, *Early Head Start Children in Grade 5:Long-Term Follow-Up of the Early Head StartResearch and Evaluation Project Study Sample* (Princeton, NJ: Mathematica, December 2010).

[6]This figure is higher than that given by the program, because it includes expenditures and contributions not included in the program's official reports. See Douglas J. Besharov, Justus A. Meyers, and Jeffrey S. Morrow, *Costs Per Child for Early Childhood Education and Care Comparing Head Start, CCDF Child Care, and Prekindergarten/Preschool Programs (2003/2004)* (College Park, MD: Welfare Reform Academy, August 2007), http://www.welfareacademy.org/pubs/childcare_edu/costperchild.pdf (accessed July 13, 2011).

[7]Michael Puma, Stephen Bell, Ronna Cook, Camilla Heid, and Michael Lopez *Head Start Impact Study: First Year Findings* (Washington, DC: U.S. Department of Health and Human Services, June 2005), http://www.acf.hhs.gov/programs/opre/hs/impact_study/reports/first_yr_finds/first_yr_finds.pdf (accessed July 12, 2011); and Michael Puma, Stephen Bell, Ronna Cook, and Camilla Heid, *Head Start Impact Study: Final Report* (Washington, DC: U.S. Department of Health and Human Services, January 2010), http://www.acf.hhs.gov/programs/opre/hs/impact_study/reports/impact_study/hs_impact_study_final.pdf (accessed July 12, 2011).

Program improvement, not program dismantling

In so many areas of social welfare, we need to do better—and should be able to do so.

Some ineffective programs should be terminated. They are simply obsolete or inappropriate to the problem they are meant to address. For example, for over five decades, the U.S. Even Start program provided "family literacy" services (including adult education, early childhood education, and parenting education) to low-income families. During that period, it went through three evaluations, including two randomized experiments, and each time the conclusion was the same: it had no statistically significant effects on children's cognitive skills, child literacy, parental literacy, and parental education.[8] After two presidents in a row, one Republican and one Democrat, recommended closure, it was finally defunded earlier this year—but only as part of the politically charged budget reduction process.

More commonly, however, the program is not hopeless. Rather, its ineffectiveness seems to stem from a weak design or poor implementation. Thus, although there are limits to what job training programs can accomplish in a weak economy, too many job training programs train participants for jobs that no longer exist or provide the wrong training for jobs that do exist. This is especially true in high-tech areas where the needed job skills are in constant flux.

Figure 2

Sometimes it makes sense to defund such programs and start afresh, but the political power of vested interests usually prevents such decisive action. Complicating matters, many seek to serve important societal problems and so should not completely abandoned. In any event, as Ronald Regan famously said: "The closest thing to immortality on this Earth is a federal government program."

Performance management

This reality establishes the central importance of performance management, which I would define as a systematic effort to

[8]Robert St.Pierre, Janet Swartz, Beth Gamse, Stephen Murray, Dennis Deck, and Phil Nickel, *National Evaluation of the Even Start Family Literacy Program: Final Report* (Washington, DC: U.S. Department of Education, Office of the Under Secretary, January 23, 1995); Fumiyo Tao, Beth Gamse, and Hope Tarr, *Second National Evaluation of the Even Start Family Literacy Program: Final Report* (Washington, DC: Fu Associates, 1998); and Robert St.Pierre, Anne Ricciuti, Fumiyo Tao, Cindy Creps, Janet Swartz, Wang Lee, Amanda Parsad, and Tracy Rimdzius, *Third National Even Start Evaluation: Program Impacts and Implications for Improvement* (Cambridge, MA: Abt Associates, 2003).

increase a program's efficiency and effectiveness. You know many of the buzz words: "total quality control," "continuous product improvement," "results-driven government," "performance-based management," "governing for results," "outcome-oriented management," and so forth.

Strong leadership and management practices are key elements of performance management, but equally needed are measures of program "efficiency" and "effectiveness." Only with efficiency studies can managers know how much the program costs (in monetary, nonmonetary, and opportunity costs) and whether it can be delivered more efficiently. And only with effectiveness studies can managers tell how well a program is achieving its goals, what elements of the program need to be fixed, and what happens when the fixes are implemented.

Measuring outputs. With some notable exceptions (such as the U.S. No Child Left Behind education program), performance measurement systems tend to focus on "outputs," that is, on a program's operations or activities—their number, intensity, cost, and quality (the latter often measured against unvalidated good practice guidelines).

Measuring outputs is critically important, and is the basis of efficiency studies: cost, cost effectiveness, and comparative cost effectiveness analyses. And, of course, they are the basis of cost-benefit analyses. Nevertheless, except when the output *itself* is the objective (such as a vaccination where other research establishes the probable outcome and impact), outputs usually tell us very little about the program's actual accomplishments, that is, about program effectiveness.

Measuring effects ("outcomes" and "impacts"). As exemplified by the U.S. Head Start program, it is all too easy to think a program "works" because its facilities look impressive, or because those who have participated seem to do well afterwards. Thus, a well-run job training program, with well-qualified instructors, attractive facilities, and satisfied trainees may have no impact on the trainee's earnings (either real or potential) compared to those who do not go through the program: the trainees might not actually learn anything, what they learn may not help them get a job, an equivalent job may be obtainable without the training, and so forth.

Figure 3

[[Many American psychiatrists make a joke of this phenomenon by saying that a third of their patients get better because of what they do, a third don't get better regardless of what they do, and a third get better on their own.

]] Hence, to judge program effectiveness, the situations of those who participated in the program need to be compared to those who did not. In the field of evaluation, this is called the "counterfactual." *So, the first point I want to emphasize is this need for counterfactuals in performance measurement.*

Impact evaluations take too long

Many people assume that estimating program effectiveness through a rigorous counterfactual requires a full-scale impact evaluation, and, certainly, in the best of all worlds, one would want to mount an impact evaluation of, say, a job training program that follows participants through the program and then for a number of years afterward to see what "impact" the program had on their future employment and earnings—compared to nonparticipants.

Such full-scale impact evaluations, however, are often difficult to mount, are sometimes quite expensive, have limited generalizability, are subject to seemingly never-ending disputes about their methods and conclusions, and, most seriously, can take years to complete—taking much too long to be useful for performance management.[9]

It took, for example, more than seven years (ten years, if you include when the thirty-month impacts were released) to complete Abt's very fine evaluation of the Job Training Partnership Act (JTPA). The JTPA study found modestly positive results for adult men and women,[10] but negative earnings effects for disadvantaged male youths and no earnings effects for disadvantaged female youth.

These findings led Congress to cut JTPA's budget for youth programs by 80 percent. By the time the results were released, however, the JTPA's youth programs had been revamped, with, among other things, the creation of a separate youth program and targeted services to those with multiple employment barriers. But none of the changes were assessed by Abt before the youth program was all but eliminated.

Thirteen years later, we are only now halfway into an evaluation of its replacement, the Workforce Investment Act (WIA). Final results are not scheduled for release until 2015—four years from now. That's more than halfway through Barack Obama's second term, assuming that there is one. Does anyone expect him to wait until then before deciding whether to put more money in the program or to radically restructure it?

---

[9]See Douglas J. Besharov, "From the Great Society to Continuous Improvement Government: Shifting from "Does It Work?" to "What Would Make It Better?" *Journal of Policy Analysis and Management,* 28, no. 2 (2009): 199–220.

[10]Bloom, Orr, Bell, Cave, Doolittle, Lin, et al., 1997, 560. Average earnings impacts per enrollee over the 30-month follow-up period were $1,837 for adult women, $1,599 for adult men (both statistically significant), but they were not statistically significant for female or male youth, with the exception of male youth arrestees, who experienced a statistically significant loss of $6,804 according to survey data on earnings.

The amount of time that it has taken to conduct the WIA evaluation is not exceptional:[11]

- *The Head Start Impact Study* (initiated in 2000). The first grade results were only released in 2010, and the third grade results are expected later this year.[12]

- *The Moving to Opportunity study* (initiated in 1994). An interim report was released in 2003, roughly halfway through the data collection process.[13] Although the data collection was scheduled to be completed in 2009, the final ten-year follow-up results have still not been released.[14]

- *The Employment Retention and Advancement evaluation* (initiated in 1998). Final results for twelve of the sixteen sites were published in 2010, but the findings for the final four sites remain unavailable.[15]
- *The Building Strong Families project* (initiated in 2002). Interim findings were published in 2010 and data collection for the final report is not expected to conclude until later this year. [16]

- *The National Job Corps Study* (initiated in 1993). The four-year findings were available in 2000, but the nine-year findings only became available in 2008.[17]

For effective performance measurement, the feedback loop has to be shorter, and much faster.

---

[11]All the dates in this summary are based on when the contracts were awarded.

[12]Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation, "Head Start Impact Study: Overview," http://www.acf.hhs.gov/programs/opre/hs/impact_study/imptstudy_overview.html" (accessed June 9, 2011).

[13]Larry Orr, Judith D. Feins, Robin Jacob, Erik Beecroft, Lisa Sanbonmatsu, Lawrence F. Katz, Jeffrey B. Liebman, and Jeffrey R. Kling, *Moving to Opportunity for Fair Housing Demonstration Program: Interim Impacts Evaluation* (Washington, DC: U.S. Department of Housing and Urban Development, 2003), http://www.rwjf.org/files/research/Moving%20to%20Opportunity-fullreport.pdf (accessed June 20, 2011).

[14]Department of Housing and Urban Development, "Moving to Opportunity for Fair Housing," "http://portal.hud.gov/hudportal/HUD?src=/programdescription/mto (accessed June 9, 2011).

[15]Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation, "Employment Retention and Advancement Project (ERA), 1998–2011," http://www.acf.hhs.gov/programs/opre/welfare_employ/employ_retention/index.html (accessed June 9, 2011).

[16]Mathematica Policy Research, "Building Strong Families: Can Well-Designed Interventions Help?" http://www.buildingstrongfamilies.info/About/index.htm (accessed June 9, 2011).

[17]Peter Z. Schochet, John Burghardt, and Sheena McConnell, "Does Job Corps Work? Impact Findings from the National Job Corps Study," *American Economic Review* 98, no. 5 (December 2008): 864–886, http://www.aeaweb.org/articles.php?doi=10.1257/aer.98.5.1864 (accessed June 9, 2011).

Less can be more

For the purposes of performance measurement, I believe that the evaluation process can be responsibly foreshortened—under well-defined circumstances—and that the result would be substantially strengthened systems of performance measurement.

Using logic models to identify reasonable assumptions. Program "logic models" (or "theory of change models" or "outcome maps" and sometimes "chains of reasoning," "theory of action," or "performance framework") have become an increasingly popular way to identify and describe the various elements of program design, management, and evaluation. They systematically map the program or policy's key components and the predicted causal links between them, portraying a theory of change and how to measure related activities and accomplishments (see figure 4).

The key to the point I want to make is the distinction between the two categories of "effects"—and the causal relationship between them (as depicted in logic models), and the use of a logic model to make reasonable assumptions about probable impacts based on measured outcomes.

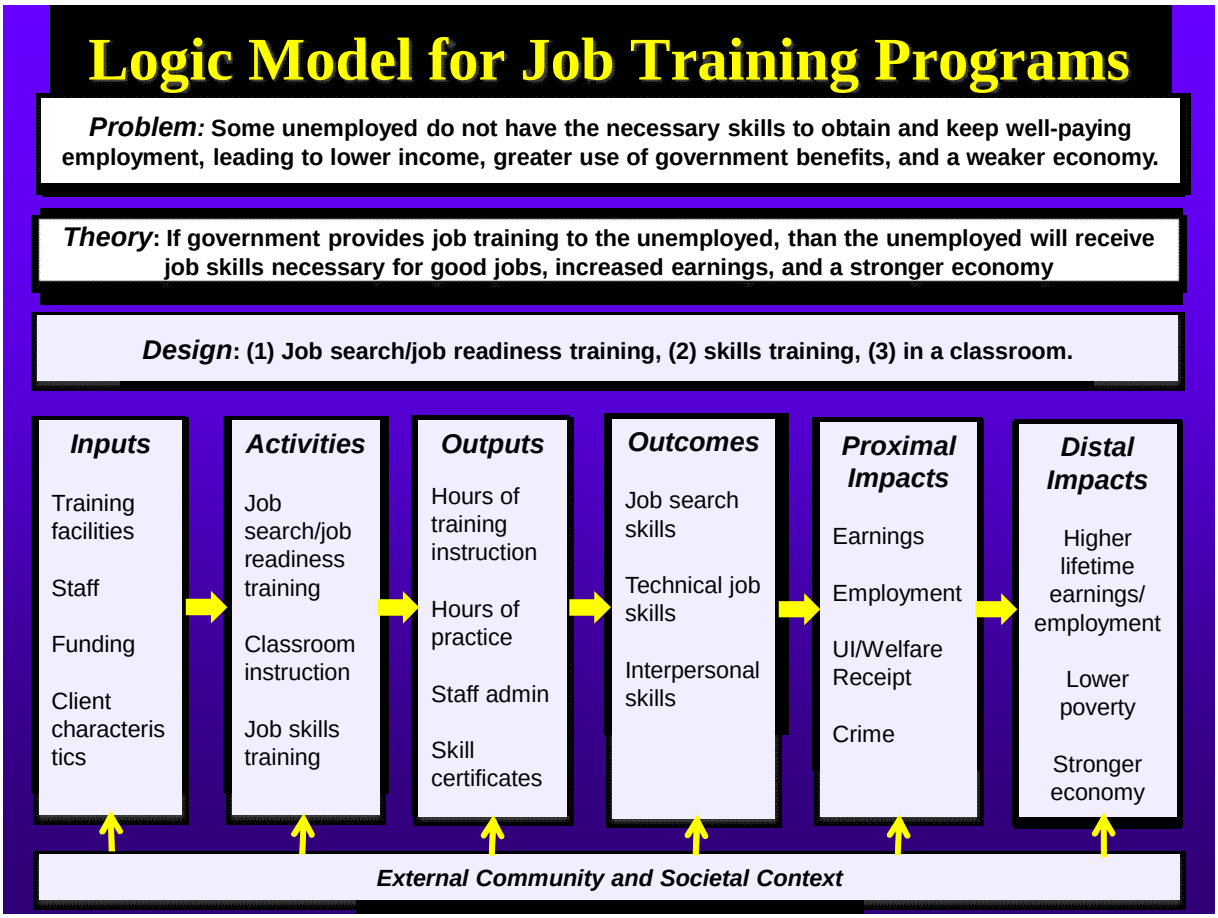
- "Outcomes" are the *immediate, often measurable, changes* to individuals, groups, or institutions caused by the program.

- "Impacts" are *the longer-term effects of those changes* on some aspect of the individual's future (such as earnings over some period of time), group, institution, or community.[18]

Explaining a logic model for remedial job training programs demonstrates what I am suggesting: An outcome might be an increase in job-related skills and the assumed impact would be greater likelihood of employment and higher earnings (for some period of time).

[18]Actually, this formulation combines "proximal impacts" with "distal impacts," being the near-term consequences of the outcomes on the individual or institution and the latter being long-term changes to the condition or situation of the individual or the community.

Figure 4



Ultimately, of course, measuring the actual, long-term impacts of a program is essential—because the desired immediate outcome may not actually have a longer-term impact on the individual, that is, the outcome may not make a difference. For example, greater skills will not translate into higher earnings if they are the wrong skills or jobs for those skills do not exist. It does little good to learn information technology (IT) skills on an obsolete computer that was last used by businesses in the 1980s.

As we saw, however, impact evaluations take a long time to conduct, largely because they need to follow participants and nonparticipants long enough to discern important impacts. (The fact that many are stand-alone demonstrations which must be organized, etc., also adds to the time it takes to complete them.) This is too long for effective performance management; in fact, it is too long for any kind of performance management.

Outcomes, on the other hand, can be gauged much more quickly—essentially as the program is operating, and require the collection of much less additional data. Thus, outcomes can be used to measure program effectiveness when the desired impact can be reasonably *predicted* to follow from the measured outcome.

In other words, *in carefully established situations*, a program's logic model may provide the grounds for a reasonable assumption—at least for management purposes—that a measured outcome is likely to have the desired impact (or not)—always remembering that this is only an assumption.

When the desired impact is reasonably predicted to follow from the measured outcome, we can assume that the program is working as intended . We assume, rightly in many cases, that certain skills (such as the ability to read) are essential in modern labor markets, and so, teaching illiterate adults to read is assumed to be a productive outcome. Thus, K-12 education rightly uses gains in reading and math ability as performance measures (subject, of course, to the question of value added).

The same can be true for learning higher-level skills (such as how to use a specific and complicated piece of equipment or the skills of a specific profession, such as registered nursing). Note that there must be an objectively determined and sufficient increase in skills—for this is the valid performance measure of the desired outcome. (Two necessary caveats, of course, are that equivalent jobs may be available that do not require such skills and that the acquisition of the skills may not result in the participant actually obtaining a job.)

Conversely, *when the program does not seem to produce any desired outcomes, it is unlikely to be working as intended.* If the program has not had any measurable effects on the participants, how can there be any long-term impacts (unless, as is often argued, the wrong outcome variables are being measured)? Defenders of Head Start programs that seem to make no improvement in children's skills, etc., call this absence of measurable outcomes but the presence of an apparent impact a "sleeper effect." That's possible, of course, but surely uncertain ground for a manager, let alone a policy maker. (In limited circumstances, the lack of change may represent the program's effect of preventing a deterioration in skills—which is why the counterfactual is often crucial to interpreting apparent outcomes.)

[[Inversely, *when an easy-to-measure impact has no other apparent cause except the program, it is reasonable to assume that the program is working as intended.* Here, for example, the trainee obtains the job, and there is no other plausible explanation except for the acquisition of skills from the program—because the job, say being a registered nurse, requires just what the program provided. Note that, in the narrow case given, the counterfactual is assumed. (This leaves open the question of the comparative effectiveness of the program compared to others serving the same purpose.)
]]

So, this is my second major point: *Carefully applied (which is often not the case), a measured outcome coupled with a logic model's theory of change—often buttressed by other evidence—can serve as a more timely and more useful performance measure than a formal evaluation of long-term impacts.*

Before leaving this subject, I would note that there are times when one need not even measure outcomes—that is, when a programs "outputs" imply its "outcomes."

- *When there is no output, so that no positive outcome can be reasonably predicted.* The program cannot be responsible for any change in the trainee's subsequent earnings if it did not actually provide services. For example, a number of the training programs in the 1980s Minority Female Single Parent Demonstration closed their doors to new participants, even though they were in the program group.[19] Participants that received no services from the program could not realistically be expected to benefit from it.

- *When the output itself (such as a vaccination) is sufficiently suggestive of a likely outcome (immunization).* A child is vaccinated (an "output"), past studies tell the likelihood of immunization (the "outcome") and, hence, the likely reduction in the disease ("impact").

[[Similarly, if a diploma, certificate, or license reliably signals the acquisition of certain skills, although technically only an output, it can be assumed to reflect what the trainee learned. (In effect, the qualification test for the diploma serves as the outcome measure, with the counterfactual being an assumption that the participant did not have those skills or knowledge before the program.) Of course, this assumption does not apply to degrees with little credibility. For example, many remedial programs award a General Education Development (GED) Diploma, but the clear research evidence is that they have no significant impact on employment or earnings.
]]

- *When the output is produced at a prohibitively high cost, so that regardless of the likely outcome, it does not meet a cost-benefit test.* Perhaps some objectives are priceless, as a U.S. television commercial claims about credit cards. Generally, however, there is a limit to what we will spend to achieve ceratin results, even saving a life. For present purposes, it suffices to mention this issue.

But, again, all these are restricted circumstances, and are easily misused (deliberately and not) by program managers—as well as politicians and program advocates. That is what makes attributing causation, that is, identifying a valid counterfactual, so centrally important in performance measurement.

---

[19]Alan Hershey, *The Minority Female Single Parent Demonstration: Process Analysis of Program Operations* (Princeton, NJ: Mathematica Policy Research, November 1988).

Identifying the counterfactual (or, how to measure outcomes)

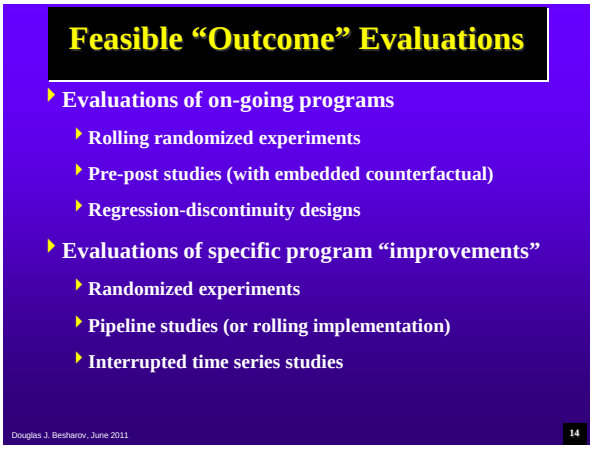
As mentioned, measuring outputs does not require a counterfactual, although comparative cost and quality are separately important. But with the exception of the narrow circumstances such as those described above, measuring program outcomes *requires estimating a counterfactual*.

How to estimate the counterfactual?

When administrative or program data for participants *and nonparticipants* are available, it is tempting to use statistical analyses (such as correlational, ordinary least squares, fixed effects, random effects, logit, and probit) and comparison group studies (such as simple differences studies, matching studies, propensity score matching studies, and difference-in-differences).

Unfortunately, such statistical analyses are plagued with unresolvable questions of selection bias and, despite the best intentions of the researcher, there is no way to know if the important observed and unobserved differences between participants and nonparticipants have been satisfactorily controlled for. They are also difficult to judge objectively—because the researcher's selection of variables and mode of analysis cannot be evaluated without actually replicating the study. As a result of this uncertainty, writes Larry Orr, such studies "inevitably shift the debate from substance to method."[20]

Figure 5



Randomized experiments are the "gold standard" of program evaluation. All other things being equal, they are the most likely evaluation methodology to achieve strong "causal validity," that is, they are best at determining the extent to which causality can be established between the intervention and the outcome (and/or impact) of interest.

This strength in causal validity, however, usually comes at a cost of "generalizability," that is, the extent to which

[20]Larry Orr, Stephen H. Bell, and Jacob Klerman, "American Lessons on Designing Reliable Impact Evaluations, from Studies of WIA and Its Predecessor Programs," (presentation, What the European Social Fund Can Learn from the WIA Experience, Washington, DC, November 7, 2009), http://www.umdcipe.org/conferences/WIAWashington/Presentations/Orr%20Bell%20Klerman%20-%20EC%20impact%20paper.ppt (accessed June 23, 2011).

the evaluation findings can be applied beyond the specific program sites studied. Given this and the other trade-offs involved in running randomized experiments,[21] alternate approaches—when carefully and appropriately applied—can also be used to determine if the training program seems to be "working." These include: pre-post studies (with embedded counterfactual); regression discontinuity studies, interrupted time series studies, and pipeline(or rolling implementation) studies.

For simplicity of presentation, I will separately discuss evaluations of on-going programs and of specific program "improvements."

Outcome evaluations of on-going programs. Here, the issue is what effect the existing program is having on participants compared to equivalent nonparticipants.

- *Rolling randomized experiments*. Larry Orr, among others, has proposed creating a rolling control group by delaying services (for the time it would take for outcomes in the program group to be measurable) to a randomly selected group of program applicants. The applicants would later be rolled into the program group, with a new control group created from new applicants.[22]

- *Pre/post studies.* These studies compare individuals or other units of analysis to themselves once at some time before and once at some time after the initiation of the intervention. A simple pre/post is based on the assumption that the individual would not have gained knowledge without the program, such as a calculus class.

  A properly normed test takes into account the possibility that other factors (such as maturation) are the cause for the difference between the pretest and the posttest (by providing a predicted posttest score independent of the program). Examples of normed tests are the U.S. National Assessment of Educational Progress (NAEP)[23] and the OECD's Programme for International Student Assessment (PISA).[24]

- *Regression Discontinuity Design studies* can be used with programs that have a

---

[21]Other problems include: denying services to those who are needful and perhaps eligible, difficulties in randomizing correctly, sometimes greater cost, problems with drop-out/no-shows and attrition, and substitution (when the control group receives similar or the same services).

[22]Larry L. Orr, personal communication with Douglas J. Besharov, 2009.

[23]U.S. Department of Education, National Center for Education Statistics, "National Assessment of Educational Progress (NAEP)," http://nces.ed.gov/nationsreportcard/about/ (accessed July 12, 2011).

[24]Organisation for Economic Co-operation and Development, "Programme for International Student Assessment (PISA)," http://www.pisa.oecd.org/pages/0,2987,en_32252351_32235731_1_1_1_1_1,00.html) accessed July 12, 2011).

"continuous" eligibility threshold, that is, one that is a numeric value on a scale ("cut-off score"). Such "cut-points" include income, birth dates, test scores, and other numeric rankings. The cut-off score is used to generate a program and nonprogram (comparison) group. For example, William Gormley used the Tulsa school district's September 1 cut-off for enrollment into prekindergarten to compare children who were able to enroll in pre-K to those children who were required to wait another year to enroll.

Outcome evaluations of specific program "improvements". Here, the issue is what effect a change in the existing program has on participants compared to equivalent nonparticipants.

- *Randomized experiments.* The idea is to randomly assign participants to the program variation or not, with the comparison being between the effect of the current system versus the change. For example, the Danish government has mounted a number of experiments where a randomly selected group of recipients are subject to different work first requirements and services.[25]

  Jacob Klerman of Abt Associates recommends a variation of this approach: randomly assign *sites* to a program group and a control group. This can provide greater generalizability if there are a sufficient number of participating sites and high-quality administrative data are available or can be collected.[26]
- *Pipeline studies* generate a program and nonprogram group based on variations in the timing of the intervention on the target population. For example, a program might be phased in over time due to funding issues. Sites that have received the program could be compared to sites who have not yet received the program but will in the future, assuming similarity across sites.

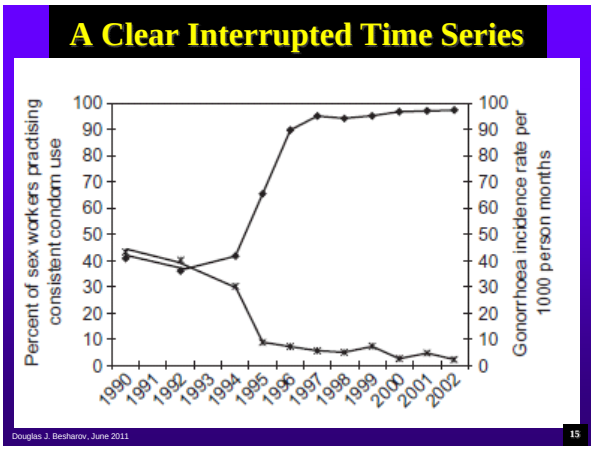
- *Interrupted time series studies* compare individuals, a changing population of individuals in the same program, or other units of analysis to themselves over an extended period of time before and after the program changes to assess its effect. For example, in Singapore, the government implemented a program to provide sex workers in brothels with education and training on sexually transmitted diseases, condom use, and negotiating techniques with the clients. After implementation of the program, condom use among newly recruited sex workers increased from about 40 percent to about 95 percent and the gonorrhea incidence rate dropped from 40 per 1000 months of sex work to about 5 per

[25]Stig Norgaard, "From Welfare to Work: The Danish Case," (presentation, Association for Public Policy Analysis and Management conference, Los Angeles, CA, November 6–8, 2008).

[26]Jacob A. Klerman, "Performance Management Systems and Evaluation: Towards a Mutually Reinforcing Relationship," (paper, Improving the Quality of Public Services: A Multinational Conference, Moscow, June 27-29, 2011).

1000 months of sex work.[27] (See figure 6.)

Figure 6



Motivating action

Before closing, I would like to raise what is to the most vexing obstacle to the implementation of this approach to performance measurement that I am proposing: the fear on the part of providers and advocates that rigorous evaluation will find that their programs are ineffectual (whether fairly or not), thus further undermining public and political support for them. Their all too common response to performance monitoring is, I am sorry to say, to circle the wagons—and oppose even reasonable performance measures, or manipulate the ones that are established.

What to do?

First, the fear of an unfair evaluation is based on sad, but too often valid, experience. Admitting that a program has weaknesses (let alone that it "does not work") can open to budget cutting. Hence, researchers and academics should work to make the techniques they use as transparent and as reliable as possible. (Here, the internet might prove extremely helpful if the agency could post information about how the performance measures were developed and what they mean.)

[27]Mee Lian Wong, Roy Chan, and David Koh, "Long-Term Effects of Condom Promotion Programmes for Vaginal and Oral Sex on Sexually Transmitted Infections Among Sex Workers in Singapore," *AIDS* 18 (2004): 1195–1199.

Second, the opposition of providers and advocates to performance measures is encouraged when there is no expectation of explicit accountability for results. Hence, policymakers and program administrators should work to create incentives, large and small, that encourage service providers to measure and publicize their performance.

The Obama administration's performance management effort, the "Quarterly Constructive Review Process," seems to address both of these concerns. The Office of Management and Budget (OMB) has instructed all federal agencies to hold quarterly reviews of agency priority goals set in the FY 2011 budget. According to Shelley Metzenbaum, the Associate Director for Performance and Personnel Management, who has primary responsibility for the effort:

> Discussions during these meetings should be guided by analyses of performance and related (e.g., problem characteristics, employee viewpoints, cost, agency skills, delivery partner capacity) data and evaluation findings relevant to the goals being discussed. They should focus on progress toward desired outcomes, explore the reasons why variations between performance targets and actual outcomes occurred, and prompt quick adjustments to agency strategies and action when needed.[28]

The process is roughly derived from models of performance management often grouped under the term "PerformanceStat."[29] Examples of this model include the New York Police Department's CompStat,[30] the State of Maryland's StateStat,[31] and Baltimore's CitiStat.[32]

After the quarterly reviews, agencies are to upload their output and outcome data to a

---

[28]Shelley H. Metzenbaum, *Performance Improvement Guidance: Management Responsibilities and Government Performance and Results Act Documents* (Washington, DC: Office of Management and Budget, June 2010), http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-24.pdf (accessed July 12, 2011).

[29]See Robert D. Behn, "PerformanceStat as a Search for Strategic Evidence," (paper, Tenth National Public Management Research Conference, cColumbus, Ohio, October 1–3, 2009), http://www.pmranet.org/conferences/OSU2009/papers/Behn,%20Robert%20D.%20PerformanceStat%20as%20a%20Search%20for%20Strategic%20Evidence.pdf (accessed July 13, 2011).

[30]See Dennis C. Smith and William J. Bratton, "Performance Management in New York City: Compstat and the Revolution in Police Management," in *Quicker, Better, Cheaper: Managing Performance in American Government* (Albany, NY: Rockefeller Institute of Government, October 2001):453–482, http://www.rockinst.org/pdf/program_management/2001-quicker_better_cheaper_managing_performance_in_american_government.pdf (accessed July 13, 2011).

[31]See State of Maryland Office of the Governor, "StateStat," http://www.gov.state.md.us/statestat/index.asp (accessed July 13, 2011).

[32]See Robert D. Behn, "The Core Drivers of CitiStat: It's Not Just About the Meetings and the Maps," *International Public Management Journal* 8, no. 3 (2005): 295–319.

new government website, www.performance.gov. (Currently, the website is only available to government agencies, but OMB plans to make it publically available.) OMB reviews the data and works with senior officials at the agency to create a priority follow-up list to be reviewed at the next quarterly review.[33]

It is too early, of course, to judge the impact, oops, outcome of these efforts, but they are much more likely to change agency behaviors that the Bush administration's Performance Assessment Rating Tool (PART) system, which tended to focus on rigorous evaluations of long-term impacts.

[33]Shelley H. Metzenbaum, "Building a High Performance Government: The Obama Administration's Performance Management Approach," (presentation, Executive Leadership Conference, Williamsburg, VA, October 25, 2010), http://www.actgov.org/knowledgebank/documentsandpresentations/Documents/Executive%20Leadership%20Conference/Track%204%20-%20Shelley%20Metzenbaum%20intro%20presentation.pdf (accessed July 12, 2011).