What kind of evidence in Evidence-Based policy?

Carol Hirschon Weiss
Harvard University

People have been talking about basing policy on evidence for over a century, and we can trace the ideas even further back, probably to Francis Bacon. But to actually do it -- to take evidence into account in policy and practice -- would mark a change from most of prior history.

Today I'm going to talk about four topics: (1) a history of empirical research and theorizing on evidence-based policy; in earlier times, we called it "knowledge utilization"; (2) why evidence does not always prevail, and why it shouldn't; (3) kinds of evidence social scientists believe decision makers should consider; (4) finally, some cautionary notes.

History of Research on Knowledge Utilization

Let me start with my own history on the subject. I began doing evaluations in the late 60s and early 70s. The U.S. government, for the first time, was giving substantial sums for social science research and for evaluation. This period saw the start of the War on Poverty, an attempt to bring poor people into the mainstream of society. Many program and policy initiatives were launched, from health and education to job training and community development. Social scientists were playing a visible role in planning the programs and later in evaluating them. It was a period of high expectations, a heady time, when many of us believed that a rich country could for the first time arrange its affairs so that the poor were no longer with us.

I was asked to do an evaluation of one program that had been funded by the government in Central Harlem in New York City. Essentially it was a domestic peace corps program, bringing black college students from colleges in the south to work with social agencies in Harlem. I did an extensive evaluation. With an assistant, I designed and carried out a three-year study. The craft of evaluation was in its early days, and the study took lots of effort, lots of time, and lots of heavy thought. We produced a long report, with an executive summary and made oral presentations to agency staff. We sent 40 copies to Washington. We had found out important things about the response of youth to training, their placement in existing agencies, and the pluses and minuses of the services they provided. After the report went out, I sat back and waited for a response -- some indication that people were paying attention to the findings.

There was nothing. Not a word, not a whisper. We might just as well have dumped the reports into the river. Why should that be? Why weren't people in the local agency or in Washington interested in what we'd found out? Why had the government spent all the money and inconvenience of having us around if they had no intention of listening to what we had to say? That's when I became fascinated with the subject of research use.

I was not alone in my concerns. Many social scientists began publishing articles about the neglect of their research -- a vast collective moan. Disenchantment was setting in. Social scientists, who had expected to help improve policy and at the same time become important in the process, were on the policy sidelines.

In my case, the neglect of the research findings was not hard to understand. The local agency administering the program was coming to the end of its grant. There was no federal money available to fund its continuation. The local staff wanted the agency to survive, so they began canvassing other government sources to learn what kinds of programs were going to be funded. They were willing to head off into new directions if that would ensure the agency's continuation. News about the past program was of little interest.

In Washington our report was relevant. A national domestic peace corps program was being activated, and its implementers might have profited from knowing some of our findings. But they were a new agency, and they were striking out in new directions. They didn't want to follow in the tracks of old
War on Poverty programs. In fact, the split between the agencies was so wide that it is unlikely that anyone at the new agency ever even heard about our report.

Meanwhile the federal government was pushing on with research and evaluation, setting up offices of R and E throughout bureaucracies -- Labor, Justice, Health, Education, Welfare, Housing, and many other departments and independent agencies. Bureaucrats were expending strenuous efforts to attract good researchers, developing RFPs (requests for proposals), reviewing proposals, demanding regular progress reports, releasing summaries, setting up conferences, and undertaking many tasks that suggested a serious commitment to research. Why were they doing all this, at considerable expense, if agencies were going to revert to their usual ruts at the end?

A number of us began doing systematic empirical studies to find out what was going on. We were not the first social scientists to become intrigued with the subject. Earlier several pioneers had been looking into knowledge utilization, as the topic was being called. Let me mention a few of these important figures: in the U.S., Ronald Lippitt and Paul Lazarsfeld; in the U.K., Albert B. Chernen. Chernen in particular had a fine awareness of the phenomena surrounding the use of knowledge for policy (e.g. Chernen 1979). He wrote, for example, that social scientists brought a sophisticated understanding of the place of knowledge in public life when they wrote about the sociology of knowledge but reverted to simplistic linear thinking when they confronted "applied research" and the attempt to move research into the policy sphere.

In the late 1970s and early 80s Nathan Caplan at the University of Michigan undertook a study of federal officials' use of research. In fact, the Institute of Social Research at Michigan established a center called CRUSK, the Center for Research on the Utilization of Scientific Knowledge. I conducted a study of federal, state, and local officials in mental health and their likely use of research supported by the National Institute of Mental Health. Along with a few others, our studies found that most research and evaluation findings did not form the basis for policy. But we also found a more diffuse use, which I called enlightenment. If research did not lead to changes in policy or practice, it did influence the climate of informed opinion. It sometimes led to changes in the aspects of an issue that people defined as important and those that declined in salience. It punctured a lot of taken-for-granted assumptions. Evaluation was particularly good at undermining accepted myths. It brought new ideas to the fore. It changed priorities. It highlighted new understandings about why things worked as they do. Not always, of course. Much research slipped through the cracks and was never heard of again. But enough research percolated into the policy sphere enough of the time for officials to value research and want to keep supporting it. It was becoming a source of news about the world out there, a kind of continuing education for officials.

Information from research and evaluation seeped into the policy arena through a number of channels. Usually the information did not come directly from reports or even report summaries. People heard about the key conclusions from conferences, conversations with other staffers, through stories in the mass media, from the agency grapevine, through informal networks of policy wonks or networks of professionals, personal contact with social scientists (not necessarily the authors of the findings of interest), from training sessions, or from friends and even their college-going children. Over time, information became known, and in some cases it became the new accepted wisdom.

The story was more positive than many disenchanted social scientists expected. Many had anticipated a bleaker landscape. The idea of enlightenment gave them a sense of hope, a sense that they were not singing in the wilderness. And study after study, in country after country, found much the same story.

But why were governments not making more direct use of research, especially the research they commissioned and paid for? Research and evaluation studies often found that public needs were going unmet, expensive programs were not making major inroads against human misery, promising policy and program ideas were withering. We could find explanations in particular cases, as I had found in my study in Central Harlem. But why was the neglect of research so pervasive?
Then came a new emphasis on the use of research in policy making, evidently stimulated by events in the U.K. There was even a new term, a good term, “evidence-based policy.” People could oppose the use of research. Research can be seen as ivory-tower, the province of eggheads. But who can oppose evidence-based policy? Maybe better times lie ahead. Maybe we are entering an era when policy makers pay greater attention to the import of data, generalizations, and theories.

The Competition to Evidence

Nevertheless, good information, however important and however central to theories of deliberative democracy, is only one input into the making of policies. As rationalists (and as self-interested purveyors of information) we like to think that good information should be the bedrock on which policy is built. But evidence is not the only contender for influence. A number of powerful competitors vie for influence in the policy process.

Some researchers dismiss these rival forces with the umbrella term "politics," as in "It's all politics." Politics becomes a pejorative term. But I have great respect for politics. It is the system we use to run our democracies. Politicians do an extraordinarily difficult job that is vital to the well-being of our nations. They have to reconcile the needs and wants of heterogeneous publics and organizations, budgetary limits, and considerations of feasibility given existing talents and resources. At the same time, they have to satisfy voters, while acting on the basis of their own principles and judgment. I readily grant that many politicians are venal or incompetent and that some politics is sordid. But I don't believe we should damn politics as the big bad wolf impeding the use of greater rationality in policy formation. Let me try to disaggregate what we generally mean by "politics" in these discussions.

When new data or research findings arrive on the conference table in councils of action, they confront four I's already sitting at the table. The I's are ideology, interests, institutional norms and practices, and prior information. By ideology, I mean people's basic values. No amount of valid data about the positive effects of abortion on the lives of women and children is likely to convince a principled upholder of the right to life to support abortion policy. Probably even good information on the numbers of disabled people collecting government benefits who could return to work will convince left-wing Laborites who fought long and hard for government benefits to support a policy for periodic re-tests of disability recipients.

Interests have to do with people's and organizations' self-interest. It may be the interest of the individual for career advancement or prestige or votes. It may be the interest of the organizational unit for greater status and larger budgets. In the U.S. we are seeing how officials with a lifelong attachment to the oil industry are willing to do irreparable damage to the Arctic wilderness in order to expand drilling for oil. Interests often trump data.

Institutions have traditions. They have made previous decisions and are set upon a certain path. They have procedures and ways of work. For example, the U.S. Congress works largely through face-to-face dealings -- hearings, discussions with constituents and interest groups, talk with staffers, conferences, committee meetings. Members do little reading; it is not part of the organizational norm. Research that arrives in written documents doesn't have much of a chance. Furthermore, data that call for abrupt changes in organizational procedures and rules have a difficult time receiving a hearing.

The fourth I is information. New data do not enter a pristine environment. Participants in decisions already have a great deal of information from a variety of sources. They have acquired a general framework for considering the issue under discussion somewhere during their education. They have experience with the issue through direct immersion and prior deliberations. They have accounts from the operating agency. They have talked to many people. New information has to fit into their general understanding of how the world works. New information has an easier time when it accords with their general scheme of things. If it is at variance with what they already know and believe, it has somehow to be strong enough to displace prior beliefs.

Signs that Evidence is Gaining Ground

The four I's are formidable rivals to the primacy of evidence. Nevertheless, I think that evidence is making headway. I can illustrate this belief with my current study.
About two years ago I wrote a research proposal to look at schools’ adoption of the D.A.R.E. (Drug Abuse Resistance Education) program in their classrooms to prevent drug use. Evaluations and meta-analysis had found that D.A.R.E. did not achieve the desired results. Youngsters exposed to the D.A.R.E. program in school were just as likely to use drugs at age 16 or 17 as youngsters in the control groups who did not receive D.A.R.E. (Dukes et al. 1997, Ennett et al. 1994, Rosenbaum and Hanson 1998). Yet some 75% of school districts in the U.S. implement D.A.R.E., as do many countries overseas. I wanted to know how school districts could continue to use D.A.R.E. in the face of overwhelming negative evidence. I wanted to find out the bases on which they made the decision to implement D.A.R.E. Didn’t they know about the evaluations? Didn’t they believe them? What counter-pressure(s) and rival factors were at work? I thought this would be a strategic research site to uncover the pushes and pulls working against evidence.

We received funding for the study and began fieldwork in the fall of 2000. We are studying four states and four school districts in each state. We have now been in the field for a year. But many things have happened in this year, and we have learned some unexpected things.

Just as we began fieldwork in 2000, news stories appeared that Salt Lake City was terminating its D.A.R.E. program. The mayor said that he had heard about the evaluations, and there was no point running a program that was ineffective in keeping kids off drugs. Then at several of the sites we visited for the study we found that they were terminating the program. Furthermore, television reports on nightly national newscasts were reporting the ineffectiveness of D.A.R.E.. Journalists like puncturing the taken-for-granted opinion that D.A.R.E. is effective. When we starting interviewing in the 16 districts we were studying, we found in several of them that somebody in the community -- a parent, the school superintendent, a teacher -- had taken the initiative and succeeded in getting the program cancelled. Just a short time ago I attended a meeting on another subject entirely which was attended by the former superintendent of schools of New York City, and in the course of an unrelated conversation he mentioned the D.A.R.E. evaluations and inveighed against D.A.R.E. as a waste of the taxpayers’ money.

Seemingly suddenly, after years of largely negative evaluations, school systems in some number were adopting alternative programs in place of D.A.R.E.. Data, it now seems clear, is making a difference.

In our interviews, a few people in schools told us that they hadn’t heard about the evaluations and their outcomes. Of those who had heard about them, some didn’t trust social science. As one person said, evaluators can make the findings come out any way they want. Some people who had heard were unconcerned because the program is so popular among students and parents. Most important may be the different basis on which they evaluated the utility of D.A.R.E.. People in the schools (and in police departments, where we also interviewed) said that they had never expected D.A.R.E. to prevent drug abuse. How could a program offered one hour for one week in fifth grade alter behavior of late adolescents five to six years later? But D.A.R.E. was good for other reasons. For example, it facilitated cooperation between schools and police departments, two public agencies that had long operated in isolation from each other. It helped students learn something about the ill effects of drugs and to see police in a non-enforcement role as regular human beings. Another reason for continuing with D.A.R.E. was that most districts are required by law to run some kind of drug prevention program, and D.A.R.E. is well known, popular with parents, and not costly to the schools.

Then to cap off the new learning, we found out that D.A.R.E. AMERICA, the parent organization, is undertaking a major revision of the program. They have been battered by charges that they are not effective, and after many years of rejecting evaluation evidence and criticizing the evaluators, they are changing the program. They are introducing more interaction between police and students, based on successful models of learning. As the program is being developed, they are having it evaluated. The University of Akron is doing the evaluation of the emerging program, with carefully developed randomized-assignment design.

It begins to seem as though the evaluations are having a major influence. It took a long time, but the news has come through, and people in the federal government, state governments, school districts, police agencies, and D.A.R.E. AMERICA itself are
changing their patterns of behavior. Evidence-based policy on parade.

Another interesting story is the use of evaluations of welfare-to-work programs in the U.S. These are programs that prepare women who have been receiving public assistance to take a paying job. Methodologically sophisticated random-assignment evaluations were done of state programs in California, New York, and Florida. Then, recently, Greenberg and his colleagues did a telephone survey of directors of all state welfare departments to find out how their states used the evaluative information (Greenberg et al. 2000). The welfare directors knew about the results of the studies. They knew that programs that placed participants directly on jobs had better success than programs that offered training. They did not believe that the results had “dramatic, decisive effects on policymaking” in their state (p. 375), but many of them did say that the results entered deliberations. Decision-makers also took into account policy moves on the federal level, political expediency, news about programs other states were implementing, and the problems that they were facing in implementation. Nevertheless, directly or indirectly the evaluations exerted an influence. Ironically, the causal attribution made possible by the random assignment designs of the studies, which the evaluators and researchers were proud of, played very little part in state welfare policy discussions. People were not interested in effect sizes but much more in the ideas generated by the studies. They found a general direction to take: less emphasis on training and more on direct placement in jobs.

Kinds of Evidence

Four kinds of evidence are having an impact on policy and program decisions: descriptive data, analytic findings, evaluative evidence, and policy analytic forecasts.

1. Description. Data on economic conditions such as the cost of living, tax receipts, stock market levels, unemployment, welfare, and trade deficits have long had a strong influence on government policy, although the influence has not always been as direct and appropriate as non-economists assume. Also influential are data on the incidence and prevalence of various health conditions, such as tuberculosis, birth defects, and smoking. More note is now being taken of figures on school dropouts, receipt of welfare assistance, juvenile delinquency. These kinds of data have a history of familiarity and utility. Officials don’t always know what to do about social pathologies, but they recognize the importance of data that show their prevalence, their location, and the direction of change. Trend data are particularly useful to show which social ills are gaining or declining in severity.

Government largely runs on the basis of good data, both internally generated records and externally produced surveys, tests, and reviews.

2. Analytic information. This category refers to research that identifies factors associated with conditions. Thus, research can indicate the factors associated with crime or with women’s entry into scientific careers or bullying in the schoolyard. In some cases the research can even demonstrate a causal relationship – which factors lead to the conditions being studied. One study currently evoking considerable controversy in the U.S. reported that small children who spend long hours in day care are more likely to exhibit anti-social behavior than children who do not spend as much time in day care (CITE). The study is controversial because a number of previous studies had found no such relationship. But if the findings are valid and hold up under replication, the public policy significance would be obvious. They would call into question, for example, current welfare rules, that require welfare mothers to take jobs even when their children are very young and therefore are forced to put their children into day care.

Findings that indicate associations among variables can give direction to efforts to reduce or increase the given behavior. But often research deals with structural factors that policy makers cannot do much about, at least in the short run. As Dilullo (1991) has written about research on crime, the studies seeking the cause of crime “tend to center on complex and remote ‘structural’ factors...As such, even public policy-minded social scientists rarely do work that has meaningful practical implications...Social scientists have established ad nauseam that a disproportionate number of predatory street crimes are committed by poor young males living in inner cities. So what? Can we cure poverty? Are we to outlaw teenage and young adult malehood? (pp. 218, 224)”
Moreover, even when research does point to practical action, it takes more than knowledge and ideas to make policy. It takes imagination and creativity to transform ideas into workable policy proposals, and it takes mobilization and political support to turn the proposals into policy.

3. Evaluation looks directly at the effectiveness of existing policies and programs. This is particularly potent information for policy making. But because the policies are already operating, any negative findings are likely to arouse waves of defensiveness – from the agencies running the program, from the legislators who authorized the program, the advocacy organizations that support it, even the beneficiaries who feel that whatever the evaluation shows about outcomes, they are at least getting attention under the current regime. Because of the general hostility toward negative findings (D.A.R.E. fought the evaluators for decades before last year), evaluation has a checkered history in affecting policy. But some people who are already favorably disposed to the position the data supports often champion it and use it to press for change. In some cases, as with school districts and D.A.R.E., the data occasionally appear to be influential. Whether advocates with data are generally more successful than advocates without data is a good empirical question. Too many other things go into policy making to expect major flip-flops. But given current trends in evaluation (almost every new program gets an evaluation), the programs with positive data may have an edge.

4. Policy analysis. Analyses of alternative future policies are attempts to foresee the future. Sometimes policy analysis is coupled with evidence-based advice. Which of the list of possible actions would likely have the best outcomes? Analysts canvass alternatives, draw on available data sources to calculate the advantages and disadvantages, and probable costs and benefits, of each of them, and report the mix of results to expect. Analysts recommend the action with the best mix of advantages and disadvantages, but they need good data to do a responsible job. Schools of public policy give their students a set of tools and strategies to predict, calculate, and report on future policies. Whether such data are influential or not depends on many factors. One factor is probably the soundness of the data on which recommendations are based. More important is where the policy analyst sits. If she is close to the action, she has more influence. Another crucial attribute is how much she is trusted by key actors. If key actors have confidence in her work, she is more influential. Because some analysts occupy positions inside government, with close and confidential relationships to political figures, it is hard to know exactly how much influence they wield. Others work outside government in think tanks, university centers, and elsewhere. As Denham and Garnett have written (1996), “The oxygen of publicity is essential to ‘think tanks’, and the ability to impress journalists with the news value of their findings can at least ensure that these are widely discussed” (p. 54). Wide discussion is a significant consequence. Of course, the ability to impress policy makers can make for even greater headway. But even when policy makers brandish policy analysis and wave reports in front of the television cameras, it is not a sure sign that the work has influenced policy. Denham and Garnett say that think tanks in the U.K. in the Thatcher period provided “intellectual legitimacy” to policy that was already decided (p. 52), but it also “helped to sustain the radical momentum of Thatcherism by reinforcing the sense of a collective crusade” (p. 52).

If we are to sum up the influence of policy analytic data, it is safe to say that it ranges from one extreme to the other. Some of the most important and long-lasting influences are invisible to the naked eye. Some of the apparent uses fade away very quickly. Still policy analysis is valuable enough to enough people to lead to a growth in the number and

---

1 Evaluation can also examine the outcomes of pilot programs before they are adopted on a large-scale basis. Such evidence can provide important information about the program is worth widespread adoption. As it happens, in the U.S. the strategy of running a pilot program before going large-scale is rarely adopted for scientific reasons. Programs are not infrequently enacted on a small-scale basis when there is a lack of agreement about the merit of the program and the pilot program is a compromise between proponents and opponents. Or a small-scale version can be run when there is insufficient money in the budget to fund full-scale implementation. Getting data on the feasibility or effectiveness of the pilot program seems like an evidence-based strategy before going into full-scale operation, much in the spirit of Donald Campbell’s “Reforms as Experiments” (1969) However, the motives behind pilot programs are rarely quite so rational, and even when evaluation data is collected, it is probably taken into account no more often than evidence on the effectiveness of other kinds of programs.
size of think tanks all over the world (NIRA 1993). This is one indicator of the potential – and sometime realization – of the power of evidence.

A Cautionary Note

My sense of the influence of evidence on policy is more upbeat than it has ever been. Many trends have converged to elevate the place of data in public policy making. But before we succumb to euphoria, let me tell you about a recent minor scandal in the U.S. A supervisor in an Oklahoma Crime Laboratory had testified in thousands of court cases about evidence that incriminated defendants, including 23 cases that resulted in sentences of death. Some of the testimony had to do with hair samples, blood samples, fiber samples, -- all the kinds of evidence that crime laboratories routinely analyze and report in court. However, with the advent of DNA testing, there are now ways to check on the veracity of crime lab reports. It turns out that Ms. Gilchrist has been giving misleading testimony. The Federal Bureau of Investigation dug into the records and examined her testimony. Of the eight cases they have reviewed, they found that Ms. Gilchrist gave invalid testimony in five.

Nor is she the only culprit. The chief serologist of West Virginia’s state crime laboratory is going on trial after the state’s Supreme Court found that he had offered false testimony in hundreds of cases. A police lab analyst in Chicago is under investigation after DNA tests showed the innocence of a defendant against whom she had testified. Two writers have been examining the series of mistakes in crime laboratories and written a crusading book (Dwyer et al. 2000, Scheck and Neufeld 2001).

What do these laboratory errors have to do with our subject, evidence for policy? Well, if there is one main precedent for reliance on evidence in making decisions it assuredly comes from the law. The whole monumental architecture of Anglo-Saxon jurisprudence rests on a respect for evidence, a search for evidence, a sharing of evidence, and a confidence in the power of evidence to guide appropriate decisions. Yet in the very sanctum of the criminal law we now have glimpses of the fallibility of ordinary procedures for analyzing and marshalling physical evidence.

We also have a good guess about the reason underlying the false reports. Partly they arise from the weakness of the analytic procedures. For example, it is evidently very difficult to match one strand of hair under a microscope with another. Another reason surely has to do with the function of crime labs. By and large, they are part of the law enforcement system. They are there not only to render impartial reports on the evidence but also (implicitly) to help prosecutors make strong cases and convict criminals. This barely acknowledged bias in the structure of crime labs suggests that when they err, they are likely to err on the side of the prosecution.

How much more vulnerable to error and bias is social science evidence. We all know that social science data do not exist “out there” in the physical landscape, just waiting for investigators to come along and gather it in. Social science data has to be constructed, with definitions, boundaries, data collection methods, analytic techniques, interpretive strategies. Many data collectors and social scientists have their own biases, and they work for organizations that have an institutionalized interest in the way the data come out. Not all of us, of course. But we are all human and ipso facto fallible.

I celebrate the heightened attention to social science data in decision-making councils. But it is worth more than a footnote to acknowledge the fallibility and time-limited character of much of the work we do.

Compatibility between Evidence-Based Policy and Public Participation in Policy Making

Finally let me say a word about another of the big trends in decision-making, at least in the U.S. It is wider public participation in decision-making. This trend is visible in both public and private organizations, where some big organizations have pared the levels of management and given more workers a say in the direction that the organization takes. Participation is on the rise in schools, where teachers are now often being given a role in school decision-making and parents are sometimes involved, too. The trend is by no means universal, and there have been ups and down over the last decade. But the rhetoric of participation is prevalent, even in business schools.

CEM Centre, University of Durham  http://cem.dur.ac.uk

290
What does this trend portend for the use of knowledge? The data isn’t in, and we can find anecdotes on both sides of the equation. For example, some parents, who had no stake in existing drug abuse prevention program in the school, were swayed by evaluative evidence to demand the end of D.A.R.E. On the other hand, in many school districts we are studying, parents are D.A.R.E.’s biggest supporters and ignore or dismiss evaluative evidence of its lack of long-term effects. My guess is that wide participation will lessen the effect of evidence because newly enfranchised groups will vote their interests first. Perhaps later they will get around to looking at the evidence, but at the start at least, they will seek to rectify their grievances. A movement toward wider participation in policy making will at least demand that researchers, evaluators, and analysts spend more time and thought on disseminating their findings to the relevant publics. Whispering in the ear of the powerful will no longer be enough. Researchers and data providers of all types will need to alter their ideas about what it takes to get a hearing for evidence and to master the arts of communication to multiple audiences. Their findings will have to inform and convince a wide swath of the public if policy is to be truly evidence-based.

References


Correspondence

E-mail: carol_weiss@harvard.edu