Reflections on the conference

Several members of the academic community (Peter H. Rossi, James Heckman, and Thomas J. Corbett) were asked to give their personal reflections on the conference, as were several participants from the policy-making community (William R. Prosser, Steven H. Sandell, Sharon McGroder, and Stella Koutroumanes).

These perspectives on the conference represent the personal views of the authors and should not be construed to represent the official position or policy of the administration, the U.S. Department of Health and Human Services, the Institute for Research on Poverty, or any other institution.

Some critical comments on current evaluations of programs for the amelioration of persistent poverty

by Peter H. Rossi, Stuart A. Rice Professor of Sociology and Acting Director, Social and Demographic Research Institute, University of Massachusetts, Amherst, Mass.

The evaluations that were at the center of attention in the IRP/ASPE conference were impressive testimony to the commitment to careful evaluation on the part of the agencies involved. Compared to even a decade ago, these evaluations almost uniformly demonstrated a high level of technical knowledge and were tackling programs of the sort that previously would have gone unevaluated or would have been approached with inappropriate research designs. Given that praise, my comments below may appear to be overly critical. It is not my intention to take anything away from the fact that the evaluations as a group represent the best of the state of the evaluation art as currently practiced by the better federal agencies. These critical remarks are aimed at improving future evaluations.

There is ample evidence in the description of the major evaluation efforts under way that sophisticated large-scale evaluation is alive and well in the United States. Especially welcome was the discovery that randomized field experiments are still being undertaken. The grand leviathan field experiments of the sixties and seventies may not be in the works in the nineties, but there will be plenty of smaller randomized experiments.

All that said, there are problems with the studies. It appears that the evaluation community may have mastered technical problems but has still to come to grips completely with substantive issues. Some of the ways in which the evaluations are falling short are discussed below.

Drawbacks of the programs and their evaluations

To begin with there is a misfit between the problem of persistent poverty, to which most of these programs are directed, and the program evaluations. The target problem is persistent poverty and dependency, with persistency defined implicitly as lasting across generations. Because the evaluations last only a few years at most, they cannot directly address the issue of whether the programs affect persistent poverty, which cannot be directly measured in so short a time. Correspondingly, the target population can only be defined as persons at high risk of being persistently poor and transmitting that poverty to their children, a tactic which depends heavily on how well risk can be defined and measured. This does not imply that appropriate short-term evaluations cannot be designed. It does mean, however, that the target population can only be fuzzily defined and the outcomes have to be proxies for persistent poverty. Selecting appropriate proxies for the long-term outcomes requires knowledge of the processes by which persistent poverty is generated and maintained. Correspondingly, knowledge is needed about the same processes in order to identify populations at risk.

The programs under discussion appear to be driven by much the same sort of policy premises: Persistent poverty is seen as a serious social problem, for which there is no known solution. Nevertheless an optimistic assumption is made that ameliorative and preventive programs exist that are both politically acceptable and efficacious. But we do not know what will be efficacious. What is politically acceptable is easier to identify. Accordingly, the programs are squarely in the mainstream as defined by the op-ed pages of our national media. Another consequence is a propensity to throw programs at problems, with the programs having the characteristic of leaving specific interventions and delivery systems to local communities to define. Not expecting that all communities will hit upon efficacious programs, this strategy leads to multisite studies in the hope that there will be some appreciable "natural" variation in programs, the analysis of which will lead to identification of effective programs. It is assumed that, in the end, a set of programs, slightly varied from site to site, will contain among them enough truly effective programs that can then be put in place throughout the country. A grass-roots democratic optimism pervades this strategy: the assumption that those who are close to the problem as it manifests itself in concrete ways in specific localities will also know best how to design ameliorative strategies.

The evaluations show some interesting features. First, although randomization is alive and well, the randomized "hothouse experiments," in which both the services and the evaluation are designed and run by experimenters, are out of favor. Instead, the services are typically designed and delivered by local organizations, and the evaluations are carried out by researchers.

A consequence is that these are "black box experiments"—experiments in which the exact nature of the treatment is not known—but with a new twist. Once the black box is constructed and used in an experimental trial, the researchers open it and examine its contents through implementation research. Whether the *post hoc* reconstruction of treatments will compensate for the disadvantages of black box experiments is problematic. I sensed that most of the researchers felt uncomfortable about the qualitative data typically collected for implementation research and had few ideas about how to integrate those data into an analytic framework.

There were other problems as well with analytic strategies. Because targets were not clearly identified, the units of analysis have yet to be specified (and in the case of some evaluations yet to be thought through). Whether the units should be parents, children, households, or families had not yet been decided.

Program goals (and hence outcomes) were also unclear: Was it the public welfare system, parent-child relations, parents, children, or their support networks—or what?—that should be affected? And what is expected to be changed by the intervention? It appeared that because parents are easiest to handle as a unit of analysis, changing the behavior of parents tended to be the program goal most easily articulated.

Finally, in many instances, the evaluation seemed to be premature. Given that any program needs some time to develop a maximum implementation, research estimating impacts should not be started until programs have begun to run smoothly. Although I believe that evaluation planning ought to be started at the same time that a program is put in operation, the actual evaluation ought not to be started until the program has been satisfactorily implemented. Otherwise the evaluation is of a program not at its best.

Perhaps the most serious deficiency in these sets of evaluations is that the programs are entirely too hastily constructed and do not appear to have been much influenced by what is already known about the problems they are expected to address, either from prior basic or applied research or from prior evaluations of similar programs. The major exception to this generalization is the planned evaluation of the JOBS program, whose design has been influenced by a thorough search of the literature on previous programs. The design of programs and their accompanying evaluations needs to be based on a thorough grounding in rich descriptive research and on analytical models of the phenomena in question. In order to design programs and their evaluations properly, we need to know how the human services system involved operates, what appears to be the source of the social problem, social and psychological characteristics of clients, the surrounding ecology in which the clients and the program must operate, and the human behavioral models appropriate to the phenomenon.

Listening to the conference presentations, I was not much impressed that either the programs or the evaluations were based on much more than the "intuitions" of local human service professionals about what might be acceptable to the funding agencies in question. I believe that this intellectual weakness arises out of the strategy of leaving program design to the intellectually weakest part of the social service system, local agencies, staffed with poorly paid, poorly prepared personnel. This is not to say that local agencies are incompetent. On the contrary, I believe that they are quite competent to carry out programs. I do not believe, however, that they have the competence to design programs based on the best knowledge we currently have available from empirical research concerning the problem in question. To have intimate first-hand knowledge about the problem is clearly essential in order to design programs, but it is not enough.

Recommendations

There are several recommendations that flow from these observations:

First, the existing evaluations can be improved by clarifying certain issues. Some thought ought to be given to how best to integrate qualitative findings from implementation into the analytic framework of the evaluations. The researchers ought to consider borrowing heavily from fields in which techniques for so doing have developed, especially the quantitative sides of anthropology, communications research, and clinical psychology. It would also be important to decide what will be the most productive units of analysis. Although it is not necessary to decide upon one such unit, it is necessary to decide which units will be used, so that the appropriate data can be collected and data management conducted accordingly.

Second, for future evaluations, I recommend that the designs of programs and their evaluations be illumined by thorough familiarity with existing knowledge. Whether this is done formally by meta-analyses or less rigorously by conventional methods of literature review need not be decided a priori. But grounding in the existing empirical literatures is necessary. It also seems to me that it is highly unlikely that local agencies have the intellectual resources effectively to access, collate, and assess the needed knowledge base. Accordingly, I believe that it is significant that the JOBS evaluation is the one most influenced by prior knowledge and is the only one that is trying to structure variation in treatment, as in its Type B design. Unless we vary treatments experimentally, we can only learn whether a given program succeeds or fails; we can learn little about how to improve it.

Third, mission-oriented agencies should appreciate more the extent to which the development of sensible and potentially effective new programs rests on the accumulation of knowledge. Although much basic research may be accomplished through the National Institutes of Health and the National Science Foundation, funds for supporting "basic applied" research are not easily available. By "basic applied" research, I have in mind rich descriptive research centered on the size, distribution, and social location of the social problem in question; longitudinal studies that describe processes of development and decline; and analytic studies that attempt to construct and test models of the social problem. The steady accumulation of such knowledge would put both the design of programs and of their evaluations on much firmer foundations.■

Basic knowledge—not black box evaluations

by James Heckman, Henry Schultz Professor of Economics and Public Policy, University of Chicago

The papers presented at this conference, taken as a whole, offer striking evidence on the folly of the current trend in evaluation research away from attempting to understand social mechanisms and the root causes of social problems and towards black box evaluations of specific social programs. Emphasis on the black box approach is a natural consequence of the currently fashionable—but factually and intellectually unsupported—belief in social experimentation as *the* method of choice in program evaluation. Advocates of social experiments seek to bypass the difficult task of understanding the origins of social problems by black box experimental analysis of specific programs.

Invoking the article of faith of experimental advocates that only *randomized* social experiments provide valid knowledge, experimentalists mimic the jargon—but not the substance—of the classical model of experimentation in agriculture. Their argument runs as follows: Randomizing persons into treatment categories and observing outcomes produces "believable" *mean* differences in outcomes. (Median differences cannot be estimated in general.)[1] There is no need to understand social mechanisms or social science—a convenient excuse for ignoring basic knowledge and for not generating it. Bombard subjects with randomly assigned treatments and out will come "convincing" "scientific" estimates without the tormenting and "unconvincing" qualifications that "mar" carefully executed nonexperimental social science.

This argument ignores a steadily accumulating body of knowledge that suggests that randomized social experiments greatly alter the programs being analyzed.[2] Even if they did not, the new emphasis on evaluating the effects of "treatments" on outcomes rather than on understanding basic mechanisms causes program evaluations of the sort presented at this conference to produce noncumulative knowledge. Each study has its own "treatments" and no attempt is made to put the treatments on a common intellectual footing so that comparisons can be made across studies or so that social problems that gave rise to a specific program can be better understood.

Many of the papers presented at this conference offer no motivation whatsoever for how the social problem addressed by the program being evaluated comes into existence. Most offer no insight into the specific mechanisms by which the proposed program will work. Because there is no attempt to step back from the specifics of the program being evaluated, no social science context is provided and no long-term knowledge is generated. The best that can be said is that some program "works" on some short-run target criterion. Basic knowledge is not produced. This is a natural consequence of the black box approach to social science fostered by those who advocate social experimentation and black box evaluations. An argument that justifies ignorance of social mechanisms can only foster further ignorance. This is a lasting—and harmful—legacy of the randomized social experimentation movement.

Millions of dollars are currently being spent on poorly planned evaluations of poorly designed scattershot social programs that attempt to solve social problems, without adding to our understanding of either the programs or the problems. Consulting firms are willing to carry out these evaluations and bureaucrats encourage their efforts, despite the dubious scientific value of their findings. There is no incentive in the current federal research contracting system to produce cumulative social science knowledge so that we can learn from these studies or understand the problems that motivated them. All we learn is whether or not the programs "worked" on some narrow—and often uninterpretable—criterion.

Vast sums are being spent on "evaluating" specific programs for which the objectives are often not clear and so the evaluation problem for them is not clearly specified. The programs that focus on child development rely on different tests administered at different ages that are not comparable for the same person and have no demonstrated relationship to adult achievement in or out of the marketplace. These programs are good examples of all that is wrong with current government human resource programs and their evaluations. Meaningless outcome measures are "evaluated" by thoughtless black box randomization methods.

The opportunity cost of this activity is the reduction in expenditure on the fact-gathering and fact-analyzing activity that produces basic social science knowledge. Knowledge of this sort is crucial for understanding the true causes of social problems and even for organizing the evidence from the "evaluations" presented at this conference. Surely the money currently being wasted on operating or evaluating these scattershot programs is better spent on collecting and analyzing basic data from sources like the National Longitudinal Survey of Youth, the Survey of Income and Program Participation, or the Panel Study of Income Dynamics, and developing a much firmer empirical knowledge base on which to conduct the study of social policy and the design and evaluation of social programs. ■

[1]Heckman, "Randomization and Social Policy Evaluation," in *Evaluating Welfare and Training Programs*, ed. Charles F. Manski and Irwin Garfinkel (Cambridge, Mass.: Harvard University Press, 1992).

[2]See the papers in Manski and Garfinkel, *Evaluating Welfare and Training Programs*.

The evaluation conundrum: A case of "back to the future"?

by Thomas J. Corbett, IRP affiliate and Assistant Professor, Division of University Outreach, Department of Governmental Affairs, University of Wisconsin–Madison

The third annual IRP/ASPE evaluation conference, "Evaluating Comprehensive Family Service Programs," likely left many observers with ambivalent feelings. On the one hand, there was a sense of challenge associated with confronting the complexities of designing and evaluating "two-generational" (and even more complex) intervention models. And some must have been comforted by the collaborative spirit apparent among normally competitive agencies and institutions in addressing those complexities. On the other hand, there must exist dismay at the primitive character of existing capacities at every level of the policy process—from program conception and inception through evaluation and institutionalization—that necessarily inhibits our ability to measure and interpret anything beyond the simplest program models.

A historical perspective

It is not difficult to imagine that conference attendees had been transported back a quarter-century or so to the heady, yet confusing, days of the last War on Poverty—particularly the period of 1962 to 1967.[1] Then, as now, the policy focus was not on income poverty, but rather on the institutional and individual correlates and causes of behavioral disadvantage. Then, as now, unidimensional interventions were seen as inadequate to the task, and complex program strategies spilled forth with dizzying celerity. Then, as now, the political imperative for solutions appeared to dominate those virtues of probity and patience that are required for sensible long-range policy/program development and testing. Then, as now (though certainly more then than now), there existed some faith that those who plied the social science trade could contribute to the doing of public policy. Then, as now, the prospects for disenchantment with the efficacy of government were high in the face of both exaggerated expectations and the crude tools for conceptualizing and evaluating outcomes. People-changing and institution-changing are once more becoming objects of public attention. The complexities of accomplishing these objectives are no less daunting now as they were then.

The degree to which social policies of the 1990s experience success relative to the 1960s depends on the extent to which theoretical and methodological improvements have, in fact, been realized. It also depends on whether public policy can move beyond the fascination with those kinds of "media sound bites" (i.e., facile solutions that play well on television) that can undermine substantive progress. Some signs are hopeful. Professional evaluators and policy analysts, who form a new cottage industry, are undoubtedly more sophisticated than they were a generation ago. A diverse audience can come together and discuss with some facility the complex trade-offs associated with high-fidelity evaluation designs (data rich/small sample designs) as opposed to low-fidelity (data poor/large sample) alternatives. And they can discuss the relative advantages of evaluating "on-the-farm" pilot programs, those which replicate typical organizational environments, as opposed to hothouse designs, which minimize contextual noise.

Some of the challenges facing the overall policy-academic community are terribly difficult. None are more apparent than the political aspects of doing policy. Normative and partisan concerns too often dominate substantive and technical foci. Answers are wanted in the short term, largely defined by political cycles, and are expected to be summative in nature. Where a slow accretion of knowledge and insight would be useful, definitive statements about impact are demanded. Complicating the situation is the fact that the hyperbole surrounding the enactment (e.g., selling) of policy makes the appearance of success less probable in the long run.

The central question of traditional evaluations is does *it* work. Increasingly, we are aware that the newer challenge is to fully understand what *it* is. Not surprisingly, the need for formative evaluations (those oriented toward developing feedback on the character of the intervention) is given as much weight as the more traditional summative forms (those designed to measure net impacts). As Robert Granger pointed out at the conference, variation across the six P's—programs, people, places, participation, processes, and payoffs—makes sorting out the operational nature of the intervention quite problematic. It is far too easy to evaluate a program label without having any real understanding of what has been examined or which of many program dimensions contribute to "net" outcomes. All the structural and intensity dimensions may be far less instrumental than the omnipresent "Q" factor—the quality factor, where competence and care contribute more to outcomes than the specifications of the formal program model. In some of the evaluations discussed at the conference it is difficult to envision how net effects would be explained given the natural (in fact, encouraged) variation that exists within and across program sites.

In short, the absence of a simply defined *it* speaks to some of those policy-making flaws evident some twenty-five years ago. We see a natural life cycle of new programs continually repeated: programs are launched with great fanfare and exaggerated claims (to sell them in the first place); the pace and scope of implementation conform more to political cycles than sober program development; outcomes are (intentionally?) unclear or overly complex, thereby difficult to operationalize and measure; the investment in program evaluation is insufficient given the complexity of underlying theoretical models and the stakes (fiscal and otherwise) at risk. Given this life cycle, it is all too easy for excitement to evolve into disenchantment and ultimately despair, not unlike the evolution from government as the solution to societal ills (the 1960s) to government as the problem (the 1980s).

Dimensions of the black box

I think we all acknowledge that more rigorous thought about the nature of the "black box" and what it takes to get inside is required. The new program models are extremely complex, involving a sequence of events and expectations tied together by a complex set of client-level decisions. Let's touch on just a few of its dimensions.

There is the factor of time (the three I's of *I*ntroduction, *I*mplementation, and *I*nstitutionalization), where there is a learning curve associated with new programs and where key structural and process variables are expected to evolve and change as lessons are learned. Process and impact evaluations must remain sensitive to the possibility that *what* is examined depends on *when* it is examined.

There is the discrepancy factor—the gap between expectation and reality. What is intended on paper is not always what happens "on the streets." These discrepancies must be fully understood and documented if an understanding of what works (or doesn't) and why something works is to be appreciated.

There is the ubiquitous cross-everything problem. Potential variation within relevant dimensions and interaction effects across dimensions (e.g., across client subpopulations, across sites, across vendors, across case managers, and so on) appears endless. Understanding these complexities is an intellectual challenge, and dealing with them methodologically is an evaluator's nightmare.

There is the transactional dilemma. What actually happens at the interface between system and client? Can we tap such dimensions as quality and intensity in any but the crudest manner? What kind of microlevel decisions are made inside the box—rule driven or professional? And if they are the latter, what can we ever know about them?

And there is the outcome conundrum at the end. What is success? Where complex outcomes are anticipated (i.e., several criterion variables of interest for each subject *and* multiple population groups of interest), it is conceivable that some measures will move in one direction while others move in the opposite direction. This makes substantive conclusions about the meaning of any set of results largely subjective in character.

Some of the answers to these dilemmas were suggested at the conference: more synthesis activities, more attention to process analyses and qualitative work, and more attention to the development of common marker variables. In the long run, however, we may have to think of a whole new way of doing business. The old form of discrete, impact-focused evaluations, awarded to firms on a competitive basis, may be counterproductive. Longer time lines, less obsession with what "works," and a more collaborative evaluation industry may be needed. The days of the short sprint—one-shot summative evaluations—may be ending. A new paradigm, where the marathon constitutes the more appropriate metaphor, may be emerging. ■

---

[1]The 1962 to 1967 period was a high point in public sector efforts to change the behaviors of low-income individuals and the institutions with which they interact. Amendments to the Social Security Act in 1962 (Public Law 87-543) dramatically inaugurated an effort to combine social services and the receipt of welfare. Among other things the 1962 amendments required a service plan for each child recipient of AFDC, based on his or her particular home conditions. The War on Poverty, which began in 1964, carried on the same emphasis, launching a set of programs designed to enhance human capital and change the communities and institutions with which the dependent poor interacted.

IRP/ASPE Small Grants Seminar

On May 7, 1992, the current winners of the IRP/ASPE Small Grants competition will present their research findings in a seminar at the U.S. Department of Health and Human Services. The public is invited to attend. The seminar will be held in Room 503A, Hubert H. Humphrey Building, 200 Independence Avenue, S.W., Washington, D.C.

The following presentations will be made:

Amy C. Butler, "The Changing Economic Consequences of Teenage Childbearing"

William G. Gale, "The Effects of Public and Private Transfers on Income Variability and the Poverty Rate"

Jerry A. Jacobs, "Trends in Wages, Underemployment, and Mobility among Part-Time Workers"

Alan B. Krueger, "The Impact of Recent Changes in the Minimum Wage: Results from a New Establishment Survey"

Susan E. Mayer, "A Comparison of Poverty and Living Conditions in Five Countries"

Charlotte J. Patterson, "Persistent and Transitory Economic Stress: Psychosocial Consequences for Children"

Mark R. Rank and Thomas A. Hirschl, "The Impact of Population Density upon the Use of Welfare Programs"

Roger A. Wojtkiewicz, "Parental Presence during Childhood and Adolescence: The Effects of Duration and Change on High School Graduation"

Reflections on demonstration evaluations: A view from the stands or the arena?

by William R. Prosser, Senior Policy Analyst, U.S. Department of Health and Human Services; visiting professor, University of Wisconsin; and co-organizer of the conference

> *It is not the critic who counts, not the man who points out how the strong man stumbled, or where the doer of deeds could have done them better. The credit belongs to the man who is actually in the arena; whose face is marred by dust and sweat and blood; who strives valiantly; who errs and comes short again and again; who knows the great enthusiasms, the great devotions, and spends himself in a worthy cause; who, at the best, knows in the end the triumph of high achievement; and who, at the worst, if he fails, at least fails while daring greatly, so that his place shall never be with those cold and timid souls who know neither victory nor defeat.* (Theodore Roosevelt)

In this piece I reflect on the different needs of scholars and government policy analysts and the problems of procuring research demonstrations. I then draw some lessons from recent work that might guide future research demonstrations.

Demonstrations are usually a messy form of field research. They involve both action—e.g., service delivery or income transfer—and evaluation research. They come about because people want to improve the state of the art of addressing social problems and aren't sure how to do it. Some demonstrations are undertaken because we know there is a problem, need to know more about it, and want to develop promising ideas. Other demonstrations are launched because we think we know something about the problem and how to lessen it, and we want to show that our ideas work. This field research involves two broad types of activity—action and assessment. Policymakers and operational people, typified by the opening quote, often place primary importance on the action aspects of the demonstration. Others, academics for example, may put more emphasis on what can be learned from the demonstrations.

Those of us in ASPE and IRP involved in planning the conference believe it was important to bring together people who have been involved in commissioning, designing, and conducting program demonstrations and evaluations, along with others who hope to use their results, to share information, encourage interagency communication and interdisciplinary social science, and to improve the state of the art of program demonstration. These people represented both the action and assessment sides of demonstration, although the assessment side clearly had a larger representation.

While I believe that ASPE and IRP currently have a very congenial, collaborative relationship, we probably have different perspectives about data needs and respond to different priorities. ASPE staff perspectives are influenced by concern for policy-making. Policy-making is often more geared to decision making and action than assessment and synthesis, although research planning is clearly a major concern and responsibility. We like to feel that we are a conduit among the action, assessment, and policy-making communities. IRP's data interests, it seems to me, are more driven by academic concerns associated with knowledge building and social science. ASPE staff pay more attention to policymakers. Both perspectives are valid. We all share one common goal: we want to help solve social problems and make our country a better place to live.

The contractors selected to design, manage, oversee, and evaluate demonstrations are often caught in the middle. They have very pragmatic requirements on cost, schedule, and technical quality imposed by federal staff. But they also care about the social problems and the social science they are undertaking. It is easy to dismiss them as "hired guns" only interested in making a buck, but such labels miss their mark. Many of these contractors are professionals who must write reports and technical papers that meet the needs of the funding agencies and the criteria of scientific journals.

Our differences seem to be most starkly displayed when it comes to demonstrations as a way to expand the envelope of knowledge to enhance our understanding of human service practice, public administration, and social science. Some scholars are skeptical that "black box" or any other form of demonstration can contribute to basic knowledge. Or, at the least, they believe that there are more cost-effective ways to further knowledge on basic social questions. My own experience from the JOBS evaluation gives me hope that demonstrations may be fruitful if designed and managed properly. When I reflect on the conference, I feel that there is still much to be learned about managing demonstrations so that they contribute to both policy and social science.

For those of us concerned about social welfare and public administration, this is not entirely an academic debate. The President in his 1992 State of the Union address suggested that states be given increased flexibility to demonstrate new ways to improve welfare. Will we use this suggestion to generate new knowledge on how—and for whom—services work, to be shared among public agencies? Will we add to the cumulative knowledge base? Or will we let one thousand flowers bloom, not knowing what kind of seed or fertilizer is used, nor the type of soil tilled?

I think we must have a better understanding of what we can and cannot gain from demonstrations—in terms of both action and assessment. Several dimensions have to be considered. First we must examine the intergovernmental dimension. Federal demonstrations generally serve three broad intergovernmental functions: (1) They develop and test new programs or modify existing ones to identify those worthy of adoption and implementation by the federal government (e.g., the Negative Income Tax experiments—NIT). Some of these experiments may even test fundamental concepts, such as the effects of transfers on the labor supply of low-income women. (2) They enable state and local government or private sector organizations to try out new ideas, supported by federal funds and within a federally mandated framework (e.g., the OBRA demonstrations and the JOBS projects). (3) They support efforts requested by local agencies to address their own specific needs (e.g., the Low Income Opportunity Board demonstrations).[1] In general, the projects presented at the conference were commissioned for the latter two reasons. They are being carried out by state and local government or private agencies to meet their own as well as federal objectives. Or, as Peter Rossi says, these demonstrations are being carried out by your ordinary American agency (YOAA).[2] Although it is not inherent in any of these three types to be more oriented toward action than assessment, it is my experience that the latter two tend to place more emphasis on action.

If we were to look at DHHS human services research and evaluation funding over the last ten or fifteen years, I believe we would find the bulk of the resources invested in demonstrations serving the second and third intergovernmental functions mentioned above. Almost no funding is going for demonstrations like the NIT, which are solely for federal policy-making. Instead, we are investing in a few large-scale, multimillion-dollar, multi-year demonstrations, often employing random assignment (like the current JOBS evaluation, the Comprehensive Child Development Program, and the Teen Parent Demonstration). A significant portion of the funding for research and evaluation also goes to a larger number of smaller demonstration projects that are much more exploratory in nature, conceptualized to examine the nature and extent of new social problems and identify best practices for dealing with them. These projects are usually designed to serve joint federal and state/local interests. I call such demonstrations action-oriented demonstrations *if* they use the bulk of their funds to provide services to ameliorate social problems *and* very little of their funds to contribute to cumulative knowledge building.

Two broad strategies are used in assessment for knowledge building: deductive and inductive. The first employs (usually large-scale) model projects, the components of which are (deductively) based on a body of earlier empirical research. These projects study large numbers of carefully selected subjects and often use random assignment and control groups. (This category would include "black box" studies.[3]) The overall purpose of such demonstrations is to provide internally valid results which can be generalized to objectively defensible public policy.

The second strategy is much more exploratory and involves inductive testing of model components or variables suggested from limited research, or, more often, current "best practices." This type of demonstration is often a first step in the isolation of important service practices that may warrant more controlled, larger-scale program development and evaluation later on. These exploratory projects are usually small-scale attempts to respond to hot national problems that cannot ethically be ignored. They attempt to initiate services based on subjective or philosophical assumptions about service strategies, client needs, and model components. Such demonstrations often have little emphasis on formal assessment/evaluation.

Reasons exist for all the demonstration types I have just discussed. In my opinion, however, the ones that have little emphasis on internal evaluation and provide the least information for dissemination need the most improvement. More effort must be made to emphasize evaluation in these projects to justify the considerable federal and state expenditures they entail. I would also like to see us do a better job of designing the evaluations of the large-scale deductive demonstrations. James Heckman seems to have little good to say about any kind of demonstration. I think that well-managed large demonstrations can contribute to social science knowledge. I agree with him that action-oriented demonstrations as currently conducted have much further to go.

I am uncertain what can be done about action-oriented demonstrations. While they do not serve a social science function, many service providers and program staff do not consider knowledge building as important as providing services. The opening quote captures their feelings quite well. We fund this type of demonstration for several reasons. Policymakers often want to accomplish something "on their watch" to improve the social welfare. Often they feel the press of time and are more comfortable with service delivery than with investing in knowledge development. Rossi correctly points out that policy time and evaluation time are in two different dimensions.[4] Policymakers want their evaluation results tomorrow, or at least this year. Evaluators know that respectable evaluations of policy demonstrations generally take three to five years at a minimum. When staff try to do demonstrations in much shorter times, they usually end up compromising social science as a result. Federal staff have limited technical skills and limited leverage to make a substantial case against such demonstrations. Sometimes policymakers also justify their skepticism of government research and evaluation based on their personal experience (and some empirical evidence) that most evaluations are really not very policy relevant. Little of the onus for this problem, in my judgment, can be laid on the Congress.

The Congress, however, may be able to play a constructive role in reducing significant investments in action-oriented demonstrations. (Although, given my experiences with congressional oversight, I am not overly optimistic.) Congress could work to establish a *constructive dialogue* with executive-branch agencies concerning the use of and results from demonstration appropriations, encourage syntheses, and use legislative language to provide a framework (without too many specifics) to foster the notion that demonstrations are for knowledge building, not just service delivery. Demonstration research funds might then be more constructively allocated by executive-branch agencies. Such a stance might give federal research staff more leverage in budget discussions and in allocating resources to research demonstrations and evaluations which appropriately balance action and assessment.

I believe that large "black box" demonstrations can contribute to social science, if properly designed and executed. Although I am concerned that we have very little to guide federal staff in designing, procuring, and managing demonstrations operated by your ordinary American agency (YOAA), we have considerable room for improvement. (A key ingredient to improve federal demonstration procurement may be the recruitment and retention of qualified federal technical staff.)

Members of the academic community have given those of us in the demonstration-evaluation procurement business some broad guides which "bound the problem" in deciding when and how to do large-scale demonstration evaluations. On the one hand, such an investment is appropriate when policymakers genuinely want the information, are in doubt about the answers, and are willing to wait for the results.[5] On the other hand, it is inappropriate when demonstrations are used to postpone decisions, to duck responsibilities, to improve public image, or solely to fulfill a grant requirement.[6] (Fortunately, I have not encountered this latter extreme in ASPE during my twenty-year tenure here.)

Michael Wiseman gives solace to those of us concerned about the policy relevance of our demonstrations. He describes how demonstrations can and sometimes do influence policy.[7] In the same collection of papers, on the other hand, David H. Greenberg and Marvin B. Mandell caution us on the limits of this influence.[8] They survey the welfare-to-work and evaluation-utilization literature and support Carol Weiss's hypothesis that three I's—ideology, interests, and (anecdotal) information—may influence whether an evaluation has much impact on decision making. That is, good evaluations seldom influence policy when there is internal consistency of these three factors before the evaluation results are in.[9] Evaluation has much more influence when there is lack of agreement among the three I's.

The principles embodied in the "Final Report of the Head Start Evaluation Design Project" discussed by Sheldon White may be generalized to other federal research/evaluation situations and might also serve as additional guidance to federal staff managing large demonstrations so that they can contribute to policy and social science.[10] I generalize the following suggestions from the final report and my own experience:

1. Develop a research strategy that has several projects rather than one large one.

2. Always make assessment an equal partner to action.

3. Use diverse methodologies and measures.

4. Identify and promote the use of a common set of variables that can be synthesized across projects. (For example, program participation has several uses and interpretations. We should encourage use of one definition or variables that can measure participation, given several definitions.)

5. Variables should cover a diverse set of outcome domains—individual, family, institution, and community.

6. Use valid techniques appropriate for the specific populations involved. (That is, do not use measures on children from low-income families that have only been tested on middle-class children.)

7. Use longitudinal designs.

8. Look at what works for whom. (I agree with critics of experiments that only compare average outcomes. We need information on the treatments and on differential subgroup impacts.)

9. Establish archives of data for secondary analysis. (The Institute for Research on Poverty is attempting to do this with data from the employment and training demonstrations.)

10. Invest in improving measures. (Development of measures is a sort of public good. As a consequence, we are probably underinvesting in this activity as a society.)

11. Utilize administrative data bases as well as other measures. (Administrative data, if reliable, are usually cost-effective in comparison to other measures. Their use often has a secondary value of improving the quality of the administrative data for administrative purposes.)

12. Periodically synthesize the results of a body of work. (*From Welfare to Work* is an example.)[11]

Many of the people involved in the two-generation strategy have been attempting to coordinate efforts in ways congruent with these principles. The JOBS evaluation also seems to be following some of these themes.

In conclusion, Weiss's three I's give me pause concerning Head Start evaluation. (Some people still consider Head Start to be a demonstration program, even after twenty-five years of operation.) The popular press and many others say that we should spend more on Head Start because it is one antipoverty program that we know works.[12] What we know is that comprehensive early-childhood programs for low-income preschool children can make a difference in educational attainment and life course and that many Head Start grantees operate programs that contain most or all of the elements of the "hothouse" programs studied and evaluated.[13] YOAA Head Start grantees might be able to emulate these results; however, "virtually no longitudinal studies of strong design have been carried out on *regular* [emphasis added] Head Start programs."[14] I believe in my heart that Head Start is a good program for these children; it is probably as effective as or more effective than the alternative uses of the funding; most of the Head Start children obtain some positive results from attending. However, evaluation research evidence from YOAA Head Start programs is long overdue and needed to bolster opinions about the efficacy of the program. When everyone around me is saying good things about a program—when the three I's are aligned, which is so seldom the case in human services programs—should a professional policy analyst say, "Hey wait a minute"? Or should he stand back quietly while the strong man struggles valiantly and spends himself in this worthy cause? ■

---

[1]The three-function framework was taken from Thomas K. Glennan, Jr., "The Management of Demonstration Programs in the Department of Health and Human Services," Rand Publication Series R-3172-HHS, March 1985. For a discussion of the Low Income Opportunity Board demonstrations, see Michael E. Fishman and Daniel H. Weinberg "The Role of Evaluation in State Welfare Reform Waiver Demonstrations," in *Evaluating Welfare and Training Programs*, ed. Charles F. Manski and Irwin Garfinkel (Cambridge, Mass.: Harvard University Press, 1992).

[2]This is an acronym first coined by Peter H. Rossi. See, for example, Richard A. Berk and Peter H. Rossi, *Thinking about Program Evaluation* (Newbury Park, Calif.: Sage Publications, 1990).

[3]The term "black box" seems to me to be used to describe a situation which includes two concepts: random assignment and limited data. The design approach and the data-gathering strategies are two separate and independent decisions. Some critics may use the term as if the two are related rather than being decided upon independently.

[4]Berk and Rossi, *Thinking about Program Evaluation.*

[5]Richard P. Nathan, *Social Science in Government: Uses and Misuses* (New York: Basic Books, Inc., 1988).

[6]Carol H. Weiss, *Evaluation Research: Methods of Assessing Program Effectiveness* (Englewood Cliffs, N.J.: Prentice-Hall, 1972).

[7]Michael Wiseman, ed., "Research and Policy: A Symposium on the Family Support Act of 1988," *Journal of Policy Analysis and Management*, 10, no. 4 (Fall 1991), 588–666. (Available as IRP Reprint no. 656.) He points to the influence of the OBRA evaluations on the 1988 Family Support Act.

[8]Greenberg and Mandell, "Research Utilization in Policymaking: A Tale of Two Series (of Social Experiments)," in Wiseman, "Research and Policy: A Symposium on the Family Support Act of 1988."

[9]Carol H. Weiss, "Ideology, Interests, and Information: The Basis of Policy and Positions," in *Ethics, Social Science, and Policy Analysis*, ed. D. Callahan and B. Jennings (New York: Plenum Press, 1983).

[10]"Final Report of the Head Start Evaluation Design Project," prepared under contract no. 105-89-1610 of the Office of Human Development Services, DHHS, with Collins Management Consulting, Inc., December 1990.

[11]Judith Gueron and Edward Pauly, *From Welfare to Work* (New York: Russell Sage Foundation, 1991).

[12]For example, see Lisbeth Schorr, *Within Our Reach: Breaking the Cycle of Disadvantage* (New York: Doubleday, 1989).

[13]Raymond C. Collins and Patricia F. Kinney, "Head Start Research and Evaluation: Background and Overview," a technical paper prepared for the Head Start Evaluation Design Project, Head Start Bureau, Washington, D.C., 1989. (Hothouse programs are programs run under ideal conditions of resources, staffing, training, and theory.)

[14]Collins and Kinney, "Head Start Research and Evaluation," p. 22.

Evaluation under real-world constraints

by Steven H. Sandell, Director, Division of Policy Research, Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services

While others have summarized or written about some of the conceptual issues discussed at the conference, I am writing from the perspective of a government research/evaluation office charged with actually implementing evaluations. I will emphasize the implications of real-world constraints in conducting evaluations.

Constraints on conducting evaluations come in all shapes and sizes. Limited knowledge, administrative and resource constraints, time horizons, and organizational and design limitations result in a substantial trade-off between obtaining information that increases scientific knowledge (about behaviors or about effective evaluation strategies) and determining how a specific program is working. The constraints force the acceptance of less than ideal evaluation designs. Researchers, who tend to emphasize problems of theoretical interest, should be challenged to find solutions for the analytic problems created by these operational constraints.

The knowledge constraint

Inadequate knowledge has an immediate impact on the design. Uncertainty about the size of the probable effect, where to look for effects, subgroup impacts, sample attrition, and control of conditions affecting the treatment and comparison groups impinges on the design of the evaluation. Learning from the first round of work-welfare demonstrations has been reflected in the structure of the JOBS evaluation. Learning from the current two-generation program evaluations will allow fine-tuning of future studies.

Gaps in social science knowledge about the expected effects of treatments limit cost-saving decisions. With knowledge about who will be affected by treatments, stratified samples can be used. Without that knowledge, samples must be larger and more universal. Knowledge about the variance of treatment effects leads to a sampling strategy that improves statistical efficiency. Findings from previous research about the time pattern for decay of treatment effects lead to evaluations designed with an appropriate length of time in mind. Without such findings, the evaluation period could be too long, wasting resources, or too short, missing important outcomes or overstating real impacts.

Administrative constraints

Administrative constraints stem from the expected interaction of human nature and the political process. Everyone wants to find evidence, as soon as possible, that a favorite program is working. No one really wants to find out that a pet program doesn't work. Is it worth spending limited evaluation dollars on a program that cannot be shown to have significant positive effects? The opposition's program should be subjected to a rigorous evaluation, but our program, which we know in our hearts works well, doesn't need it. Often program legislation is designed with evaluations mandated, but with requirements that militate against developing scientifically optimal research designs.

Limited budgets and limited time

Academics, and even government policy analysts, easily offer suggestions on how specific evaluations can be improved. These suggestions often fail to take into account real-world budget constraints and trade-offs. Lengthened time periods to observe treatment effects are almost always useful but costly. Increasing the sample size conflicts with use of the resources for longer surveys or other data collection. Discussion at the conference was useful because these constraints were (at least implicitly) taken into account.

Time constraints in evaluations have several dimensions. First, results are usually desired by policymakers at a specific time, often stipulated in legislation. Sometimes funding and reauthorization decisions, which depend on legislative calendars, are dependent upon evaluations. Because programs evolve over time (reflecting changes in purpose, external factors, funding levels, and personnel), the time period for an evaluation can affect the results. Speedy evaluation of new programs that require shakedown periods may give premature and incorrect answers to important questions.

Organizational limitations

Complex programs often have multiple sponsors and service deliverers. Organizational perspectives affect the defining of evaluation questions as well as the evaluation itself. Programs with multiple goals, sponsors, clients, and outcomes require that priorities be established in developing an evaluation design.

Design limitations

Finally, the benefits of experimental designs are limited by the treatments that are controlled. The point of random assignment determines the nature of the questions the experimental design can directly address. Effects that take place before or long after the point of random assignment must be scrutinized using the same techniques used in nonexperimental analyses. The superiority of experimentally designed evaluations depends on the importance of the question(s) that are treated experimentally. If there are several important questions and only one can (practically) be treated experimentally, then it is somewhat misleading to label the results with respect to those other outcomes as experimental.

Conclusions

Discussion at the conference not only confirmed the existence of these constraints, it crystallized my thinking to deal realistically with them. First, all good things cannot be accomplished in a single evaluation: Constraints require making choices among all scientific and policy goals. Second, it is likely that under some circumstances (because of the juxtaposition of several constraints) a useful evaluation cannot be conducted. It is important to be realistic about what can be accomplished under specific circumstances. If, for example, owing to inadequate samples or budgets, a credible impact evaluation cannot be carried out, it is helpful to recognize that fact early and conduct instead a decent process evaluation.

Notwithstanding my emphasis on constraints in this short article, I came away from the conference with a positive outlook. Under most circumstances, a useful evaluation can be conducted, despite programmatic, budgetary, and other conditions that circumscribe the options. The scientific paradigm of building on previous research can be applied to evaluation strategies and should lead to increased subject-area and evaluation knowledge, as well as to the required program-specific information. ■

Comprehensive family service programs: Evaluation issues

by Sharon McGroder and Stella Koutroumanes, staff members of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services

The major purpose of the IRP/ASPE conference, "Evaluating Comprehensive Family Service Programs," was to help define critical evaluation issues associated with evaluating multifaceted social programs. Additional objectives were to bring together evaluators and researchers from different fields to promote familiarity with current efforts in these fields and to help government agencies conceptualize and structure future evaluation research.

We believe that the conference was very successful in accomplishing these objectives.

The conference presented state-of-the-art programs and demonstrations aimed at assisting families through an array of coordinated services. The consequential challenges to evaluation research became clear. Our comments here will summarize our impressions of key evaluation issues raised.

Issues raised

Limitations of experimental design. The evaluations presented at the conference employed a variety of methodologies. Both the JOBS evaluation and the Comprehensive Child Development Program Impact evaluation, for example, use experimental designs—mandated in federal legislation—to determine program impacts. The Youth Opportunities Unlimited Initiative (YOU), on the other hand, is not proposing any control groups or comparison sites with which to compare the effects of the intervention; consequently, it is unclear how program impacts will be ascertained.

It became immediately clear that traditional welfare research methodology—the experimental design—may not be sufficient in some instances or necessary in others to evaluate comprehensive family service programs. First, the federal government designed these family service programs to be flexibly implemented in order to respond to the particular needs of families in a particular community. Consequently, the federal government does not prescribe any specific model of how services should be delivered nor, in some cases, which services should be delivered. Thus, unlike traditional research in welfare economics, which often tests the effectiveness of a program model, describing the "treatment" in comprehensive family programs is difficult.

Moreover, even if random assignment to "program" and "comparison" groups yields differential impacts, experimental designs do not explain what it was about the "treatment" that produced these results. Was it a certain subset of services? A particular delivery mechanism? Was the overriding contributor to success a specific philosophy or an energetic program director? For this reason, there is a current trend in social service research to look beyond the question of "did the program work?" to explore "what worked, for whom, under what circumstances?" This trend reflects the multiplicity of components within a comprehensive family service program, recognizes the heterogeneous population being served by these programs, and acknowledges that one "treatment" may not be equally effective in every circumstance. Answers to "what works for whom?" yield the kind of information program planners and policy analysts need if they are to design and target effective programs and policies.

So while questions on overall program impact can be answered by comparing relevant outcomes for the experimental and control groups, questions on "contributors to impacts" cannot be answered by an experimental design. Ascertaining which program components contributed to impacts can be better explored with nonexperimental techniques, most notably, multivariate analyses.

Integrating qualitative data. A discussion of qualitative data and methodologies was particularly lively. Conference participants agreed that process studies, case studies, and use of ethnographic and other qualitative data can yield additional information about why or how an intervention was successful. We concluded from this discussion that researchers need to integrate qualitative and quantitative evaluation approaches to more fully describe program impacts.

While there was agreement on the need to explore the roles of case-study approaches, qualitative measures, and process evaluations in designing evaluations, there was concern about the general lack of "rigor" in applying these measures and methodologies. James Heckman commented that most evaluation research tends to be atheoretical, lacking conceptual frameworks and behavioral models from which research questions should be derived and the appropriate methodologies employed.

The need for a conceptual framework. Conference participants also observed that an analytic plan for the data generated from an evaluation is often not developed until well into program operations and data collection. Without a conceptual framework or model to guide inquiries, evaluators sometimes resort to "fishing" through the data to see what interesting relationships emerge. This procedure may be acceptable in cases where very little is known about the topic and researchers are navigating unknown waters—say, in basic academic research. But if the purpose of an evaluation is to answer particular questions—which is usually the case in policy research and program evaluations—then it is unacceptable to design an evaluation and gather data without first proposing a conceptual framework, specifying hypotheses to be tested, and designing the appropriate analysis plans which address these key research questions.

Measurement issues. Some important measurement issues were also raised at the conference. A recurring theme was the need for more basic research on ways in which to measure impacts and to specify which outcomes we want to measure. Standardization of measures for use across projects is an urgent need; there is little agreement on, for instance, the measurement of program participation. A coherent set of common baseline and outcome measures, of process and participation measures, would be of immense benefit. Moreover, since interventions are often aimed at ameliorating problems faced by both parents and children, this raises questions on who is the unit of analysis: Is it the child? For what outcomes? Is it the parent(s)? For which outcomes? Is it the parent/child relationship and broader measures of family functioning? Researchers will need to struggle with these issues resulting from the trend toward more comprehensive family service programs.

Major developments in the design of evaluation research

Over the years, we have observed three major developments in the field of evaluation research design which converged at the conference. First, we have witnessed the incorporation of qualitative and quantitative evaluation approaches to more fully describe program impacts. For example, the Comprehensive Child Development Program has on-site ethnographers to document patterns of service utilization. It is hoped that their reports will shed light on why certain outcomes were or were not achieved.

Second is the recognition of the need to describe the process through which a program has impacts. For example, the JOBS evaluation contains a process and implementation study, which will explore individuals' patterns of participation in JOBS, given their baseline characteristics and specific site attributes, and how this relates to outcomes. Exploring the dynamics of the black box through process evaluation and implementation studies is an important aspect of these family service programs.

Third is the tendency to not explicitly state formal hypotheses. We believe this results when little is known about a particular area. Initially research focuses on descriptive information using case studies and ethnographic methods to provide an overview of the issue and suggest hypotheses for further study. As patterns emerge, conceptual frameworks are derived and hypotheses developed, from which targeted research questions are designed. For example, the YOU demonstration is intended to have impacts on the community which in turn will improve outcomes for individuals. Little research is available, however, to suggest hypotheses on how this can be done. Consequently, it is acceptable that hypotheses are not explicitly stated, because of the exploratory nature of this demonstration. On the other hand, the JOBS evaluation relies on a history of research from which current hypotheses are formed on the relationship between education and employment programs and self-sufficiency. In this case, it is necessary to rigorously test clear hypotheses in order to answer important policy questions.

The conference impressed upon us the fact that evaluation of comprehensive family service programs is in its infancy; as a result, hypotheses are not explicitly stated and analytic plans are not specific. We believe there must be some tolerance for this ambiguity, as long as researchers strive to incorporate findings into a growing knowledge base.

Consequently, we believe that researchers in every social science discipline have a role to play in refining conceptual frameworks, developing interdisciplinary hypotheses, and specifying research questions in the area of comprehensive family services.

These three major developments have led to a new and visionary approach to evaluation. The report "Head Start Research and Evaluation: A Blueprint for the Future"[1] has led the way to rethinking how to evaluate multisite national programs. We view this as containing three steps. The first step consists of outlining the scope of the evaluation by framing the issues, clarifying the analytic plan, and specifying a common set of input and outcome measures. The second step consists of allowing the local program to operate as usual, with local evaluators collecting the process and impact data. The last step consists of drawing conclusions on major themes within and across programs in order to help explain variations in outcomes as site and program characteristics vary. At this point, research findings can be translated into practice and policy.

Next steps

It is precisely because of the difficulty in evaluating comprehensive family service programs that it is so important to conduct research systematically and begin to build upon previous work in order to push forward the field of research on family service programs. This task entails conducting a synthesis of research activities and disseminating the findings to researchers, policymakers, and analysts. To facilitate this process, ASPE and IRP should consider options for follow-up to the conference. Activities could include commissioning monographs or sponsoring technical working groups to address some of the methodological issues and recommendations that emerged at the conference. IRP should be actively involved in developing methodology and in structuring future evaluations. Such technical assistance would encourage researchers to both draw upon and add to the existing knowledge base of social science research. ■

Because of an error in weighting data from the October Current Population Survey, Figures 7 and 8 are incorrect in Robert M. Hauser, "What Happens to Youth after High School," *Focus* 13:3 (Fall and Winter 1991). The correct figures are shown below. The correction does not change major trends and differentials. However, corrected rates of college entry are lower than those originally estimated in each racial-ethnic group.

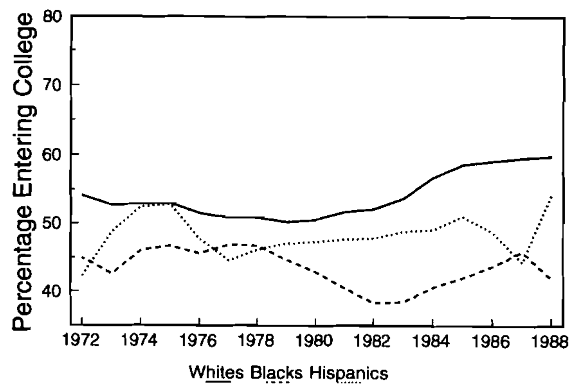

Figure 7. College Entry among Recent High School Graduates: White, Black, and Hispanic Men, 1972–1988

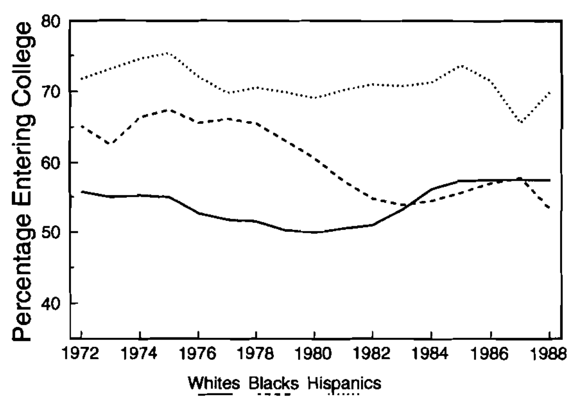

Figure 8. College Entry among Recent High School Graduates with the Average Social Background of Whites: White, Black, and Hispanic Men, 1972–1988

[1]"Recommendations of the Advisory Panel for the Head Start Evaluation Design Project," prepared under contract no. 105-89-1610 of the Office of Human Development Services, DHHS, with Collins Management Consulting, Inc., September 1990.