University of Wisconsin–Madison
Institute for Research on Poverty

Focus

Volume 14

Number 1

Spring 1992

ISSN: 0195–5705

Measuring poverty

by Patricia Ruggles

Patricia Ruggles is a senior research associate at the Urban Institute and a member of the National Advisory Committee of the Institute for Research on Poverty. She recently published *Drawing the Line: Alternative Poverty Measures and Their Implications for Public Policy* (Washington, D.C.: Urban Institute Press, 1990).

A great deal has been written about the measurement of poverty in the United States over the past several decades. As that literature demonstrates, poverty is ultimately a normative concept, not a statistical one. Although this article focuses on a set of statistical issues in the measurement of poverty, in the final analysis setting a poverty level requires a judgment about social norms. While analysis of statistical data can be very helpful in providing some basis for judgment, such a judgment cannot be made on statistical grounds alone. As Adam Smith put it more than two hundred years ago, poverty is a lack of those necessities that "the custom of the country renders it indecent for creditable people, even of the lowest order, to be without." Such necessities cannot be identified in some neutral, scientifically correct way—they do indeed depend on the "custom of the country," and some notion of what that custom requires must enter into their selection.

The United States has a set of official poverty thresholds, established more than twenty-five years ago. Although these standards may have represented a reasonable social minimum in 1963, normative standards change over time, and norms such as the poverty line must therefore be reassessed periodically. It is appropriate to start this reassessment by discussing our concept of poverty and considering why we might want to measure poverty in the first place.

What is poverty and why should we care?

Probably the most basic questions we can ask about poverty are what is it and why should we care about it? There are of course many different possible answers to these questions, as past debates over poverty and antipoverty policies have amply illustrated. Some writers have defined poverty to include any type of major deprivation, whether material—lack of specific goods and services—or more intangible—lack of access to good jobs, lack of appropriate role models, and so forth. And membership in the "underclass," which at least overlaps with the population in poverty, is often defined on the basis of behavioral factors like teen pregnancy or low attachment to the labor force, rather than on the basis of material deprivation alone.

Writers on poverty measurement for program and policy purposes, however, most often focus on measures of material well-being rather than on behavioral factors or more intangible forms of deprivation affecting the poor. Measures of access to goods and services—typically based on the income needed to support a minimal level of consumption—are used by such writers to measure material well-being. It is not that other types of deprivation are seen as unimportant; it is simply that the first goal of antipoverty programs is typically to provide minimally adequate levels of material well-being, and a measure with a material focus is needed to gauge the need for and success of such programs.

Even economists who are interested in these measurement issues sometimes treat poverty as if it were merely a special subset of the problem of inequality, however. In fact, the literature on the measurement of inequality is both much broader and more comprehensive than that on poverty measurement, and some writers seem to feel that good measures of inequality may preclude the need for a separate, specific measure of poverty. After all, if there were no inequality, presumably there would be no poverty either. For many such writers, any consideration of poverty or of antipoverty policies automatically translates into a consideration of the distribution of income and wealth.

From the point of view of the policymaker, however, a concern about poverty does not necessarily imply any interest at all in broader issues of distribution. Many policymakers start instead with the idea that intuitively formed the basis for the War on Poverty of the 1960s: that there is some minimum decent standard of living, and a just society must ensure that all its members have access to at least this level of economic well-being.

Typically, policymakers who express concerns about poverty either are thinking of some basic notion of "decency" of this type and/or are worried about the impacts of very low levels of consumption on the future needs, abilities, and behavior of those who are poor. Either type of concern lends itself most readily to a poverty measure that is defined in terms of some specific level of consumption that is considered to represent a necessary minimum—a "minimum decent standard of living."

Although the concept of a minimum standard can be made operational in a number of different ways, all of the poverty standards that might do so have two important features in common. First, they focus on economic well-being, not behavior, beliefs, or general levels of satisfaction or happiness. This is not to imply that these other things are unimportant—it's just that few policymakers consider them directly relevant to the policy goal of providing a decent standard of living, which is typically thought of in material terms. For the purpose of assessing programs that are explicitly designed to improve economic circumstances, a poverty measure that focuses on economic resources rather than on these other factors will be most useful. And second, of course, such a poverty standard must focus on those members of society whose command over goods and services is most limited. It is this feature that distinguishes a poverty measure from broader measures of inequality.

How should we measure poverty?

These features narrow the field of possible poverty measures a little, but there are still many measures that could be constructed that would meet both of these criteria—including the official U.S. poverty measure. Although this measure is somewhat flawed, it is inevitable that any alternative proposal will be compared to it. It is therefore helpful to outline its major features before turning to any discussion of alternatives.

The official U.S. poverty standard grew out of a series of studies undertaken by Mollie Orshansky for the Social Security Administration in the mid-1960s.[1] Orshansky faced the same problem that statistical agencies today face in developing poverty measures—statisticians are typically very uncomfortable with the idea of making normative judgments about how much people "need." Orshansky addressed this problem by starting with a set of minimally adequate food budgets for families of various sizes and types that had been calculated by the Department of Agriculture and that therefore had some claim to "scientific" accuracy.

To obtain a poverty line, she simply multiplied these minimum food budgets by a factor of three, on the assumption that food typically represented about one-third of total family expenditures. This one-third estimate, in turn, came from a 1955 food consumption survey, and was probably already outdated in 1963 when Orshansky first used it—consumption data from 1960–61 indicate that food consumption was closer to one-fourth of the typical budget by then. Nevertheless, according to Orshansky's scale, any family whose income was less than three times the cost of the minimum food budgets of the Department of Agriculture was classified as poor.

In 1969 a slightly modified version of the Orshansky scale was mandated by the Bureau of the Budget as the standard poverty measure to be used by the government statistical establishment as a whole. Since 1969 the Orshansky poverty scale has been subject to considerable criticism, but, with relatively minor changes, it still forms the basis for our official poverty measures. The original Orshansky measure has been updated for changes in prices since the 1960s, but no adjustment has been made to take account of any other changes in needs or consumption patterns that have occurred over this time.

This official poverty measure thus consists of a set of dollar amounts—called thresholds—that vary by family size. If a family of a given size has an income below the threshold for its size, the family is considered poor. Families with incomes over the threshold are counted among the nonpoor. Income, for the purpose of measuring poverty, consists of money income before taxes. It does not include noncash forms of income such as food stamps and Medicaid. Table 1 shows these official poverty thresholds for families of different sizes.

These poverty thresholds provide a fairly crude measure—families with incomes only a few dollars apart aren't really that different, though one will be classified as poor while the other is nonpoor—but this measure serves to give some indication of major changes in the size of the poverty problem. The raw poverty count, or the percentage of the population in poverty, is often supplemented by other measures, such as the "poverty gap," which measures the aggregate amount by which families with below-poverty incomes fall below the line.

Table 1

Weighted Average Poverty Thresholds in 1988

Size of Family Unit	Threshold (Dollars per Year)
One person (unrelated individual)	$6,024
15–64 years	6,155
65 years and over	5,674
Two persons	7,704
Householder 15–64 years	7,958
Householder 65 years and over	7,158
Three persons	9,435
Four persons	12,092
Five persons	14,306
Six persons	16,149
Seven persons	18,248
Eight persons	20,279
Nine persons or more	24,133

Source: U.S. Bureau of the Census, Current Population Reports, Series P–60, no. 166, *Money Income and Poverty Status in the United States: 1988*, p. 88.

Note: The official income and poverty estimates are based solely on money income before taxes and do not include the value of noncash benefits such as food stamps, Medicare, Medicaid, public housing, and employer-provided fringe benefits.

In my view, the thresholds that make up our official poverty measure are now quite outdated as indicators of real family needs, however. To understand why, it is helpful to think about why poverty thresholds must be adjusted at all.

Adjusting poverty measures for change over time

The most obvious reason for adjusting thresholds over time is because prices change. An amount of money that was adequate for a family in 1967 would have bought far less ten years later. Even if one thinks of poverty as resources below some "absolute" level of consumption that is not expected to change in real terms over time, it is still necessary to adjust for these price changes—in other words, to make the real purchasing power of the standard the same over time. As mentioned earlier, this is the one adjustment made annually to the thresholds.

Prices, however, are not the only things that change over time. People's incomes and family structures also change, and so do the goods and services that are available for consumption. Since 1955, for example, when the consumption data underlying our official thresholds were collected, major changes in consumption patterns have occurred.

Some goods commonly consumed today did not even exist in 1955, and others were relatively rare—for example, according to consumer expenditure data the average family did not have a telephone. Most families with children could count on the services of a full-time homemaker in 1955, and many fewer children lived with only one parent, so few families faced child care expenses. And the relative price of housing, in particular, is very different today from its price in 1955; in 1955, the average U.S. family spent about one-third of its income on housing, and today it spends about 42 percent. This change especially affects the poor, who spend a much larger share of their budgets on housing than do typical families. All of these changes, and others like them, contribute to changes in minimum family needs over time.

Thus, the most obvious problem in adjusting an absolute standard only for price changes is that over the very long run the goods available to be consumed will change almost beyond recognition—and these changes in turn will affect our perception of needs. A century ago, for example, few households had indoor plumbing or electricity. A set of minimum consumption needs established in 1890 and indexed for changing prices alone would today exclude such goods, therefore, even though they are now considered basic needs.

Further, as long as there is some continued real growth in the economy as a whole, incomes will generally rise relative to prices (although during recessions price gains may temporarily exceed wage increases). As a result, if poverty

thresholds are adjusted only for prices, they will fall farther and farther behind average standards of living.

Relative versus absolute measures of poverty

Relative measures of poverty, typically based on some fixed relationship to aggregate or average income, are often advocated by economists to correct this problem. Such measures do capture at least those changes in minimum acceptable living standards caused by rising real incomes. Under this approach, as incomes rise in general, poverty thresholds are adjusted upward by a similar percentage. The most commonly proposed relative poverty measure is a threshold set at some specific percentage of the median income—most often, 50 percent.

Historically, the earliest thresholds calculated by Orshansky, which were for 1959, had a four-person standard that was equivalent to about 49 percent of the median income for that year. Because growth in incomes substantially outstripped growth in prices between 1959 and 1967, however, by 1967 the four-person standard had already declined to about 43 percent of the median income for families as a whole. By 1988 this standard had declined further, to about 37 percent of median family income. Table 2 shows poverty thresholds for a three-person family in 1988 under a variety of different poverty measures—the relative standard is in column 2.

Opponents of the relative income or consumption approach to poverty measurement argue that it presents too much of a "moving target" for policy assessment, and that it is in some sense not fair to judge our antipoverty efforts against such a standard. Indeed, this type of standard will rise most rapidly in periods of rapid economic growth, when most people, including most of the poor, are likely to be experiencing a growth in their real incomes and consumption opportunities. Even though low-income families may consider themselves better off under such circumstances, they would not be judged less poor under a relative poverty measure unless their income or consumption levels actually rose more than did the median level for society as a whole.

To put it another way, poverty cannot decline under a relative poverty measure without some change in the shape of the income distribution as a whole. It is much more difficult to design (let alone enact) policies to carry out such a major redistribution than it is to design programs to improve the consumption opportunities of the poor.

One problem with the use of relative income as the basis for indexing thresholds over time, therefore, is that relative measures may be more closely tied to changes in income distributions or inequality than to changes in minimum needs. If the major policy purpose of a poverty line is to set a standard of "minimum adequacy" to be used in program and policy assessment, a standard that, for example, falls in real terms during recessions is less than ideal, since presumably the real needs of the poor do not fall similarly. More broadly, a measure based on relative income, while involving fewer apparently arbitrary judgments of needs than an absolute standard, is correspondingly less closely linked to the basic concept of minimum adequacy.

Table 2

Poverty Thresholds for a Family of Three in 1988 under Alternative Poverty Measures

Family of Three	Official Measure	Relative Measure at 50% of Median[a]	Measure Based on Housing Consumption[b]	Measure Based on Updated Food Multiplier[c]
Threshold in dollars	$9,435	$12,737	$14,530	$15,850
Ratio of measure to official measure	1.0	1.35	1.54	1.68

Source: U.S. Bureau of the Census, Current Population Reports, P–60, no. 166, *Money Income and Poverty Status in the United States: 1988*; and Patricia Ruggles, *Drawing the Line* (Washington, D.C.: Urban Institute Press, 1990).

[a]Poverty threshold for four-person families set at 50 percent of the median income, and all other thresholds adjusted accordingly, using equivalence scales implicit in official thresholds.

[b]Based on Fair Market Rents and Housing Affordability Guidelines used in the Section 8 Subsidized Housing Program. See Ruggles, *Drawing the Line*, for details on the method of calculation.

[c]Calculated using the same general methods as the original Orshansky standard, but with a multiplier updated to reflect the changing share of food in family budgets. See Ruggles, *Drawing the Line*, for general discussion and details on the method of calculation (in Appendix A).

A poverty standard that is not increased as real incomes and consumption levels rise, however, runs the risk of becoming increasingly unrealistic over time. Clearly, over time "the custom of the country" changes, and our definition of necessities must change with it.

If our definition of minimum adequacy does not keep up with social norms for consumption, those whose incomes and consumption levels fall under the poverty line are increasingly likely to be out of the economic mainstream in other ways as well. For example, in the last decade alone the proportion of adult, nonelderly household heads in poverty who work full time has fallen from about 43 percent to under 36 percent. As real wages rise and the poverty line remains fixed in real terms, it is increasingly unlikely that someone who works a significant number of hours will remain poor, at least under the official definition. As a result, the poverty population comes to exclude most low-wage workers.

On the one hand, many such workers (and others among the near poor, such as retirees) may still experience real economic hardships, in the sense of being unable to afford those goods that the custom of the country deems neces-

sary. And on the other hand, those who are still poor under the absolute scale even as it declines in relative terms are in some sense a much more "hard core" poverty population than were those who were judged poor under this scale when it was first established in the mid-1960s. Because the line is so much farther from the norm for our society, people who fall under it are more likely to be those with particularly severe problems, or perhaps even multiple problems—the disabled, young single mothers, those with very little education and/or very low job skills. It is indeed a challenge to design programs that will help those with such major problems to become more self-sufficient.

Further, if the measurement-related aspects of this shift are not well understood, some analysts may misinterpret the evidence of increases in these problems among the poverty population. They may erroneously assume that our existing antipoverty programs are backfiring and actually creating a more severely handicapped poverty population over time. Or one less likely to want to work!

Alternative measures of poverty

If price-indexing an absolute standard isn't satisfactory because it doesn't reflect real changes in minimum needs, and indexing by relative income changes isn't satisfactory because income fluctuates too much and also isn't directly related to minimum needs, what *should* we do?

If minimum adequacy is indeed our major concern, a more direct approach is to re-estimate the market basket of "minimum needs" at regular intervals—such as every decade. This is essentially the approach now used by Statistics Canada, for example.[2] The specific updating methodology used in Canada is somewhat mechanistic, however. An alternative approach, which I would advocate for the United States, would be to update our set of absolute poverty thresholds for changes in needs and consumption standards over time by calling upon some set of "experts" to set normative standards of consumption for a market basket of specific goods, and then to revise those standards for changes in consumption at some set interval such as a decade.

Many commentators have argued that expert opinion as to family needs is in reality just as arbitrary and just as subjective as any other opinion—there is no scientific way to determine just how much of what goods any particular type of family really needs. In some abstract sense, this is true. In a broader sense, however, the same constraints that operated when Orshansky set her original thresholds would presumably continue to operate when thresholds were revised—estimates that were extremely far from a social consensus as to real family needs would meet with substantial criticism and would be unlikely to be adopted.

Family budgets that detailed projected spending on a variety of different goods would be particularly likely to be criticized by advocacy groups interested in specific goods if estimates for those goods were truly unrealistic. Criticism from housing advocates, for example, resulted in substantial revisions in proposed changes to housing subsidies in the early 1980s. To facilitate this process, however, certain safeguards would be appropriate—proposed revisions by expenditure category might be published in advance, for example, with provision for public comment, and analyses of actual spending patterns at various income levels might also be required for purposes of comparison.

A subsistence standard cannot simply be based on the actual consumption patterns of the poor, since presumably those consumption patterns have already been constrained by a lack of resources and may therefore be inadequate in important respects. A normative market basket should not exactly mirror the consumption of middle-income families either, however, since such families may spend more on "luxuries" than would be consistent with minimum adequacy. Presumably, most categories of consumption should fall somewhere between these two sets of consumption standards, and proposed standards that fail to do so deserve to be suspect.

FOCUS is a Newsletter put out three times a year by the

Institute for Research on Poverty
1180 Observatory Drive
3412 Social Science Building
University of Wisconsin
Madison, Wisconsin 53706
(608) 262–6358

The Institute is a nonprofit, nonpartisan, university-based research center. As such it takes no stand on public policy issues. Any opinions expressed in its publications are those of the authors and not of the Institute.

The purpose of *Focus* is to provide coverage of poverty-related research, events, and issues, and to acquaint a large audience with the work of the Institute by means of short essays on selected pieces of research. A subscription form with rates for our Discussion Papers and Reprints is on the back inside cover. Nonsubscribers may purchase individual papers from the Institute at $3.50 for a Discussion Paper and $2.00 for a Reprint.

Focus is free of charge, although contributions to the U.W. Foundation–IRP Fund sent to the above address in support of *Focus* are encouraged.

Edited by E. Uhr.

In other words, even though there is no one "right" bundle of consumption needs for the poor that all experts would agree on, we do know enough to eliminate a very large number of clearly wrong answers. In this sense, an expert-determined market basket need not be seen as essentially arbitrary, even if we concede that an exact determination of needs is not really possible. While experts who work for the government are likely to be under some political pressure to come up with poverty lines that are as low as possible, any consumption-based standard is still likely to exceed a standard that has been adjusted only for price changes over a very long period of time.

Setting normative consumption standards for a wide range of basic goods is indeed a job for experts and was clearly beyond the scope of my study.[3] As a substitute, however, I have considered two much more limited consumption-based measures: one that, like Orshansky's original thresholds, is tied to food consumption, and one that is based on housing needs. These two standards both indicate that a consumption-based approach would be likely to result in substantially higher thresholds than those found under the official measure. Thresholds for a three-person family under these two measures for 1988 are shown in columns 3 and 4 of Table 2.

Like Orshansky, I balked at the idea of making up my own consumption norms and instead chose some that were already in use, at least in some form. The "housing consumption" standard is based on data on "fair" rents and standards for housing as a share of total budgetary needs established under the Section 8 Subsidized Housing Program, the basic rental subsidy program for low-income families in the United States.[4] Details on the derivation of this standard are given in my study, but basically the norms are derived from program guidelines, which, unlike our poverty thresholds, recognize that housing prices in the United States have risen substantially as a share of budgets over the past fifteen years. Unfortunately, the program—and the rental housing cost data it uses—has only been in existence since 1975, so fully comparable estimates for the earlier years cannot be computed. Rough estimates based on other housing data, however, imply that poverty thresholds would have been in the neighborhood of the Orshansky standard in 1963, although, as Table 2 shows, by 1988 they would have been about 1.54 times the official standard.

The other consumption-based standard shown in Table 2, the updated food multiplier standard, is computed using an even simpler methodology. In this case, Orshansky's original approach of multiplying a basic food-need standard by the inverse of the share of food in the average family budget has been duplicated exactly, but with updated data on budget shares. Average families in the United States today spend about one-sixth of their budgets on food. As a result, today's multiplier would be about six rather than the estimate of about three that Orshansky used. Again, the updated multiplier approach produces thresholds well above the official ones—in 1988, for example, they would have been 1.68 times the official level.

It is worth reiterating that both the housing consumption and updated food multiplier thresholds are only rough proxies for a standard based on a more complete market basket of necessary goods. Further, it is possible that a set of thresholds based on a broader survey of minimum consumption needs would not have risen as much relative to the official thresholds as did the two alternatives considered here. On the other hand, the housing standard may actually understate needs, since housing costs make up a very large share of the budgets of most poor families, and the use of the relatively conservative estimates of the Department of Housing and Urban Development (HUD) of the appropriate budget share for housing may understate the impacts of housing cost increases on the needs of poor families.

In summary, a detailed examination of changes in the costs of a complete market basket of necessary goods would be the preferred approach to constructing a good estimate of changes in the needs of the poor since the mid-1960s. In the absence of such a study, however, the housing consumption standard developed here, which has been designed to be relatively conservative in its estimates of changes in needs, should provide a reasonable, if perhaps slightly low, estimate of the current minimum consumption needs of poor families.

Before turning to an examination of the impacts of alternative poverty measures on our perceptions of poverty, one further issue should be mentioned. All of the poverty measures discussed so far, including the official measure, consist of two parts: a measure of family needs, such as a set of poverty thresholds, and a measure of family resources—for example, family income. The official poverty measure uses pretax cash family incomes as the basic resource measure that is compared to the thresholds to determine whether or not a given family is poor. While this resource measure is far from ideal, the major focus of this article is on the measurement of needs rather than resources, and so for the sake of consistency pretax cash incomes have also been used as the resource measure here.

How much difference does using this rather limited resource measure make in examining either the incidence of poverty or changes in poverty rates over time? Ideally, one should be using a resource measure that does a better job of actually measuring the family's spendable resources. For example, taxes should be subtracted from resources, because families can't actually use the money they pay in taxes to purchase the goods and services included in the need standard. On the other hand, noncash benefits such as food stamps should be included in income, because these benefits do increase access to minimally necessary consumption. In practice, excluding taxes and including food stamps would have very little impact on either income or measured poverty rates in total, although this improved measure would change the specific families that appear to be poor.

Noncash benefits such as medical care are more controversial—while access to medical care is a necessity, counting the value of that medical care as if it were income can be misleading. Because medical care prices are so high, some families who are eligible for Medicare or Medicaid theoretically receive resources above the poverty line in medical benefits alone! In other words, even with no cash income at all, these families would not be counted as poor, even though they could not pay rent or buy food to eat. Clearly, a poverty standard that will not cover the costs of medical care does not reflect total needs, and treating medical care as if it were cash exaggerates the resources available to such families to meet other needs. For this reason, I would advocate excluding medical benefits, which are not fungible, from the basic income measure, and instead having a second need standard for health care. Under this system, families might be judged poor on the basis of their "cash-like" incomes, or they might be medically needy (or both or neither), but medical benefits would not be counted against nonmedical needs.

Trends in poverty over time: What progress have we made?

So far I have focused on ways to measure poverty, and have argued that our official poverty measure seriously understates the extent of the problem. How important is this understatement? Does it really change our perception of the size of the poverty problem or of the progress that we've made in combating poverty since the 1960s?

As Table 3 shows, the answer to these questions is yes—the poverty definition used can have very big impacts on our perception of poverty. Even under the official measure, poverty rates are very high today relative to the past; today's poverty rate is still above the level seen at any point during the 1970s, for example. But under the alternative poverty thresholds, poverty rates are not only much higher in every year, but the trend is also less favorable in the recent period.

In general, the higher the threshold, the greater the number of people who will be counted as poor. Because income is not evenly distributed, however, a given percentage increase in the poverty threshold does not necessarily translate into a proportional increase in poverty rates. In fact, because so many families have incomes in the neighborhood of the poverty line, changes in poverty thresholds almost always have a more-than-proportional effect on measured poverty rates.

For example, using a relative threshold set at 50 percent of the median implies an overall poverty rate of almost 20 percent, compared with the official poverty rate of 13 percent. Also, because median income has risen since the 1982–83 recession, the trend since 1982 looks worse under the relative measure than under the official estimates (see Table 3, column 2).

Table 3

Poverty Rates for Selected Years under the Official Measure and Three Alternatives

Year	Official Measure	Relative Measure at 50% of Median[a]	Measure Based on Housing Consumption[b]	Measure Based on Updated Food Multiplier[c]
1972	11.9%	17.9%	NA	17.3%
1977	11.6	17.4	20.7	18.0
1982	15.0	18.9	25.1	25.5
1987	13.5	19.7	23.4	25.9
1988	13.0	19.5	23.0	25.8

Source: Calculated from the Current Population Survey for years shown. Data for 1972–1987 from Patricia Ruggles, *Drawing the Line* (Washington, D.C.: Urban Institute Press, 1990), Table 3.4, p. 55. Data for 1988 from "Falling Behind: The Growing Income Gap in America," Joint Economic Committee, U.S. Congress, Washington, D.C., December 1990.

[a]Poverty threshold for four-person families set at 50 percent of the median income, and all other thresholds adjusted accordingly, using equivalence scales implicit in official thresholds.

[b]Based on Fair Market Rents and Housing Affordability Guidelines used in the Section 8 Subsidized Housing Program. See Ruggles, *Drawing the Line*, for details on the method of calculation.

[c]Calculated using the same general methods as the original Orshansky standard, but with a multiplier updated to reflect the changing share of food in family budgets. See Ruggles, *Drawing the Line*, for general discussion and details on the method of calculation (in Appendix A).

The two consumption-based estimates of the poverty rate are even higher; they imply that 23 to 26 percent of the population are poor. These measures also imply a worse trend than does the official measure. At best, they indicate that there has been little improvement in the proportion of the population in poverty since the recession of 1982–83.

These results are pretty shocking. Poverty rates of this magnitude imply that serious need is a much more widespread phenomenon than we are used to thinking, and they also imply that we have actually lost a substantial amount of ground against poverty since the early 1970s. It is tempting to think that poverty rates this high could only result from unrealistically high thresholds. The evidence is otherwise, however. The official poverty cutoff for a typical three-person family in 1989 would still be only $9885—or only about $824 per month. Under the Department of Agriculture's Thrifty Food Plan, which is an index representing a short-term, subsistence level of spending on food, such a family could expect to spend $260 per month on food alone. That leaves less than $565 for everything else—rent, medical expenses, child care, transportation, clothing, etc. This is not too realistic; rent alone would

consume most of that remainder. The national average fair market rent, as calculated by HUD, would be almost $500 per month for a two-bedroom apartment—and rents are of course even higher in the large cities where many of the poor live.

Given these costs for food and housing alone, a higher poverty threshold seems warranted. Even the housing consumption standard calculated above, for example, would only imply an income of about $1270 per month (1.54 x 824), or about $15,225 per year, for such a family to be considered out of poverty. Such a family would still need to spend about 60 percent of its income to meet its most basic food and housing needs, but some income would remain for things like Social Security taxes, child care costs, transportation, clothing, and other work expenses.

In summary, alternative poverty measures can change our view of the long-term trend in poverty as well as of the absolute number of people who are currently poor. We are all familiar with the story told by the official statistics, that poverty rates have fallen significantly since the mid-1960s, although even under official estimates poverty rates rose sharply during the 1980s, and even before the current recession we had not seen a return to the levels of the 1970s.

But the story appears even worse if we look at alternative measures of poverty as well. Using either an adjustment for rising incomes or for changes in consumption patterns, we find that today's overall poverty levels are comparable to those seen when war was declared on poverty in the mid-1960s.

Who are the poor today?

What kinds of people are included in today's poverty population? Who are these millions of people who are seen to be poor under these various definitions?

Just as striking as the differences in total poverty rates under alternative thresholds are the impacts of these alternatives on the composition of the poverty population, shown for 1988 in Table 4. Because the distribution of income varies across population groups, relative poverty rates will also vary depending on the level of the poverty threshold. This can be seen most clearly by comparing the poverty rates for the elderly to those for all persons under the various thresholds.

Under the official price-indexed thresholds, the poverty rate for the elderly is below that for all persons--12 percent for the elderly compared to about 13 percent for all persons. Under the relative-income-adjusted threshold, however, the rate for the elderly actually exceeds that for the population as a whole--22.9 percent for the elderly, compared to 19.5 percent for the population as a whole.

As poverty thresholds rise, the proportion of the elderly shown as poor rises even more relative to the proportion for the population as a whole. Under the updated food multiplier approach, for example, about 32 percent of the elderly would be counted as poor, compared to about 26 percent of the population as a whole.

This shift in relative poverty rates has important implications for public policy. One of the great antipoverty success stories of the past two decades has been the decline in poverty rates for the elderly population. Almost 30 percent of the elderly were poor in 1967 under the official thresholds, but by 1988 only 12 percent were. In contrast, the official 1988 poverty rate for the population as a whole is much closer to the 1967 level: 13 percent in the later year, compared to 14.2 percent in the earlier one. The official

Table 4

Poverty Rates for Selected Population Groups under the Official Measure and Three Alternatives, 1988

Poverty Rates for Various Groups	Official Measure[a]	Relative Measure at 50% of Median[b]	Measure Based on Housing Consumption[c]	Measure Based on Updated Food Multiplier[d]
All persons	13.1%	19.5%	23.0%	25.8%
Persons under 18	19.7	26.6	31.3	34.6
Persons aged 65 and over	12.0	22.9	28.6	32.4
Persons in female-headed families	32.8	43.7	48.2	51.3
Whites	10.1	15.9	19.3	22.1
Nonwhites	31.4	40.6	46.1	49.4

Source: Calculated from the March unrevised 1989 Current Population Survey, which provides data on family incomes in calendar year 1988. Taken from "Falling Behind: The Growing Income Gap in America," Joint Economic Committee, U.S. Congress, Washington, D.C., December 1990. Methods used to derive figures are discussed in Patricia Ruggles, *Drawing the Line* (Washington, D.C.: Urban Institute Press, 1990), Chap. 3 and Appendix A, and comparable figures for 1987 are given in Table 3.5, p. 57.

[a]Based on unrevised data. The Census Bureau has just released figures for 1988 which show the official poverty rate for all persons at 13.0 percent, but detailed data tapes containing these revised data are not yet available.

[b]Poverty threshold for four-person families set at 50 percent of the median income, and all other thresholds adjusted accordingly, using equivalence scales implicit in official thresholds.

[c]Based on Fair Market Rents and Housing Affordability Guidelines used in the Section 8 Subsidized Housing Program. See Ruggles, *Drawing the Line*, for details on the method of calculation.

[d]Calculated using the same general methods as the original Orshansky standard, but with a multiplier updated to reflect the changing share of food in family budgets. See Ruggles, *Drawing the Line*, for general discussion and details on the method of calculation (in Appendix A).

poverty rate for the elderly fell below that for the general population for the first time in 1982 and has remained below the overall poverty rate since then. Some analysts have argued that as the relative position of the elderly has improved—even as federal budget constraints have become tighter—a smaller proportion of our resources should be directed into programs serving the elderly. The data shown in Table 4 make it clear, however, that the degree of improvement in the relative status of the elderly is quite sensitive to the specific set of thresholds used.

Although changes in the relative poverty status of the elderly under alternative thresholds are the most dramatic examples of the impacts of the level of the threshold on the composition of the poverty population, the relative poverty status of other population subgroups can also be affected. Poverty rates for children, for those in female-headed families, and for nonwhites, for example, are always well above those for the population as a whole, but the gap does narrow slightly (at least in percentage terms) as thresholds rise.

In general, as poverty thresholds rise, the population seen as poor comes to resemble more closely a cross-section of the population as a whole—although obviously, under any of these definitions, children, those in female-headed families, and nonwhites are still far more likely to be poor than is an average white adult. Conversely, as discussed earlier, a threshold that is fixed in absolute terms, and which thus tends to fall relative to median income, will come to identify a narrower subset of the population as poor over time. This will occur even if there is no change in the overall distribution of income across demographic subgroups within the population as a whole.

Of course, being identified as "poor" or "not poor" does not make the individuals involved any better (or worse) off, but such a shift may have some political consequences. As the characteristics of the poverty population diverge farther from those of the "typical" family, the poor are likely to become more isolated politically and to be seen as an underclass whose problems are principally caused by their own aberrant behavior. This perception may in turn undermine support for programs designed to combat poverty.

Conclusion

The specific poverty measures that we use have played an important role in shaping our perceptions both of the extent of real economic need and of the characteristics of those who are most deserving of our help. Ultimately, different measures may well lead to different priorities in setting antipoverty policies. Probably the single most important aspect of a poverty measure, in terms of its impact on public policy, is the proportion of the population that it suggests to have inadequate levels of consumption. For that reason, this discussion has focused on setting poverty thresholds, and the implications of those thresholds in defining the poverty population.

Relative poverty measures appeal to many economists because they depend only on a fixed relationship to median income, and so one can set thresholds while avoiding the awkward and obviously value-laden process of defining need, except in some very global sense.

Ultimately, however, the relative measure is not a practical way to set poverty standards for the purpose of policy analysis. The basic flaw in this approach is that the concept of poverty that most people normally use does in fact imply some fairly specific value judgments—and these judgments are not consistent with the view that only people's relative levels of consumption, rather than their actual consumption, matter in assessing poverty. Not all needs or desires are generally considered equal in judging whether or not someone should be counted as poor. The need to eat regularly and to have someplace warm and dry to sleep is widely recognized; the need to own a particular brand of sneakers or jeans, while deeply felt by many teenagers, is rarely considered of equal importance by policymakers.

More generally, social and political concerns about poverty arise from many different causes, but almost all of them have to do either with basic notions of fairness and justice or with concerns about the impacts of very low levels of consumption on future needs, abilities, and behavior. In either case, these concerns are likely to be much stronger with regard to some types of consumption than others, and it is appropriate, in a policy context, to weight those types of consumption more heavily in determining need.■

---

[1]See Orshansky, "Counting the Poor: Another Look at the Poverty Profile," *Social Security Bulletin*, January 1965, pp. 3–26.

[2]See M. C. Wolfson, and J. M. Evans, "Statistics Canada's Low Income Cut-Offs: Methodological Concerns and Possibilities," Statistics Canada Discussion Paper, Ottawa, 1989, for a detailed description of the methodology used to compute the Canadian low-income cutoffs.

[3]Ruggles, *Drawing the Line: Alternative Poverty Measures and Their Implications for Public Policy* (Washington, D.C.: Urban Institute Press, 1990).

[4]See Ruggles, *Drawing the Line*, Appendix A, for details, both on fair market rents and their use in housing subsidy programs and on the calculation of the specific standard discussed here.

Evaluating comprehensive family service programs: Conference overview

Since 1986 a number of federal agencies have initiated large-scale demonstration programs designed to relieve the deprivation of parents and children in impoverished families. The evaluation of such programs formed the subject of a conference held in Washington on November 14–15, 1991. The conference was one of a series jointly sponsored by the Institute and the Office of the Assistant Secretary for Planning and Evaluation (ASPE) in the U.S. Department of Health and Human Services.

This conference series began in 1989 with a one-day workshop in Washington to provide ASPE staff and other members of executive departments with expert counsel on practical approaches to evaluating the programs created by the Family Support Act. The second meeting, more academic in tone, consisted of a two-day national conference in 1990 at Airlie House, Airlie, Virginia, where federal representatives, evaluation professionals, and academic researchers examined the assessment of welfare and training programs.[1]

The 1991 conference advanced into the more complex realm of projects offering services for disadvantaged parents and their children. Represented among its 120 participants and observers were academicians, professional evaluators, federal staff involved with program planning, and members of private foundations and service organizations. The programs discussed at the conference are those sometimes referred to as "two-generation interventions":

> A potentially powerful new strategy for assisting families in poverty is being tested in a set of new program models that target welfare-dependent women with young children. These models vary in several respects, but have a common strategy: they help families attain economic self-sufficiency through education and job training while also providing other services, such as parenting education and high-quality child care, that support children's healthy development. As two-generation interventions these programs show promise of addressing both immediate and long-term impediments to healthy development and educational success for poor children.[2]

By attempting to improve simultaneously the circumstances of parents and the life chances of their children, these programs span welfare and employment efforts on one hand and child development, child welfare, and social service efforts on the other, with the result that those operating the programs as well as those evaluating them represent a variety of disciplines and professions.

The evaluation projects for seven major programs were presented and discussed. Three of the programs were authorized by Congress (the Job Opportunities and Basic Skills Training Program, JOBS; the Comprehensive Child Development Program, CCDP; and the Even Start Family Literacy Program). Two originated in federal executive agencies (the Teenage Parent Demonstration and Youth Opportunity Unlimited, YOU). One is a state initiative (the Washington State Family Independence Program, FIP), and one (New Chance) is privately sponsored. In addition, programs still in early stages were briefly discussed, and the evolution of Head Start evaluations over the past twenty-five years formed the subject of a special presentation. Capsule descriptions of the programs and their evaluations accompany this article, and main features of the seven large projects are compared in Table 1, pages 14 and 15. The conference agenda appears on page 21.

The conference had four principal purposes: to summarize the state of evaluation methodology, to identify the key issues in assessing these complicated programs, to permit evaluators in different fields and disciplines to pool their knowledge, and to help ASPE structure future evaluations in the area of family services. The consensus, upon conclusion, seemed to be that while the meeting moved forward on all four dimensions, a significant contribution lay in providing the opportunity for evaluation contractors to exchange information concerning the nature, problems, and accomplishments of their projects.[3] Also important was the opportunity for federal staff members from the legislative branch (Senate and House, General Accounting Office, Congressional Budget Office) as well as the executive branch (Housing and Urban Development, Education, Labor, Agriculture, several agencies within Health and Human Services, the Census Bureau, and the Office of Management and Budget) to attend the deliberations and gain knowledge bearing on their own work.

Several themes materialized during the presentations and the vigorous discussions that ensued. The following summary attempts to capture major points.

The time and place for evaluations

Martin Gerry, Assistant Secretary for Planning and Evaluation in DHHS, noted in his introductory remarks that formal evaluation of social programs has taken on greater importance in recent years, amid growing concern over learning what works, and how well. Evaluation in the 1980s

of experiments by several states with welfare reform directly influenced the Family Support Act and encouraged Congress to include evaluation requirements in the authorizing legislation for two other programs (see capsule descriptions).

Conference participants noted the advantages that can accrue from a congressional mandate for evaluation. It strengthens the hand of government researchers who want to analyze the effects of public policy on individual behavior. It may open doors to funding by government agencies that would otherwise remain closed. And the specification of a particular form of evaluation, exemplified in the requirement that random assignment be used for the JOBS evaluation, can help researchers convince reluctant program operators that there is reason to assign clients to different forms of treatment.

The problem with this congressional attention, pointed out in other comments, is that it may impede evaluation design. The federal procurement process that is set in motion by a congressional mandate for evaluation sometimes occurs too early, before a program has been clearly developed—before there is certainty concerning what is to be evaluated. Allowance must therefore be made for changing the evaluation design as the program matures and alters. This may be accomplished by explicitly permitting and encouraging redesign as a program progresses.

In the case of programs whose effectiveness is contested and controversial, as is true of those involving family preservation services, it may be desirable first to step back and assess the feasibility of an evaluation before proceeding to design one. In other cases, evaluation can profit from prior experience and move to a second generation of effects, comparing not just the average effect of Treatment A among all those who receive it versus those who do not, but the relative effects on different subgroups of Treatment A versus Treatment B. In this way the JOBS evaluation benefited from the years of experience that preceded it, when the Manpower Demonstration Research Corporation (with support from private, not public, funds) evaluated state experiments in welfare reform.

Such experience is sorely lacking in other program areas, especially in the complicated realm of family services. Conference members agreed that careful thought is required in advance to identify the subject of evaluation, the variables to be defined, and the measures to use. And yet, as one participant commented, too much delay in formulating an evaluation may mean that it never gets off the ground.

The design of the evaluations

The opening remarks that described the charge of the conference called attention to the fact that the seven major evaluation projects share several design features. All but one (YOU) use random-assignment designs to measure effects on parents. Of these all but one (FIP) measure effects on children as well. In an effort to determine what dimensions of a program make a difference—what works for whom—increasingly complex experimental designs are being used, such as the random assignment scheme of JOBS. For similar reasons, most of the evaluations are collecting a large amount of baseline information prior to random assignment. This information often goes beyond simple demographic variables to include, as do New Chance and JOBS, measures of depression, baseline literacy, and self-confidence.

All of the evaluation designs include cost-benefit analyses. This is an especially difficult exercise when programs provide benefits that are hard to quantify. How can one give a monetary value to benefits that children obtain from the education and training of their parents?

Another universal feature of these evaluations is that they contain extensive studies of implementation: that is, they closely observe what services are delivered to which clients and how the service delivery system is organized. This scrutiny of what goes on "inside the black box" to learn about the intensity of services, the structure of services, and the details of staff-client interactions should reveal not only how programs shape people, but how people shape programs.

On the other hand, certain design characteristics are unique to individual projects. FIP matches treatment and comparison sites, rather than randomly assigning clients within sites. YOU allocates funds to neighborhoods rather than to service projects. CCDP assigns special observers to record program implementation at each site. Even Start focuses on adult and child literacy and their interconnection. Three projects—the Teenage Parent Demonstration, JOBS, and New Chance—have embedded more detailed, qualitative substudies within the larger evaluation. Some of the projects require clients to participate; others are voluntary.

Particular design issues that were discussed extensively at the conference include (1) designating the unit of analysis, (2) determining the appropriate follow-up period, and (3) taking account of "transactional analysis," defined below.

What is the focus of analysis?

The problem of defining the unit of analysis is endemic in these two-generation programs, owing to the many actors involved. The teen-parent intervention directly concerned mothers and their children, but also affected the lives of others—parents of the mother, other relatives, boyfriends, the children's fathers. The question becomes which units to track in the course of evaluation. In some of the other studies, a "focus" child within a family is chosen for in-depth examination. But could we not gain rich information by examining siblings as well? Other family members? The questions remain open.

How long should an evaluation last?

Determining an appropriate follow-up period is also difficult. The two-year follow-up for the teen-parent impact analysis means that the average age of sample members will then be 20, yet the transition from school to work usually covers ages 18 to 24. Would it not be preferable to extend the follow-up to a longer span of time? This is an expensive proposition, and adequate resources may not be available. In the case of programs such as Even Start that involve early childhood education, we would like to know what happens to the children as they progress through elementary and secondary school. In the case of programs intended to improve parenting skills, like CCDP and New Chance, we want to learn what kinds of parents the children themselves become, one generation later. The time horizon stretches on.

Can we learn more about behavioral changes?

Reference to the interaction of case managers and their teen-parent clients prompted a recommendation from psychologists at the conference that evaluation of these programs should give consideration to transactional analysis, a term referring to study of the succession of modifying interactions that take place in the course of a program—between managers and clients, between mothers and children, among the various agents involved in the process of a program. This form of analysis is dynamic, going beyond observation of single individuals at fixed points in time. Economists in the audience noted the parallels between this type of study and that described in the job-search literature, which focuses on the sequential decisions made by job seekers who solicit and receive a series of job offers. In the same way, transactional analysis follows a conditional-probability strategy—examining a particular event in the light of events that preceded it—to track the quality and cumulation of program effects.

Qualitative and observational research

Common among these evaluations is the specification of an ethnographic or observational component, a topic that received particular attention at the conference. A special study within the Teenage Parent Demonstration, funded by private foundations and about to be fielded, will examine parent-child interactions to determine the effects of the demonstration on parenting skills and child development. Its data include videotapes, interviews, and surveys of home environment. The JOBS evaluation contains a substudy of a group of mothers and children to examine family environment and dynamics. It also proposes to videotape mother-child interaction. For its process evaluation CCDP assigns to each site "project ethnographers" charged with providing descriptions of the dynamics and natural history of the unfolding projects. Even Start measures a parent's ability to teach a child by observing a particular "task": while the parent reads a simple book to the child, a trained observer uses a precoded rating form to record aspects of their interactions. YOU calls for periodic, intensive field visits by trained ethnographers to describe the nature of community life, problems encountered in delivering services, and the experiences of youth in the program.

The value of this kind of information was underscored by conference participants. It offers us a closer look into the black box of program implementation, providing another layer of explanation concerning a program's operation and effects. It illuminates differentials in treatments, helping us discern when a program is well managed or when its clients are ill served. It permits appreciation of the richness and complexities of the experiences of staff and clients in these multifaceted programs. Not least important, it offers accessible, even colorful, information to the program sponsors, members of government at all levels, and the concerned public. This type of data sustains interest in a project until outcome data are available, which often takes three to five years.

Some critics took issue with this form of research. Terminology was one target: "ethnography" in its strictest sense refers to a branch of anthropology dealing with systematic description of human cultures according to prescribed procedures. This is not necessarily the sense in which the term is applied in the evaluations, even though it appears in their descriptions. "Observational research" may be a more accurate term, but its results are just that—observations made by individuals, potentially carrying a subjective element, no matter what pains are taken to reduce that element through careful training of the observers and use of standard protocols for observation.

The utilization of qualitative data is fraught with difficulties. Many of the research contractors and government project officers acknowledged that they face a formidable task as they attempt to merge process data with outcome data to gain understanding of what works for whom, and why. General agreement prevailed that information of this nature has value and purpose but must be collected and used with care and precision.

The need to extend basic research and disseminate its results

Prominent themes in the discussions included the need for standardization of measures, for "meta-analyses," and for syntheses of research results.

The multiplicity of units and variables factored into these evaluations means that further research should be devoted to ways in which to measure effects and to specify which effects we want to measure. Standardization of measures is a basic requirement if we are to draw generalizable conclusions from these assessments. There is little uniformity across programs, for example, on measurement of program participation. Is it a specified percentage of time spent in

program activities over a specified calendar period? Should it include a measure of intensity of participation? How does one gauge intensity? A large challenge to the JOBS evaluation lies in formulating measures of participation that will permit comparability across sites in order to meet the required performance standards.

A consensus emerged that a coherent set of common baseline and outcome measures, of process and participation measures, would be of immense benefit. More particularly, it was recommended that analysts attempt to designate "marker variables"—basic definitions and measures common to diverse programs—to help move evaluation methodology forward by permitting convergence of analytic concepts and tools.

Meta-analysis has been defined as "the use of formal statistical techniques to sum up a body of separate (but similar) experiments."[4] As a scientific tool it has proved controversial. Its advocates argue that it can illuminate the nuggets of truth lying under a mountain of sometimes conflicting research results. Its detractors rejoin that only under severely restricted conditions can such analysis be performed well enough to be convincing. If it is indeed possible to succeed with this form of study, these complex programs offer unusually fertile ground for its application.

Several participants emphasized the need to synthesize and disseminate the results of evaluations of previous programs before launching major new efforts. An example cited was the publication of *From Welfare to Work*, a summary of the results of state experimentation with welfare reform prepared by the Manpower Demonstration Research Corporation, which provided the basis for the JOBS evaluation. (Preparation of the summary was required under the JOBS evaluation contract, as a result of a recommendation at the 1989 IRP/ASPE workshop, mentioned earlier.)

Summaries of this nature would promote dissemination of findings and provide the opportunity for evaluators to take time to think about basic issues before moving ahead. Evaluators expressed the desire for government agencies such as ASPE to commission more syntheses destined for two separate audiences: the policy community, including federal, state, and local government staff, legislative staff, advocacy groups; and the academic community. The first audience needs summaries of results for its immediate purposes. The second can use them to promote accumulation of a body of knowledge and to further the development of social science. Needed for this purpose also, it was felt, are public use tapes from the evaluations, which will facilitate secondary analyses and additional academic research.

Afterword

The conference offered testimony both to the advances that have been made in evaluating antipoverty programs and to the distances that remain to be crossed. The personal views presented below (see pp. 22–34) provide more detail concerning these achievements and deficiencies. The reflections of the three members of the academic community point to the remarkable degree of technical competence revealed by the evaluations and to the pressing need to bring to them more basic knowledge and research. The comments of members of the policy community tell us of the practical problems inherent in assessments of this nature and possible means to deal successfully with those problems.

It is hardly surprising that evaluations of two-generation interventions contain shortcomings, in view of the scope and complexity of these programs. What might be considered surprising, however, was the strength of personal concern and professional commitment expressed by virtually all conference participants. Evaluators and project officers alike repeatedly gave evidence of their solicitude for, and determination to alleviate, the circumstances of troubled families. Given that level of commitment, as well as the intellectual resources apparent in the conference room, one might conclude that we have grounds for optimism concerning efforts to overcome barriers to evaluation of complex social programs. ■

---

[1]A selected set of the papers was subsequently edited and published as *Evaluating Welfare and Training Programs*, ed. Charles F. Manski and Irwin Garfinkel (Cambridge, Mass.: Harvard University Press, 1992). See display box, page 36.

[2]Sheila Smith, "Two-Generation Program Models: A New Intervention Strategy," *Social Policy Report* (of the Society for Research in Child Development), 5:1 (Spring 1991), p. 1.

[3]Evaluation contractors are the private firms that conduct evaluations under contract with government agencies and private foundations. For a discussion of their role, see the Introduction to *Evaluating Welfare and Training Programs*.

[4]Charles Mann, "Meta-Analysis in the Breach," *Science*, Vol. 249, August 3, 1990, p. 476.

Table 1

Characteristics of Family Service Programs Being Evaluated

	JOBS	Comprehensive Child Development Program (CCDP)	Even Start	Teenage Parent Demonstration	Youth Opportunities Unlimited (YOU)	Washington State Family Independence Program (FIP)	New Chance
Status	Ongoing program	Demonstration	Demonstration	Demonstration	Demonstration	Demonstration	Demonstration
Coverage	Broad, affecting large segment of the welfare caseload (but with specially targeted subgroups)	Broad: family must have income lower than poverty level and a newborn infant or pregnant woman	Selective: family must have an adult eligible for Adult Basic Education, a child between ages 0–8, and live in a Chapter 1 attendance area	Broad, focusing on teenage custodial parents with only one child (or pregnant with first child)	Broad: affecting all youth within designated target areas of 25,000 population	Entire public assistance caseload, alternative to AFDC, in 5 sites	Selective within a highly targeted segment of the welfare caseload (parents aged 16–22 who are dropouts and gave birth by 20)
Participation Requirement	Mostly mandatory; likely to be substantial variation across sites	Voluntary	Voluntary	Mandatory; noncompliance results in a sanction that is lifted only when teen comes back into compliance	Voluntary	All families eligible for ADFC must enter FIP instead. All may then participate in employment and training (E&T)	Voluntary in most locations
Level of Disadvantage of Participants	Mixed: some are short-term recipients; others are highly disadvantaged	High: poor	High: low-literate and poor; 78% high school dropouts, 71% incomes under $10,000	High: all are teenage welfare recipients in inner-city areas. Even though one-third had completed high school and another one-third were in school, basic skills levels were very low	Mixed: depending upon community, which must have at least 30% of population below poverty	All are recipients of public assistance (AFDC-eligible)	High: nearly all young mothers without diplomas who are dropouts
Participation Rates	Modest levels of participation anticipated due to normal welfare dynamics and limited state resources for services and follow-up	Participation is voluntary; expected to vary across sites	Participation is voluntary; expected to vary across sites	Fairly high levels of at least initial compliance, but also fairly high levels of sanctioning	Modest levels initially; design intention is to reach the needs of all youth in the target area	All participate in income assistance part of FIP. Over half of those voluntarily participate in E&T part of FIP	Fairly high levels of participation due to rich services and voluntary nature of program in most sites
Structure (agencies involved in administering the program)	Program coordinated through welfare agencies	Grantees are community-based organizations, hospitals, local education agencies, universities	Grantees are local education agencies	Program coordinated through the welfare agencies in Chicago and in Camden and Newark, N.J.	Coordinated through a lead agency (SDA or PIC); to link with a wide range of organizations and programs; operating out of a site located within the target area	Income maintenance case coordination and supportive services administered by welfare agency; E&T admin. by employment security agency	Program offered through community-based organizations, schools, and municipal organizations
Mode of Service Delivery	Mixed, with heavy emphasis on off-site education and training through referrals to existing community services	Coordination of and referral to existing services; direct provision of a mix of in-home and on-site services; extensive reliance on case workers	Coordination of and referral to existing services; direct provision of a mix of in- home and on-site services; some reliance on case workers	Mixed off-site and on-site. Referrals to existing schools, GED programs, skills training programs; all sites offered workshops and GED instruction on-site (eventually discontinued in one site due to low enrollment)	Mixed, on-site in the community-based project site; coordinated through other agencies located in the target area; some off-site	Interagency arrangements with schools, community colleges, JTPA, CBOs, etc.	Education and personal development services primarily on-site, specially designed with target population in mind; skills training primarily off-site

Uniformity across Sites (in administration, service delivery)	Low: considerable local discretion	Low	Low	Moderately high, with variation primarily in the method of delivering on-site workshops, caseload sizes, and availability of community resources	Low; considerable local discretion	High uniformity of program regulations, guidelines. Moderate variations in client interactions, priorities, E&T services	High, prescriptive model
Services	Education, skills training, work experience, job search assistance, case management, child care, transportation assistance	Health, early childhood education, employment training, life skills, case-work, parent education, literacy skills	Adult basic education, early childhood education, life skills, parenting education	Education, job search, skills training, summer employment, case management; workshops on family planning, motivation, wide range of life skills	Employment and training, education, recreation and sports, counseling, health care, social services (including drug prevention), etc.	Assessment, case coordination, special services for pregnant teens, education, job search, trng., voc. trng., OJT, parenting skills, child care, transitional child care and Medicaid. Cash incentive bonuses above welfare grant for partic. in education, training, or if employed	Education, skills training, work experience, employment preparation, career exploration/ counseling, life skills instruction, family planning and health education, personal and group counseling, pediatric and maternal health care, and parenting education
Provision of Child Care	Financial support, referrals to providers; variability in quality anticipated	Coordination with Head Start, other local preschools and day care programs, direct provision of day care or preschool services	Coordination with Head Start, other local preschools, some child care provided to enable parents to participate	Financial support, counseling, referrals to providers; on-site care at two sites; considerable variation in quality	Mostly as a supplemental service to a program	Extensive funds for child care while in FIP and for 1 year after leaving FIP owing to employment	Mostly on-site or arrangements in developmentally oriented programs
Age of Participants' Children at Intake	Usually 3 to 17, but sometimes younger	0 (prior to birth) to 12 months	0 to 8 years	0 to 3 at intake; 80% had child under 1; some participants enrolled while pregnant	Not applicable	0 (prior to birth) to age 18	0 to 5, mostly at younger end
Number of Sites	8	1989: 23 1990: 24	1989: 76 1990: 123 1991: 240	3	7	5 with FIP; 5 non-FIP	16
Evaluation Design: Random Assignment or Other	Random assignment	Experimental evaluation in all projects. Random assignment	Descriptive survey of all projects and participants; experimental evaluation in 10 purposively selected projects. Random assignment in 5 of the 10	Random assignment	Process and outcome; highly qualitative	Matched comparison sites (5 and 5)	Random assignment
Number of Subjects (experimentals and controls when random assignment)	40,000–50,000	2,500 Es, 2,500 Cs	Descriptive: 3,000 families Experimental: 1,200 Es, 1,200 Cs	6,091 (1,281 in Camden, N.J.; 1,348 in Newark; 3,462 in Chicago)	NA	Approximately 15,000 FIP, 15,000 non-FIP	2,320
Start Date and Expected End Date	October 1989–September 1997	April 1990–March 1995	January 1990–October 1993	1986–1992	July 1990–June 1995	July 1988–June 1993	January 1989–September 1995
Expected Total Evaluation Budget	$15 million[a]	$10 million	$2.9 million	$3.9 million	$1.69 million	Approximately $3 million	$12 million[b]

Source: Originally prepared by Robert Granger, MDRC, and modified for the conference.

Note: For program descriptions, see accompanying summary.

[a]Includes payments to sites to offset research-related costs.

[b]Includes site payments and site development costs.

The family service programs and their evaluations: Capsule descriptions

1. *Programs Authorized by Congress*

The Job Opportunities and Basic Skills Training Program
Sponsor: U.S. Department of Health and Human Services
Evaluator: Manpower Demonstration Research Corporation

The Family Support Act of 1988 required that its centerpiece, JOBS, be evaluated to determine the effectiveness of different approaches to help welfare applicants and recipients increase self-sufficiency through education, training, and support services. The evaluation plan subsequently developed calls for an impact analysis, an implementation and process study, and a benefit-cost analysis, plus a special study of a subgroup of mothers and their young children. Eight sites—counties, cities, or combinations of both—representing a variety of regional attributes will participate. Their selection is nearing completion and enrollment activities are beginning. The evaluation will cover 48,000 people, randomly assigned—as required by the Act—to control or treatment groups.

The impact analysis will examine effects on employment and earnings and on receipt and amount of AFDC and Food Stamps in all evaluation sites. In three sites where surveys will be administered, effects on income levels, educational levels, literacy, basic math, and child development will be considered. In four of the sites, random assignment to treatment (JOBS) or to control status (the regular AFDC program) will be followed; in the others, assignment will be to a control group and to one of two types of treatment groups: the regular JOBS program or a variant created to test alternative approaches—e.g., education and training, or direct job placement, or use of different case management strategies. The impact analysis will utilize administrative data on earnings, employment, and welfare receipt for five years after program entry, and follow-up surveys will be conducted in three sites where detailed baseline data are collected. Impacts within sites and across sites will be analyzed, the latter to assess the relative effectiveness of the various program approaches.

The implementation and process study will examine the ways in which various programs are put into operation, documenting resource levels and funding sources, organizational structures, links among agencies involved, operating procedures, targeting strategies, staff levels and caseload ratios, case management practices, and messages conveyed to clients. Data sources include field research, staff surveys, automated program tracking systems, and case file records. The U.S. Department of Education is supporting a special study at three sites of the implementation of adult education, to provide information not previously available on the nature and quality of the education provided to welfare recipients.

The benefit-cost study will estimate the total costs of the various programs at each site as well as the costs of particular activities or components within the programs. These expenditures will be compared with the benefits estimated in the impact study.

The analysis of the subgroup of mothers and children, subcontracted to Child Trends, Inc., will explore maternal and child development. It involves 2,500 pairs of mothers with children aged 3–5 in three sites, randomly assigned to control and treatment groups. Basic demographic and work-welfare history data will be taken from the intake information; the mothers will take a literacy test and be assessed for attitudes toward work, welfare, training, and child care, as well as feelings of depression and mastery. For a subset of 600 of these families at baseline, personal interviews will be conducted in the home and the quality of the mother-child relationship will be observed. These families will be included in the follow-up survey samples to learn what changes occur in their lives, how the interaction between mother and child affects the mother's participation in JOBS, the nature of the child's development, and, with anticipated funding from the U.S. Department of Education to support collection of school data, the child's school attendance and behavior.

The Comprehensive Child Development Program
Sponsor: U.S. Department of Health and Human Services
Evaluators: CSR, Inc., and Abt Associates, Inc.

Authorized by the 1988 amendments to the Elementary and Secondary Education Act, CCDP is a demonstration program conducted under very general federal guidelines to explore the effectiveness of intensive health, social, and educational services to young families in poverty. Eligible families are those that include a pregnant woman or child under one, have incomes under the poverty line, and agree to participate in program activities for five years. A competitive proposal process was used to fund a variety of agencies—universities, hospitals, public and nonprofit organizations, and school districts—at 24 sites around the country, 18 in urban areas and 6 in rural locations, involv-

ing 2,500 families. Although the form of service can vary, all projects are required to intervene as early as possible in children's lives, to involve the entire family, to serve the special needs of infants and young children, to promote parents' ability to contribute to their children's development and their own self-sufficiency, and to offer continuous services until the child that determined the family's eligibility (the "focus" child) enters elementary school. Project activities began in 1990. Case managers play an important role in assessment and coordination of needed services.

Like JOBS, this program carries a legislative mandate for evaluation, which DHHS divided into two parts: study of the feasibility and implementation of the projects, and a national impact evaluation. CSR is conducting the first; Abt Associates, the second.

The purpose of the implementation evaluation is to determine whether and how these complex projects can be successfully launched. It is examining program start-up, organization of service delivery through interagency agreements, costs of delivery, utilization of services, and program changes over the course of the demonstration. Its five sources of data include the project proposals, quarterly progress reports and other project documents, reports by special observers, reports from site visits, and quantitative data from the automated management information system installed at each site.

For the impact evaluation, a randomized design was achieved in each site through the program requirement that projects deliberately recruit more families than could be served, and then assign eligible families to program and comparison groups, the latter to receive whatever social services would normally be offered in the absence of CCDP. Objectives are to assess the impact of the program on the development of children, parents, and families; to determine whether the CCDP concept that an agency can coordinate a comprehensive set of services is feasible and effective; and to search for practices that can be used to improve comprehensive, early-intervention projects for low-income families. This evaluation is longitudinal: it will measure attributes of the families over time, focusing on the child of interest and the mother. The feasibility of administering measures to the fathers as well is being studied. Baseline demographic information concerning the families is being collected, and the families will be contacted every six months for assessment by means of a parent interview and tests administered to the child. The evaluation data are collected at each site by a two-person team, consisting of a permanent Abt staff member and a person hired for the child testing. The tester will not know whether the family is in the treatment group or the control group.

Even Start Family Literacy Program
Sponsor: U.S. Department of Education
Evaluator: Abt Associates, Inc., with a subcontract to RMC Research Corporation

This demonstration program offers educational services to both child and parent through an integrated program of early child education, adult basic skills training, and parent training. A family is eligible if it contains an adult who needs basic skills training, a child between the ages of 1 and 8, and lives in a Chapter 1 (low-income) elementary school attendance area. Four-year grants are offered to school districts, which provide the services directly or arrange for them through existing community programs. Even Start began with 73 grants in 1989; their total is expected to reach almost 250 this year.

The 1988 legislation that authorized the demonstration (amendments to the Elementary and Secondary Education Act) requires annual independent evaluation of its programs. The evaluation contract, awarded in 1990, has four parts: (1) construction of a large-scale data base, the National Evaluation Information System, which contains a common set of data from each project and most participants—descriptive statistics on, for example, the nature of the project, services provided, progress in adult basic skills and children's school readiness; (2) an in-depth study of ten projects, half with randomized experimental designs, to complement the broad-based data with small-scale, detailed analysis of the relationship between services received and short-term outcomes; (3) other local evaluation studies as desired by individual grantees, provided that they first receive approval from the Department of Education; and (4) submission by individual grantees of evidence of

their program's effectiveness to the Department of Education's Program Effectiveness Panel.

The national evaluation contractor worked with projects to define the national information system and provide technical assistance to project managers, who are responsible for data collection. The national evaluation contractor then analyzes this information, sends it back to the projects, and incorporates it in annual reports. The in-depth study was designed by the evaluator, with input from the local managers. The five randomized projects are small, each involving about forty families, half assigned to Even Start and half to a control group. An intensive measurement battery will examine a number of hypothesized outcomes to gain a closer look at the program's effectiveness.

2. *Programs Initiated by Federal Executive Agencies*

Teenage Parent Demonstration
Sponsor: U.S. Department of Health and Human Services
Evaluator: Mathematica Policy Research, Inc.

Formally known as the Demonstration of Innovative Approaches to Reduce Long-Term AFDC Dependency among Teenage Parents, this project originated in DHHS and lasted from 1986 through mid-1991. At three sites, Camden and Newark, New Jersey, and the south side of Chicago, Illinois, all teenage parents who began receiving Aid to Families with Dependent Children (AFDC) for themselves and their child were required to attend an intake session and were then randomly assigned to treatment or control status. The treatment consisted of participation in appropriate education, training, or employment programs as long as AFDC was received. Failure to participate could result, after warnings, in sanctions—reduction of the AFDC grant until the parent complied. Services to program participants included case management, child care assistance, allowances for transportation and other expenses, and workshops to promote motivation, life skills, and the ability to pursue continued education, training, or employment. Those assigned to control status could not receive the program services but were free to pursue training and education on their own. About 3,000 teenagers took part in the demonstration programs, and another 3,000 teenagers received regular services.

The evaluation has four components. The *implementation analysis* has assessed program delivery by observing operations, interviewing staff members, and studying program records and documents. The *impact evaluation*, nearing completion, compares the experiences of treatment and control group members over a two- to four-year post-program period. It uses information obtained at intake concerning personal characteristics and basic skills test scores; administrative data obtained through March 1992 concerning welfare payments, earnings, and child support; and information obtained two years after program completion through a follow-up interview and basic skills retest. Outcomes of interest are school completion and performance, basic skills growth, employment and earnings, welfare dependence, fertility, child-rearing practices, and child support received. The *cost-effective analysis* assessed direct and indirect administrative and service costs and compared them to benefits from the point of view of governments, society, and participants. Finally, an *in-depth analysis* used qualitative data from focused group discussions, personal interviews, conferences with project staff, and case-tracking data on program participation and outcomes. This component extended our understanding of the backgrounds and circumstances of participants and their responses to the opportunities and requirements of the program.

Three ancillary studies were also conducted: a survey of the child care available and patterns of use by parents in the demonstration sites, a survey of the child care needs and actual use among the welfare-dependent teenagers in the evaluation sample, and a special study funded by the Rockefeller Foundation and the Foundation for Child Development to examine interactions and developmental processes between the mothers and their children and the relationships between those interactions and processes and developmental outcomes for the children.

Youth Opportunities Unlimited (YOU) Initiative
Sponsor: U.S. Department of Labor
Evaluator: Academy for Educational Development

This demonstration program was created by the Department of Labor to test ways of improving the long-term employability of youth in neighborhoods of about 25,000 people where the poverty rate is 30 percent or more. Its guidelines are general, allowing local flexibility. It is being conducted at seven urban and rural sites by the local governing boards for the Job Training and Partnership Act program in the communities. They can use any of four core models of service: learning centers (residential or nonresidential, community centers or schools, where basic skills and vocational training are offered); alternative high schools operated by local school districts, offering intensive remedial reading; construction projects in which skilled craftsmen train youth while rehabilitating dilapidated housing; and, in rural areas, initiatives to increase enrollment in postsecondary schooling by establishing two-year work-study colleges or setting up satellites of community colleges. In addition, one or more complementary programs are to be offered, including apprenticeship programs with unions or firms, employability programs for teen parents, summer training and education programs, alternative schools run by community colleges, and community youth centers offering counseling, recreational and cultural opportunities, and job market information.

Each of the seven initiatives began with a planning grant, out of which the successfully funded proposal was developed. The programs began operating in mid-1990 and will continue with federal support for three years. The federal

funding represents half of the support for each YOU program, the rest to be matched by local funds and resources. The goal is for each program to be self-supporting by the end of the demonstration period.

The evaluation lasts from 1990 to 1995 and has three parts. A two-person team conducts periodic, intensive site visits to monitor the development, implementation, organization, and management of each program. Trained observers also visit the sites periodically to document the nature of community life, problems of and services offered youth, and the ongoing experiences of program participants. Finally, an information system consisting of public documents and administrative records is used to track five outcome measures: school attendance, dropout, teen parenthood, welfare dependency, and juvenile delinquency.

3. *Program Initiated by a State*

Washington State Family Independence Program
Sponsor: State of Washington
Evaluator: The Urban Institute

The Family Independence Program (FIP) supplements AFDC by offering special incentives for recipients to gain employment and training. At five welfare sites within the state all eligible applicants for AFDC enter FIP instead, which provides them the option of receiving supplemental services that include financial bonuses; an assessment of needs made jointly by client and staff; case management; aid in budgeting, family planning, and parenting; assistance in obtaining resources from other agencies; education, occupational training, and employment services; child care; and medical care. The last two services are continued during the first year of employment.

After an extensive planning period, FIP was put into operation in 1988 and will continue until 1993. Implementation of the JOBS program in 1990 brought many of FIP's features to the AFDC program throughout the state. The main differences between JOBS and FIP are that the latter offers financial incentives, that it cashes out food stamps, and that it provides more extensive child care.

The evaluation of FIP involves the five treatment sites and five matched comparison sites that maintain the usual AFDC program. Both treatment and comparison groups comprise about 15,000 recipients each. The first of four parts of the evaluation is a *net impact analysis*, which focuses on estimation of the effect of FIP (as compared to AFDC) on employment, earnings, duration of welfare receipt, and return to the rolls. The effect of the food stamps cashout is also being assessed. The impact analysis uses administrative data as well as interviews with participants. The second part of the evaluation examines *program implementation and operations*. Its data are taken from interviews with administrators and staff, questionnaires completed by the staff, observations of group activities, and program documents and records. A *cost-benefit analysis* will compare the cost of administering FIP with that of AFDC, will contrast benefits paid under FIP with those under AFDC, will estimate the likely long-term savings from FIP for both state and federal governments, and will assess the costs and benefits to participants. It will utilize the results of the impact analysis and administrative cost records. Finally, the evaluation will *synthesize and interpret* all of these results to identify successful program features and operational practices and to describe ways in which unsuccessful parts of FIP might be improved.

4. *Program Privately Initiated*

The New Chance Demonstration
Sponsor: A consortium of private foundations and the U.S Department of Labor
Evaluator: Manpower Demonstration Research Corporation

Designed and managed by MDRC, New Chance is directed toward young AFDC mothers who have dropped out of school. It offers comprehensive services to promote the economic self-sufficiency and parenting skills of these mothers and the social and emotional development of their children. Services are delivered through either schools or community organizations, are intensive (30 hours a week of classroom and other activities) and last for 18 months, after which follow-up services are offered for a year. Services include basic education and GED preparation, employment readiness, health care, counseling in life management and decision making, pediatric health services, child care designed to foster child development, and case management. Most of the services, including child care, are offered at a single project site. The demonstration began in 1989, lasts until 1995, and covers sixteen sites in ten states that together represent a mix of economic conditions, welfare grant levels, and ethnic groups. The program is deliberately small in scale, owing to the intensity of services: each site is expected to serve about one hundred women.

Selection of a research sample of 2,300 mothers, two-thirds in a treatment group and one-third in a control group, was completed in July 1991. Process, impact, and benefit-cost analyses will be conducted. The process study examines various implementation strategies to determine which seem to be most conducive to program success. Modes of service delivery, patterns of participation, and choices made by program operators are observed. This study uses both quantitative data, obtained through a special automated management information system installed at each site, and qualitative information drawn from site visits, field reports, and memoranda by the evaluator's staff. The impact study will gauge program effectiveness in terms of the mother's education and employment; parenting practices and health; welfare dependency; and improvement in the cognitive, behavioral, and health status of the children. Data for this analysis will be collected by in-person interviews at 18 and 36 months after entry into the sample. The cost-benefit

analysis, still in the process of formulation, faces the technical difficulty of valuing a broad array of possible program effects.

5. *Other Programs in Early Stages*

The last session of the conference briefly reviewed four evaluation projects that are in developmental phases. The first three originated in federal departments; the last is a private initiative that has some federal support.

Feasibility of Evaluation of Family Preservation Programs
Sponsor: U.S. Department of Health and Human Services
Evaluator: James Bell Associates

The intent of family preservation programs is to avoid the need for foster care by delivering intensive, short-term welfare services to troubled families. Concern over recent increases in the number of children in foster care has prompted the introduction of several bills in Congress that would fund such programs and require their evaluation. Because of controversy and disagreement concerning the effectiveness of these programs and methods for assessing them, DHHS awarded a contract for an "evaluability assessment," an exercise designed to produce a reasoned basis for proceeding with an evaluation that will benefit both practitioners and policymakers. The assessment will attempt to identify the critical design and policy issues surrounding family preservation services and will gauge the feasibility of conducting valid and useful evaluations of these programs. The methodological issues it will try to resolve include appropriate measures of program success, appropriate control or comparison groups, the effect of voluntary participation on differences in outcomes, and barriers to data gathering and analysis posed by the need to obtain adequate sample sizes and to observe laws protecting the privacy of participants.

WIC Child Impact Study: Field Test
Sponsor: U.S. Department of Agriculture
Evaluators: Abt Associates, Inc.; Johns Hopkins University; Westat, Inc.

Although the Special Supplemental Food Program for Women, Infants, and Children (WIC) has grown rapidly and gained a strong base of support since it began in 1972, little is known about the impact of the program on children. The USDA has sought to address this issue, using a successive-stage approach. First, the Department reached a cooperative agreement for a design feasibility study with the University of North Carolina at Chapel Hill and the Research Triangle Institute. The cooperators determined that a study was feasible and recommended a quasi-experimental design using WIC and non-WIC infants identified through state birth records. Second, the Department awarded a contract for a field test of the recommended quasi-experimental design and an alternative design developed by the evaluators listed above. The alternative was an experimental design that calls for the recruitment of WIC-eligible but unserved pregnant women, with random assignment to a treatment or a control group. The field test was completed in November 1991. The results will be used by the USDA to decide how to proceed with a WIC child impact study.

The Head Start Family Service Center Demonstrations
Sponsor: U.S. Department of Health and Human Services
Evaluator: A consortium of local evaluators

Competitive grants have been awarded to 33 local Head Start agencies to provide extended services to families of children participating in the Head Start program. The intent is to demonstrate how the agency can work with other community agencies and organizations, public and private, to deal with problems of substance abuse, illiteracy, and unemployment among the parents. It is hoped that the demonstrations will help construct and test innovative ways in which to identify family problems, motivate family members to move toward self-help, link families with appropriate community services, and support them as they work out solutions to their problems. The Head Start Bureau in DHHS will provide coordination, technical assistance, and analysis of common data elements to produce an integrated summary of the process and impact evaluations that are being conducted by local evaluators.

The Young Unwed Fathers Demonstration
Sponsors: A consortium of private foundations, with additional support from agencies within the U.S. Department of Labor and U.S. Department of Agriculture
Evaluator: Public/Private Ventures

Low-income men aged 16–25 who have fathered children out of wedlock and are unemployed form the clientele of this pilot program, which is being tested in six sites around the country. Conducted by a variety of community agencies ranging from Goodwill Industries (Racine, Wisconsin) to the Pinellas Private Industry Council (Clearwater, Florida), the program provides access to employment and training opportunities; counseling referrals to other forms of support; education and training services; and classes in parenting values and skills. Fieldwork began in 1991 and will last 18 months. The research component of the project includes studies of project implementation at each site, the effects on participants across sites, a cost analysis, and a qualitative study at four sites, designed to provide information on the lives and experiences of the young men. ■

Agenda

Third Annual IRP/ASPE Evaluation Conference: Evaluating Comprehensive Family Service Programs Washington, D.C., November 14–15, 1991

November 14

Welcoming remarks: Martin Gerry, Assistant Secretary for Planning and Evaluation (ASPE); Robert M. Hauser, Institute for Research on Poverty (IRP)

Introduction and Overview: William R. Prosser

Session 1. Comprehensive Programs for Mothers and Children. Chair: Charles F. Manski

Teen Parent Demonstration
Rebecca Maynard, Mathematica Policy Research
JOBS Evaluation
Barbara Goldman, Manpower Demonstration Research Corporation; and Nicholas Zill, Child Trends, Inc.
New Chance
Robert Granger, Manpower Demonstration Research Corporation

Report of the Head Start Evaluation Advisory Group
Sheldon White, Harvard University

November 15

Session 2. Comprehensive Programs for Mothers and Children (continued). Chair: Steven H. Sandell

Comprehensive Child Development Program Outcome Study
Jean Layzer, Abt Associates
Comprehensive Child Development Process Study
Ruth Hubbell, CSR, Inc.
Even Start
Robert St. Pierre, Abt Associates
Youth Opportunity Unlimited
Manuel Gutierrez, Academy for Educational Development
Washington State Family Independence Program
Lee Bawden, Urban Institute

Session 3. Projects in Early Stages. Chair: Ann Segal

Family Preservation Evaluability Assessment Project
Elyse Kay and Jay Bell, James Bell Associates
WIC Child Impact Study
Joan McLaughlin, U.S. Department of Agriculture
Young Unwed Fathers Demonstration
Catherine Higgins, Public/Private Ventures
Evaluation of Head Start Family Service Center Demonstrations
James O'Brien, Head Start Bureau, Administration for Children and Families, DHHS

Summing Up
Rapporteurs: Peter H. Rossi and James Heckman

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 demon-

stration, 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 sug-

gested 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, experi-

mental 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.

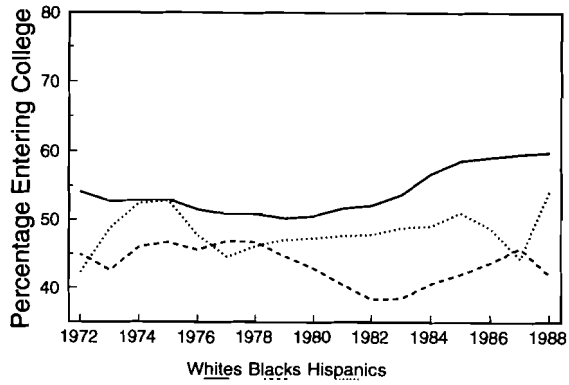

Figure 7. College Entry among Recent High School Graduates: White, Black, and Hispanic Men, 1972–1988

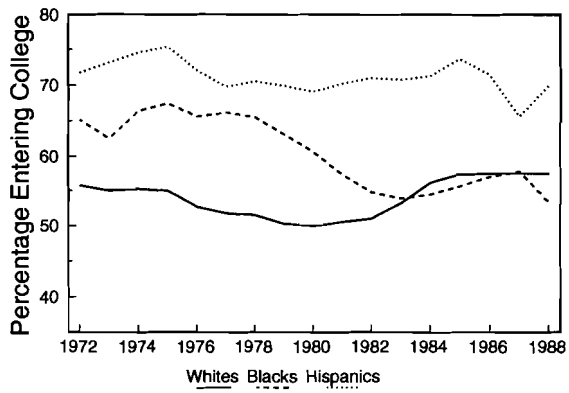

Figure 8. College Entry among Recent High School Graduates with the Average Social Background of Whites: White, Black, and Hispanic Men, 1972–1988

[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.

Evaluating Welfare and Training Programs

Charles F. Manski and Irwin Garfinkel, editors

The purpose of government-sponsored welfare and training programs is to bring disadvantaged citizens into the economic mainstream. How best this can be accomplished is not known, and the programs enacted to date reflect a variety of assumptions about what works best and why. The purpose of evaluation is to learn from past experience so that we may improve the effectiveness of programs in the future.

> It may seem self-evident that social programs should regularly be assessed and revised in the light of lessons drawn from experience. Nevertheless, systematic program evaluation is a recent development. Modern evaluation practice is generally agreed to have begun in the middle 1960s, when attempts were made to evaluate the impacts of programs proposed as part of the War on Poverty. Earlier efforts were largely limited to descriptions of how enacted programs were administered.
>
> Concern with program evaluation has spread rapidly since the 1960s. Today almost every substantial social program is subjected to some form of evaluation. Findings from evaluations not only fill many professional journals but are reported routinely in the media, where they presumably influence public thinking on social policy.
>
> Evaluation requirements now appear in major federal statutes. Evaluation is prominently featured in the recently enacted Family Support Act of 1988, which revised the Aid to Families with Dependent Children (AFDC) program. In Title II of this statute, Congress mandated separate implementation and effectiveness studies of training programs initiated by the states under the new Job Opportunities and Basic Skills Training Program (JOBS). Taking unusually specific action, Congress even stipulated the mode of data collection for the effectiveness study: "A demonstration project conducted under this subparagraph shall use experimental and control groups that are composed of a random sample of participants in the program."
>
> Charles F. Manski and Irwin Garfinkel, "Introduction," p. 1.

Given the self-evident need to evaluate welfare and training programs, this volume addresses the methodological questions that arise in carrying out such evaluations. In the Introduction the editors examine the domain of an evaluation (what part of a program should be subjected to evaluation), controversies regarding evaluation methods (such as reduced form vs. structural evaluation; the selection problem), and some of the special problems related to the evaluation of social programs.

The chapters in Part I describe evaluation practice during the past decade and report findings from some notable recent evaluations, such as the demonstrations under the Omnibus Budget Reconciliation Act (1981) and the programs under the Job Training Partnership Act (1982). Part II explores methodology and, in particular, the role of social science in evaluation. In Part III the various institutions that administer social programs are examined.

The chapters were commissioned for this volume and were presented at a national conference on evaluation sponsored jointly by the Institute for Research on Poverty and the Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services. It was held in April 1990. (For contents of the volume and information on how to obtain it, see box, p. 36.) The third annual conference on evaluation is described in this issue of *Focus*. ■

EVALUATING WELFARE AND TRAINING PROGRAMS

Edited by
Charles F. Manski and Irwin Garfinkel

Harvard University Press, 1992

$39.95

Contents:

To obtain *Evaluating Welfare and Training Programs* write to Harvard University Press, Customer Service, 79 Garden Street, Cambridge, MA 02138 (617-495-2480 or 495-2577)

Notes on Institute researchers

Adam Gamoran is serving as Associate Chair of the Department of Sociology at the University of Wisconsin–Madison. During 1992–93 he will be a Fulbright Fellow at the Centre for Educational Sociology, University of Edinburgh, Scotland.

Michael Gerfin of the Volkswirtschaftliches Institut at the University of Bern, Switzerland, will be visiting IRP during 1992 to conduct research on microeconometric aspects of labor supply and econometric methods. He has a postdoctoral grant from the Swiss National Science Foundation.

Arthur Goldberger was elected a Foreign Member of the Royal Netherlands Academy of Sciences. He continues to serve on the Commission on Behavioral and Social Sciences and Education at the National Research Council.

Linda Gordon spent five weeks in February–March 1992 in residence at Bellagio, Italy, where she worked on her own book on the history of welfare thinking in the United States, 1890–1935, and on a co-authored book with philosopher Nancy Fraser on the language and assumptions of contemporary welfare debates. She will travel to Sweden and Denmark in May to speak on welfare history and family violence, and to Hungary in August to give a paper at an international conference on the rise of the middle class. She serves on the Editorial Board of the *American Historical Review* and the Executive Board of the Organization of American Historians. She is presenting papers on the history of the underclass for a Social Science Research Council conference and volume and on the history of teenage pregnancy for a MacArthur Foundation conference and volume.

Peter Gottschalk presented testimony on changes in inequality in several industrialized countries to the House Ways and Means Committee of the U.S. Congress in February 1991. He also served as a consultant to the Joint Center on Budget Priorities, which helps state social service agencies adjust to budgetary cuts.

David Greenberg recently served on a panel for the state of Maryland that reviewed mandated insurance benefits for the treatment of alcohol and drug abuse. He is spending the 1991–92 academic year as a visitor at the Institute for Research on Poverty and the Robert M. La Follette Institute of Public Affairs at the University of Wisconsin–Madison.

Robert M. Hauser was elected to the Commission on Behavioral and Social Sciences and Education, National Research Council. He testified before the House Subcommittee on Census and Population in March 1991 on statistical needs for the future U.S. labor force. In May 1991 he presented a paper on trends in college entry among blacks, whites, and Hispanics to a National Bureau of Economic Research conference on the Economics of Higher Education. He was appointed Director of the Institute for Research on Poverty in July 1991.

Robert Haveman is spending the 1991–92 academic year as a Visiting Scholar at the Russell Sage Foundation in New York City. While there, he and **Barbara Wolfe** are writing a monograph based on their Institute-supported research on the "Economic Determinants of Children's Success." During 1991, he made presentations at the Association of Public Policy and Management (APPAM), the American Economics Association, and Cornell University, Michigan State University, Grinnell College, and Columbia University. He is President-Elect of the Midwest Economics Association.

Karen Holden is Graduate Chair of the Masters program in Family Economics in the Department of Consumer Science at the University of Wisconsin–Madison. She has been appointed to the editorial advisory board of *New Directions for Program Evaluation* and is a member of two expert advisory groups that are planning the first interview of the Health and Retirement Study, being undertaken by the Institute for Social Research, University of Michigan.

Thomas Kaplan has recently joined the Institute as a researcher and will soon be appointed Assistant Scientist. He has been the Planning Director of the Wisconsin Department of Health and Social Services and on the faculty of Waynesburg (Pa.) College.

John F. Longres is Chair of the Publications and Media Committee of the Council on Social Work Education. In 1991 he was awarded an honorary membership in Phi Kappa Phi.

In March 1991, **Charles F. Manski** testified before a U.S. Senate Finance Committee Subcommittee on Social Security and Family Policy on welfare dependency. He recently served on the National Research Council Committee on the Federal Role in Education Research and joined the Board of Overseers of the Panel Study of Income Dynamics. In December 1991 he became editor of the *Journal of Human Resources.*

Robert Mare is Director of the Center for Demography and Ecology, University of Wisconsin–Madison.

Margo Melli has been elected a Vice President of the International Society on Family Law and has been invited by the Office of Legal Adviser of the U.S. State Department to help review draft articles for a Hague Convention on Intercountry Adoption. She recently received an award from the U.W. System for Outstanding Contributions to Advancing the Status of Women in Higher Education.

Daniel Meyer joined the Institute as an affiliate in 1991 after spending a year at the Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services. He was coordinator of "Paternity Establishment: A Public Policy Conference," held in February 1992 in Washington, D.C.

Robert Moffitt gave a series of invited lectures on taxes, transfers, and labor supply at the University of Stockholm. He has also been named a coeditor of the *Review of Economics and Statistics.*

Robert D. Plotnick is associate dean of the Graduate School of Public Affairs, University of Washington, and an associate editor of *Demography.*

Gary Sandefur is serving on the Board of Overseers of the Panel Study of Income Dynamics until 1993, the National Science Foundation Sociology Panel also until 1993, and the Social Science Advisory Board of the Poverty and Race Research Action Council. He will be the invited keynote speaker at the April 1992 Conference on Graduate Education for Minority Students, sponsored by the Committee on Institutional Cooperation.

Nora Cate Schaeffer has been appointed to the editorial boards of *Public Opinion Quarterly, Sociological Methodology*, and *Sociological Methods and Research.* She will be spending a year at the Center for Survey Methods Research at the U.S. Bureau of the Census beginning in August 1992.

Judith A. Seltzer became Associate Director for Training in the Center for Demography and Ecology, University of Wisconsin–Madison, in July 1991 and was elected to the Council of the Family Section of the American Sociological Association last August. She is also an invited member of a panel sponsored by the Russell Sage Foundation, the Ford Foundation, and the Foundation for Child Development that is evaluating the effects of the child support provisions of the Family Support Act.

Marsha Seltzer was elected a Fellow of the American Association on Mental Retardation in 1992. She was appointed to the Executive Committee of the Academy on Mental Retardation and serves as Associate Editor, *American Journal on Mental Retardation.*

Karl Taeuber is chair of the 1990 Census Advisory Committee for the Inter-University Consortium for Political and Social Research. He is Resident Director for 1991–92 of the Wisconsin Year Abroad program at the University of Warwick, England.

James Walker is 1991–92 Robert Eckles Swain National Fellow of Domestic Policy at the Hoover Institution, Stanford University. He was recently named coeditor of the *Journal of Human Resources.*

Michael Wiseman completed his term as Associate Director of the Robert M. La Follette Institute of Public Affairs in December 1991. In June 1990 and 1991 he was a Visiting Professor at the Center for Social Policy at the University of Bremen, Germany. In 1991 he was appointed a Vilas Associate by the Graduate School of the University of Wisconsin–Madison.

John Witte is Professor of Political Science, affiliated with both the La Follette Institute of Public Affairs and the Industrial Relations Research Institute. He was recently appointed to the Advisory Panel for the Study of School Choice of the U.S. Department of Education.

Barbara Wolfe is a Visiting Scholar at the Russell Sage Foundation for the 1991–92 academic year. She will serve as Chair of the Scientific Committee for the August 1992 Meetings of the International Institute of Public Finance in Seoul, Korea.

Lawrence Wu is spending academic year 1991–92 as a fellow at the Center for Advanced Study in the Behavioral Sciences, Stanford, California.

Order form for FOCUS NEWSLETTER and INSIGHTS (free of charge)

Send to: FOCUS
Institute for Research on Poverty
1180 Observatory Drive
3412 Social Science Building
University of Wisconsin
Madison, WI 53706
(Fax: 608-262-4747)

Name: _____

Address: _____

_____
City State Zip

(Multiple copies of any issue: $1.00 each)

❑ I wish to receive INSIGHTS

Order form for Institute DISCUSSION PAPERS and REPRINTS

Prepayment required. Make checks payable to the Institute for Research on Poverty in U.S. dollars only.

SUBSCRIPTIONS: July 1991–June 1992
❑ Discussion Papers and Reprints ($50.00)
❑ Discussion Papers only ($40.00)
❑ Reprints only ($25.00)

INDIVIDUAL PUBLICATIONS: (Please fill in number or title and author)

Discussion Papers ($3.50) _____

Reprints ($2.00) _____

Special Reports (prices vary) _____

Send to: Institute for Research on Poverty
1180 Observatory Drive
3412 Social Science Building
University of Wisconsin
Madison, WI 53706

Name: _____

Address: _____

_____
City State Zip

Focus

1180 Observatory Drive
3412 Social Science Building
University of Wisconsin–Madison
Madison, Wisconsin 53706

Nonprofit Org.
U.S. Postage
PAID
Madison, WI.
Permit No. 658

UNIVERSITY OF
WISCONSIN
MADISON