University of Wisconsin–Madison
Institute for Research on Poverty

Focus

Volume 18

Number 3

Spring 1997

ISSN: 0195–5705

Evaluating comprehensive state welfare reforms: An overview

Thomas Kaplan

Thomas Kaplan is Associate Scientist at IRP.

In August 1996, Congress and the president replaced the 60-year-old Aid to Families with Dependent Children (AFDC) program with a block grant, Temporary Assistance to Needy Families (TANF), permitting states to experiment with new forms of assistance to low-income families. States seem likely to try increasingly ambitious reforms as they gain experience under their TANF flexibility, at least if they respond as they did to federal permission to waive AFDC requirements. When President Bush first announced the new AFDC waiver options, states sought approval for only one or two proposals at a time. In the last months of AFDC, however, state waivers grew much more comprehensive. North Carolina, for example, secured a waiver that would allow the state to

- introduce an individualized personal responsibility contract;
- start a Learnfare initiative;
- require teen parents to live in a supervised setting;
- change the 100-hours rule in the Unemployed Parents program;
- increase the asset eligibility limits for AFDC recipients;
- expand the number of families required to participate in a JOBS program;
- increase sanctions for noncompliance with JOBS;
- impose time limits on receipt of AFDC benefits;
- require parents to obtain routine immunizations and medical examinations for their children.

Wisconsin Works (W-2): A brief description

The W-2 program will eliminate AFDC and replace it with cash assistance available only through work or participation in worklike activities. Under the program, which is scheduled to start on September 1, 1997, participation requirements will begin when the youngest child is 12 weeks. Families will be assigned to a Financial and Employment Planner, who places them on a level of a "self-sufficiency ladder" and helps them move up the ladder to greater independence, as indicated in the grid in the table. Small loans, which can be paid back in cash or community service, will be available in order to assist participants to find and keep work. Although some W-2 recipients will technically receive a grant, all cash income in the program will be disbursed on an hourly basis in return for each hour of work or program participation. Failure to participate will reduce income. Unlike the current AFDC program, the income will not depend on family size, but only on the case head's hours of participation and level on the W-2 self-sufficiency ladder. Also unlike AFDC, W-2 is not statutorily identified as an entitlement. Under W-2, there is no provision for subsidized formal education.

W-2 participants will receive assistance with child care costs; this assistance will require copayments from recipients. Working families with incomes below 165 percent of the poverty line at program entry will be eligible for child care subsidies. Wisconsin has also requested, but not at this writing received, a federal waiver to replace the current Medicaid program for W-2 recipients with a W-2 medical program that would provide benefits generally similar to those available under Medicaid to a somewhat different population (some expansions and some contractions in eligibility), contingent on the payment of a premium. All W-2 services will be available to both single- and two-parent families with children under 18 and with incomes below 115 percent of the poverty line. W-2 will also eliminate the current practice under which child support income beyond the first $50 in a month goes to public agencies to reimburse welfare expenditures. W-2 participants will be able to keep all the child support paid on their behalf.

Two significant early implementation phases of W-2 have been in place since March 1996:

Self-Sufficiency First. This is a diversion program which requires applicants for AFDC to complete an interview with a financial planning resource specialist before applying for AFDC and then to participate in the state JOBS program during the 30-day AFDC application processing period. In these 30 days, the applicant must devote at least 60 hours to JOBS participation, including 30 hours of direct employer contact. If needed, child care is provided during the applicant's hours of JOBS participation.

Pay for Performance. This is an intensive JOBS program requiring 20–40 hours per week of participation from AFDC case heads and imposing heavy sanctions on those who fail to comply. For each missed hour of required JOBS participation, a penalty equal to the federal hourly minimum wage is imposed, first on AFDC and then on Food Stamp benefits. Recipients who participate in JOBS for less than 25 percent of scheduled hours receive the full penalty, which reduces the AFDC grant for the next month to $0 and the Food Stamp benefit to the federal minimum of $10. Subsequent participation in JOBS restores benefits for future months.

Level of W-2	Basic Income Package	Time Required of Recipients	Program Time Limits	Est. Child Care Copays ($/mo.)	
				Licensed Care	Certified Care
Unsubsidized employment	Market wage + Food Stamps + EITC	40 hrs/wk standard	None	$101–$134	$71–$92
Trial Job (W-2 pays maximum of $300/mo. to the employer)	At least minimum wage + Food Stamps + EITC	40 hrs/wk standard	Per job: 3 mo. with an option for one 3-mo. extension; total 24 mo.	$55	$38
Community Service Job (CSJ)	$673 per mo. + Food Stamps (no EITC)	30 hrs/wk standard; and up to 10 hrs/wk in education and training	Per job: 6 mo. with an option for one 3-mo. extension; total: 24 mo.	$38	$25
W-2 Transition (placement contingent on assessment by the state Vocational Rehab. agency)	$628 per mo. + Food Stamps (no EITC)	28 hrs/wk work activities standard; and up to 12 hrs/wk in education and training	24- mo. limit, but extensions permitted on a case-by-case basis	$38	$25

Sources: K. F. Folk, "Welfare Reform under Construction: Wisconsin Works (W-2)," *Focus* 18, no. 1 (special issue 1996): 55–57, and presentation materials created by the Wisconsin Department of Workforce Development.

Notes: The income package and child care copayment are based on Governor Thompson's proposals in the 1997–99 biennial budget. Estimated child care copayments are for a three-person family with two children receiving no child support payments. Department of Workforce Development materials express child care copayments on a weekly basis; the monthly copayments shown in the table assume 4.2 weeks per month. For the purpose of estimating child care copayments, the Trial Jobs position is assumed to pay minimum wage, which, after October 1, 1997, will be $5.15 per hour, or $858 per month, and the unsubsidized-employment package is assumed to range from $6–$7 per hour, or $1,000–$1,170 per month.

The comprehensiveness of such reforms and the potential for still broader change in the future create at least two challenges to evaluation. First, because components of a broad reform could be synergistic, the impact of a comprehensive reform may surpass the sum of the impact of each individual reform within it. If so, evaluators would understate the impact of a reform component *in a particular context* if they evaluate only the one component. Yet assessing a comprehensive reform in its entirety, especially if its elements change over time, can be harder than assessing components of it.

Second, comprehensive reforms create special difficulties for evaluations that utilize an experimental design, a standard way of determining cause and effect. The difficulties arise because comprehensive welfare reforms often bundle new requirements for school attendance, work, or a daily activity that leads to work, along with modified policies for adjusting family income according to family size, into packages designed to generate social messages that discourage dependency on public assistance, promote work, and influence family-formation decisions. It is possible that such messages would be less effectively communicated in an environment of random assignment to control and treatment groups than in a more universal program. It is also possible that the control group would be influenced—and thus rendered invalid as a true control—by the messages conveyed to the treatment group.

Researchers have increasingly remarked on these and other problems of evaluating comprehensive welfare reform. But reaching consensus on evaluating such programs has proved elusive. In November 1996, IRP held a conference on the evaluation of the Wisconsin welfare reform program (Wisconsin Works, or W-2), in an effort to use one state welfare reform plan as a laboratory for evaluation. Although Wisconsin is not the only state with comprehensive reform, W-2 is among the most detailed and ambitious of the state plans. (Its basic features are summarized on p. 2.) IRP affiliates and staff believed that the program offered a worthy test of our ability to plan an evaluation of comprehensive welfare reform.

Intensive IRP work on the evaluation of W-2 started in the summer of 1995 and was supported by financial assistance from the Joyce Foundation and the Charles Stuart Mott Foundation. Some of the issues that researchers confronted are discussed below.

The distinction between monitoring and evaluation

Monitoring involves tracking income, earnings, family composition, and other social and economic indicators as a new program is created. Evaluation goes beyond monitoring, to determine if a causal link exists between the reform and any changes observed in the indicators. Experimental evaluation designs, controlled so that the only difference between the experimental and control group is the new program, offer the standard method of determining how much the changed policies altered the relevant indicators. But if experimental design cannot be used, will any other design allow for evaluation, or is close monitoring of the indicators (a useful task in itself) all that will be possible?

Members of the IRP working group agreed that experimental design was inappropriate for evaluation of W-2, but differed over whether reasonable causal inferences could be made in the absence of experimental design. The majority maintained some optimism that at least limited causal attributions would be possible without experimental design. Comprehensive welfare reform, after all, is not the only policy intervention for which experimental evaluation designs are unsuitable. The effects on the economy of Federal Reserve Board policies, the deterrent effects on potential aggressors of new weapons systems, and the impacts of local government efforts to encourage economic development are also unknowable through formal experimental design. Yet useful causal knowledge can accumulate concerning these and similarly complex interventions.

Nonexperimental counterfactuals

In an experimental design, the causal link between the intervention and observed changes in indicators arises from the experimental structure: people or families are randomly assigned to the new treatment and to the status quo. If we have grounds for believing (1) that the people in each group did not differ in relevant characteristics upon their entry into the program, (2) that the treatment and control groups did not "contaminate" each other during the experiment, and (3) that the two groups were equally affected by other changes occurring at the same time, then any post-treatment differences between the two groups in the relevant indicators can be attributed to the interventions.

But if experimental designs cannot be used in the new welfare reform evaluations, a different "counterfactual" is required if we are to have any sense of what would have happened without the reforms. IRP researchers discussed seven other counterfactual possibilities, which we later determined to fall broadly into two approaches: *pre-post designs*, in which post-treatment indicators are compared to pretreatment indicators, and *cross-state designs*, in which Wisconsin indicators are compared to indicators in other states with different interventions. The advantages and disadvantages of these approaches are discussed in papers by Robert Haveman, Glen Cain, and Irving Piliavin and Mark Courtney.

Key indicators

Regardless of whether or how causality is established, the populations and outcomes to be studied must be identified. All members of the working group favored a concentration on the broad low-income population—those with incomes below 200 percent of the federal poverty line—not just on participants in the W-2 program. The reason for looking beyond program participants is the possibility that the program will have entry and exit effects—that is, the program itself may induce some people to avoid, enter, stay in, or drop out. In that case, its full impact could not be inferred from observing only program participants. There was less agreement on the possible outcomes of interest. Some members of our group favored concentrating on earnings, hours worked, and receipt of welfare. Others favored attention to a broader set of concerns, including the effects on family-formation decisions, health status, participation in health care programs, the incidence of child abuse and neglect, and the use of foster homes and other out-of-home placements. All of us wanted to give attention to "process indicators," those indicators that would help program managers judge the quality of implementation and make timely policy adjustments based on early program experience. We did not reach consensus on what those process indicators might be, although papers by Thomas Corbett, Michael Wiseman, and Karen Holden and Arthur Reynolds narrowed our differences (the last two are summarized in this issue of *Focus*).

Other issues

In the course of our discussions, several other issues were raised, though at less length. How long should W-2 operate before it was subjected to impact evaluation? How dependable were the various data sources? What does and does not qualify as a W-2 intervention? (We followed the evolution of the W-2 program over the course of months, and considered whether—especially for pre-post analysis—W-2 changes should be deemed to include major policy modifications that the state of Wisconsin made within AFDC during 1996, or just the establishment in 1997 of W-2 itself.)

Our deliberations generated the papers that formed the backbone of the November conference, and that are summarized in this issue of *Focus.* With them we include invited comments from participants in the conference, to give some sense of the intense and directed discussions that took place. We invite readers, in their turn, to move the discussion forward.■

The proceedings of the conference upon which this *Focus* issue reports have been published in
Evaluating Comprehensive State Welfare Reforms: A Conference
IRP Special Report no. 69, March 1997. 273 pp.

An electronic version is available without charge through the IRP Web site at http://www.ssc.wisc.edu/irp/
Printed copies may be purchased from IRP Publications for $14.50
(order form is on p. 75).

A related publication is "Informing the Welfare Debate: Perspectives on the Transformation of Social Policy,"
IRP Special Report no. 70, April 1997. 164 pp. $10.00.

The next generation of welfare reforms: The challenge to evaluation

Thomas Corbett

Thomas Corbett is Assistant Professor in the School of Social Work at the University of Wisconsin–Madison and Acting Director of IRP.

Welfare devolution will profoundly affect how we evaluate policy and program innovations. The existing standards for doing impact evaluations based on classic experimental techniques, which work best for limited changes within stable program environments, may no longer be feasible nor warranted in all instances. Standards for conducting implementation and process analyses have never been well developed, yet they are increasingly critical given the complex character of recent reform. The IRP conference upon which this issue of *Focus* reports was convened, in essence, to think through the challenges that the evaluation community must address and to begin a dialogue that eventually will identify those evaluation tools and strategies appropriate to the next generation of reform. The conference focused upon Wisconsin Works (W-2), a dramatic reform being introduced in Wisconsin that seemed emblematic of what we might expect under welfare devolution.[1]

The federal legislation enacted in August 1996, the Personal Responsibility and Work Opportunity Reconciliation Act (PRWORA), transfers substantial responsibility over what had been the Aid to Families with Dependent Children (AFDC) program to the states. Key provisions of the act convert public assistance to poor families with children from an open, individual entitlement, in which the fiscal cost of providing cash support is shared by the federal and state governments, to an entitlement to the states, in which future assistance is time-limited and the federal contribution is a fixed or capped amount ("block grant").

Under the new legislation, states will theoretically have greater flexibility in designing and managing their support programs for poor families with children.[2] But the greater flexibility brings increased responsibility and greater risk. Under PRWORA, states and local governments are likely to bear the full fiscal risk of policy decisions *at the margin*, after the fixed federal fiscal contribution is exhausted. Thus, if policy makers assume certain behavioral responses to a reform and guess wrong, they could easily incur substantial costs or be required to ration services and benefits as welfare budgets come under increased scrutiny.

In such a devolved policy environment, program evaluations become more useful as the decisions vested in local governments become more complex and consequential. But though the *value* of knowledge is greater, the *price* of obtaining it will also increase. The federal government will no longer mandate that programs be evaluated, nor ensure that certain methodological standards be maintained, although it will continue to play a role (see note 7). Many jurisdictions will hope to be free riders, letting others incur the fiscal and other costs of doing evaluations while still taking advantage of the results.

Evaluations of the *next* generation of state reforms will also be challenging, because emerging programs and policies emphasize varying combinations of what Lawrence Mead has termed a "new paternalism."[3] These programs, which often require school attendance, work, or a daily activity that leads to work, are bundled into packages designed to discourage dependency on public assistance, influence family-formation decisions, and promote work. They often call for radical alterations in agency culture which, among other things, will decentralize decision making, rendering management control more problematic.

Devolution and the challenge to evaluation

Devolution has appeared in two main guises. *Structural or legislated* devolution calls for the transfer of control over welfare to the states. It terminates the individual entitlement status of selected income maintenance and service programs; groups related programs into broad program areas, defined either by common target populations or common service technologies; converts federal contributions for programs from matching formulae into closed-end block grants so that, on the margin at least, costs are no longer proportionately shared among local and federal governments; and eliminates or reduces the federal role in rule making, evaluation, and technical assistance.

With passage of PRWORA, this vision was partially realized. The law is the result both of widespread support for reform and a public perception that states have the recipe for making welfare reform work. This belief is in part the product of widespread state innovations in welfare program operation that are known as Section 1115 waivers after the section of the Social Security Act that permitted them. Waiver-based innovation therefore, constituted the second welfare reform strategy, sometimes referred to as *incremental reform.* Beginning on a small scale in the early 1980s, and spurred on by the rhetoric of the 1992

campaign to "end welfare as we know it," state waiver requests exploded. By the time that the welfare bill was signed by President Clinton, devolution of program and fiscal responsibility was an established fact, and any remaining pretense to a national welfare policy, other than the procedural requirements associated with the cost neutrality and evaluation requirements, had largely disappeared.

The outpouring of investigation associated with waiver-based demonstrations promised to enhance our understanding of the dynamics of public assistance and to identify program and management improvements. Prior to PRWORA, both Republican and Democratic administrations encouraged states to incorporate an explicit evaluation scheme in their requests for waivers. The cumulative evaluation results were seen as the social policy equivalent of a *Consumer Reports* special issue on techniques of welfare reform. They might possibly allow programs and policies to be compared and might serve as the welfare equivalent of the Consumer Union "best buy."

This happy state of affairs may not transpire, for several reasons. First, many of the experiments in progress contain serious flaws in content and assessment strategy that, should they be completed, will diminish both the management utility and the internal and external validity of the outcomes. Second, PRWORA changes federal and state incentives for experimentation, diminishing the likelihood that most ongoing experiments will be completed. Third, the character of future reforms may be so different from the changes examined in the past that the new demonstrations will not provide much useful instruction about how to proceed in the future—a true discontinuity in policy making.

Waiver-based activity may, nonetheless, provide clues to future policy directions. First, the *scope* and *pace* of waiver activity demonstrated just how extensive devolution was, even before it was formally legislated. Over 90 percent of all states and the District of Columbia had at least one approved waiver. The new legislation provides that existing waivers may, at the discretion of the state, continue in effect for the duration of the originally granted time.[4]

The *complexity* of state-based welfare demonstrations also increased rapidly. In the early days, a state would request permission to modify a few provisions of the Social Security Act in order to implement one or two new ideas. In recent years, the number of major changes to program parameters contained in a single waiver request would be in double figures. States increasingly "borrowed" ideas from other jurisdictions and bundled them together in complex reform packages (see the list of North Carolina reforms in the article by Thomas Kaplan, this issue, p. 1). The political popularity of the reforms could not be ignored by the nation's governors. Whether justified or not, claims of state successes gained widespread media attention while national solutions, such as President Clinton's Work and Responsibility Act of 1994, sputtered in Congress and quickly disappeared from sight.[5]

Perhaps the most important trend in recent state demonstration activity is the stress on changes designed to alter critical personal and interpersonal behaviors—to help (or obligate) people to play by society's rules: get a job, get married, make responsible fertility decisions, be a good parent, and obey the law. As late as 1992, for example, only a couple of states had shown interest in the family cap concept, under which benefits would not be raised if mothers had more children while on AFDC. By the summer of 1996, about 40 percent of all states had actual or proposed family cap provisions, and some 18 states had (or had proposed) provisions that required minor parents on assistance to live with their parents or in a supervised setting as a condition of eligibility.

Evaluation methods and the next generation of reforms

In the old welfare world (just a few years ago), change was limited, incremental, and linear. The tools for learning from these experiments had been fairly well developed and standardized by a handful of large evaluation firms.[6] But the discontinuity in policy making imposes new expectations upon the conventional methods of the evaluation community.

Will program devolution stimulate or inhibit innovation and knowledge building? States might continue to experiment with new policies and program forms, given that they are freed from most federal regulations. Or they might become more cautious, given that additional fiscal risk is shifted to them. Or they might initially engage in innovative behavior, given that many will experience a fiscal windfall in the short term, but be more conservative as the federal contribution declines. Nor will welfare demonstration activity necessarily lead to more theoretical and practical knowledge. The federal government will no longer mandate evaluations or require rigorous evaluation designs, and states may have neither the will nor the budget to undertake them (see the comment by Robert Lovell, p. 9).[7]

If we assume that past waiver activity is a guidepost to the future, the next generation of reforms will be:

behavior oriented rather than *income oriented.* By extension, the dominant treatment modalities will shift from income support strategies to service technologies, broadly understood. It is likely that there will be a shift toward case management, crisis management, and counseling activities that resemble the social work functions that marked the provision of public assistance some three decades ago.[8]

dynamic rather than *static*. They will actively work toward changing participants' behavior and attitudes. W-2, for example, explicitly talks about participants ascending multiple tracks (or tiers) built into the system and "graduating" into the labor market and mainstream society.

longitudinal in character. Participants will be viewed as being in a *process.* They will be subject to time limits, both within program components and overall. The workers and the system must remain sensitive to where participants are relative to these temporal constraints.

craft oriented rather than *routinized*. Under the old welfare, some of the decisions were fairly complex, but the basic intent was to treat all participants alike—a rough justice. The new generation of reforms are designed to treat participating families individually. Many involve the negotiation of personal "social contracts" or "individualized employment plans."

multidimensional rather than *unidimensional*. Participants will not proceed through the welfare experience in lockstep but are likely to be tracked along different paths. Differential *tracking* suggests that important decision points exist where "triage" will occur and participants will be sent along distinct program trajectories that explicitly recognize diversity within the welfare population.

characterized by complex, *discretionary* decision making that will require a good deal of professionalism. There are three basic forms that administrations might take. One is to control the behavior of frontline technicians through organizational structure (pyramid shape, vertical communications), management style (detailed manuals, strong supervision), and training methods (encourage professional standards). The second is to recruit professionals, who bring in their own standards and expertise and are likely to resist highly controlling organizational regimes. The third is to move to performance-based management or to engage in privatization.

labor intensive for the organization that implements them. The old welfare involved repetitive, routine decision making, with an emphasis on efficiency and accuracy.[9] Participants who wanted help were referred to other systems. Not surprisingly, administrative costs often were less than 10 cents on the dollar of benefits issued. W-2 and similar reforms will require intensive case management and a very active participant-worker interaction.

The paradigm shift that these changes represent will inevitably be accompanied by differences in management and oversight. Standardized, routinized rules and procedures lend themselves to stable policy environments and vertical, hierarchical management structures. The new welfare policy environments are neither very stable, nor amenable to top-down or vertical control.[10] It will become more difficult for states to centrally manage and prescribe all aspects of agency operations. Consequently, intrastate variation may begin to rival interstate variation as a descriptor of the next generation of reform.

Wisconsin Works: The case study

Articles in this issue use Wisconsin Works (W-2) as the policy example through which to explore the evaluation issues posed above. Arguably, no other state reform so fully captures the attributes and spirit of the next generation of welfare reform. Because W-2 is so complex and ambitious, we have no way of predicting in advance its net effects on the well-being of low-income families in Wisconsin and on their communities, labor markets, and the systems that provide critical services. We believe that it is a suitable, though far from perfect, laboratory for evaluating process and program implementation questions.[11]

Future directions for evaluation

The conventional approach to evaluation has been to change one or two parameters of the existing welfare program, randomly assign participating families into either an experimental or a control group, and examine "net" outcomes on a limited number of measures that virtually everyone agreed were important. But now, policy designers want to communicate a whole new set of messages to low-income communities regarding the work ethic and family values. From the parochial perspective of the states implementing the reforms, there is little reason to look beyond whether welfare rolls decline and whether fewer children are being reared out of wedlock. From a broader social view, we ought to adopt a more comprehensive and longer-range research and evaluation plan.

There are many puzzling questions that evaluators must confront in the future, and not all of them lend themselves to simple technical solutions. How does one establish a *counterfactual?* How are the correct *criterion variables* selected? How does one agree upon which *target groups* to examine? What is the appropriate *unit of analysis*, individual or case or agency or county? How does one go about determining *overall* and *component* effects? Should the implementation analysis be used to shape and refine the program, if that increases policy instability and confounds the impact analysis? Should local discretion and flexibility be curtailed, so that the character of the intervention may be better understood? As the need for good empirical information increases, the cost and difficulty of obtaining those answers increase commensurately.

Any future evaluation agenda should not only assess the success of the reforms in meeting their central objectives, but should also review other possible consequences and the mechanisms through which both intended and unintended consequences occurred. For example, the next

generation of reforms is intended to affect communities, not just program participants. Population and entry effects are, therefore, of considerable importance. So too is the context for the findings, insofar as this affects both interpreting them and applying them in other settings.

As a convenient way of viewing some of the evaluation challenges in this "age of policy discontinuity," we can locate a number of the issues along three axes: (X) major methodological questions; (Y) substantive policy changes; and (Z) alternate foci of concern (or units of analysis).

X axis. The major methodological questions are typically organized in three parts. *First*, we want to assess how closely operations reflect policy intent, particularly with complex interventions requiring complicated sequencing of tasks carried out by varied actors representing different institutions. If we do not, we wind up with "black box" evaluations that are difficult to interpret. *Second*, we want to assess benefit/cost or cost/effectiveness ratios. Do benefits exceed costs or, at a given cost, do alternative strategies produce differential outcomes? *Third*, we want to assess "net" outcomes or impacts by comparing outcomes for those exposed to the intervention with an appropriate comparison group (the counterfactual), created through random assignment of individuals or sites, or through statistical procedures designed to account for heterogeneity between groups.

Y axis. The new generation of social policy innovations are intentionally designed to affect several aspects (or domains) of the lives of those exposed to the program. W-2, for example, is designed to effect changes in the following domains: labor supply, skill or human capital development, health status, child care arrangements, fertility decisions, family formation and functioning, and child development. Each area raises questions about what outcomes are critical, how the outcomes can be operationalized, what data sources exist, and how good they are. Multiple domains also raise questions about interpreting results that may go in contrary directions.

Z axis. The new generation of reforms raises questions about the appropriate unit of analysis. As the demonstrations become more ambitious, the anticipated effects are likely to be felt at many levels. IRP affiliates have long been concerned about this, and organized a national conference in 1991 to discuss the complexity of moving from one-generation evaluations (adult-only) to two-generation (adult and child) evaluations.[12] Along this axis, different issues will arise as we measure effects on the adult caretakers(s); the family as a whole; the child, community, and labor market effects; and institutional and service-provider effects.

Illustrations of each area above must suffice. Process (or implementation) evaluations have traditionally been the stepchild of impact evaluations. They have become more important as welfare-to-work evaluations done in the 1970s and 1980s raised questions about what really was being tested (if anything). But the craft of doing these well is still in its infancy, and there are few accepted protocols for collecting data and reporting findings on complex operational systems. The objectivity and comparability of process evaluations must be enhanced.

Impact evaluations, on the other hand, raise important questions about how to establish a counterfactual (no exposure to the program) when the state wants to saturate the county or state. States often have good reasons for doing so (interest in changing community norms, or worries about administrative complexity, or migration effects, etc.), but statewide programs do complicate the task of making causal inferences. There is legitimate confusion about whether the AFDC program, now technically defunct, can serve as the counterfactual to a new program. Perhaps instead we should be comparing new competing models for assisting low-income families.

Recognizing that effects may appear in different domains is one thing. Measuring changes in different domains is quite another. Collecting and interpreting high-quality data become more problematic as we move toward measures beyond the conventional outcomes of welfare utilization, economic well-being, and labor force participation. Some evaluation issues cut across any classification scheme. At any level of analysis, critical data may not exist, or exist in the right form, or be credible and reliable.

Conclusion

This article has overviewed both the importance and the difficulties of doing high-quality evaluations of welfare reform in the future, and has suggested basic principles that might inform future evaluations. The difficulties are many, but perhaps the greatest challenge to future evaluations lies in the possibility that states, no longer dominated by federal evaluation requirements, will substantially abandon rigorous, dispassionate evaluation activity. The trenchant observations of Michigan State official Robert Lovell (page 9) make it clear just how serious this risk may be.[13]

The irony is that although devolution promises a veritable explosion in innovation and social policy knowledge building, we may well wind up learning very little from this infusion of creativity and energy. The transformation of the social safety net upon which many states are now embarking carries with it great opportunities and great risks. As noted, some of those will fall to the states, who will incur a greater share of program costs over time. But much of the risk falls upon the most vulnerable segment of society—children. In recent years, one child in four under the age of six has lived in poverty. If some of the new programs improve the well-being and life prospects of disadvantaged children, we need to know that, and we need to know what the mechanisms of suc-

Invited comment: A skeptical view of evaluation

The Institute's November conference on evaluating comprehensive welfare reform in general, and Wisconsin's program in particular, highlighted for me the range of difficulties we face in applying the social science tools developed in the last fifty years to the problems faced by states in the next five years. This is not rocket science; it's much more difficult. The rocket engineer chooses among cost, weight, and reliability, has a very successful theory of physics to predict results, knows the goal exactly, and can test each component individually before assembling the product. Social welfare evaluation has none of these advantages: our tradeoffs among cost, reliability, timeliness, protection of subjects, and threats to validity are more complex, our theories have only weak predictive power, we have as many goals as the programs we study, and we seldom have the luxury of testing each aspect of our designs separately.

Much of the discussion at the conference was about evaluation approaches expected to guard against this or that important threat to validity. This is thought to be very important in the highly political environment in which welfare policy is now debated; we want our results to be the subject of discussion, not our methods. I fear, however, that the search for a "bulletproof" design is ultimately doomed. Just as they can claim success merely by implementing a reform, policy advocates can successfully dismiss research results by attacking the methods, or the bias of the researcher, merely by making a plausible-sounding argument. Subsequent analysis of the argument becomes insider stuff, of little interest to the press or public. For example, a major study of the termination of general assistance in Michigan, showing that few former recipients were working, was dismissed because the survey-takers located only half of those in the sample. Almost surely, the other half were faring less well, but this counter-argument did little to blunt opposition to the study's findings.

Also doomed as welfare policy study tools, I think, are both the longitudinal study extending over two or more years and the national sample studies based on the Survey of Income and Program Participation, the National Longitudinal Study of Youth, the Panel Study of Income Dynamics, and similar databases. With devolution of nearly all policy authority to the states, you can almost feel the legislators' and lobbyists' excitement. Each of them with a pet project or tax cut in need of funding will be tempted to finance it through small reforms or reductions in welfare. Many will succeed, and the result will be to make welfare reform an annual event in many states. Thus, anyone launching a study which takes two or more years to complete risks buying an expensive irrelevance, and few states will be able to afford this risk. National databases seldom have state-level samples of those below the poverty level which are large enough to reach reliable conclusions. This was an acceptable fault when AFDC provided a rough national standard but, with fifty-one different programs in place, determining the policy implications of results will be difficult or impossible.

For now, Michigan policy makers and evaluators are placing their faith in studies based on available administrative databases, quick-turnaround surveys (particularly those focused on special subpopulations), and longer-term studies of single policy options like full-grant sanctions for failure to participate in job training programs. The comprehensive evaluation of our waiver policies will be completed this year, and will be our last for a while.

Robert G. Lovell, Director
Staffing and Program Evaluation Division
Michigan Family Independence Agency

cess are. If the lives of these children and families deteriorate in some places, not only the children but also the community will pay a heavy price. Without a serious and sustained evaluation undertaking, efforts at program and policy improvement and correction may well be random, ill-informed ventures.■

---

[1]The paper upon which this article is based appears in full in "Evaluating Comprehensive State Welfare Reforms: A Conference," IRP Special Report no. 69, University of Wisconsin–Madison, March 1997. The author is grateful for the contributions of Michael Wiseman.

[2]States may have *less* flexibility now than they did under prior welfare law and a liberalized waiver policy. PRWORA is a very inconsistent piece of legislation in which devolution of authority over welfare to states is offset by many prescriptive provisions respecting program objectives that must be met and expenditures of federal dollars that are prohibited. Moreover, even assuming modest inflation rates, the $16.4 billion capped federal commitment will decline in value by 20 to 25 percent, in inflation-adjusted terms, over the course of the legislation.

[3]L. M. Mead, "Welfare Policy: The Administrative Frontier," IRP Discussion Paper no. 1093-96, University of Wisconsin–Madison, 1996. For fuller discussions of the new federalism policies, see *Focus* 18, no. 1 (special issue 1996).

[4]For a discussion of welfare waivers, see E. Boehnen and T. Corbett, "Welfare Waivers: Some Salient Trends," *Focus* 18, no. 1 (special issue 1996): 34–37.

[5]President Clinton's proposed national welfare reform was submitted to several committees on June 21, 1994, after more than a year of development. The proposed legislation, including explanatory text, was 622 pages in length. It was too long and complicated to get a fair hearing in a Congress that was increasingly concerned with mid-term elections. The 1994 congressional election results buried any hope for the president's proposal and dramatically shifted the debate.

[6]Basically the Manpower Demonstration Corporation, Mathematica Policy Research, Abt Associates, Urban Institute, and a handful of others. They refined basic approaches that had been developed partly at IRP in the late 1960s and early 1970s, when the methods for

evaluating the Negative Income Tax and Supported Work proposals were developed.

[7]PRWORA sets aside $7.5 million per year to support state evaluations of their welfare reforms. An RFP to help select recipients of these resources was issued in fall 1996, with responses due January 15, 1997. Some 30 states submitted 43 separate proposals totaling about $20 million.

[8]Social work services were an integral part of welfare systems when it was argued that recipients might be rehabilitated or counseled out of dependency. This was particularly true in the 1960s, though the rehabilitation theme can be traced back to the Scientific Charity movement of the 1880s. Income maintenance functions were formally separated from service functions in the early 1970s.

[9]Particularly in states like Wisconsin, little discretion remained in the system. For example, the Wisconsin Computer Reporting Network (CRN), developed in the 1970s, automated virtually all core decisions regarding the determination of eligibility and the calculation of benefits. Permissive language in the welfare manuals—the use of words such as "may"—was replaced by the use of nondiscretionary words such as "shall."

[10]In the last decade, Wisconsin, for example, has obtained 15 relatively major waivers and has launched other changes not requiring waivers. The contract has just been awarded to evaluate two major precursors to W-2, Self-Sufficiency First and Pay for Performance (see this issue, p. 2), but the state is already considering how to fold them into the next generation of reforms, to be implemented in less than 18 months.

[11]For an excellent treatment of administrative issues, see Mead, "Welfare Policy."

[12]*Focus* 14, no. 1 (Spring 1992):10–34 reports upon the conference.

[13]Dr. Lovell, who has been involved in most of Michigan's evaluations for over a quarter of a century, is retiring from state service at the end of June 1997.

Making Ends Meet: How Single Mothers Survive Welfare and Low-Wage Work

by Kathryn Edin and Laura Lein

Making Ends Meet offers compelling evidence that in the present labor market, unskilled single mothers who hold jobs are frequently worse off than those on welfare, and neither welfare nor low-wage employment alone will support a family at subsistence levels.

Kathryn Edin and Laura Lein interviewed nearly four hundred welfare and low-income single mothers from cities in Massachusetts, Texas, Illinois, and South Carolina over a six-year period. Their budgetary analyses reveal that even a full range of welfare benefits—AFDC payments, food stamps, Medicaid, and housing subsidies—typically meet only three-fifths of a family's needs, and that funds for adequate food, clothing and other necessities are often lacking. Leaving welfare for work offers little hope for improvement. Jobs for unskilled and semi-skilled women provide meager salaries, irregular or uncertain hours, frequent layoffs, and no promise of advancement. Mothers who work not only assume extra child care, medical, and transportation expenses but are also deprived of many of the housing and educational subsidies available to those on welfare. Regardless of whether they are on welfare or employed, virtually all these single mothers need to supplement their income with off-the-books work and intermittent contributions from family, live-in boyfriends, their children's fathers, and local charities.

Almost all the welfare-reliant women interviewed by Edin and Lein made repeated efforts to leave welfare for work, only to be forced to return when they lost their jobs, a child became ill, or they could not cover their bills with their wages. Mothers who managed more stable employment usually benefitted from a variety of mitigating circumstances such as having a relative willing to watch their children for free, regular child support payments, or very low housing, medical, or commuting costs.

Kathryn Edin is Assistant Professor in the Department of Sociology and Center for Urban Policy Research at Rutgers University. She is an IRP associate. Laura Lein is Senior Lecturer, Department of Anthropology and Senior Lecturer and Research Scientist, School of Social Work, at the University of Texas at Austin.

Available from: Russell Sage Foundation, 112 East 64th St., New York, NY 10021. 320 pp. $19.95 paper, $42.50 cloth.

A pre-post design for state-based evaluation of national welfare reform

Robert Haveman

Robert Haveman is John Bascom Professor of Economics and Public Affairs at the University of Wisconsin–Madison and an IRP Affiliate.

The 1996 federal welfare reform legislation that has replaced Aid to Families with Dependent Children (AFDC) with Temporary Assistance for Needy Families (TANF) has far-reaching implications for the well-being of the low-income population.[1] The new law has several features that make it difficult, perhaps impossible, to reliably evaluate both its overall national effects and the particular effects of any state's implementation of the law.

First, a key provision of the new law is that the federal government will provide funds in block grant form to each state, allowing it to design its own system of cash support largely free of federal requirements regarding benefits or administration. A national welfare system characterized by some degree of coherence and similarity will be replaced by 50 quite disparate systems, constrained only by the few mandates in the national legislation.[2] This fact has important implications for efforts to provide a nationwide assessment of the new law or to design an evaluation of any particular state's policy.

Second, the embedding of state policy changes within a large-scale national policy shift poses an especially difficult research challenge. We may be able to compare the behavioral and administrative changes occurring over time in one location (site or state) to those occurring in another location. This simple comparison does not, however, *evaluate* the effects of the policy change, because underlying social and economic conditions in the different locations may also change in different ways and at different rates. Or we may assess changes in the variables of interest that are attributable to a particular state's policy change—a true "evaluation." But to do this we must know what has changed in the state's economic and social environment, apart from changes that are due to the policy. In both cases, research is hindered by the fact that, in a "general equilibrium" world, what happens in one state will be affected by the policy changes that are adopted in other states.

Third, some states seem prepared to implement on short notice a new and radically different program that is generally consistent with the new law; others are likely to make changes far more slowly. A successful evaluation of the policy change adopted by a particular state must take into account the speed of implementation within that state and in neighboring states.

Finally, the range of policy changes that states will undertake in response to the new law is potentially enormous. The populations eligible for support under any state's old and new systems will be quite different. Whereas the prereform system in all states had the same objective—supporting the income of eligible people—the new systems will seek to enforce work on a different pool of eligible citizens and to make assistance conditional on work. The financial incentives that states will offer to their low-income populations under the new law will be quite different from those that existed under the old law.

These considerations pose difficulties for program evaluators attempting either a national or a state-specific assessment. Some involve large questions of evaluation design, others involve practical questions regarding the outcomes (that is, the variables of interest) to be studied. Whereas a national evaluation might seek to measure general impacts at the national level, ignoring those that are specific to states or regions, this article addresses a different level of evaluation—the assessment of the effects on a state's population of the particular version of the policy change implemented in that state.[3] I discuss the choice of design strategies available for evaluating a particular state's initiative. I conclude that all the possible designs have fundamental problems, and that choosing a design will depend on both cost and the relative weights assigned to the problems associated with each.

An "ideal" evaluation of a state-based welfare reform measure

Some assumptions

In discussing the characteristics of an "ideal" evaluation, I make some assumptions, most or all of which will be violated in any real-time evaluation. In a subsequent section, I discuss evaluation strategies that may be feasible in a complicated world in which the assumptions do not hold.

First, in designing a state-based evaluation of welfare reform, one question seems to me central:

What is the impact of the policy change on the economic activities and well-being—work effort, family structure changes, health and nutrition changes,

changes in the care and nurturing of children—of those individuals and families that are the most likely to be affected by the policy change?

This question has two main elements. One is "impact": the study should seek to distinguish those changes in particular variables that are attributable to the policy from changes that may be due to other factors. The second is its emphasis on individual people and their well-being. A public policy is efficient only if the benefits that it conveys to people—its positive effects on their lives, living conditions, and well-being—exceed the costs that it imposes on them.

If this fundamental question regarding net effects on well-being cannot be answered reliably, one should seriously question the wisdom of devoting substantial resources to studying many other questions that might be asked about the changes that will follow the policy shift. Given this perspective, I conclude that a well-designed, longitudinal sample survey of households, and the measurement of outcome variables observable in survey responses, must form the core of a reliable analysis. However, apart from—and in addition to—such a survey, some useful evaluative information can be obtained through administrative data and time-series information on aggregate effects; these data can be studied separately from data obtained from a survey of households.[4]

Second, I assume that the policy change from the pre- to the postreform systems will be discrete, and that the characteristics of both systems can be clearly identified and described. But in some states, major reforms have been undertaken prior to—and in anticipation of—passage of the new law, so that defining and measuring the prereform system will be problematic. It is also likely that real-time policy implementation will be slow and uneven, so that in the years following passage of a state's reform law all that the analyst will be able to observe and assess will be some unknown combination of the prereform and postreform systems—a "policy change process" rather than a discrete policy change.

Third, I assume that the postreform system will develop in response to the 1996 legislation, and that no major changes in federal law will be made in subsequent years. In any real-time evaluation, this assumption is also problematic.

Fourth, I assume that the policy change undertaken by states in response the national law will be designed to "change the culture" in the state—that is, one of the objectives of the policy will be to change citizens' perspectives regarding the responsibility of the public sector to provide income support and the need for individuals to accept responsibility for their own financial well-being. Hence, virtually all citizens with low permanent incomes will be affected by the legislation; evaluation should not be limited to the population of current program recipients, or to those who would have been eligible for the current program.

Fifth, I assume that each of the designs considered is "feasible," in the sense that no state's legislation would forestall any particular approach and that obtaining the necessary data is possible, even though it might require data collection and household surveys in other states.

Some basic principles

A reliable evaluation of a state-based policy change embodied in the new welfare law presupposes at least two basic principles:

1. Clear identification of both the "counterfactual" and the "factual" policies. The analyst needs a clear description of the nature of the state's welfare system as it existed prior to the new welfare reform law (the counterfactual), and as it exists (or will exist) with the new law (the factual).

2. Quantitative comparison of the levels of variables of interest under both factual and counterfactual states of the world. We are interested in the *difference* between the observed level of a variable of interest (e.g., single mothers' work hours or earnings) under the **n**ew policy (the factual value of the variable—for convenience, L^n) and its observed level had the **p**rior policy remained in effect (the counterfactual level of the variable, L^p). Only by observing and reliably measuring both L^n and L^p can the effect of the policy change—L^n minus L^p—be obtained. If the new policy is imposed as a replacement of the prior policy, the measurement of L^n is, in principle, straightforward; it is the level of the variable observed over time, given the imposition of the new policy. Analysts cannot, however, directly measure the value of L^p, because it reflects the value that would have existed if the prior policy had been in effect, when in fact it has been replaced. They must, instead, attempt to identify a counterfactual that will yield an environment and behavioral incentives (and hence, levels of the variable of interest) as close as possible to those that *would have existed* if the prior policy had been in effect.

In general, there are three ways of measuring L^p, the counterfactual level of the variable:

1. Establish an *experimental design*, so that a randomly assigned sample (a control group) of those in the entire group affected by the policy change continues to operate under the rules of the prior (or counterfactual) policy. In this case, L^p can be directly measured for the control group and compared with L^n, measured for those confronting the new policy.

2. Establish a *comparison-site design*, by defining a comparison group of individuals who are not randomly assigned, but who confront behavioral incentives and state-of-the-world conditions that are as close as possible to those of the prior policy. L^p can then be measured for individuals in this "control site," and compared with L^n.

3. Establish a *pre-post design*, by defining a comparison group of individuals who are not randomly assigned but who, in fact, confronted the prior policy. L^P can then be measured as it existed for the comparison group, and compared with L^n.

The principles and the choice of a design

An experimental design

The "social experimentation" technique has become a popular method for evaluating the impact of proposed changes in social policy measures. Individuals in the target group (say, a state's lower-income citizens) are randomly assigned to a treatment and a control group; the former group is subjected to a "treatment," in the form of a policy environment that differs in some well-specified way from another environment in which that policy is not put in place. The kind of policy to which this design has been applied is, for example, a new job training program that is viewed as a replacement for an existing training program. Simultaneous observation of the two groups over time may reveal a difference between them in the level of a particular variable of interest—for example, the amount of work or the level of earnings after the training is completed. This procedure assumes that the new (or "treatment") policy does not affect the underlying "state of the world," perhaps because it is a limited or small change. Thus the hours of work or the earnings of the control group will reflect the state of the world if the new policy had not been introduced, and any difference in the values of these variables between the experimental and control groups can safely be attributed to the new policy. This experimentation technique can proceed if it is feasible to isolate the randomly assigned control group from the treatment group. The control group becomes the counterfactual against which the treatment group is compared.

The trade-offs. The experimental design can measure the outcomes for the control and experimental groups contemporaneously, hence securing constant external circumstances for the two groups. But it has two serious (and, perhaps, fatal) flaws. First, in the face of a policy that seeks to change the "culture" of public income-support expectations within the state, it will be difficult or impossible to isolate a within-state control group from the incentives of the new policy. Second, to implement an experimental design, a state would have to permit some citizens to continue to rely on the prereform system, and mandate that program administrators work with these clients under the terms of the previous system. In addition, a separate, pre-post design must be developed if evaluators are to use administrative and aggregate data to study the effects of change. By definition, such data are available only for civil jurisdictions or administrative agencies and only in time-series form. The different conceptual bases for these evaluations—the contemporaneous observation of samples of control- and treatment-group members in the experimental design and before-and-after administrative information for the pre-post design—could pose problems of interpretation.

A comparison-site design

In order to identify a counterfactual, this design requires a control group of individuals in an environment where both external circumstances and behavioral incentives are as similar as possible to those existing under the prior policy. As with the experimental design, the counterfactual group must both match and be isolated from the group of individuals who are subject to the incentives of the new policy. In the experimental design, random assignment assures that the control group "matches" the characteristics of the treatment group. There is no such assurance in forming the comparison-site group. In this design, a "statistical match" of individuals in the two groups must be made.[5]

The trade-offs. This design confronts two difficulties. The first is securing comparability between the treatment and control (comparison-site) groups in those external circumstances that are *not* associated with the policy change. Statistically matched individuals located in a comparison site must face economic and social conditions as close as possible to those that prevail in the site that is subject to the new policy. Second, the incentives and constraints of the policy in place in the comparison site must be as close as possible to those of the prior policy in the site (or state) whose policy change is being evaluated. But prereform policy differences between any two states or sites are likely to be substantial. Moreover, all states must respond to the policy mandate of the 1996 legislation. Thus it is unclear that any state will meet the requirements for a comparison site.[6] Finally, administrative and aggregate information available in the comparison site will inevitably differ in subject matter, coverage, and definition from that available in the evaluation site. As with the experimental design, a separate pre-post strategy must be developed for securing reliable and comparable data in two separate jurisdictions and perhaps distant geographic locations.

A pre-post design

In both the experimental and the comparison-site designs, evaluation of the policy change requires *contemporaneous* measurement of the variables of interest for groups representing those to whom the policy has been applied and those to whom it has not (respectively, the treatment vs. control and the comparison-site vs. policy-site groups). In the pre-post design, however, measurements for the control group must be made *before* the new policy is implemented. Two groups of individuals must again be designated, one subject to the prior policy and the other subject to the new policy. As with the comparison-site design, individuals in the two groups must be statistically matched.

The trade-offs. This design has the important advantage of being entirely "within site," and hence the effects of the policy on families within the state, obtained through a longitudinal household survey, can be made consistent in both time and coverage with the impacts measured by administrative and aggregative data. But because an evaluation using the pre-post design is based on measurements at two points in time, the analyst must attempt to secure comparability of the external circumstances—changes over time in demographic, social, and economic variables—that are *not* associated with the change in policy. In the absence of such similarity, it will be necessary to develop techniques to statistically adjust for the effects of differing conditions on those household variables and market changes that are central to the evaluation.[7] Securing an accurate representation of the prereform policy would seem in principle straightforward, but the evolution of state policy may in reality preclude a clear delineation of the environments before and after the policy change.

A pre-post evaluation design for Wisconsin Works (W-2)

What are the constraints and the environment with which an evaluation of W-2 will have to cope?

First, although successful evaluation of W-2 requires that the nature of the policy change be clearly defined, some of its elements have already been implemented in several counties, and all counties are being urged to seek employment for applicants prior to offering income support benefits.[8] Nor will any state with an economic and social environment similar to Wisconsin have in place a welfare system similar to the prereform Wisconsin system.

Second, it seems unlikely that the state of Wisconsin will be willing to exempt some sites (counties) from the W-2 legislation or to maintain the prereform AFDC system in order to make an experimental design feasible.

Third, Wisconsin has already embarked upon "changing the culture of welfare," pursuing a major and discrete change in the social expectations regarding work and individual responsibility of low-earnings-capacity individuals (especially, poor single mothers). This culture change will also influence the behavior of citizens in other states in which changes in policy are occurring, even if these states have not implemented a change as drastic as W-2.

Finally, the economy of the state in 1998 and after may not be so robust as it is now. Because of W-2, the low-wage labor market will experience an increase in labor supply (and hence downward wage pressure), relative to conditions before W-2 was implemented.

Given these considerations, I conclude that:

A pre-post evaluation design offers the best prospects for securing a reliable assessment of W-2, not because of its inherent superiority, but because the requirements of the experimental and the comparison-site designs and the constraints imposed by the policy environment severely limit their feasibility.

What steps, then, would be necessary to implement a pre-post evaluation design for W-2?

First, as soon as possible, a longitudinal household survey should be developed and implemented to elicit information on citizens' behavior and well-being under the existing welfare system and, later, under W-2. The survey should be at least annual and should be fielded for, say, 6–8 years, in order to obtain reliable information over time on the changes in well-being and behavior that are the object of the reforms. The variables included in this survey would emphasize the employment and work activities of the adults in the family, family income sources, information on children and their well-being, location, family structure, housing arrangements, and health and health care for family members.

The survey should concentrate its sample on "low-permanent-income" families (perhaps, families in the bottom quintile of the permanent income distribution).[9] Such a sample would include all current welfare recipients as well as virtually all those in the state who might, in the future, apply for support and be affected by labor market developments caused by the implementation of W-2. This suggested procedure assumes that observing these families from the present (or the earliest date for fielding a survey) to a date at which W-2 is judged to be "implemented" would yield enough reliable information to form a picture of the prereform (or counterfactual) levels of the variables of interest, L^p.[10] Observing these same families at a point in time after which W-2 has been fully implemented can provide an accurate picture of their postreform behavior and well-being, L^n.

In the survey, Milwaukee County should be heavily oversampled, because the circumstances addressed by the 1996 legislation are largely those present in urban areas with high numbers of minority, inner-city welfare recipients. It also seems likely that the implementation of W-2 will be slower in Milwaukee than in other parts of the state; hence, the possibility of securing reliable prereform information on well-being and behavior may be the greatest there.

This discussion has rested upon a basic assumption, that the evaluation of W-2 should focus on its effects on the well-being—as measured by income, work, and family structure—of low-income families in Wisconsin. There are, however, other important questions about the effect of W-2 on Wisconsin citizens. These include, for instance, the effect of the policy change on the access of citizens to other sources of income (especially nonem-

ployment income), or on the provision and availability of public and private services. These are important elements of the economic and community environment in which low-income citizens live. The possible effects of W-2 on many program services may better be captured by program and administrative data than by a survey. Prominent among these services are the availability, quality, and price of child care services; foster care placements throughout the state; the availability of family planning and abortion services (see Maria Cancian and Barbara Wolfe, "Outcomes of Interest," in this issue). Also of concern are the incidence of eviction from public or subsidized housing due to families' inability to meet rental payments; the adequacy of public and private emergency housing stocks to meet greater need; the prevalence of nutritional deficiencies and behavior problems in the local schools; and the ability of private food pantries to meet increased demands upon their services.

These considerations suggest the importance of designing a pre-post evaluation in conjunction with a strategy for collecting administrative and aggregate socioeconomic data (see Kaplan and Meyer, "Toward a Basic Impact Evaluation of Wisconsin Works," in this issue). The nature of the findings to be expected from the two coordinated, but independent, evaluation efforts should be clearly delineated at the outset to avoid duplication of effort in gathering pre-post information.

Some final thoughts

In this discussion, I have not considered the alternative of a combined comparison-site/pre-post design. Such a design would make possible a "difference within differences" analysis framework which, by exploiting the effects of the policy change both across time and across sites, could provide additional observations and reliability.[11] Designing such an evaluation would be exceedingly complex and costly; moreover, it would not eliminate the need to obtain statistical control for the effect of changes in demographic, social, and economic factors unrelated to the policy change. The most concrete example of this problem is the need to adjust pre-post results from the household survey and from administrative and aggregate information for changes in the underlying state of the economy. The relevant questions here take the form: "How would the work status (income, child care needs, etc.) of the individuals in the household survey have changed from the earlier to the later period, if underlying demographic, economic, and social conditions had changed, but no policy change had been implemented?" If these kinds of questions can be answered, evaluators can adjust the observed changes in the variables of interest to take account of changes extrinsic to the reforms.

Except for the studies presented at the November conference, little thought has been given to this issue, although there have been major recent advances in statistical modeling of the determinants of changes in time-series data. Given the importance of the question, and the nature of the policy changes at issue, it seems worthwhile to mount the research efforts necessary to develop reliable statistical models for forecasting without-policy changes that would assist in evaluating state welfare reform.■

---

[1]The paper upon which this article is based appears in full in "Evaluating Comprehensive State Welfare Reforms: A Conference," IRP Special Report no. 69, University of Wisconsin–Madison, March 1997.

[2]Prior to the new law, each state had a support system consisting of a state-specific AFDC program that provided income support to primarily single-parent families and that had to meet a detailed set of federal requirements and specifications; a federal Food Stamp program that provided food-based assistance on a national, uniform basis to families supported by the AFDC program in a state and to other low-income families; and a state-specific Medicaid program that provided health care support to AFDC-supported families plus other low-income families that met certain criteria, and that also had to meet a detailed set of federal requirements and specifications. This prereform set of programs varied by state in benefit levels and accessibility but retained a semblance of a coherent national system through the uniform national Food Stamp program and the uniform federal requirements for the AFDC and Medicaid programs.

[3]Assessing the national effects of the policy change may well be more feasible than assessing the effects of a particular state's new law. National data sets, such as the Survey of Income and Program Participation (SIPP), the National Longitudinal Survey (NLS), and the Current Population Survey (CPS), provide the basis for assessing the changes that the legislation has had over time on the populations of interest. Moreover, the Bureau of the Census has designed a nationally representative, longitudinal survey, the Survey of Program Dynamics (it is described by Daniel Weinberg in this issue). The Urban Institute project, "Evaluating the New Federalism," intends to make use of some of these databases in its impact evaluation of the reform, and to undertake special surveys in a selection of states.

[4]A state-based evaluation avoids some of the problems that would confront an evaluator attempting to measure national effects but creates others. Perhaps the largest is that created by the potential migration of citizens in response to the policy change. In discussing the ideal evaluation, I ignore this difficulty.

[5]The loss of comparability associated with a statistical match—as opposed to random assignment of a state's permanently poor residents to old and new policy regimes—is a serious limitation on the reliability of results based on the comparison-site design.

[6]In principle, it might be possible to establish a within-state comparison-group design. This would require that some part of the state maintain the prior policy regime, and consciously administer that regime as if no state policy change had occurred. (For example, it has been suggested that, in Wisconsin, counties that border Minnesota and that are dominated by Minnesota newspaper, radio, and television media should be administratively mandated to maintain the pre-W-2 AFDC program.) It seems unlikely that a segment of a state could, in fact, be kept immune from a state-based policy change with implications as fundamental and far-reaching as those required by the federal legislation.

[7]In evaluating state changes in welfare, policy analysts seeking statistical controls for the effects of changed social and economic conditions on variables of interest confront the same challenge faced by those who have attempted to model or "forecast" change over time in state welfare caseloads: How will the underlying changes in the environment affect the number of welfare recipients, as reflected in entry and exit rates of welfare programs? The forecasting success of such studies is decidedly mixed. See U.S. Congressional Budget

Office, *Forecasting AFDC Caseloads, with an Emphasis on Economic Factors* (Washington, D.C.: Congressional Budget Office, 1993); Steven Garasky, "Analyzing the Effect of Massachusetts' ET Choices Program on the State's AFDC-Basic Caseload," *Evaluation Review* (December 1990): 701–10; and Robert Plotnick and Russell M. Lidman, "Forecasting Welfare Caseloads: A Tool to Improve Budgeting," *Public Budgeting and Finance* (Autumn 1987): 70–81.

[8]For example, the Pay for Performance program (see this issue, p. 2) is in effect in several counties in Wisconsin, and rather large numbers of families have been sanctioned. This program has more in common with W-2 than it does with the conventional, pre-policy-change AFDC program.

[9]The New Jersey Income-Maintenance Experiment used such a "low-permanent-income" criterion as the basis for selecting the sample for this experiment, which may offer a guide in defining the sample for a W-2 evaluation. See the three-volume report, *The New Jersey Income-Maintenance Experiment*, Vol. 1, *Operation, Surveys, and Administration*, ed. D. Kershaw and J. Fair; Vol. 2, *Labor-Supply Responses*, ed. H. W. Watts and A. Rees; Vol. 3, *Expenditures, Health, and Social Behavior; and the Quality of the Evidence*, ed H. W. Watts and A. Rees (New York: Academic Press, 1976, 1977).

[10]Such a survey may not, in fact, be able to yield an accurate picture of behavior and well-being under the prior policy, AFDC, because changes designed to move AFDC toward W-2 are already being undertaken. Irrespective of the evaluation design, the evaluator needs a reliable estimate of well-being and behavior of the low-permanent-income population before welfare reform is implemented. To the extent that what can be observed in the state is not prereform AFDC but some combination of AFDC and elements of W-2, neither an intrastate experimental design nor a pre-post design will be able to measure the impact of the policy change. A comparison-site design may be able to measure the impact but only in the somewhat unlikely circumstances that a comparable non-Wisconsin site is available.

[11]For an explanation of the "difference within differences" method, see Cain, "Controlled Experiments," in this issue.

Postdoctoral fellowships, University of Michigan

The University of Michigan's Research and Training Program on Poverty, the Underclass, and Public Policy offers one- and two-year postdoctoral fellowships to American minority scholars in all the social sciences. Fellows will conduct their own research and participate in a year-long seminar under the direction of Sheldon Danziger, Professor of Social Work and Public Policy, and Mary Corcoran, Professor of Political Science, Public Policy and Social Work. Funds are provided by the Ford Foundation. Applicants must have completed their Ph.D. by August 1, 1998. Application deadline is January 13, 1998. Contact: Program on Poverty, the Underclass, and Public Policy, 540 E. Liberty, Suite 202, University of Michigan, Ann Arbor, MI 48104.

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
Fax (608) 265-3119

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 Jan Blakeslee.

A Survey of Program Dynamics for assessing welfare reform

Daniel H. Weinberg

Daniel H. Weinberg is Chief of the Housing and Household Economic Statistics Division, U.S. Census Bureau.

The Personal Responsibility and Work Opportunity Reconciliation Act of 1996 (PRWORA), enacted as P.L. 104-193, is a comprehensive piece of legislation with far-reaching implications for many programs. In particular, though, the law eliminates the open-ended federal entitlement program of Aid to Families with Dependent Children and creates a new program called Temporary Assistance for Needy Families, which provides block grants for states to offer limited cash assistance.

Specific program evaluation needs can be served by a series of focused single-purpose surveys or experiments. But if the research community were to rely solely on highly focused data collection, there would inevitably be major gaps. Although single-purpose approaches to data collection are useful, an omnibus data collection vehicle can provide the basis for an overall evaluation of how well welfare reforms are achieving the aims of the administration and the Congress, and meeting the needs of the American people. This requires a survey that casts a wide net, one that simultaneously measures important features of (1) the full range of welfare programs, including both programs that are being reformed and those that are unchanged, and (2) the full range of other important social, economic, demographic, and family changes that will affect the effectiveness of the reforms. Further, such a survey should be in place before the reforms are enacted, to allow adequate assessment of baseline circumstances.

Section 414 of PRWORA specifically directs (and funds) the Bureau of the Census to:

> continue to collect data on the 1992 and 1993 panels of the Survey of Income and Program Participation [SIPP] as necessary to obtain such information as will enable interested persons to evaluate the impact of the amendments made by Title I of the Personal Responsibility and Work Opportunity Reconciliation Act of 1996 on a random national sample of recipients of assistance under State programs funded under this part and (as appropriate) other low income families, and in doing so, shall pay particular attention to the issues of out-of-wedlock birth, welfare dependency, the beginning and end of welfare spells, and shall obtain information about the status of children participating in such panels.

To implement this directive, the Bureau established a team to carry out this survey effort, which we call the Survey of Program Dynamics (SPD).

As directed by the legislation, data already collected in the 1992 and 1993 SIPP panels will provide extensive baseline (background) information from which to determine the effects of welfare reform. SIPP is a longitudinal survey of households, each of which was interviewed at least nine times, at four-month intervals, and followed if they moved. The SIPP collects more detailed data than any other national survey regarding program eligibility, access and participation, transfer income, and in-kind benefits. Regarding economic and demographic data, the 1992 and 1993 panels collected very detailed data on employment and job transitions, income, and family composition. By interviewing the same households in the SPD, analysts would then have data for the baseline prereform period, the reform implementation period, and the medium-term postreform period. These data are required to assess short-term and medium-term consequences and outcomes for families and individuals. The use of both panels will also double the size of certain groups of interest, subject of course to our ability to recontact households in the two panels and their willingness to participate. Because the funding provided is not sufficient to interview all households in both panels past 1997, we will have to subsample after 1997.

The Census Bureau has also worked closely with policy agencies to develop and field topical modules that enhance the value of the basic SIPP data. Modules of special interest here include those on (1) education and training, (2) marital, fertility, and migration histories, (3) family relationships within the home, (4) work schedules, child care, child support, and support for non–household members, (5) medical expenses and utilization of health care services, and (6) child well-being.

Current plans are for data to be collected for each of the six years from 1996 through 2001, providing panel data for ten years (1992–2001) when combined with the 1992 SIPP data. Our original plans were to have an instrument ready to field concurrent with welfare reform. Because the legislation was vetoed twice during 1995, plans were put on hold. We were unable to pretest the SPD questionnaire and could not field the survey we had designed in 1997. Nevertheless, we felt it critical to fill the data gap between the end of the SIPP observations and the start of

the basic SPD observations. Consequently, the survey has been designed with three fundamental sections: (1) the "bridge" survey which will provide the link between the 1992 and 1993 panels of the SIPP and SPD; (2) the 1998 SPD which will use the core instrument already developed to collect annual retrospective data starting in 1998; and (3) the 1999 SPD Child Well-Being Module, to be administered starting in 1999, though its content may vary from year to year.

SPD "bridge" survey. It is very important to collect income and program participation data in spring 1997 for calendar year 1996 from as many of the 1992 and 1993 SIPP households as we can find. Data for 1996 will be collected in April–June 1997 by administering a modified version of the annual March 1997 Current Population Survey (CPS) demographic supplement, with a few new questions designed to collect summary 1995 data for the 1992 SIPP panel (who were last interviewed in January 1995). Finding people who move is critical to the success of any longitudinal survey, particularly one as focused on the low-income population as SPD. This is particularly crucial to the SPD, given the time that has elapsed since the last interview. We will also be testing the use of a $20 monetary incentive for low-income households in an attempt to reduce nonresponse to the bridge survey; the Census Bureau demonstrated that such an incentive was successful in reducing nonresponse to wave 1 of the 1996 SIPP panel.

1997 pretest of the 1998 SPD. We are using 1997 to complete work on the 1998 SPD questionnaire. The University of California at Berkeley is authoring this instrument for a computer-assisted personal interviewing (CAPI) environment; UC–Berkeley is the developer of the CASES authoring language that is used for most computer-assisted Census Bureau surveys. We plan a pretest for September, using about 500 retired 1996 CPS households in four of our regional office locations. From this test we will have a good idea of how well the instrument does in an operational environment. We also will test the use of a Self-Administered Adolescent Questionnaire using audio cassettes to obtain information from youths 12 to 17 years old.

1998 SPD. Using the fully developed CAPI instrument, data will be collected once each year in March–May, with annual recall for the preceding calendar year. There will be a set of retrospective questions for all persons aged 15 and older that focuses on topics such as jobs, income, and program participation. Also included will be a section focusing on children in the household; it addresses topics such as school status, activities at home, child care, health care, and child support. Average interview length would be about one hour per household. The adolescent survey (still being developed) will probably include issues such as family conflict, vocational goals, educational aspirations, and crime-related violence.

1999 SPD and later. Work has begun on identifying the topics for the 1999 child supplement. We will focus in particular on elements of child assessment using clinically tested assessment scales.

There are four technical issues we still need to resolve:

1. Subsampling. It is clear we cannot interview all households in the 1992 and 1993 SIPP panels in 1998. How much we will subsample depends on the response rate to the 1997 SPD bridge survey. Further, we have yet not decided which groups should be overrepresented, but we expect to use the welfare reform law as a guide.

2. Weighting. Current thinking involves weighting the interviewed population to represent the April 1997 U.S. population, releasing a cross-section file, and also developing a longitudinal file weighted to represent the 1993 population (the basis for the SIPP weights). Differential attrition of the low-income population is a serious concern in developing appropriate weights.

3. Database and product development. Users of the longitudinal data will have a hard time figuring out how to use data from three separate surveys (SIPP, CPS, and SPD) simultaneously in a longitudinal analysis. The challenge is to create a longitudinal data set with annual data from the SIPP and the CPS in a format consistent with the way data will be provided from the 1998 SPD, so that users can develop familiarity with the data and be ready for the first wave of SPD.

4. Supplementary data. We have contracted with the University of Wisconsin to create a complementary data base of state and county welfare program characteristics that we could then match to the SPD data. (County-level matches would have to remain confidential and researchers would have to work on that matched data set at the Census Bureau to maintain respondents' confidentiality.) We may also be able to match other administrative data provided to us electronically (e.g., tax returns, welfare program records) to the survey data; again, these would need to be accessed only at secure Census Bureau locations.

For further information, contact the SPD Team Leader, Patricia Johnson, at (301) 763-8199.■

Outcomes of interest, evaluation constituencies, and the necessary trade-offs

Maria Cancian and Barbara Wolfe

Maria Cancian is an Assistant Professor in the LaFollette Institute of Public Affairs and School of Social Work, University of Wisconsin–Madison, and an IRP Affiliate. Barbara Wolfe is Professor of Economics and Preventive Medicine at the University of Wisconsin–Madison, and the Director of IRP.

How will we know if the new generation of state-based, comprehensive welfare reform is a success? This paper suggests why we need broadly focused evaluations and begins to identify the outcomes of interest and the different constituencies for evaluations of the new welfare programs. It explores trade-offs that evaluators will face, and outlines their implications for some representative outcomes. We use the Wisconsin Works (W-2) plan as a point of reference because it is substantially more developed than are the plans of most states, but we expect that most of the issues raised will also apply to other states.[1]

There are many reasons that evaluators should look beyond the effects of the new programs on participants and their families only.

First, under the block-grant policies put in place by the Personal Responsibility and Work Opportunity Reconciliation Act of 1996, states will have more freedom to design programs for needy families and will more fully absorb the financial consequences of their decisions. These changes raise a host of issues regarding the *fiscal and administrative responsibilities* of federal, state, and local governments that evaluators will have to consider.

Second, under the new policies there is no individual entitlement. States are free to determine which families receive assistance and under what circumstances. This raises questions of *accessibility and intrastate equity*. There is no longer an assumption that state residents are all guaranteed equal benefits, given equal income and family size.

Third, under the new legislation, program structure and eligibility standards will vary even more substantially than they did under Aid to Families with Dependent Children (AFDC), also raising issues of *interstate equity*. And because the five-year lifetime cash benefit limit prevails across states, participants in a low-benefit state not only will receive lower benefits, but also will reduce their opportunity to receive benefits in a higher-benefit state at a later time.

Fourth, the scale and timing of state reform efforts may have large effects on state *labor markets and social institutions*. In Wisconsin, for example, the vast majority of current AFDC participants are expected to be in community service jobs or private-sector jobs, subsidized or unsubsidized, by September 1997. Some may not succeed. What will be the consequences for the low-wage labor market, for the child care market, and for public and private organizations that provide housing or social services?

Finally, the new state welfare programs emphasize work, support by both parents of their children, provision of requested or needed services only, and reliance on market and performance mechanisms in implementing programs. Proponents of Temporary Assistance to Needy Families (TANF) and W-2 have argued that welfare reform will encourage favorable *behavioral changes*. If participants do make substantial behavioral changes, what will be the consequences for them, and for the community at large? If such changes are not evident, should elements of the reform be reconsidered?

The constituencies for an evaluation and the consequent trade-offs

Different constituencies have different questions, and hence are likely to be interested in outcomes based on widely differing units of analysis. For example, analysts of welfare reform are interested in the effects on the population formerly eligible for AFDC, on all low-income persons, and on the general community. Program participants have an interest in the effects on their quality of life. The fiscal and administrative aspects of welfare reform do not directly speak to family effects, but are clearly of interest to citizens and to government officeholders. Program administrators are likely to be interested in outcomes that capture the consequences of the decisions they make at the local level. And other constituencies have direct and indirect interests in the outcomes of reforms: private charities, employers, schools and teachers, health care providers. In assessing data collection priorities, therefore, it may be useful to recognize a number of trade-offs.

Trade-off 1: The populations of interest

Past (and future) families eligible for welfare form only a small part of a state's total population, and very few of them appear in existing, representative national or state data sets. For those interested in the consequences of reform for the eligible or potentially eligible population, then, it would be preferable to generate a special survey

sample or to substantially oversample current welfare users, with some addition of low-income families and young teens at high risk of pregnancy.

The sampling issue is even more critical for the study of smaller populations—for example, families with infants or with disabled children, or poor immigrants. For these groups, a household survey will include very few observations. Yet it is necessary to sample all low-income households, participant and nonparticipant, to learn what proportion of the entire population uses any services, how the labor market performs, what unmet needs may exist, and especially what changes may be occurring in all these areas. Hence the direct trade-off: the smaller the proportion of any group of interest (e.g., disabled children), the greater the need to oversample it, the higher the unit cost of sampling, and the lower the value of the oversampled cases to the entire analysis.

Trade-off 2: Individual and community-based analysis

In different communities, the underlying conditions are likely to differ in ways that may significantly influence the effects of welfare reform; such conditions include, for instance, unemployment, types of firms, housing stocks, and child care and health care services. Communities may also allocate different resources to programs and implement and administer them differently. Thus an evaluation that considers community conditions and program components is likely to be better and more comprehensive. What, for example, are the effects in a community in which there is nearly full employment? in rural areas or small towns? Because collecting these data is expensive, not all communities would be included in such a survey. Yet to determine statewide effects—for example, whether the low-income population in Wisconsin is better off as a result of W-2—we need statewide representative data. And concentrating data collection in a few communities clearly limits our ability to analyze the impact of community differences on the population(s) of interest.[2] Thus there is a trade-off between collecting data that can better inform us about implementation and impact in a particular locality and data representative of the state.

Trade-off 3: Evaluation, cross-community comparisons, and monitoring

To *evaluate* the effects of a program, we need measures of a counterfactual—what the outcomes would have been if the program had not changed. If, for example, we can measure outcomes before and after the reforms by means of administrative or other data, we will have some ability to evaluate welfare reform. But what if such data do not exist for potentially significant outcomes? In those cases, comparing sites or monitoring outcomes can tell us about program implementation and effectiveness and indicate areas where policies may require further change—and such information is very important for a variety of constituencies. Hence the trade-off: even though evaluation remains our primary goal, we may also wish to collect information on outcomes that cannot, strictly speaking, be evaluated.

Trade-off 4: Short-term and long-term outcomes

Some outcomes, such as labor market and income measures, will be useful immediately. Some, such as the impact of reduced parental time on young children, may require many years for a full assessment. Moreover, some outcomes may differ in the short and longer run. For example, sanctions may be immediately harmful to a family, but we need to know interim and long-range effects. Outcomes may also change over time, because of changes either in the program itself or in external circumstances, such as the state of the economy. In allocating resources, a trade-off exists between funding an immediate evaluation and setting aside resources to collect data for a more comprehensive one in the future.

Outcomes of interest

To demonstrate these relations among outcomes and trade-offs, we briefly examine a representative set of outcomes: (1) work requirements, (2) child care, (3) child health, and (4) family formation (Table 1 contains a comprehensive list).

The effects of work requirements

Labor market outcomes include, using W-2 as an example: (1) the distribution of participants across four job tiers (see this issue, p. 2); (2) the probability and timing of movements toward (or away from) unsubsidized employment and the stability of such employment; (3) earnings, benefits, and work-related expenses; (4) the administrative efficiency and cost of subsidized employment; (5) changes in wages or job availability for low-wage workers not participating in W-2; and (6) the availability, productivity, and cost of labor for firms.

The populations of interest. We must know how these primary outcomes differ for individual participants but also across participant groups in order to assess the adequacy of the program's evaluation, placement, and support services. The new work requirements for mothers of young infants suggest that we may also wish to pay close attention to this relatively small group. But the timing and scale of work requirements suggest that there may be effects on the general labor market, especially for low-wage workers, that require evaluators to consider the trade-off between collecting a broad sample and oversampling particular groups.

Individual and community-based analysis. Some work outcomes such as total earnings and benefits are of interest statewide and for individuals. But for many outcomes, particularly those relating to work requirements and planning and support services, both individual and community-level data are important. For example, differ-

Table 1
The Outcomes of Interest

For each outcome, it is necessary to consider trade-offs among the populations of interest, between individual and community-based data collection strategies, between evaluation and monitoring, and between short-term and long-term outcomes.

I. Well-Being of Participants
  A. Economic
    1. Work experience
    2. Earnings/compensation
    3. Other sources of income & benefits in kind
    4. Total income relative to needs
    5. Dependency: transfers as a proportion of income
  B. Noneconomic
    1. Health status
    2. Education/training
    3. Stability of interpersonal ties

II. Well-Being of Low-Income Families
  A. Child well-being
    1. Health
    2. Education
    3. Child care
    4. Nutrition
    5. Maltreatment
  B. Family structure/formation
    1. Family structure
    2. Time with parents
    3. Paternity establishment
    4. Birth rate; out-of-wedlock birth rate by age group
  C. Family Human Capital
    1. Parents' education
    2. Parental investments in children
  D. Housing
    1. Homelessness
    2. Density
    3. Geographical mobility

III. Labor Market Outcomes
  A. Firm
    1. Vacancies
    2. Net cost per low-skilled employee
    3. Worker productivity
  B. Other low-wage workers
    1. Unemployment and underemployment
    2. Earnings/compensation

IV. Externalities
  A. Crime
  B. School quality
  C. Child care
  D. Health care
  E. Migration
  F. Supply of auxiliary service providers

V. Fiscal Effects
  A. Total costs
  B. Allocation between federal, state, and local

ences in the type of initial job placement and the probability and timing of transitions between tiers are likely to be influenced by local labor market conditions.

The feasibility of evaluation, cross-community comparisons, and monitoring. In Wisconsin, we can trace earnings by linking administrative data from AFDC, W-2, and unemployment insurance (UI), by county, before and after implementation of W-2. We might then use a pre-post design to evaluate the effects of W-2 on, say, earnings and employment stability among participants and other workers. But evaluating the effect on the eligible population as a whole will be more difficult, owing to the absence of information on individuals who may be discouraged from applying for W-2 services—perhaps, particularly, younger persons just finishing their schooling.

Short-term and long-term outcomes. W-2 requires a large and immediate transition to work. To assess the feasibility of this approach and the possibility of administrative and labor market incompatibilities, we should, for example, examine the extent to which wages and earnings increase over time and work experience alone leads to improved economic status, particularly if the economy should falter. This is especially important in light of the retreat from long-term training and education. The employment patterns of W-2 participants with a long history of AFDC receipt and of new entrants are expected to be substantially different, further suggesting the importance of long-term evaluation.

The effects of child care

Work requirements under the new welfare programs will increase demand for child care and the time children spend in care. In Wisconsin, W-2 changes eligibility for state-subsidized child care and the cost of care to participants. To increase the numbers of providers, W-2 introduces a "provisionally certified" category (see this issue, p. 2), with less stringent licensing requirements. This, plus haste to increase the supply of regulated providers, may result in lower-quality child care for participants and nonparticipants alike. If programs require copayments that vary with the cost of care, parents will have an incentive to place children in lower-cost, possibly lower-quality care.

The populations of interest. Child care has diverse constituencies, including participant families, the general population of working parents, and those concerned with effects on children's well-being, on W-2 employment transitions and costs, and on the working conditions and wages of child care providers. The outcomes of interest include the availability and quality of child care for all families with children, for W-2 participants themselves, and for selected groups, such as children with special needs and very young children. As with work requirements, evaluators will face a trade-off between collecting a statewide representative sample of households to assess the impact on child care among all families or oversampling particular groups.

Individual and community-based analysis. Individual data will be needed to evaluate the effect of care on child well-being, measured by such indicators as total time in care, ratio of providers to children, stability, use of informal care, and the probability that older children are left without supervision. The availability, cost, and convenience of child care bear upon the ability of parents to work and to meet their expenses. Community-level data

will be needed to evaluate such outcomes as the supply of child care by type, or the success of public and private agencies, employers, and community organizations in coordinating child care services.

The feasibility of evaluation, cross-community comparisons, and monitoring. In Wisconsin, there are reliable baseline data on the regulated child care market in each county, and administrative data on the use and cost of subsidized child care among low-income parents. A pre-post design might be used to evaluate the impact of W-2 on the regulated care, and cross-county comparisons could also be made. The absence of systematic preimplementation data for unregulated care and of state baseline data on individual outcomes limits our ability to do an impact evaluation. Those outcomes, however, may well be sufficiently important to warrant collecting data to monitor effects.

Short-term and long-term outcomes. An adequate supply of child care in the short term is important, especially given the focus in W-2 on timely transitions to work. Information about short-term outcomes is needed to determine if there should be changes in child care administration, coordination, and the new category of providers. However, many outcomes are long term in nature—the effects on child well-being, school readiness, and later life, for example. Other outcomes may change over time; for example, the number and quality of child care providers or child care workers' wages may change in response to the increased demand for care. The evaluation of child care will, therefore, involve outcomes with a variety of time frames.

Effects on child health

Some factors may influence health directly, and some measures of health may capture other changes; we look at both here. The outcomes of interest to different constituencies include child health status, nutrition, and ability to participate in school; the availability of parents to monitor their children's health, take them to the doctor, and provide support to children with special needs; changing demand for health services; and access to and costs of regular and emergency care.

The populations of interest. Children in low-income families are more likely to be exposed to environmental hazards (such as lead paint), to injury or violence, to poor nutrition, and to emotional stress. The work and time demands on parents in families participating in W-2 may reduce provision of preventive medical care or healthy food for their children; sick children may be sent to child care or school, thus exposing other children to illness. Some groups, such as, infants and children with disabilities, may be especially vulnerable to the impact of new work requirements on their mothers. The health status of children in W-2 families should therefore be monitored, and this requires individual data. Measuring the effects of W-2 on the population at large requires statewide data on the proportion of persons by groups (especially children) who have health insurance coverage, the proportion who report that their children have excellent, good, fair, or poor health, the proportion of children aged 1–4 who are fully vaccinated, the proportion of low-weight births, and accident and injury reports from hospitals or police.

Individual and community-based analysis. The health of children can be measured at the individual level by anthropometric indicators, by surveys of parents, and by medical records, including Medicaid records, if they can be made available.[3] Community-level data can provide information on supply, such as the ratio of providers relative to the population, and the use of paraprofessionals, counselors, and other support personnel by provider groups (clinics). Other community-level outcomes include changes in the demands made upon health care providers, such as higher need for emergency care or changes in hours of operation, and the providers' responses to those changes. These outcomes are likely to affect utilization, if not health itself.

The feasibility of evaluation, cross-community comparisons, and monitoring. Most of these measures of children's health could be collected for children in the pre-W-2 period by using medical records, vital statistics, or existing surveys. Measures such as days of school missed or the proportion of children with a diagnosed physical, mental, or learning disability may be obtainable from administrative or school records for the preimplementation period. Birth records, hospital records, and anthropometric data are expensive to collect; yet even if we lack the resources for a pre-post evaluation, certain health outcomes, such as emergency room use, are so important that they should be monitored.

Short-term and long-term outcomes. Short-term outcomes that will give us information on the immediate effects of W-2 on child health include changes in days of school or of child care missed, in the proportion of children in these settings who are sick, in the rate of injuries, and in the use of medical care in communities that serve a high proportion of W-2 families. Short- and longer-run outcomes such as the effects of increased stress on children could be gathered in surveys of parents, teachers, and child care workers. Only over a much longer time should we expect to see effects that reflect nutritional changes, which are indicated by anthropometric measures. Similarly, the proportion of children with no or incomplete immunizations may gradually change over time. And the proportion of the low-income population with health insurance may change as parents change employment tiers or eligibility for Medicaid is lost.

Effects on family formation

One principle behind welfare reform is to encourage parents to be responsible for their children. Another principle, explicit in the federal reform bill, is to reduce pregnancies and births among unmarried women, especially teenagers. The national legislation requires teen-

age mothers to live at home or with a responsible adult in order to receive any benefits. Other program changes in state welfare reforms and under TANF may bring unpredictable change to family-formation incentives.

Outcomes of interest with regard to family formation include living arrangements: whether married or not, whether cohabiting in a stable relationship, whether living with one's parents or other adults; subsequent fertility, such as delayed or forgone births; and noncustodial parents' relations with their children.

The populations of interest. The population most directly affected includes single mothers with limited assets and income, teenagers who are potentially eligible for services, noncustodial parents, and selected subgroups of the general public. A primary variable of interest is the rate of out-of-wedlock fertility among the "at risk" teenage population—those who have already given birth as an unmarried teen and younger teens living in low-income communities. Statewide data are readily available over a long period of time for in- and out-of-wedlock fertility for the entire population, including teenagers, and for subpopulations of the state. These data would also serve as a type of control for changes in fertility rates that might be due to other events and changing circumstances within the state.

The greater involvement of nonresident parents, especially fathers, in their children's lives is considered a particularly important outcome. Measures of this might include support paid and time spent with the children. We would concentrate on the fathers of children who are eligible for W-2 but, if society is successful in increasing fathers' involvement, we might expect to find this success reflected in the broader population of single-parent families.

Individual and community-based analysis. We may wish to compare individual data on marital status, childbearing, and living arrangements of W-2 participants with data for past AFDC recipients. In communities, zip-code areas, or census tracts that have high rates of participation in welfare-related programs, we can compare rates of out-of-wedlock childbearing by age, over time. TANF includes substantial funding for abstinence education and financial incentives for states that reduce nonmarital childbearing. To evaluate their effectiveness, we need to collect indicators of educational and attitudinal changes at the community level.

The feasibility of evaluation, cross-community comparisons, and monitoring. Administrative data from birth certificates are the easiest data to collect and go well back in time. They contain information on mother's age, race, marital status, and previous live births. Residence data allow aggregation to the community level. Data on marriages, divorces, and annulments, recorded by age of the individual and by county, may also include information about children and custody. In Wisconsin, data from administrative records and surveys of parents are available from the Wisconsin Court Record Database (WCRD); they have some limitations, but might serve to construct a baseline.[4] Both the WCRD parent surveys and the Panel Study of Income Dynamics allow us to estimate the approximate amounts of time spent by nonresident parents with their children before W-2. To answer such questions after W-2, survey data would be necessary.

Short-term and long-term outcomes. We can observe short-term effects on teen fertility, marriage rates, divorce rates, and the parenting role of the noncustodial parent. But in all of the family-formation measures, the short-run response may be smaller than, and even in a different direction from, the long-run response. If, for example, the teenage nonmarital birth rate is reduced in the short run by targeted programs, both social mores and contraceptive knowledge may change, further reducing such births. If fathers allocate more time to their children, their relationship may become closer in the short run, but if financial obligations or children's demands are viewed as too great, time spent may decline in the longer run.

Conclusions

The resources for evaluation are clearly limited, and different constituencies would make different choices. We emphasize (1) those aspects that can be evaluated rather than simply monitored, although not to the absolute exclusion of important effects for which we have no prereform data; (2) those impacts that can be evaluated using administrative data, which will of necessity emphasize the target population rather than the overall population or near-eligibles; and (3) broad measures of well-being in the overall population that can be drawn from existing, large-scale datasets such as the Current Population Survey and the Wisconsin Family Health Survey. The major remaining choice in evaluation design is, we believe, between a panel survey of the low-income population of the entire state and a selective, community-based survey of that population. Because community differences are potentially important in assessing the availability of services, job opportunities, and other community-level factors, we believe that a selective panel survey is the better choice.■

---

[1]The paper upon which this article is based appears in full in "Evaluating Comprehensive State Welfare Reforms: A Conference," IRP Special Report no. 69, University of Wisconsin–Madison, March 1997.

[2]One approach that has been used in a number of studies, particularly in the health area (e.g., the Rand Health Insurance Study, the Epidemiological Catchment Area Study), has been to select communities as the first basis for data collection. The use of data based in communities allows study of the impact of services provided, the labor market, etc., on outcomes of interest but limits understanding of the broader consequences throughout the state. It is not clear if community studies can be pooled for analysis. If there are only a limited number of

Indicators of child well-being: An update

On April 21, 1997, President Clinton issued an Executive Order, "Protection of Children from Environmental Health Risks and Safety Risks." Noting that children suffer disproportionately from such risks, the order required that each federal agency make it a high priority to identify and assess environmental health and safety risks to children and to ensure that its policies, programs, and standards address such risks. It established a four-year, cabinet-level task force to recommend federal strategies for children's environmental health and safety and to prepare a biennial report on research, data, and other information that would enhance the government's ability to respond to risks to children.

The order also required that the Office of Management and Budget convene an interagency forum on child and family statistics. Its purpose is to produce an annual report on the most important indicators of child well-being in the United States. Specifically, the forum is to "determine the indicators to be included in each report and identify the sources of data to be used for each indicator. . . provide an ongoing review of Federal collection and dissemination of data on children and families, and . . . make recommendations to improve the coverage and coordination of data collection and to reduce duplication and overlap."

The president's Executive Order can be seen as the culmination of steadily growing interest in developing indicators of child well-being. Such interest has been a natural response to the discouraging statistics that have spelled out a rise in child poverty since the 1970s. IRP has taken a prominent part in the endeavor to create objective indicators that will obtain broad public and political acceptance. In the early 1990s, it began producing annual reports on the well-being of children in Wisconsin, as part of the KIDS COUNT project funded by the Annie E. Casey Foundation. Realization of the inconsistencies and inadequacies of existing data led to a systematic effort to improve the range and quality of statistical information on children's well-being. In November 1994, the Institute was a cosponsor of a conference on Indicators of Children's Well-Being that was convened to begin the process of creating a statistical system capable of monitoring the well-being of children over time and space and across social groups.[1] At about the same time, a federal Interagency Forum on Child and Family Statistics was created to foster coordination and integration of federal data collection and reporting.

In December 1995, a one-day workshop on State-Level Indicators of Children's Well-Being (IRP was again a sponsor) brought together over 50 members of state and federal agencies and others concerned about the well-being of children.[2] Spurred in part by the expansion of state responsibility for welfare policy and programs that had begun through federal waivers, the workshop was considered less an end in itself than an impetus toward subsequent activities. (The WELPAN network described in this issue of *Focus* had its origins in the December workshop.)

Another such activity is the twelve-state Project on State-Level Child Outcomes, which first met in November 1996. Sponsored by the Administration for Children and Families (ACF) in the federal Department of Health and Human Services, it brings together state officials from California, Connecticut, Florida, Illinois, Indiana, Iowa, Michigan, Minnesota, Ohio, Oregon, Vermont, and Virginia, ACF officials, and representatives of research and policy organizations such as Child Trends, IRP, and the Chapin Hall Center for Children, at the University of Chicago. Its goal is to establish a consistent terminology, set of child outcomes, and measurement tools that will allow states and program evaluators to find common ground as they review the changing practices of fifty states. The project is developing a conceptual model that will be both comprehensive and flexible enough to meet very different state needs and has begun to identify populations of interest and potentially useful survey tools.■

---

[1]Sponsors of the conference were IRP, Child Trends, Inc., the Office of the Assistant Secretary for Planning and Evaluation in the U.S. Department of Health and Human Services, the National Institute of Child Health and Human Development, and the Annie E. Casey Foundation. *Focus* 16, no. 3 (Spring 1995) reports upon the efforts to develop child indicators and summarizes the conference. The proceedings were issued as a three-volume IRP Special Report, *Indicators of Children's Well-Being* (SR 60A-C, May 1995). A volume based on the conference papers is being prepared for publication in fall 1997 by the Russell Sage Foundation.

[2]See *Focus* 18, no. 1 (special issue 1996) for a review of the workshop.

Cancian and Wolfe notes, continued

sites—say four to six sites—there would be too few degrees of freedom to characterize the system. And if data were pooled, we would need to have accurate weights on individuals in the community-based samples to represent the state. It is not at all clear what would serve as the basis of these weights.

[3]Existing surveys such as the Survey of Income and Program Participation and the Wisconsin Family Health Survey now collect some of these data, although these surveys do not contain a large sample of low-income children.

[4]The database contains records of divorce and paternity cases from 21 Wisconsin counties and has only a small sample of AFDC-eligible cases.

Controlled experiments in evaluating the new welfare programs

Glen G. Cain

Glen G. Cain is Emeritus Professor of Economics, University of Wisconsin–Madison, and an IRP Affiliate.

This article reaches two main conclusions.[1] The first is that a controlled experiment is not a practical method for evaluating Wisconsin's new welfare system, Wisconsin Works (W-2), and is unlikely to be practical in any other state. The second is that controlled experiments can be used to evaluate component parts of the new welfare programs, and that by analyzing the experimental method we can better choose alternative methods for the majority of cases in which such experiments are not feasible.[2]

The question of whether any new system is working well must be posed in conjunction with a second question: Compared to what? In most evaluation studies that use controlled experiments, the alternative to the experimental program—the counterfactual—is the status quo. I doubt, however, that state officials view a return to the old system as an alternative to the new system even if the 1996 federal welfare legislation would permit them to. Instead, comparisons between or among states are more relevant to the issue of improving the new welfare system in a particular state. What would the alternative welfare program be in a controlled experiment, if the old system is ruled out?[3] The question has received no attention by state government officials because none (to my knowledge) has proposed using a controlled experiment.[4]

Below, I illustrate the potential power of the controlled experiment method and the standards it sets for achieving a good evaluation, discuss specific problems in using controlled experiments to evaluate the new state welfare programs, and review the problems in evaluating a state's new welfare program with cross-state observational data.

The controlled experiment as the evaluation design that sets a standard

In its conventional application, the controlled experiment tests a new program to allow a comparison of its outcomes with the outcomes from the existing system. Consider experiments to determine whether a training program increases the employment, earnings, and incomes, and affects other outcomes for welfare mothers. Typically, a sample of welfare mothers, including new recipients, would be selected and randomly assigned to either a treatment group eligible to receive training or a control group that continues in the existing welfare system without access to the training program. The goal of the experiment is to measure its effects if the experimental program were made a permanent part of the welfare program, available to all.

Assume this experiment has the following three positive features. (1) The training program in the experiment is similar in content, quality, and implementation to what the program would be if it were extended to the entire population of welfare mothers. (2) The treatment and control families are followed for long enough to determine the post-training and long-run outcomes, using household interviews and other sources of data, such as administrative records. Both (1) and (2) are necessary to achieve unbiased estimates of the relevant outcomes of interest—the full (or long-run) benefits and costs of the permanent training program. (3) The proposed training program is voluntary. Those who choose not to participate or who drop out of the program remain in the treatment group for purposes of comparison with the control group. The voluntary nature of the training program removes a possible ethical objection, which is that mandatory training could make some of those randomly assigned to the experimental group worse off than the control group. (Note that if the permanent program were mandatory—requiring participation in the training program to receive welfare benefits—then the experiment, to be a true test, would also have to be mandatory.) This issue arises below in discussing controlled experiments for evaluating the new state welfare reforms.

This controlled experiment for evaluating the training program has two potentially important weaknesses. One is the possibility that, if the program were made a permanent part of the welfare system, the population of new entrants would change. If the training program were notably rewarding, some women who would not have entered the old welfare system might decide to enter the new system. If it were mandatory and onerous, some women who would have entered the old welfare system might not enter the new system. The experiment would not detect these outcomes, because the training was not imposed on all new entrants to the welfare program.

A second weakness is that a full-scale training program, unlike a relatively small experimental program, is likely to have an adverse effect on the low-wage labor market by substantially increasing the supply of low-wage work-

ers. Most low-wage workers are not on welfare, and the experiment could not determine the effects for this population of a full-scale program.

Neither weakness implies that the experiment would be futile. It can answer some, but not all, important questions. It offers persuasive evidence of the ability of a replicable training program to affect the earnings capacities of welfare mothers. The experiment's central questions and its intended policy decision are well defined. How do the outcomes from the training program compare with those of the existing system (a fundamentally scientific question)? Should the existing system be changed to include the training program (a fundamentally political question)? The causes of the measured change in outcomes may be assessed, and information needed for a benefit-cost analysis of the training program collected.

If the experiment were to contain three components—training, child care, and health insurance—then all the outcomes being studied should, in general, be attributable to the package of the three components, but it would be difficult to assign the causal effect of any particular outcome to any one component. Such individual attribution of causality is unnecessary if the full-scale, permanent program matches the experimental program.

Evaluating a complete change in a program by a controlled experiment: A pessimistic appraisal

In our hypothetical training program, a clear policy decision was at issue: whether to adopt the new training program or keep the old. There is no comparable decision at stake in evaluating a new state welfare system; the old system was part of a national program that has been repealed and replaced.

There are valid scientific reasons to wish to compare the social impact of the new program with the old, but there are also technical and logistic problems facing a controlled experiment in which the old system is the treatment program and the new system the control.

The first, and least serious, is that using a controlled experiment to evaluate the new welfare reform may raise ethical objections, if the random assignment to the treatment group would make the families worse off than participating in the new welfare program.

Second, the experiment would have to be relatively long, say five or six years, to represent behavior that matches the behavior of those who would be in the old welfare program if that program had continued or if it were reestablished. Long-duration experiments are more difficult to carry out and more expensive than short-duration experiments. A time period of five to six years for a controlled experiment in which the old program is the treatment group is a drawback that is probably exacerbated by the fact that neither policy makers nor participants seriously consider the old program as a permanent alternative to the new program.

Third, perhaps two years or so may elapse before the new welfare system settles into its normal, long-run operational status. This is a challenge to any evaluation strategy, but the problem seems more serious if the evaluation is a controlled experiment that compares the old and new systems. An evaluation of programs such as the Wisconsin reforms based on an observational study of other state programs has many problems of design (see below), but it does compare programs that are in similar stages of development.

Fourth, in Wisconsin, components of W-2 have been under way for over a year. Assigning a treatment group of families to the old system will involve families that are already altered ("contaminated") by exposure to the new program. Is there a sufficiently large group of current welfare recipients who have been unaffected by the new programs and who, therefore, can be assigned to a treatment group in the old system and to a control group that remains in W-2?[5] (Some states have made no changes in their current welfare system, and others have made minor changes that have affected only small numbers of families.)

The fifth, and most serious problem with a controlled experiment that randomly assigns welfare families to the old welfare system is that the state's economic environment is expected to change under the new system, and the outcomes of the treatment families in the old welfare plan are unlikely to represent what they would be if the old system had been maintained statewide. Assume that under the old welfare system a substantial fraction, say 25 percent, of welfare mothers find a job that takes them off welfare. However, under the new welfare system, the pressure for welfare mothers to find jobs will be great, making employment more difficult and perhaps less rewarding for those who are in the treatment group (under the old system). Assume that only 10 percent, rather than 25 percent, of these treatment-group women find jobs and that 40 percent of welfare mothers in the new program become employed and leave welfare. The new system would thus seem to show a 40 percent success rate compared to a 10 percent success rate. But the new system's superior outcome is upwardly biased; the (hypothetically) true difference is 40 percent compared to 25 percent.

In summary, the obstacles to using controlled experiments to test new welfare programs against the old programs seem insurmountable. If the technical problems could be overcome, there is, indeed, intense scientific interest in the comparison between the two systems, but the old programs are not the relevant counterfactual be-

cause they are not realistic alternatives to the new programs. The best chance for a useful controlled experiment to test the new program in its entirety against the old program exists in states that (1) have new programs that are decidedly less generous than the old program, (2) can carry out an experiment for at least five years, (3) have new programs that are relatively simple and can be operating at their steady-state level quickly, and (4) can begin the experiment before the new program gets under way.[6] The value of controlled experiments that test some reform or component of an ongoing program that has reached a state of normal operation remains, and there will surely be ample opportunities for these experiments in the future.

Comparisons among state programs: An alternative evaluation strategy

In evaluating a state's new welfare program, the programs in other states appear to have more policy relevance—although less scientific interest—than the old system in the particular state. Cross-state evaluations will not produce a rigorous evaluation: they will not obtain unbiased estimates of the causal effects of even the state program as a whole, let alone isolate the causal effects of the various components of the program. Ballpark estimates are more likely. Nevertheless, the idea that the programs in other states offer alternatives, in whole or in part, to a given state's program is realistic and is in the spirit of the old tradition of viewing variation in state programs as "experimental laboratories." Here I summarize some pertinent aspects of cross-state evaluation.

A preferred strategy for gathering the data for the evaluation requires surveys across states before the new programs get under way (or soon after) and a subsequent longitudinal survey design. With these data, the analytic model commonly called "difference in differences" can be used.[7] Surveys, unlike administrative data, are uniform and easily comparable across states, cover a wider range of behavioral outcomes, and cover groups who are not participating in or do not appear in the records of the welfare system.[8]

The longitudinal surveys should continue over a relatively long period, say six years. Many outcomes of interest occur after several years elapse, such as the effects of the mandatory terminations of welfare after a period of up to five years. Other behavioral outcomes, like family composition and fertility, may take several years to appear. And a long-duration study avoids several biases that are likely to appear in the short run. The new programs will improve over time in their performance, as administrators learn what works. An opposite bias is that the program's performance will appear to decline because the administrators have every incentive to make the program look good in its first or second year by "skimming the cream." They will tend to find jobs first for those on welfare who are the easiest to place, and place their clients in jobs that are easiest to fill, such as low-level public jobs or jobs in subsidized nonprofit agencies.

Cross-state comparisons are natural, because the programs are legislated to vary across states. In contrast, if the evaluation design is based on within-state variation, these variations must be imposed as exceptions to the state's law. Or, more generally, the models used to explain variation in the outcomes among, say, counties would have to take into account the reasons there are differences in the county programs. Intercounty and, less frequently, interstate migration will also complicate the interpretations of these comparisons.

The diversity in state programs has good and bad consequences for evaluation. Wide variation in the program variables is useful in predicting and explaining outcomes. But unless the underlying variations in the states' economic and social conditions are measured and controlled for, biases in the estimates of program effects can be severe. For example, the large stock and flow of immigrants in California are not only features of the state's economic conditions but will also partly determine its welfare program. In Wisconsin, however, there is less attention to immigrants as a group that will be affected by W-2, although there has been considerable attention to welfare-induced migration from Illinois.

A persuasive program evaluation based on cross-state comparisons is basically difficult because it ultimately depends on statistical controls not only for good measures of program differences but also for the levels and changes of the states' economic and social environments. Picking similar states to compare, controlling for state differences in observable environmental factors, and applying the "difference-in-differences" model are ways that can only lessen, but not eliminate, the basic methodological problem. Some degree of reliance on a priori reasoning is needed. For example, if a state provides higher benefits and longer time limits for its welfare recipients, its job placements are likely to be lower; if the state offers a more generous child care program, then its job placements are expected to be higher. Finally, the state programs may be ranked by the level of their generosity, and even if specific explanations for the differences in outcomes cannot be attributed to particular components of each program, there is value in knowing the relation between various outcomes and the overall levels of program generosity. I end, therefore, with the modest proposal of evaluation by the traditional means of a theoretically based observational study. ■

---

[1]The paper upon which this article is based appears in full in "Evaluating Comprehensive State Welfare Reforms: A Conference," IRP Special Report no. 69, University of Wisconsin–Madison, March 1997.

[2]Controlled experiments are seldom conducted to evaluate even small-scale social programs. None have been used for such large-scale economic programs as federal minimum wage laws, federal tax legislation, health care programs, or even for state-level versions of these programs. One exception is the U.S. Department of Labor's evaluation of the Job Training Partnership program, which combined both experimental and nonexperimental designs, beginning in the early 1990s. The success of these evaluation designs is still under debate.

[3]Note that in evaluating the new state programs by observational rather than experimental studies, the same issue arises: a comparison with the old system offers the advantage of a clearly defined counterfactual, and it is one with considerable scientific interest, but the counterfactual of greatest policy relevance—some "viable alternative welfare program"—must be defined.

[4]Wisconsin officials who are planning the state's evaluation have explicitly rejected this method, as is indicated by the research plan contained in the Wisconsin Waiver Proposal, submitted to the federal government on May 28, 1996.

[5]The precise nature of such contamination needs to be examined. One version is that the new welfare program promises to "change the culture," which implies that the experimental group under the old welfare system will behave differently than they would if the new welfare system had not been implemented. The same criticism could be made of past experiments. For example, if the negative income tax had replaced welfare, would the subsequent cultural changes have been so extensive that, say, the labor-supply outcomes of either the treatment or control groups were incorrectly estimated? Economists would view this form of cultural change as a change in workers' tastes (or preferences) for market work. Is that likely? Perhaps the cultural change in the wake of the new state welfare system takes the form of certain institutional changes; for example, an expanded role by private charities. Still another type of "contamination" is that people in the randomly assigned treatment and control groups will behave differently just because they are aware of the existence of each other. Again, this problem was not raised, as such, in the negative income tax experiments. The famous Hawthorne effect was considered, but the program was considered sufficiently unobtrusive to avoid this problem.

[6]Two other papers presented at the conference, those by R. Haveman and by T. Kaplan and D. Meyer, also point out technical difficulties in carrying out a controlled experiment. See "Evaluating Comprehensive State Welfare Reforms" for full versions (abridged versions appear in this issue of *Focus*).

[7]Consider that a comparison of the *levels* of two states' outcomes is likely to reflect not only the differences in the states' programs, but also the states' long-standing differences in institutions, population composition, histories, and so forth. Now assume that these long-standing differences are reflected in the two states' differences in their preprogram measures of an outcome, such as the proportion of families on welfare. By measuring the *change* in this outcome that accompanies the change in the welfare system, the investigator may be able to attribute the outcome change to the change in the welfare programs. The latter measure is called a difference in differences.

[8]A more optimistic view of the scope and coverage of administrative data is that of Kaplan and Meyer (see note 6).

Recent IRP reprints

Hoynes, H. *Welfare transfers in two-parent families: Labor supply and welfare participation under AFDC-UP.* 1996. 38 pp. R748.

Robins, P., and Fronstin, P. *Welfare benefits and birth decisions of never-married women.* 1996. 23 pp. R749.

Wu, L. *Effects of family instability, income, and income instability on the risk of a premarital birth.* 1996. 21 pp. R750.

Reynolds, A., Mavrogenes, N., Bezruczko, N., and Hagemann, M. *Cognitive and family-support mediators of preschool effectiveness: A confirmatory analysis.* 1996. 22 pp. R751.

Wolfe, B., Haveman, R., Ginther, D., and An, C. *The 'Window Problem' in studies of children's attainments: A methodological exploration.* 1996. 13 pp. R752.

Wiseman, M. *State strategies for welfare reform: The Wisconsin story.* 1996. 32 pp. R753.

Haveman, R. *Reducing poverty while increasing employment: A primer on alternative strategies, and a blueprint.* 1996. 36 pp. R754.

Wiseman, M. *Welfare reform in the United States: A background paper.* 1996. 54 pp. R755.

Wolfe, B., and Perozek, M. *Teen children's health and health care use.* 1997. 23 pp. R756.

Haveman, R., Wolfe, B., and Peterson, E. *Children of early childbearers as young adults.* 1997. 28 pp. R757.

Kim, R., Garfinkel, I., and Meyer, D. *Is the whole greater than the sum of the parts? Interaction effects of three non-income-tested transfers for families with children.* 1996. 12 pp. R758.

Interstate comparison of welfare reform programs

Irving Piliavin and Mark Courtney

Irving Piliavin is Emeritus Professor of Sociology and Social Work and Mark Courtney is Assistant Professor of Social Work, University of Wisconsin–Madison. Both are IRP Affiliates.

Can interstate comparison of the new welfare programs offer a valid basis for impact evaluation?[1] Defensible impact evaluations require at least three components: specification of a set of potential program effects; identification of a methodology for measuring the hypothesized changes; and—because some change might have occurred without the program—a procedure for estimating what would have happened in the absence of the program.

In this article we examine alternative designs for measuring hypothesized changes based on interstate comparisons.[2] We propose, for reasons discussed below, to use two information sources and three potential survey designs in several states. The two information sources are (a) merged agency administrative data, and (b) interviews with agency staff members and program participants.[3] The potential designs include (a) single-wave cross-sectional studies, (b) multiwave cross-sectional studies, and (c) multiwave panel (longitudinal) studies. We assume that data dealing with prereform experiences may be available from agency databases, but that information supplied through program participant interviews will not.[4] The various forms of investigation that are possible from different combinations of these data source and design possibilities are represented in the cells of Table 1.

Strengths and weaknesses of data sources for interstate comparisons

The public institutions that make use of administrative databases (e.g., work programs, child welfare services, corrections, schools, housing services, and mental health programs) generally use them to monitor program participation, including reasons for beginning and ending participation. Such information can be useful to evaluators, particularly when program participation or service provision are important "inputs" or "outcomes" of interest.

However, administrative data have three important disadvantages as the sole bases for evaluating welfare reform efforts, and particularly for cross-state evaluations:

1. Because the comparability and range of data from state administrative data systems differ, cross-state comparisons beyond the dates of program participation and the provision of benefits are often impossible.

2. Because administrative data fail to provide information on the experiences, motivations, and relationships among current or past program participants, the consequences for families of participating in the programs cannot be ascertained.

3. Since administrative data have low utility to those who generate them (i.e., line workers in social agencies) and the quality of data entry is often not monitored, information from this source is at least questionable.

Nevertheless, program participation and transition information supplied through administrative data sources pro-

Table 1
Potential Designs for Cross-State Evaluation of Welfare Reform Based on Administrative and/or Interview Data

Data Source	Design		
	One-Wave Cross Section	Multiwave Cross Section	Multiwave Panel
Merged administrative data	1	4a Post- only administrative data	7a Post- only administrative data
		4b Pre- and post- administrative data	7b Pre- and post- administrative data
Interviews	2	5	8
Merged administrative data and interviews	3	6a Post- only administrative data	9a Post- only administrative data
		6b Pre- and post- administrative data	9b Pre- and post- administrative data

vide an important complement to information concerning details of life experiences among present and past program participants. The best sources for information about these experiences are family members, and the best vehicles for tapping these sources are in-person survey interviews. The primary disadvantages of survey interviews are practical and financial, and involve decisions about whom to interview, and when; determination of the sample sizes required to make meaningful decisions about program effects across jurisdictions; and formidable logistical problems in contacting and interviewing family members and other informants. The cost will be high for any cross-state evaluation that relies significantly on data directly obtained from program participants.

Selected cross-state evaluation designs

Cell 9 investigations

These studies, the most detailed class of those represented in Table 1, use panel designs, employ individual families as the unit of analysis, and generate data through interviews and administrative records. Within this class, the most common investigation follows a sample of families who apply to participate in the welfare program in each of several states over the same period of time. Policy-relevant family-composition changes and experiences can be recorded and compared for sample families from states that implement distinct welfare reforms. Changes in family circumstances can be related to *specific* program characteristics, conditioned on family and state attributes. Investigations of this type may also follow more than one sample entry cohort in each state, making it possible to assess the relative consequences of different welfare reform efforts as they mature over time. This is an appealing prospect, given the uncertainty surrounding many program changes being contemplated by states.

A major advantage of this class of investigations is their ability to provide data on the experiences of sample families when they are no longer served by agencies for which administrative data are available.

Cell 8 investigations

Investigations in Cell 8 are also able to provide data on families who have left the program. But because Cell 8 data are limited to interview responses from family members and other informants, the information on agency transactions may contain unknown biases—a clear weakness relative to Cell 9 studies. The strength of Cell 8 studies relative to the remaining investigations in Table 1 is their ability to collect data on changes in family composition and experiences. These data are beyond the capability of panel designs using agency databases (Cell 7) or are simply not available in cross-sectional study designs (Cells 1 through 6).

Cell 6 investigations

This class of studies comprises investigations based on multiwave cross-section designs, utilizing data from interviews as well as agency administrative databases.[5] Because the studies provide detailed information on the present status of current participants in (or applicants to) welfare reform programs, we can compare the circumstances of families in different programs. They also allow us to study how these circumstances change across cross-sectional samples over time, and whether these changes vary across distinct programs. However, we cannot investigate within-family change over time as a function of experiences in the program, nor the circumstances of families who formerly participated in, or applied to, aid programs.

Cell 4 investigations

This class of studies makes use of information from linked agency databases over time, information on the substance of agency programs, and statewide demographics to provide limited information about the effects of pre- and postreform programs. Using some areawide controls, these data allow us to compare the character of agency caseloads at given points in time. Over time, this information will provide data on the numbers and attributes of participants in different welfare reform efforts and on their involvement in programs of other social service agencies. No information would be available on people not served by these agencies.

One form of investigation not included in Table 1 has occurred sufficiently often that it deserves mention: the comparison across states, perhaps over time, of *unlinked* agency administrative data that consist simply of percentages of people served by various social agencies. These numbers may be used to argue for the failure or success of welfare programs. A simple case is the current finding among agencies in Milwaukee and Madison, Wisconsin, that the numbers requesting help from homeless shelters dramatically increased after the early spring of 1996. Agency administrators argue that the reason for the increase, at a time when requests have traditionally decreased, is the gradual implementation in Wisconsin of W-2 precursor programs. There are, however, no systematic data showing that those seeking help were either former participants in AFDC or people rejected from or not qualified for the W-2 program. Supporters of W-2 may argue that the increased homelessness reflects a sudden and recent increase in migration to an economically flourishing area by relatively poor people who have not yet obtained employment. Again, there are no supporting data. Except under very unusual circumstances, investigations based on this design will fail to provide the information necessary for even minimally convincing evaluations.

The projects we have represented in Table 1 assume, within each state, that we can draw samples from the

Invited comment: The value of interstate comparisons

Wisconsin and other states are moving rapidly to transform their traditional welfare and income transfer programs into work and family self-sufficiency systems. The conference accurately reflected both the sense of urgency for evaluating the new systems and the challenges that evaluators will face in carrying out the evaluations. To ensure that there are adequate data and information for evaluation, conferees discussed the need to develop data and reporting systems that serve the needs of both program managers and program evaluators. If we want to have accurate and reliable data with which to conduct evaluations and other research, we should invest some effort now to develop such systems. As public agencies move more towards performance measurement and monitoring, it may be more feasible than ever before to integrate evaluation data needs with management data needs. Specific state and local evaluations and studies will increasingly depend upon data from administrative information systems. It seems wise to assume that funding for large-scale program evaluations will be much more limited than in the past decade or two. It makes sense, therefore, to plan for research and monitoring in advance, as part of ongoing program management, rather than simply assume that eventually a costly external evaluation will be imposed upon the new systems.

One area not covered in much detail at the conference concerns the need to conduct cross-state and national research as well as discrete state evaluations. While it has always been difficult to incorporate comparative cross-state analyses into national studies, that challenge is even greater as the nation moves towards a more decentralized system of benefits, services and program structures for aiding the poor. National policymakers could spearhead such an effort, even if they no longer can provide as much funding as in the past. For example, federal agencies could support (and possibly fund) a coordinated effort by encouraging a number of states to maintain certain core data items that could be useful for both management and evaluation. This does not mean imposing more reporting requirements on states, but rather facilitating a cross-state cooperative research initiative that would be defined by state, rather than federal, officials.

Demetra Smith Nightingale
The Urban Institute, Washington, D.C.

population of families which have had some form of systematically recorded contact with welfare reform programs and possibly with their AFDC forerunners. The samples, depending on the interests of evaluators, may be composed of program participants or, for more thorough analyses, program applicants. An important population not tapped by these samples consists of families who, although potentially eligible, fail to apply for welfare programs.[6] The analyses that would be necessary to capture this population must be based on state probability samples of those eligible for but not participating in the program.

Clearly, we view the cross-state evaluations represented in Cells 8 and 9 of Table 1 as offering distinct advantages over the alternatives. The data gaps in evaluations based solely on administrative data are so great that valid program evaluations using only such data seem impossible. Since reliable estimates may require comparisons across at least eight to ten states, the costs of the evaluations represented in Cells 8 and 9 may appear to be prohibitive. Yet in the not-too-distant past, there were multimillion-dollar evaluations of programs (for example, the National Supported Work Demonstration) that were much more limited in scope and potential effect than the various versions of welfare reform being introduced throughout the nation. The costs to be incurred in implementing even the most elaborate evaluation efforts outlined in Table 1 are relatively small compared to the magnitude of the possible impacts of welfare reform.

The costs of the approaches identified in Table 1 vary significantly. In general, designs that rely solely on administrative data are much less expensive than designs that require direct collection of data from informants. In some states, data managers will be able, at relatively low cost, to generate reports on the program participation of cross-sections of the population (necessary for Cells 1, 3, 4, and 6 of Table 1). But they generally have less experience in merging data on program participation or in analyzing event histories across service systems. Thus designs represented by Cells 7 and 9 of Table 1 will certainly require the active participation of experts in complex data management, matching, and analysis.

These complications notwithstanding, the costs of analyzing administration data will be small in comparison to the costs of gathering data directly from households. The University of Wisconsin Survey Research Center has estimated the direct costs of undertaking panel surveys of program participants, though not of interviewing program operatives—in part because we have not yet determined what information these officials might usefully provide—and has come up with ballpark estimates, based on some assumptions. First, we assume initial interviews will be 1.5 hours in duration and that subsequent interviews will be 0.75 hours in duration. Second, we assume that the initial sample of applicants and participants will be 1,000 in each state, and that the sample attrition will be 10 percent between consecutive waves. Third, we assume that the panel in each state will be interviewed six

times over a period of 2.5 years. The costs of collecting data from the panel in each state, given these conditions, will be slightly more than $1 million. Costs of administration, data preparation, and analyses must be added to this sum. If brief phone interviews are undertaken to obtain details of major changes in family circumstances, and if there are about three such changes per family, another $125,000 should be added to the total for each state's panel costs.

Constructing counterfactuals

In a comparative study such as our proposal, the question under analysis is not so much "What are the impacts of welfare reform in State X, compared to what would have happened if the state had continued under the status quo?" The relevant question instead is "What are the impacts of welfare reform in State X (and Y and Z) compared to the impacts of other policy choices made by other states?" The use of other states as the counterfactual in this way demands that:

1. The states being compared are alike in all relevant background characteristics, such as the economy, key demographic variables, and managerial structure and skill. Although this is an unlikely occurrence, it should be possible to minimize differences. For example, states can be compared only within their own region, because economic change appears to affect regions similarly. Also, states in which welfare programs are administered at the state level can be compared only with one another, and states in which welfare programs are administered by counties can form a separate universe of comparison.

2. The differences in state characteristics can be compensated for through statistical adjustments. Robert Haveman notes that no easy and obvious compensations exist for different rates and kinds of economic change occurring in different states.[7] But if perfection is not achievable, rough adjustments based on, say, differing rates of unemployment may be possible. At least the evaluators will be able to identify the adjustments they did make and invite suggestions for alternative adjustments from secondary researchers.

3. Some policy variables may appear obviously to be critical because they seem to exert the same influence in all of the states that tried them, regardless of background characteristics. It is possible that no policies will be so potent as to overcome all the other differences that exist among states and exert universally apparent impacts. But it is also conceivable that states that restrict public assistance to, say, less than six months would show quite different work patterns and child and family well-being impacts than do states that allow public assistance for five years, and that these patterns would become apparent despite the many other differences among states. Certainly such patterns will be missed entirely if we do not look for them through cross-state comparisons.

In all probability, a combination strategy will be necessary that (a) limits analysis to particular regions, (b) compensates for background characteristics through statistical adjustment, and (c) searches for exceptionally potent policy variables. Although no a priori certainty of useful findings under this design is possible, the strategy identified here seems no less likely than others to identify causal results in an environment in which experimental design is not a useful evaluation technique.■

---

[1]The paper upon which this article is based appears in full, under the title "Prospects for Comparing Wisconsin Works to Welfare Reform Programs Outside the State," in "Evaluating Comprehensive State Welfare Reforms: A Conference," IRP Special Report no. 69, University of Wisconsin–Madison, March 1997.

[2]We do not consider potential program effects because they are discussed at length in this issue by Cancian and Wolfe, "Outcomes of Interest, Evaluation Constituencies, and the Necessary Trade-offs").

[3]For the sake of consistency we use the term "interview" throughout the paper when referring to data collected directly from individuals for the purpose of a cross-state evaluation.

[4]It is possible that data dealing with prereform experiences may be obtained from postreform interviews, but recall problems and program exposure effects may limit their usefulness.

[5]It may seem odd to discuss "cross-sectional" administrative data, but in many cases administrative databases retain only snapshots of program participation (e.g., whether a person received a particular service in a given month) and are not suited to the construction of actual event histories.

[6]R. Blank and P. Ruggles, in "When Do Women Use Aid to Families with Dependent Children and Food Stamps?" *Journal of Human Resources* 31, no. 1 (1996): 57–89, estimated that single mothers fail to enroll in AFDC during 30–38 percent of the months for which they are program eligible. This failure occurs among sample members who are less needy (e.g., higher proportion employed, smaller proportion receiving other forms of assistance), are more advantaged (higher education, less likelihood of disabilities), and anticipate benefits insufficient in size or duration to warrant their program participation.

[7]R. Haveman, "Potentials and Problems of a Pre-Post Design for State-Based Evaluation of National Welfare Reform," in "Evaluating Comprehensive State Welfare Reforms: A Conference," IRP Special Report no. 69, University of Wisconsin–Madison, March 1997.

Toward a basic impact evaluation of Wisconsin Works

Thomas Kaplan and Daniel R. Meyer

Thomas Kaplan is Associate Scientist at IRP and Daniel R. Meyer is Associate Professor of Social Work at the University of Wisconsin–Madison and an IRP Affiliate.

The Personal Responsibility and Work Opportunity Reconciliation Act of 1996 overturns 60 years of federal welfare policy, eliminating the federal structure of Aid to Families with Dependent Children (AFDC) and replacing it with a block grant which states can use to design and operate programs for the poor.[1] Wisconsin has received federal approval for its planned use of block grant funds, called Wisconsin Works (W-2). The full W-2 program is scheduled to start on September 1, 1997. The two transitional programs that have been in place statewide since March 1996, Self-Sufficiency First and Pay for Performance, will, however, complicate comparison between W-2 and its predecessor, AFDC.[2]

In evaluating social programs, it is difficult enough to measure the impact of one discrete change when related policies remain constant. W-2 presents many major changes occurring continuously over a long period, and evaluation will be correspondingly more complex. In this article, we outline a basic impact evaluation of the Wisconsin program and briefly review evaluation strategies for those effects that we consider to be potentially most critical: effects on income and poverty, on dependency, on child care and child welfare, on health status, and on the living arrangements and family structure of low-income households.

The evaluation plan

The scope of the evaluation

The hopes of influencing the work and family lives of present or potential welfare recipients is a distinguishing feature of the emerging versions of comprehensive state welfare reform.[3] To reflect the aspirations of program designers and the fears of program critics, evaluations must cover multiple dimensions, but they must also be small enough to be manageable. Thus we propose a basic impact evaluation built around the six primary domains of income, dependency, child care, child welfare, health status, and living arrangements and family structure. These were chosen because they are central to the purposes of the reform and to the well-being of the families affected. They also offer a high likelihood that evaluators will find a measurable effect using data that are now available or that could be available with minimal future investment.

W-2 may affect many groups in addition to W-2 participants themselves: those who would be eligible but do not participate; adults in the low-wage labor market, including those with no minor children, who are not eligible for the program; and the wider community of employers, child care providers, schools, social service agencies, governmental units, and taxpayers. To keep this proposed evaluation manageable, we focus our attention on individuals rather than employers, institutions, or entire markets. We do not propose a benefit-cost analysis, which would require a major and complex effort, particularly in determining such potential indirect costs as, for example, the increased (or decreased) costs borne by the educational system if children are less (or more) prepared for school. We do not seek to identify an idealized evaluation, but to think about the difficult decisions one must make if one is to conduct a limited evaluation, with a limited budget. But we do believe that our strategy, using a single counterfactual and a similar approach to evaluating each potential impact, may answer the most critical questions.

The counterfactual

We propose a primary comparison between outcomes under W-2 and outcomes in Wisconsin before the implementation of W-2—a pre-post design. Other options, such as an experimental model or a comparison between states, appear either weaker or unsuitable in this context.[4] Experimental designs that randomly assign participants work best when (a) only a few program elements are undergoing change (because it is then administratively feasible to operate an experimental and control program), (b) the control group can easily be isolated from "contamination" introduced by the experiment, and (c) the intervention is not expected to generate community feedback effects. W-2 meets none of these conditions.[5] Random assignment also requires the cooperation of the relevant administrative agencies, and this may not be forthcoming.

Comparing results under W-2 to results from another state would be most helpful if differences in the well-being of low-income households or the status of the low-wage labor force could reasonably be attributed to different public assistance strategies in the two states. But many other policy and economic variables could influence observed differences among states—a recession in a particular industry, a natural disaster, or state policy toward related public programs (elimination of General Assistance, removal of public schools from the property

tax, etc.). Moreover, states have very different administrative record-keeping systems, and we propose to rely heavily on administrative data in the basic evaluation.[6]

The most serious problem with a pre-post time-series design is the possibility that something else is changing at about the same time that AFDC gives way to W-2.[7] Clearly, some major factors *have* changed: the minimum wage increases in 1996 and 1997 and the 1990 and 1993 changes in benefits under the Earned Income Tax Credit (EITC) make evaluation of the wage and employment effects of W-2 for low-wage workers quite problematic. The ending of AFDC, changes in the Food Stamp program, and changes in the Supplemental Security Income (SSI) program—all components of the federal welfare reform passed in 1996—will take effect contemporaneously with W-2, further complicating the simple comparison of W-2 and AFDC. Below we identify some strategies for dealing with these changes, recognizing that none are perfect.

Basic data source

We propose to depend primarily on administrative data to examine effects, partly because of the expense of obtaining data through surveys over time. Capturing the effects of an intervention on income growth, for example, would require panel studies measuring the same people over a period of years; we can follow the same family's income tax records much more efficiently.

Cost is not the only reason to prefer administrative data. The first wave of any survey would be mounted well after much of W-2 had gone into effect. Thus the best way to assess the full impact of W-2 is somehow to capture the situation that existed before it began. Administrative data in both electronic and paper form do extend back into the past and offer some reliability. Further, it may take some time before W-2 is operating smoothly. In order to account for different effects between the first and subsequent years of operation, a panel survey would need either a very large initial sample (in order to ensure that there would be a sufficient number of new W-2 recipients in later years) or a new sample each year, both expensive options. Finally, the implementation analysis may reveal that W-2 is implemented very differently in some counties than others. If administrative records are being used as a sampling frame, researchers will be able to oversample in some counties several years *after* program differences have been detected.

Nevertheless, we believe that surveys of low-income families could provide valuable additional information: for instance, a simple survey gathering demographic information on individuals in the administrative databases, variants of the community-specific surveys suggested in the article in this issue by Cancian and Wolfe, or supplementary analyses using the Urban Institute's New Federalism Household Survey.[8] Analyses using ethnographic data would also add an important dimension to the primarily quantitative analyses that we propose in the basic evaluation.

Time period

If the time period for an evaluation is too long, the results may be too late to influence the development of the program, and the counterfactual becomes more difficult. In a pre-post evaluation, other significant events may occur during a lengthy evaluation period that create their own effects; in a cross-site design, significant changes are more likely to occur in the "control" site also. Yet a program evaluated too early, for too short a period, may tell us nothing about the way the program would work when it is more mature; and programs that, like W-2, seek to affect community norms may need time to take hold.[9] In W-2, there is another complication: the time limits built into some components (see this issue, p. 2). Any evaluation that is concerned about the effect of the program's five-year limit must extend far enough to examine outcomes for those who exhaust their eligibility. And W-2 proponents recognize that, in the short term, income may decline for some participants, but they hope for longer-term increases in income from earnings in the unsubsidized work force.

Therefore, we propose a two-part evaluation: (1) a three-year examination focusing on short-term effects and (2) a seven-year examination focusing on longer-term effects; in seven years some families will have exhausted benefits and the time period should also be long enough to capture longer-term wage increases. Further, we propose no impact evaluations for the first year, only process evaluations, allowing at least some time for the program to change, develop, and reach a steady state.

Impact domains in a basic evaluation

In this article, space precludes a detailed examination of the methodology for the six primary domains already noted. As an example of our approach, we examine in some detail our proposed strategies for the analysis of the effects of W-2 on the primary domain of income and poverty, and highlight differences and difficulties in the issues and strategies for the five other primary domains.

Primary potential impact 1: Income and poverty

Hypothesis 1.1: W-2 will have an impact on incomes among low-income families with children.

Because some factors are likely to result in increased incomes and others in decreased incomes, we do not specify the predicted direction of effect, but propose that it is important to identify (a) whether incomes of low-income parents as a whole increase or decrease, and (b) the types of families in each category.

Income increases could occur through five primary mechanisms:

1. Some low-income families who do not currently have the job skills to compete in the labor force may eventually obtain the skills (and/or the work experience) to enable them to earn more than they could have earned in the past or would have received from AFDC.

2. For recipients who have some job skills, but would receive wage offers too low to make working worthwhile, W-2's requirement to work, combined with child care and (perhaps) health insurance subsidies, may shift the incentive toward employment. These individuals may eventually earn more than they would have under the old AFDC system.

3. Adults in some low-income families do not currently work outside the home because AFDC allowed them to receive benefits with minimal effort. Obliged to work and confronted with time limits on cash assistance, these individuals may eventually achieve higher incomes than under AFDC.

4. Many husband-and-wife low-income families are not currently eligible for assistance because of the limitations in the AFDC-U program and the severe funding restrictions upon child care assistance. W-2 subsidizes child care assistance and (perhaps) health assistance to such families, and may also enable some single-earner families to become dual-earner families, in both ways increasing their incomes.

5. By allowing families receiving assistance to keep the entire amount of any child support paid, W-2 directly increases incomes and potentially increases the total amount of child support paid.[10]

W-2 may, however, reduce income for five groups of families who used to receive AFDC.

1. Some families will exhaust the W-2 time limit and stop receiving benefits.

2. W-2 is not an entitlement, and during periods of recession and state budget shortfall some families who would have been participants may not receive benefits.

3. In the W-2 Transitions, community service, and trial job components of W-2, large families who participate will receive less income than they would have received under AFDC.

4. Families who are sanctioned for failing to comply with various provisions of W-2 will have lower benefits.

5. Some families eligible for W-2 may choose not to join the new program because of its work requirements.

For each family that no longer receives benefits, an increase or decrease in total income depends on its response to this new regime.

Finally, W-2 will provide some families with assistance at a lower level than AFDC. One key group is families with two or more children. Under AFDC, for instance, families with one adult and three children received $617/month; under W-2 Transitions they will receive $628/month, but child care copayments and perhaps health care premiums will reduce their net incomes to an amount well below the former AFDC grant. For larger families, the disparity between W-2 and AFDC is greater. Again, eventual total income depends on a family's response to the lowered benefits.

Data, comparison group, and analytic approach

For the AFDC period, we have the ability to examine both the *earnings* of AFDC recipients before, during, and after receipt, through the quarterly earnings records of the Department of Industry, Labor, and Human Resources (DILHR), and their *family income* through an extract from the tax records of the Wisconsin Department of Revenue (DOR).[11] Our proposed approach is built around three main comparisons: (1) comparing income changes of W-2 and AFDC recipients; (2) comparing pre- and post-W-2 income changes of a broader sample of low-income families; (3) comparing pre- and post-W-2 income changes among higher-income families.

1. For *the W-2–AFDC comparison*, we propose selecting two samples of W-2 recipients in 1998 (or another year after full implementation of W-2)—a sample of those who enter the program during the year and a sample of those who were participating at the beginning of the year. Using the social security numbers of the adults, we propose merging tax, earnings, W-2, and food stamp records to calculate the total income of these families in each year from 1999 through 2005. For the comparison AFDC group, we would select two parallel samples of recipients drawn from the administrative AFDC records for 1988 and again calculate their family incomes in each year from 1989 through 1995 using similar records. We propose drawing separate samples of entrants and current recipients because many analyses show that these two groups are quite different.[12]

Our main longer-term evaluation is a comparison of the later family incomes (in 2005 and 1995, respectively) for these samples of entrants and recipients. We estimate a multivariate equation in which later family income is the dependent variable and the key independent variable is whether the family was an AFDC recipient (in the early period) or a W-2 recipient (in the later period). As control variables we incorporate, for example, features of the local labor market, individuals' educational levels, and family size and structure. We will also explore the effect of sociodemographic characteristics to determine if W-2 had different effects for different groups of people. Our main short-term evaluation compares the 2001 incomes of the 1998 W-2 recipients with the 1991 incomes of the 1988 AFDC recipients, using a similar analysis.

2. All these analyses test whether the later incomes of W-2 participant families are different from the later incomes of AFDC recipient families. But W-2 may also have entry effects that influence whether low-income families participate or not. Thus our second major comparison looks at *low-income families with children during the W-2 and AFDC periods*. We propose drawing two additional samples from Wisconsin tax records for 1998 and 1988, of families with children who have income less than twice the poverty line (by chance, some welfare families may also be included in these samples). We then propose to examine income in 2001 and 1991 (short-term effects) and 2005 and 1995 (longer-term effects), through analyses that parallel the analyses of recipients. Tax data might be supplemented with survey data to gain a full sample of low-income nonrecipients.

3. One method of increasing our confidence that any observed change in incomes is due to W-2, rather than to other factors, is to include a variety of control variables. Another method is to *compare incomes over time among higher-income families with children* (families with taxable income above 300 percent of poverty). W-2 is likely to have little effect on the incomes of upper-income families. Thus if the family incomes of the low-income sample in 2001 are higher than in 1991, but the family incomes of the higher-income sample have not changed, we can have greater confidence that the effects we observe are the result of W-2 rather than of economy-wide changes.[13]

Issues and limitations

Here we note two main difficulties; others are discussed in our conference paper (see note 1).

Defining income. Should we use the traditional definition employed by the Census Bureau—pre-tax cash income only—or a concept of disposable income that includes near-cash in-kind benefits and also deducts taxes? The latter is preferred by many analysts. In the basic evaluation we propose, which relies only on administrative data, we will not have measures of employer-provided health insurance, housing assistance, or a good measure of home ownership; this limits the measure of income we can construct. For those filing tax returns, we propose to calculate income by summing taxable income, AFDC or W-2 benefits, food stamps, and the federal EITC, and subtracting federal and state income taxes and payroll taxes. For those not filing tax returns but with in-state AFDC or W-2 income or in-state earnings, we will estimate income by summing earnings, AFDC or W-2 benefits, and food stamps, and subtracting estimated payroll taxes. We will also consider including estimated out-of-pocket child care expenditures; this information will, however, be more available under W-2 than under AFDC.

Defining a W-2 participant. Individuals who are placed in trial jobs, community service jobs, or W-2 Transitions are clearly participants. We do not propose to count families that receive child care subsidies as W-2 recipients because we generally lack information on families that received only child care assistance in the AFDC period. If there is information on individuals receiving only Medicaid in the AFDC period, we could include both them and the corresponding W-2 group (those receiving health insurance subsidies only, if a W-2 health plan is implemented). Another difficult group includes those who come for help and are diverted to a private-sector unsubsidized job, receiving no other services. W-2 considers these individuals to be participants and appropriately will count them as successes, but we believe that they cannot be included in the sample of "W-2 clients" because no comparable circumstance is available for the AFDC period. The analysis of income changes for the entire low-income sample may enable us to estimate effects on this "referral-only" population.

Limitations of the proposed strategy. The proposed evaluation strategy has several limitations regarding information on income. First, Wisconsin's administrative records of income miss some individuals—for example, those who move out of state. Second, tax records also miss information on those who do not file taxes. Wisconsin, however, offers a refundable credit for low-income renters and home owners, increasing the likelihood that a low-income family will file a tax form.[14] Incorporating information from earnings and welfare records increases the coverage. Analyses conducted with data from the Wisconsin child support project suggest that income estimates should be available for about 90 percent—perhaps over 95 percent—of a low-income sample.

The measures of family income are also limited. Taxable income does not include various income sources, including transfers and child support.[15] Earnings records report only earnings in *covered* employment, not those in the informal sector. Work-related expenses are not fully available. Tax records provide only limited information on family composition and offer no way to identify other adults in the home (cohabitors, roommates, relatives); such information is needed to determine economic well-being. For many families with gross income under 130 percent of the poverty line, Food Stamps administrative data can fill in some of this information.

Administrative records of earnings and taxes paid contain very limited information on families, leaving us without basic data such as race and educational level. This information will be available for any families who ever received AFDC or W-2, but even for these families we will not know whether they speak English nor anything about their family of origin, and we will have only sketchy information on work experience. We consider these limitations potentially quite problematic; one potential solution is a short survey gathering basic data on individuals who are in the administrative records.

Hypothesis 1.2: W-2 will decrease incomes among families without children.

Three provisions of W-2 may have a direct impact on the earnings (and therefore income) of adults without children. First, W-2 will force many low-skilled parents into the labor market. The general supply of low-skilled people wanting jobs will increase, presumably decreasing wages for low-skilled jobs and making employers less likely to retain marginal current employees. Second, W-2's provision of employer subsidies for hiring from a targeted class may discourage the hiring and retention of nontargeted employees, decreasing their incomes. Finally, W-2 creates many community service jobs in nonprofit organizations and local governments; these organizations may also, therefore, forgo hiring other individuals and be less likely to retain marginal employees, both of which lead to lower incomes.[16]

Data, comparison group, and analytic approach

The approach to evaluating this question follows that for the first hypothesis: draw a sample of low-income families without children in 1998 and 1988, calculate their incomes in 2001/1991 and 2005/1995; examine whether the level of income in 2005/2001 is lower than in 1995/1991; and examine whether individual changes in income from 1998 to 2005/2001 are lower than they were from 1988 to 1995/1991.

Issues and limitations

The definitional and data issues discussed above are also relevant here. In addition, the data from DILHR and the Department of Workforce Development include only information on quarterly earnings, not on hours worked or weeks worked. Thus a basic evaluation scheme can identify those with low earnings and determine whether their proportion increased, but cannot differentiate those who worked longer hours to keep their total earnings roughly comparable. If W-2, for example, causes hourly wages to fall from $10 to $5 an hour, individuals who, under AFDC, were working 250 hours quarterly (for total earnings of $2,500) may, under W-2, keep total earnings constant by working 500 hours quarterly.

Hypothesis 1.3: W-2 will change the proportion of families with children who are poor, and will change the poverty gap (the aggregate amount needed to bring everyone up to the poverty line).

Conventional measures of poverty in the United States are merely measures in which family income is compared to a threshold that varies by family size. Thus the issues, strategies, and limitations discussed under Hypothesis 1.1 are relevant here. Because of the well-known limitations in the official governmental measure of poverty, we propose to use a variety of such measures, including the official measure and a measure that is as close as our data will allow to that proposed in 1995 by the Panel on Poverty and Family Assistance of the National Research Council.[17]

Assessment of poverty status makes the need for accurate information on family composition even more critical. The lack of information on child care and other work expenses is also particularly troubling for measures of poverty status that use the concept of disposable income rather than gross cash income. Still, there are no better alternatives; no information exists on work expenses during the AFDC period.

Primary potential impact 2: Dependency

Hypothesis 2.1: W-2 will lead to a decreased reliance on means-tested transfers among low-income families with children.

There are several possible ways to conceptualize "dependency" and "reliance." We suggest focusing quite straightforwardly on the proportion of income that comes from means-tested transfers (programs specifically for the poor).[18]

Hypothesis 1.1 above suggested mechanisms through which earnings and child support, and thus income, would eventually increase, and the share of income from means-tested transfers would decline automatically. Recipients who lose transfer income because of time limits, lack of entitlement, or sanctioning will also see their transfer ratio decrease. Thus the main evaluation in this domain has to do with whether *recipients* are able to become less dependent, and these constitute our prime population of interest. But W-2 may also affect whether low-income families *become* recipients; because eligibility is opened to two-parent families, more people may receive benefits. To evaluate this possibility, we will conduct a parallel analysis of the sample of low-income families with children described under hypothesis 1.

Looking at the percentage of income derived from means-tested transfers, however, adds new difficulties to those described under hypothesis 1. What, for instance, is included as a means-tested transfer? Some benefits are clear: food stamps, the AFDC benefit, SSI, and W-2 Transitions are all means-tested transfers. Presumably the EITC and public subsidies for child care and health care (both Medicaid and the public portion of the premium under W-2) are also transfers. But within the W-2 program, how should we view trial jobs? From a taxpayer's perspective, the subsidy to the employer is a benefit available only to the poor; thus only wages above the subsidy should count as earnings. But recipients perceive themselves as *working* for the entire amount; thus none of it should count as an "unearned" benefit. Community service jobs create a similar problem, although it is clear that the state treats these as grants, rather than earnings. Finally, what if private not-for-profit organizations provide vouchers or cash to individuals who have

exhausted their benefits or who have been sanctioned under W-2? These "gifts" represent, in effect, a transfer of dependency from public assistance to private charitable sources, though the data will not exist to evaluate them.

Primary potential impact 3: Child care

Child care arrangements for low-income families are likely to be affected by W-2. We believe the most important effect is child care quality, because it potentially has long-term consequences for child development.[19] But reaching agreement on direct measures of child care quality is difficult. We suggest two indirect measures for which data exist.

Hypothesis 3.1: W-2 will cause (a) an increase in formal complaints against child care providers made to the office that licenses the providers and (b) an increase in substantiated cases of child abuse and neglect by child care providers.

W-2 imposes higher copayments for licensed than for certified child care; this provision may lead families to select lower-cost child care or to leave their children unsupervised. W-2 also establishes a new category of provisionally certified child care providers for whom training requirements and regulation are limited and who may provide lower-quality care. Finally, the requirement that all recipients work may cause demand for child care to outstrip supply, at least in the short run, both increasing the cost of care and causing some parents to accept child care arrangements of lower quality than they would have liked.

The most serious allegations against child care providers, those for abuse or neglect, are part of the child welfare record-keeping system, although county staff investigating such cases may not always enter data reliably. We propose to examine pre- and post-W-2 records in counties that do have adequate records to determine rates of substantiated abuse or neglect by child care providers among the entire potential population of child care users, not merely among welfare recipients, because the entire child care market is likely to be affected by W-2.[20] The number of children in Wisconsin who are in child care is unknown, making it difficult to construct rates of this form of abuse and neglect before and after W-2; it may be possible to broadly estimate them from data on child care use in the Survey of Income and Program Participation (SIPP).

This measure of quality is clearly quite limited. The number of complaints can be affected by whether people know where to complain, by the perceived threshold of dissatisfaction (the point at which one is justified in making a complaint), and the likelihood of corrective action; all of these could change substantially over several years, especially if new populations who may not previously have worked must now send their children to child care providers. Parents with other child care options may remove a child without lodging a complaint of abuse or neglect. Parents with few other child care options may not want to risk losing the arrangement they have. Nonetheless, substantiated incidents of abuse and neglect are so serious that we believe they are important to track and to try to evaluate.

Better measures of child care quality—child/staff ratios, the size of child groups (regardless of ratios), the child care training and formal education of providers, and the frequency of turnover of child care providers—are simply unavailable.[21] It would be appropriate to start collecting such data for licensed, certified, and provisionally certified providers, even if formal pre-post analysis and causal inference are impossible.

Primary potential impact 4: Child welfare

"Child welfare" here refers not to general child well-being, but to the official concerns of the formal child welfare system in Wisconsin—child abuse and neglect and the placement of children in substitute care (primarily foster care) when parental care is unavailable or considered inadequate.[22]

Even if the actual rate of child abuse or neglect remains unchanged, reports may increase because of greater public attention to the issue or changes in the law concerning who is required to report suspected cases. Because actual abuse and neglect are unambiguously bad, however, it is important to try to assess whether major social policies such as welfare reform influence the direction of the imperfect indicators available to us.

On the one hand, the W-2 program might cause child abuse and neglect reports and substantiations to increase. In the first place, parents who are sanctioned for failure to participate in W-2 lose income and may neglect the basic health and physical needs of their children.[23] Second, parents who, under AFDC, would have worked for few or no hours each week but who participate for 40 hours per week under W-2 may find the pressure of full-time work combined with single parenting to be overwhelming, and they may behave inappropriately to their children as a result. And finally, new officials—the Financial and Employment Planner or the Supportive Services Planner—are paying close attention to families. On the other hand, reports might decrease because parents who successfully participate in W-2 may then also more competently fulfill their parenting roles.

The placement of children in foster care could increase or decrease for largely the same reasons as reports of abuse and neglect. Because only a small percentage of any population ever enters the formal child welfare system owing to abuse, neglect, or out-of-home placement, a panel study, though highly desirable, would require a

very large sample. It would also be complicated by the need to obtain accurate answers to retrospective questions about these difficult experiences. We thus turn again to administrative data for the basic evaluation design, although an additional targeted evaluation of child welfare issues, probably using survey data, also has merit.

It would be possible, although not simple, to use the basic pre-post samples described above to obtain some information about the direction of change in child welfare indicators. The tax information provides the county of residence and a name or social security number, and it would be necessary to ask selected counties if, in the particular year at issue, any child in the family was reported or substantiated as abused, neglected, or placed in out-of-home care. Once these matches were accomplished, the basic comparison could be similar to the income/dependency comparisons described above: is there a change in the low-income samples before and after W-2?[24]

There are many difficulties. Because of the small numbers of cases involved, any change might not be statistically significant, even if it would have enormous practical significance. Data from the AFDC period would be available in only a limited number of counties that may not be representative of the state. It might be especially difficult to separate the effects of W-2 from the effects of other changes in the child welfare environment—notably the strong trend over time toward increased out-of-home placements and greater reporting and substantiation of child abuse and neglect.

Primary potential impact 5: Health status

As already noted, the state still awaits a federal Medicaid waiver that would enable it to replace the Medicaid program for W-2 participants with a new W-2 health plan.[25] Even without this change, however, other provisions of W-2 might affect health care, for instance:

Low income is at least a predictor and possibly a cause of poor health among children, and family incomes could rise or fall under W-2.[26]

It is not yet clear whether eligibility for W-2 will confer automatic eligibility for Medicaid, or that a joint application process will exist. Health status could be affected if the process of achieving eligibility for Medicaid becomes more time-consuming and complex.

Parents who are working more hours under W-2 than they would have under AFDC may cut back on routine primary health care for their children because of their busier schedules, or, feeling less dependent and more empowered with respect to the health care system, may increase their use.

W-2 could influence the level of prenatal care received by women expecting their first child. Under the AFDC program, women with no other children who meet eligibility criteria become eligible in the seventh month of pregnancy, whereas under W-2, pregnant women will not be eligible for W-2 work programs until after the child is born. The lack of cash benefits throughout the last months of pregnancy may reduce the use of prenatal care.

W-2 could also affect the prenatal care of women with other children who are enrolled in W-2. These women must remain in a W-2 work program to receive benefits, except for the first 12 weeks after the child's birth. Alternatively, the greater self-reliance imposed by W-2 could positively influence the care a pregnant woman takes during her pregnancy.

Perhaps the most direct and sensitive indicator of the use of routine care, if the basic Medicaid program is not converted to a W-2 health plan, would be the percentage of Medicaid children below age 6 (children below that age are eligible for Medicaid up to 165 percent of the poverty line) who have received standard immunizations and Healthcheck screens (Healthcheck is the Wisconsin equivalent of the Early and Periodic Screening, Diagnostic, and Treatment program). Both indicators are available from Medicaid administrative records. We propose limiting evaluation of the use of basic and preventive services to Medicaid recipients before and after implementation of W-2 because data are available within the Medicaid program. A significant weakness of this approach is that takeup rates among low-income children may change as a result of W-2.

Assessing the impact of W-2 on the health status of newborns is possible for broader populations.[27] The most widely accepted indicator of the health of a newborn is birth weight, which appears on birth certificates, along with the mother's social security number. Thus it should be possible to compare the proportions of births to low-income women (with incomes below 200 percent of the poverty line) that have normal, low, or very low birth weights before and after the inception of W-2.

One significant limitation in a pre-post comparison of health care utilization among Medicaid children is that the state has increasingly emphasized the provision of routine health care. It will be hard to disentangle the effects of W-2 from these new administrative emphases. It may also be necessary to perform the birth weight comparisons separately for the major racial and ethnic groups identified on the birth certificate, since a change in the racial or ethnic composition of mothers in the state could influence birth weights independently of W-2.

Primary potential impact 6: Living arrangements and family structure of low-income households

W-2 could decrease the number of children in mother-only families for three reasons. First, if AFDC, as has

been claimed, has encouraged nonmarital births by providing financial support to parents without requiring work or worklike activities, W-2, by requiring work activities, may decrease them. Second, if AFDC has encouraged divorce or separation or discouraged marriage because eligibility has been easier for single parents, W-2's offer of services contingent only on the income and assets of the family, without regard to whether it contains one or two adults, could decrease divorces and increase the rate of marriage. Third, by eliminating the practice of providing benefits based on family size, W-2 could reduce the fertility of welfare recipients. On the other hand, if W-2 is perceived as offering attractive assistance to enter the labor force, the program could encourage women to have one child, whether in or out of wedlock, to obtain W-2 eligibility.

Determining the actual influence of W-2 on these and related possibilities will be difficult. W-2 administrative records will contain data on the number of people in the household, because child care copayments and (if implemented) premiums for the W-2 health care program will depend on household income relative to the poverty line. But these calculations do not require information on the actual composition of the family (the number of adults and their relationship to each other, the number of children, and the relationship of the children to the adults), and it is not certain that such information will be included in administrative records.

The likelihood of good administrative data on births is much higher. Wisconsin birth certificate documents record the birth order of the child and the marital status, zip code, and social security number of the mother. This information can be matched to income tax and wage records, allowing us to compare the percentage of nonmarital births to low-income women or in low-income areas, pre- and post-W-2. We are also able to compare later fertility—subsequent births—among recipients of AFDC and W-2.

Administrative records on marital status—other than the mother's status at the time a child is born—are unlikely to be available under W-2. Two possibilities for determining whether W-2 has any effect are (1) a brief survey of respondents' recent and current marital status and their marriage and divorce histories (to obtain information for the AFDC period) or (2) a match of the social security numbers of AFDC and W-2 case heads against social security numbers in divorce records, to gain a sense of the relative impact of AFDC and W-2 on divorce.

Because W-2 could affect childbearing and marital decisions by women not on the program, the full impact of W-2 on living arrangements and family structure is discernable only through a consideration of the broader low-income population, not just of those on AFDC or W-2. Whatever data are used, the sample would need to consist of families with minor children and with incomes below, say, 200 percent of the poverty line. Attributing any observed changes to W-2, rather than to broader cultural and environmental factors that influence marital and fertility decisions, will be challenging.

Secondary potential impacts

W-2 is a very broad intervention, and it could affect a wide variety of domains and populations. We view several potential consequences of W-2 as slightly less significant for the evaluation than the primary domains that we have nominated. They are generally less central to the purposes of the reform, likely to affect fewer people, or, as with health insurance coverage, may change little from the status quo after federal waiver possibilities are clarified. Among these secondary domains we would include homelessness, residential mobility, health coverage, and the living arrangements of children with disabilities. Space precludes their discussion in this article, and readers are referred to our extended analysis (see note 1). But these by no means exhaust the range of possible effects. Here, we list a few other possibilities for careful evaluation. These are, for adults:

- increased asset accumulation, because W-2 has different asset provisions from AFDC;
- decreased employment stability, because employers may be less likely to retain marginal employees;
- increased dependency among low-income families without children because, if their wages and employment stability decline, their food stamp usage might increase (food stamp usage for childless families that are not employed or participating in an employment program is limited to three months.)

For children, we should consider school readiness among entering kindergartners; the educational achievement of older children; and the rate and severity of juvenile crime.

Evaluations of any policy change of the magnitude of W-2 in a complex and dynamic environment will be imperfect. The approach we suggest contains several limitations, but so do other approaches. We believe this basic evaluation strategy can provide reasonably clear information on the impacts of W-2, and at a reasonable cost.■

---

[1]The paper upon which this article is based appears in full in "Evaluating Comprehensive State Welfare Reforms: A Conference," IRP Special Report no. 69, University of Wisconsin–Madison, March 1997.

[2]W-2 and the two transitional programs are briefly described in this issue of *Focus*, p. 2. In September 1996, 5,182 AFDC recipients in Milwaukee County, out of a caseload of 31,000, received initial sanctions under Pay for Performance. G. Schuldt, "AFDC Sanctions

Rise in County," *Milwaukee Journal Sentinel,* September 12, 1996, p. 3b.

[3]See also M. Cancian and B. Wolfe, "Outcomes of Interest, Evaluation Constituencies, and the Necessary Trade-offs," in this issue of *Focus.*

[4]For a more comprehensive review of these other options, see the articles by R. Haveman, by I. Piliavin and M. Courtney, and by G. Cain in this issue of *Focus.*

[5]See V. J. Hotz, "Designing an Evaluation of the Job Training Partnership Act," in *Evaluating Welfare and Training Programs,* ed. C. F. Manski and I. Garfinkel (Cambridge, MA: Harvard University Press, 1992); T. D. Cook and D. T. Campbell, *Quasi-Experimentation: Design and Analysis Issues for Field Settings* (Boston: Houghton Mifflin, 1979); I. Garfinkel, C. F. Manski, and C. Michalopoulos, "Micro Experiments and Macro Effects," in Manski and Garfinkel, ed., *Evaluating Welfare and Training Programs.*

[6]In particular, different states have different levels of income at which taxes must be filed, and the extent to which low-income families file tax returns probably differs substantially across states.

[7]Cook and Campbell, *Quasi-Experimentation.*

[8]We do not rely on the Urban Institute survey as a primary data source because we believe it is desirable to have annual data on the same families over a period of more than five years. The Urban Institute project is described briefly in *Focus* 18, no. 1 (special issue 1996), pp. 4–5

[9]D. T. Campbell, "Can We Be Scientific in Applied Social Science?" in *Evaluation Studies Review Annual,* ed. R. F. Conner, D. G. Altman, and C. Jackson, 9(1984): 26–48.

[10]Ethnographic evidence suggests some nonresident parents only paid $50/month of child support because additional amounts did not benefit the family. See K. Edin, "Single Mothers and Child Support: The Possibilities and Limits of Child Support Policy," *Children and Youth Services Review* 17 (1995):203–30.

[11]Federal income tax data have historically been very difficult for researchers to access because of confidentiality concerns, but researchers in Wisconsin have been given access to Wisconsin state income tax data. For analyses using Wisconsin tax data, see E. Phillips and I. Garfinkel, "Income Growth among Nonresident Fathers: Evidence from Wisconsin," *Demography* 30 (1993):227-41; D. R. Meyer, "Supporting Children Born outside of Marriage: Do Child Support Awards Keep Pace with Changes in Fathers' Incomes?" *Social Science Quarterly* 76 (1995): 577–93. M. David discusses a related approach in more detail; see "Monitoring Income for Social and Economic Development," in "Evaluating Comprehensive State Welfare Reforms."

[12]M. Cancian and D. R. Meyer, "A Profile of the AFDC Caseload in Wisconsin: Implications for a Work-Based Welfare Reform Strategy," IRP Special Report no. 67, University of Wisconsin–Madison, 1995; D. Friedlander and G. Burtless, *Five Years After: The Long-Term Effects of Welfare-to-Work Programs* (New York: Russell Sage Foundation, 1995).

[13]This analysis enables the researcher to control for factors that affect individuals of all income levels, but not to distinguish between W-2 and other changes that affect low-income individuals only.

[14]In 1995, a family with one child with earnings of less than $8,250 and rent of $300/month that does not include heat would receive $724 by filling out the one-page form. If this family owned a home and paid $1,500 in property taxes, they would receive $1,160. Full-year AFDC recipients are not eligible for this credit.

[15]Administrative records of child support income for most families will exist for the W-2 period. Unfortunately the child support administrative data system that existed during the AFDC period had several problems, limiting the comparability of these data across the two periods.

[16]Because public employee unions will be paying close attention to employment trends in local government, any potential impact may be more pronounced in nonprofit organizations than in local government.

[17]See C. F. Citro and R. T. Michael, eds., *Measuring Poverty: A New Approach* (Washington, D.C.: National Academy Press, 1995) for a critique of the official measure and a description of the NRC measure. *Focus* 17, no. 1 (Summer 1995): 2–28 has an extended report upon the NRC measure.

[18]See D. R. Meyer, "Child Support and Welfare Dependency in Wisconsin," unpublished Ph.D. dissertation, University of Wisconsin–Madison, 1990; M. Smiley, "Private Needs and Public Welfare: Rethinking the Idea of Dependency in a Democratic Culture." Unpublished ms., University of Wisconsin–Madison, 1996; P. Gottschalk and R. A. Moffitt. "Welfare Dependence: Concepts, Measures, and Trends," *American Economic Review* 84, no. 2 (1994):38–42; D. R. Meyer and M. Cancian, "Economic Well-Being Following an Exit from AFDC," paper presented at the Association for Public Policy Analysis and Management Research Conference, Pittsburgh, October 1996.

[19]See also the discussion by K. F. Folk, "Evaluation of Child Care Services under W-2, Wisconsin Works Program," in "Evaluating Comprehensive State Welfare Reforms."

[20]Any increase in raw numbers could merely denote more children in child care, not child care quality.

[21]On measures of child care quality, see W. T. Gormley, Jr., *Everybody's Children: Child Care as a Public Problem* (Washington, D.C.: Brookings Institution, 1995).

[22]We consider the health status of children below. Researchers are only beginning to examine the effects of welfare reform on measures of child well-being that go beyond the traditional measures of health status, child abuse and neglect, substitute care, and educational achievement; see A. Collins and J. L. Aber, "State Welfare Waiver Evaluations: Will They Increase Our Understanding of the Impact of Welfare Reform on Children?" Working Paper of the National Center for Children in Poverty, Columbia University School of Public Health, 1996. See also M. Courtney, "Welfare Reform and Child Welfare Services," in "Evaluating Comprehensive State Welfare Reforms."

[23]Wisconsin statutes officially prohibit findings of neglect solely for reasons of poverty, but a finding of neglect because a parent did not take full advantage of all opportunities available in W-2 would be possible.

[24]In this case, comparing increases among low-income samples potentially affected by W-2 with increases among middle-income samples may not be very informative, because the child welfare system predominantly affects low-income people.

[25]Evaluation issues surrounding such a change are discussed in Kaplan and Meyer, "Toward a Basic Impact Evaluation," in "Evaluating Comprehensive State Welfare Reforms."

[26]P. L. Geltman, A. F. Meyers, J. Greenberg, and B. Zuckerman, "Commentary: Welfare Reform and Children's Health," special report included in *Health Policy and Child Health* 3, no. 2 (Spring 1996).

[27]The issue is discussed by G. Sandefur and M. Martin in "Evaluating the Impacts of W-2 on Family Structure and Maternal and Child Health," in "Evaluating Comprehensive State Welfare Reforms."

Invited comment: Difficulties with a pre-post design

The papers written for the conference were of very high quality. The conference proceedings were quite interesting and potentially of great use to all those interested in evaluating welfare reform. I felt, however, that the central thrust of the papers which provide proposals for evaluation designs was misguided. The major papers advocated designs which were essentially of the pre-post or interrupted time-series type of evaluation. Much attention was given to, and considerable imagination shown in the development of, use of combinations of Wisconsin's unique tax data base and administrative records in order to develop estimates, based on the pre-W-2 period, of what would have happened to low-income persons' employment and earnings had W-2 not been instituted. These predicted outcomes would then be compared to the actual experience under W-2 in order to estimate its impact.

These proposals overlook one of the most important observations provided in Glen Cain's article in this issue: an evaluation of impact requires a counterfactual (what would have happened to participants in the absence of the program). The old AFDC program does not appear to be a realistic alternative to which to compare W-2, since the new federal law ends entitlement programs in this domain. It is highly unlikely Wisconsin will return to a program structure like the old AFDC program, so that the elaborate reconstruction of experience in Wisconsin under AFDC would not provide information relevant to the policy evaluation objectives.

A further problem with the pre-post or interrupted time series designs is their unreliability as a means of providing a counterfactual.[1] The problem is that these types of evaluation designs depend on a period of strong, stable trends prior to the program initiation, and/or an ability to model the process over time which has been shown to be highly reliable in predicting further outcomes. The record of using this type of model to predict AFDC caseloads is horrendously bad. To understand why this type of caseload modeling is problematic one needs only look at the recent caseload record in Wisconsin, as well as elsewhere in the country: sharp rises in the early 90s and sharp falls in the last two years.

Putting these two previous points together, the evaluation designs proposed would use a highly unreliable method—pre-post or interrupted time series—to generate a counterfactual employment and earnings estimate, which, even if it could be accurately estimated, would not represent a realistic, relevant alternative to W-2; the result would be highly questionable estimates of an irrelevant alternative. Such an evaluation exercise would not lead to policy improvement.

I suggest consideration of an approach to evaluation of W-2, and welfare reform efforts in other states, which has two general elements: (a) monitor the economic and social well-being of the low-income population; and (b) design and implement one or several random-assignment experiments to test critical components of the W-2 package. Let me expand briefly on each element.

Monitoring. If there is significant deterioration in the economic and social circumstances of the low-income population, there will be a call for the state or federal government to take steps to alleviate the problems. This will be the case whether the worsening conditions are caused by general economic factors (a recession) or by deleterious effects of changes in the welfare system. Well-designed monitoring will help to focus on the precise nature of the problems, e.g., inability to get jobs or jobs lost because of economic downturn, transportation barriers, child care limitations, or fundamental inability to hold a job. The "New Federalism" project being run by the Urban Institute is attempting a multistate monitoring effort, and Wisconsin is one of the states on which it will gather data. This will be a good base to build on, and efforts are already being made to supplement the sample for Milwaukee.

Testing critical components. The W-2 reform is a complex package of program elements. There are a whole host of questions regarding the choices made for each element of the package, and papers at the conference outlined many of these. For almost any one element of the package, critical issues could best be addressed through systematic variation in the program characteristic and random assignment of participants to one variant or another. For example, the expansion of child care benefits is one of the most significant, and potentially expensive, aspects of W-2. By systematically varying the level and character of the child care subsidy and using a random assignment design, following participants and their children, one could learn a good deal about the costs and benefits of child care. Another area ripe for experimental evaluation is the subsidy to employers for the trial jobs segment of the program. Information derived from these types of rigorous evaluations of program elements would be useful not only to Wisconsin but to all the states struggling with welfare reforms. In contrast, an assessment of the W-2 package as one piece, compared to the pre-welfare-reform period—even if it could be effectively done, which I seriously doubt—would be of limited usefulness to others and would provide little guidance to Wisconsin on how it might alter its program to be more effective.

Robinson G. Hollister
Joseph Wharton Professor of Economics
Swarthmore College

---

[1]For a review of experience with these types of designs and their problems see R. Hollister and J. Hill, "Problems in the Evaluation of Community-Wide Initiatives" in J. Connell, A. Kubisch, L. Schorr, and C. Weiss, ed., *New Approaches to Evaluating Community Initiatives* (Aspen, CO: The Aspen Institute, 1995). For concrete evidence on the persistence of bias using such methods in examples of welfare reforms see D. Friedlander and P. Robins, "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods," *American Economic Review* 85, no. 4 (Sept. 1995): 923–37.

The actors, decisions, and complexities of welfare reform: The W-2 example

Elisabeth Boehnen

Elisabeth Boehnen is Database Administrator at IRP.

The next generation of welfare reforms breaks new ground in program design and management, and evaluating them demands a larger assemblage of analytic tools and more rigorous methods than in the past.[1] One important method is a process analysis. For an evaluator who seeks to understand why a program succeeds or fails, or for a state agency that wishes to replicate a program, it is essential to look carefully at the program's structure and principal actors—to describe its operation fully, to identify problems and their potential solutions, and to document the qualifications, training, and job descriptions of personnel. This article offers a highly selective overview of program design and management structure using Wisconsin Works (W-2) as an example. A large number of people and activities are involved in a program as complex as W-2; this is an attempt to narrow the focus to describe only those points where, we believe, key decisions are made between the program representatives and the participants. It does not try to set any additional priorities for evaluative purposes nor suggest any formal evaluation design.

W-2 and similar reforms are best conceptualized as a series of consecutive interactions, or episodes, between the participant and the agency. These events, activities, and decisions are not independent; the outcome of earlier episodes determines, or should determine, the participant's consequent experience in the program. If problems arise in the execution of events or in the appropriateness of the decisions being made, the effects can ripple throughout the system. Queues (waiting lists) can develop, participants can be allocated to the wrong tracks, and services may not be delivered or coordinated in a timely fashion (or at all). Conversely, if participants go through the steps as they were laid out by the program designers and if services are provided promptly, then the system can be viewed as working in the way it was intended.

The decisions made by the program actors about, or with, program participants are the glue that holds together the substance of W-2, as they would any similarly complex reform. Who makes the decisions and how are they made? Are they routine and made according to rigid protocols, or are decision makers accorded full professional discretion? First we describe the roles of some of those responsible for operating W-2 and then lay out the critical interactions that affect both the institutional actors and the participants.

Key institutional actors

What we've learned, in the handful of counties where we've implemented the changes, is that you just can't say 'domine, domine, *you're all family independence specialists now.' —Robert Lovell, Michigan Family Independence Agency*

The six roles discussed below are those primary positions where agency staff makes a decision with or about a participant.[2] In each W-2 agency, their exact responsibilities and the flow of participants through the system may look a little different. Figure 1 locates these positions in the flow of program activities and identifies several decision points associated with each.

Front desk receptionist

The receptionist performs the initial gatekeeping and "triage" function of the W-2 agency, which in some places is called a "Job Center." This staff member is responsible for greeting "customers" and directing them to the appropriate place in the Job Center to "complete their business"—for example, job-seeking customers are referred to a self-service "Job Net" computer. From those who are seeking W-2 services, the receptionist also gathers initial demographic data, creates a database tracking report, provides preliminary employment information, screens for eligibility for an expedited appointment for food stamps, and directs the customer to others in the agency or to resources outside the W-2 agency.

This is the first program representative whom the potential participant encounters at the agency. The title suggests a clerical support position. However, there appears to be a substantial discrepancy between the term "receptionist" (defined in the *American Heritage Dictionary* as "an office worker employed chiefly to receive callers and answer the telephone") and the extraordinary responsibilities of this person. In planning a process analysis, one would need to conduct a thorough examination of the position, responsibilities, and effects of the receptionist on potential participants. How much discretion and latitude does the receptionist have in deciding who moves where in the program? How much employment information is presented and how complete is it? This is the first step in tracking the potential participant; is it clear what information should be gathered? How much pressure is placed on the receptionist to divert clients from participating further in the program, and to whom is he or she accountable?

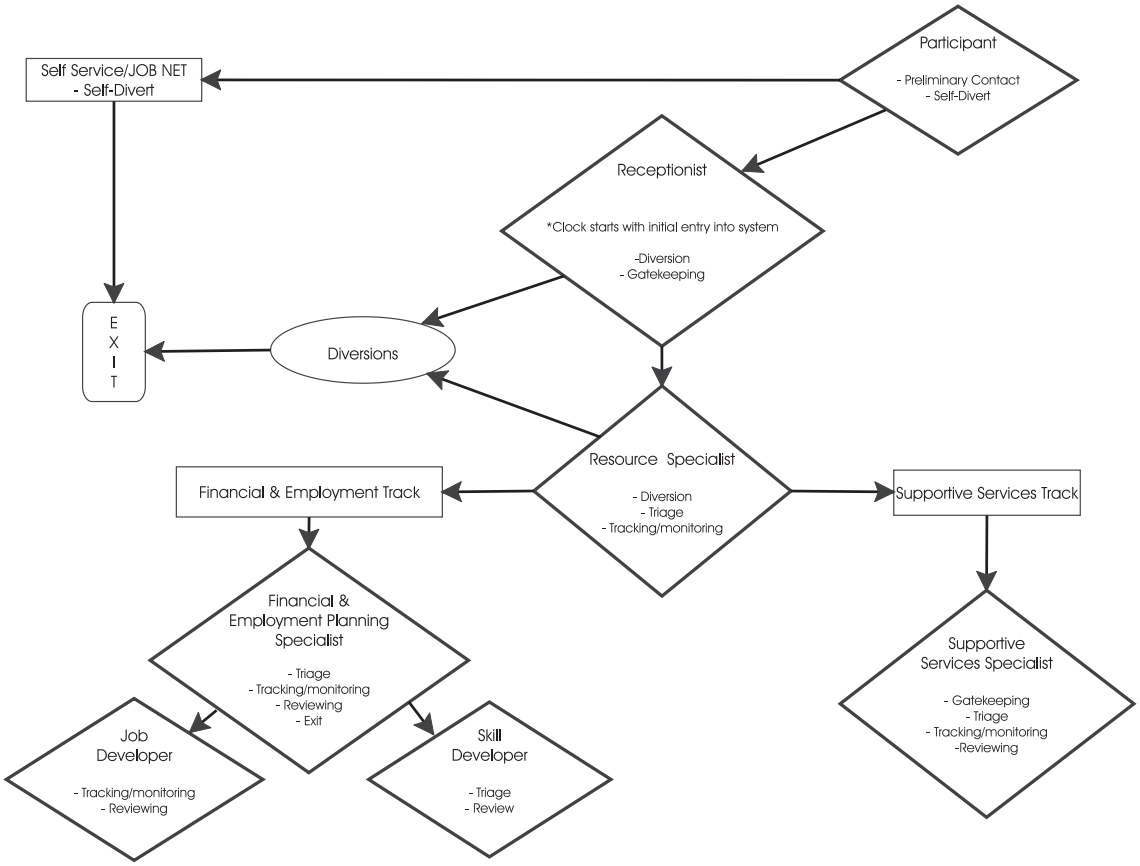

Figure 1. Key decision points for W-2 actors.
Note: For more complete information on the complexities of the W-2 program, see Thomas Corbett, "Wisconsin Works: A View from the Ground," in "Evaluating Comprehensive State Welfare Reforms: A Conference," IRP Special Report no. 69, University of Wisconsin–Madison, March 1997.

Resource specialist

The resource specialist is the first agency representative with whom the participant interacts in a substantive manner; in this encounter, the experience with the program will be shaped. This person also performs key gatekeeping and triage functions—for example, referrals, diversion of potential participants away from W-2 to other resources and opportunities as appropriate, and preliminary evaluation for W-2 services.

Given the critical location of this role in the W-2 program, the quality of the interaction between the potential participant and the agency representative is crucial, and should receive much attention in any process analysis. The resource specialist exercises considerable professional latitude and judgment. For example, the diversion function can be conducted with varying degrees of diligence and sensitivity. Some may approach the task as an opportunity to divert only those who evidence a clear capacity for independence, thus erring on the side of caution; others may aggressively divert potential participants as a way of deterring frivolous use of public assistance, thereby erring in the other direction. The same person may meet the responsibilities of the position in different ways, depending on organizational needs, aggressively diverting people when caseloads exceed capacity and doing so less aggressively at other times.

Financial and employment planning specialist (FEP)

The primary role of the financial and employment planner is to "explore with the potential applicant the options available to meet employment goals using personal, family and community resources: to make a preliminary determination of whether or not W-2 services are appropriate and, if they are, to facilitate the maximum degrees of self-sufficiency."[3] The FEP is responsible for many case management functions: determining eligibility, completing initial screening of employment skills and job readiness, helping the participant to develop a financial plan, setting short- and long-range self-sufficiency goals, referring to community services and other employment opportunities, monitoring performance, and making tracking decisions that determine the participant's route through W-2.

Again, this is a position in which there is much responsibility and latitude in decision making. The FEP is at the heart of a program such as W-2. It is the FEP who is responsible for getting the participant employed and off the caseload, for exploring the "potential" of the participant, and providing him or her with employment options, training, and other needed services. At the same time, the FEP must also communicate the absolute seriousness of the work obligation and the consequences of noncompliance. From the standpoint of a process analysis, there are

many questions regarding the planners and the decisions that they make with, and about, the participant, in order to determine whether the program is actually working as it was intended. How are these staff trained to deal with the wide variety of people they will encounter, and the diverse roles they must play? What criteria have been established to assist them in identifying such problems as substance abuse or severe child and family problems? If the caseload of the FEP is too large, how does that affect the quality of the time and information afforded each participant?

The three other roles described here are ancillary to the FEP, but are important in examining the program as a whole. The *supportive services planner, job developer, and skill developer* require rather similar abilities. These program representatives must be able to work with diverse people and systems—service providers, employers, and participants. They must be able to motivate, plan and even adjudicate. Above all, they must be able to negotiate among individuals. The more successful will use innovative techniques in maximizing available resources and employment opportunities.

Supportive services planner

Supportive services specialists determine eligibility for services such as child care, the W-2 health plan, advanced EITC payments, food stamps, child support enforcement activities, transportation, and job access loans. Using a case management approach, the supportive services specialist links participants and community service providers.

Job developer

The job developer builds relationships and partnerships with community businesses to promote W-2 services and participants, identifies employment opportunities (subsidized and unsubsidized), and negotiates contracts with businesses. Job developers are responsible for assisting participants in identifying their career goals, educational skills, employment history, job opportunities, and potential barriers to employment. They also monitor participants and programs.

Skill developer

Staff in this position develop and implement training both in the classroom and at training sites. They make individual placements of W-2 participants, or place them in workshops or in a combination of workshop, community service job, or trial job. They survey employers to develop a training curriculum that not only meets employers' hiring and retention needs, but also helps the participant in such areas as pre-employment skills, life skills, occupational skills, attitudes, appearance, self-esteem, and parenting.

These descriptions are brief and simplified, but suggest that the people performing the jobs will require skills in different areas of expertise in order to perform their duties competently. Furthermore, each position allows great independence and flexibility of decision making, with potentially very large consequences for the experience of the participant and the efficiency of the agency.

Key procedural steps and decision points

One can evaluate the actors, gather data, and make assumptions regarding the program from the perspective of the service deliverers. But to fully understand the process, one must also be able to identify the points at which important decisions are made regarding each participant and where things may go in directions not intended by program designers and operators. The kinds of questions we ask here about W-2 should be asked about any complex and ambitious state welfare program in which there is a potential for great local variation in practices and operations.

First contact

What is the substance of the message that the potential applicant for welfare or employment first hears, possibly even before he or she walks in the door? What is its quality—is it hostile or supportive? These messages are an early determinant of whether the participant undertakes the program.

Gatekeeping

Gatekeeping is essentially the responsibility of the receptionist, and encompasses the participant's first experiences at the agency. This is an extremely hectic period, with much information being exchanged and difficult decisions being made. How are expectations established, data gathered and verified, and the appropriate set of services put together? As with first contact, gatekeeping sets the tone for subsequent experiences. For example, the participant's experience will be very different if the receptionist provides helpful information and has the time to assess her potential needs rather than shunting her off to the Job Net computer without much additional assistance.

Diversion

"Diversion to work" is a prominent feature of many recent welfare reform programs in which attachment to the labor force and diversion from cash assistance are primary goals. Diversion points (see Figure 1) are those places in the program where the staff is urged to encourage the participant to be self-sufficient and use resources other than those provided by the agency. This is admirable in theory. But a drop in caseload activity does not mean that participants are being served appropriately. They may be diverted by the agency staff, or in some cases do not bother to apply because they hear from friends or neighbors that it is "too much trouble." Diver-

sion is a process that needs to be thoroughly examined. Those participants who are diverted by staff should be followed up, to see if they are obtaining the services they need from the community. Those who are eligible for agency services, but are not taking advantage of them, should also receive attention.

Triage and plan development

This is a very complex set of decisions related to the determination of eligibility, the assessment of participants' current and potential needs and abilities, the movement of the participant into an appropriate track or group, and the organization of a specific set of activities and experiences. Basically, triage is the act of deciding "who gets what, when they get it, where they get it, and how they will get it." All of the agency positions described above engage in some form of triage (see Figure 1). The receptionist makes initial decisions regarding where to send the potential participant. The FEPs have a great deal of latitude in making decisions regarding eligibility, employment potential and track assignments, and the provision of additional services. The supportive services planners are also involved in eligibility determination and deciding the types of services the participant actually needs. The job developers and skill developers play a very large part in determining where a participant will be, and in what type of job or training.

Tracking, monitoring, and transitions

The participant's basic experience in W-2 will be the employment activities encompassed in one of several tracks—trial jobs, community service jobs, and W-2 Transitions. The decision about where a participant will start in W-2 has great consequences. Slotting participants in a track that is too demanding may lead to failure and employer dissatisfaction, while assigning them too low may not set high enough expectations. As participants are "sorted and tracked" according to policy, what happens if they are misallocated? Staff discretion, in conjunction with the agency goal of diversion, gives even greater importance to the decisions regarding the job track. The transition between tracks, between jobs, and even from one form of time limit to another invokes questions about those who are empowered to make such decisions and about procedures to assure that transitions are smooth and consistent.

Monitoring of participants is central, since benefits are tied to work or activity compliance, and other supportive services such as child care and, potentially, health care may depend on continuing copayments by participants. Except for the receptionist, all of the other W-2 staff discussed here are responsible for some type of monitoring or tracking. Will it be clear to frontline workers what data they need to obtain from participants, and will they have the time to track them as they are required to do? How much *leakage* will there be—how many participants will "fall through the cracks," fail to move from one worker to another or one component to another, or not receive the services they need or are entitled to? Where will the *queuing problems* occur? If participants are not efficiently monitored, program administrators will fail to see where bottlenecks in the program are developing.

Review and adjustment

Under W-2, all activities and assignments are time-limited (two years in a given track, with an overall lifetime limit of five years). Participants may have their status reviewed and rereviewed as they are reassigned to different program tracks or as they move up the ladder toward self-sufficiency. If problems occur, or if they do not comply with program rules, they may face more reviews and intensive case management. Here again workers' decisions regarding extension of time limits, continued eligibility, or terms of compliance are discretionary.

Exit, follow-up, and recidivism

Participants can exit at any time. It is still unclear how recidivism will be minimized and what other supports families will receive in the labor market (for instance, access to health care). Some exits may be more desirable than others (employment, as opposed to giving up), and the availability and quality of work support services may determine recidivism rates or well-being over time. Agency staff must have the time and the incentive to monitor and follow up, particularly after a participant is placed in a job.

The complexity of the W-2 program makes it very difficult for the state to prescribe an administrative approach. Frontline workers will have to deal with the resulting ambiguity and vagueness on their own. Given the many points that allow for nonroutine, discretionary activities, some participants may be overlooked, may not know where they are supposed to be or what they are supposed to be doing, and may not be receiving the appropriate services.

A participant's-eye view

Any substantive change in the culture of an agency or the organization and delivery of a program will be reflected in participants' experiences. What are the qualities that appear to be most important if a W-2 participant is to experience the program as its originators intended?

Clarity. The participant should understand exactly what the program options are. Are these messages being communicated clearly and concisely to participants? Is there ambiguity, or inconsistency, or role confusion among personnel that would lead to miscommunication?

Celerity. If the agency does not provide the participant with services and information as quickly and efficiently as possible, the meaning of the program may become unclear and its seriousness be questioned. Furthermore,

the participant's situation may change, causing further problems in the types of services required and in any further interactions with the program.

Consequences. There must be clearly explained consequences if either agency or participant fails to perform. Agencies must be held accountable for making sure that participants receive the information and services they need. Participants need to know exactly what is expected of them and what will happen when they make particular choices.

Seamlessness. The program should be seamless for the participant, who should not be shunted from one office to another or from one placement track to another without clearly understanding what is happening and why. Not only should agency staff know what each staff person is required to do, but case management should be tracking participants efficiently.

Accuracy. Most decisions in W-2, whether prescribed or discretionary, have solid factual content. Questions such as "What track am I on? How much do I have to pay for child and health care? Am I in compliance with the rules?" should receive accurate and unambiguous answers.

Quality. For child care, job search assistance, and job development and matching, it is possible to set at least minimal levels of quality. Participants should not have to question the quality of any services that are provided them after they walk in the front door.

Monitoring progress

If administrators and policy makers are to prevent or correct the problems that may occur with complex programs, they must know the numbers. Are the caseloads dropping? How many people are in each track? They must also know what is happening to participants. Are they working; what kinds of jobs do they have? Are the supportive services adequate? Are participants able to handle the additional responsibility of work plus family? Some of these are "soft" issues, and answers cannot be derived from administrative data, where most program analyses will start.

For administration and management, a well-designed, automated case management system capable of tracking individuals and families over time and across programs is essential. Currently, no state has a system with the ability to handle complex programs such as W-2, but California, Oregon, Massachusetts, Indiana, Illinois, and a few others have some components in place. Wisconsin is also working to upgrade the state system. In the interim, it is essential that special case-tracking efforts be made. Generic surveys may have to be put in place to keep track of key dates, outcomes, decisions, and service delivery—perhaps on a sample basis, for reasons of cost and feasibility.

Some decision points in the process may be so important that we should consider constructing surveys or questionnaires around those points. If the distribution of new participants across the W-2 tiers radically deviates from prior expectations, we will want to find out why.

Another approach to examining program efficiency and operation is to conduct interviews with front-line operations personnel. Structured interviews can document how staff view the program and their role in it, and what problems they perceive. Unstructured interviews with management permit exploration of problems unforeseen by the evaluators.

Finally, the evaluators (or program designers) should establish expectations about what is supposed to happen and how it is supposed to happen. How quickly, for instance, should participants be able to find a child care slot? A set of expectations creates norms against which actual performance can be compared and adjusted. As the actual practice of running programs for disadvantaged families becomes more diverse, the need to accurately describe program intent and how operations match that intent will increase.

As this deconstruction of W-2 should make clear, a comprehensive program evaluation does not merely ask "does X work?" It requires serious political dialogue to determine what might constitute success in the new program and a thorough examination of its implementation to determine whether the actual program reflects the original intent. Equally as important is whether the participant's experience reflects the intentions of program designers and administrators. Too often and too easily, evaluators have thought they understood what was being implemented and ended up assessing labels rather than realities.■

---

[1]The author is grateful for the material assistance of Thomas Corbett in putting together this article, which draws heavily upon their joint paper, "Wisconsin Works: A View from the Ground," in "Evaluating Comprehensive State Welfare Reforms: A Conference," IRP Special Report no. 69, University of Wisconsin–Madison, March 1997.

[2]This is in no way an attempt to present all the roles that are necessary to administer a complex program such as W-2. These job descriptions are based on the Fond du Lac County, Wisconsin, Work-Not-Welfare program (the precursor program to W-2; Fond du Lac is one of the first counties where W-2 has already been implemented), and the Wisconsin Works Plan prepared by the Dane County Department of Human Services.

[3]Dane County Department of Human Services, Wisconsin Works Plan, Section 28 (October 1996).

Process evaluation for state welfare reforms

Karen C. Holden and Arthur Reynolds

Karen C. Holden is Professor of Public Affairs and Consumer Science and Associate Director of the La Follette Institute of Public Affairs and Arthur Reynolds is Assistant Professor of Social Work at the University of Wisconsin–Madison. Both are IRP Affiliates.

A process evaluation attempts to illuminate the administration of a program and the behavior of its clients from its initial implementation to the observation of its outcomes.[1] The evaluation documents the way the program is put into practice. It measures the form and duration of the program, examines variations among service agencies and client groups, and identifies intervening and confounding activities. The systematic documentation and analysis of this process enable evaluators to draw causal links between variations in the mobilization and delivery of program services and measured outcomes.

An evaluation's process component provides policy makers with the guidance they seek from an outcomes evaluation—whether and how to extend or modify programs in time, space, or coverage. Should the program as a whole or an individual component be continued as it is, be modified, or be terminated? How can it be extended to different places—for example, should other states adopt welfare program components similar to those of Wisconsin's welfare program in expectation of similar results? Should it be extended to different populations—for example, to poor noncustodial parents? To measure only outcomes—simply assuming that the full program was implemented as intended and regarding implementation as an unobserved "black box"—provides no guidance on modifications and extensions that might improve those outcomes.

Key issues in process evaluation

In a process evaluation, three main questions are addressed:

1. Are the administrative services and resources that are central to the success of the program in place?

2. To what extent are program services being delivered to the target population in the manner specified in the program design?

3. How do program services and administration differ across program sites, and how do these differences affect service delivery and coverage?

For at least three reasons, evaluating implementation is essential to assessing social programs and services:

1. *To validate a fundamental assumption of impact (outcomes) evaluation.* A crucial reason for conducting a process evaluation is to learn whether the program has been delivered as intended to the target population, lest the magnitude of the outcomes be erroneously attributed to diminished program effects rather than accurately attributed to deficient implementation.[2]

Because implementation problems usually result from incomplete delivery of programs or the implementation of the wrong program for the targeted population, outcome evaluations may often unknowingly underestimate effectiveness. For example, the evaluation of the initial Head Start Program for preschoolers and their parents greatly overestimated, owing to a lack of process documentation, the extent to which health services, family services, and staff training were available and provided. The authors of the evaluation report found no program effect, concluding that "it was impossible to know in detail the actual program that these children experienced."[3] Evaluations of the implementation of Title I educational block grants illustrate a complementary problem—that, even if resources are available and services are delivered, they may be delivered nonuniformly and may fail to reach the population that is most in need.[4]

2. *To help explain why a program worked or did not work.* There are three general reasons that programs do not show their intended effects: inadequate program design or theory, poor program implementation, and inadequacies in the evaluation research design or measures. Process evaluations help distinguish among these reasons, for example, enabling investigators to clarify the precise program elements and features that were the source of the effects of the program. This is especially important for complex, multiple-component programs in which implementation may vary across geographic areas.

3. *To promote program replication and utilization of the evaluation*. If the workings of a program and its elements in different sites are understood and documented, its essential operative features can be identified and disseminated for use in other settings. And because a major goal in evaluation is the utilization of findings, implementation evaluations also encourage collaboration between evaluators, program designers, and other stakeholders in the reform endeavor.

Evaluating innovative, large-scale social service programs

In innovative and untried social service programs, it may take many years to fully quantify the program's imple-

mentation and its effectiveness for families and children. Such gradual implementation raises questions about evaluation design. Should the emphasis be on process evaluation alone, impact evaluation alone, or some combination of the two? On the one hand, program administrators would like quick and early feedback, in order to alter program services as soon as possible to address unexpected results and unwanted side effects. On the other hand, particular interest groups would like evaluation resources devoted to a "true" test of a fully implemented program. We propose a middle ground: to conduct a long-term outcome evaluation and to give attention to documenting the timing and type of services received by program clients in a way that also allows for short-term feedback to program managers. Although this strategy may alter program administration early in the evaluation period, it also assures that the outcome evaluation is examining the program intended.

It is also very difficult to conduct comprehensive outcome evaluations of large-scale programs without a careful mapping of the character of the treatment across program sites. In statewide social service programs, there is not a single regime introduced at one moment in time, but many different treatments introduced across counties or agencies, begun simultaneously and then gradually modified.

Evaluations of federal block grant programs have indicated that global impact evaluations that collapse analysis across sites are not only less likely to measure a significant program impact but are often misleading, because the program is administered differently in different sites (often legitimately, due to local needs).[5] One useful approach is to study variation in both program implementation and response to the program at each of a number of sites. The evaluations may then be pooled using standard meta-analysis techniques to provide a general picture that also identifies the effect of intersite differences such as different populations and service components.

Process data can also illuminate unintended program consequences.[6] Many proposed state welfare programs are time-limited transition-to-work programs. Although the threatened loss of program services may move clients into the labor force more effectively than time-unlimited programs do, several unintended consequences may occur:

If the supply of good child care providers is very limited, program incentives to enroll children in the least expensive care available could place many children at significant risk of developmental problems.

The rapid transition into full-time work may discourage parents from enrolling their children in compensatory programs, some of them part-day programs, that have proven effectiveness (e.g., Head Start, special education interventions).

Program participation may affect the amount and quality of parents' participation in school and in children's educational outcomes.

Employers may not provide sufficient training for program participants if the amount of training an employee needs goes beyond the tax incentives provided to the employer.

Examining process in Wisconsin Works

Wisconsin Works (W-2) is far more complex in its administration and the services received by participants than was Wisconsin's Aid to Families with Dependent Children (AFDC) program. (The basic features of W-2 are outlined in this issue, p. 2.) Also in contrast to AFDC is the intent of W-2—to change the culture of the welfare population and, as Lawrence Mead describes it, "to control the lifestyle of the adult participants."[7] W-2 does this in part by setting conditions for the receipt of aid and imposing strong administrative suasion and sanctions on participant behavior. Continued eligibility requires, for example, that W-2 clients cooperate in efforts to establish the paternity of the dependent child, furnish the W-2 agency with information the agency determines to be necessary, and make a "good faith effort," in the agency's eyes, to obtain employment. W-2 agencies are likely to operationalize these requirements differently, with different behavioral consequences for participants and different program outcomes.

W-2 is also a block-grant program whose effectiveness will depend in large part on the availability and coordination of local community resources. Local agencies managing program components may not be under the direct control of program managers. They may be private or public, those traditionally engaged in job search (e.g., temporary service agencies) or those offering educational or family services (e.g., Head Start agencies). These agencies may organize administrative tasks differently, implement program components in different ways and on different schedules, and have different levels of expertise in providing job counseling, training, and placement services, child care, or health benefits.

The latitude allowed W-2 agencies in administering the program is likely to result in wide variation in program attributes and sanctioning policies. This variation is both a challenge and an opportunity to evaluators. On the one hand, the variation among counties means that individuals participating in W-2 over the same calendar period will in fact be participating in different programs. In a statewide evaluation, program effects in some counties may be obscured. On the other hand, geographic and time-related variation provides a laboratory in which to evaluate the relative importance to outcomes of individual program components, promising stronger outcome effects and more useful policy recommendations.

Attributing success under Wisconsin Works. Even under the optimistic assumptions of complete and successful program implementation and appropriate identification of comparison groups, changes in the health, employment, or family status of program participants cannot be attributed to particular program components without some understanding of their structure, coverage, and extent. Changes that are attributed to W-2 will be the result, in part, of the resources available locally and mobilized for particular participant groups, and of the success of participants in using those resources to move toward greater self-sufficiency. One purpose of a successful outcome evaluation design is to separate the availability of resources from the client's willingness to use those resources. Implementation of W-2 may stimulate the expansion of community resources to meet the greater demand, a program outcome that is likely to influence the behavior of other groups in the community. Process evaluation can provide insight into the role that each of these factors may play in the differences in success among sites and participant groups.

Process and outcome evaluation

The ultimate goal of a process evaluation is to be able to link outcomes with program implementation, coverage, and type and duration of treatment.[8] Data should make it possible to assess (1) the extent to which program administrative components are implemented, (2) the coverage and types of services provided to clients, and (3) the key components of difference from a particular counterfactual, such as the previous policy or programs experienced by other comparison groups. The method of gathering the data and what data must be gathered will depend in part on the design of the outcome evaluation.

In the classic randomized experiment, program elements to which individuals are assigned are generally well defined, individuals are assigned to treatment or control groups at a specific point in time, and their participation is monitored (see also the article by Cain, in this issue). An experimental design integrates process and outcome evaluation, since the delivery of program services and client selection is designed explicitly to meet evaluation needs. Causal inference is based on measured outcomes and known treatment differences between the randomly assigned groups.

A cross-state comparison proposes one or a set of states as the counterfactual, seeking to attribute differences in an outcome measure to differences between welfare programs in the state under study and the counterfactual states (see also the article by Courtney and Piliavin, in this issue). To determine the cause of differences in outcomes requires information that will distinguish the programmatic differences. This design demands that cross-state agency data be obtained to describe and compare substantive program elements.

A pre-post evaluation measures program effects as differences in the value of selected outcome variables between a period defined as prereform and another defined as postreform (see also the articles by Haveman and by Kaplan and Meyer, in this issue). The power of this design (the probability of finding statistically significant effects, if they exist) depends on selecting comparison periods that are truly representative of the intended counterfactual "preprogram" years and the "postprogram" years. This design requires, perhaps more than others, a great deal of information on the implementation process, since the timing of program implementation must be well specified.

Data for cross-agency comparisons

One approach to evaluating the implementation of programs such as W-2 would be to conduct an "evaluability assessment," a pre-evaluation effort designed to clarify program intent, the stages and feasibility of implementation, and the likelihood of improving program performance from the point of view of policy makers and interest groups.[9] Such an assessment is especially appropriate at the beginning of a new program, to identify its main attributes, the services that will be provided, and the services that managers are most interested in evaluating. Evaluability assessment can also be seen as delineating those factors in a program's implementation that may determine its effectiveness.

Gathering data on those services identified in the evaluability assessment makes it possible to measure the strategies adopted to provide services in individual agencies, to compare services for consistency with the intent of the program, to understand the flow of participants through the program and how they leave it, to understand the process by which participants are matched with appropriate services and their shifts among services, and to chart differential selection into, persistence in, and attrition from program services. Answering such questions across service delivery areas is, we believe, the most valuable purpose of outcome and process evaluation of programs such as W-2. Even in the long run, W-2 service agencies are likely to adopt different programmatic arrangements and impose sanctions differently, providing the opportunity to test how these differences are associated with program outcomes. These comparisons are essential for welfare policy making in a world in which federal and state legislation has eliminated the former AFDC regime as a policy option.

The ideal evaluation data system would allow the individual participant data required by the outcome evaluation to be linked to process data that summarize the character and services of the agency responsible for coordinating participant services at the particular time. This link would make it possible to identify, in a program like W-2, the timing of program entry by each participant—a

standard item in any outcome evaluation—and the administrative environment in which that participant was served. In any process evaluation design, administrative data are likely to be a main source of information, but understanding of the perceived and actual roles of administrative staff must depend upon data gathered by observation of selected service sites and administrator surveys. Program administrators, adjusting initially to a completely new welfare system, may be reluctant to fully cooperate with a survey effort outside the normal demands of their administrative duties. For this reason, careful thought should be given to the needs of process evaluation in designing intake and client flow forms.

Under W-2, for example, administrative data from each linked site (see the article by Wiseman in this issue) and site surveys would provide information on administrative structure and services in the aggregate. Selecting a cohort of participants every 6–12 months for an outcomes study would provide information on how new cohorts moved through the system. These cohort data, linked to additional program data and site-provided data that are periodically updated, would allow a comparison of the effects of program implementation and service systems on participants over time. Note that, if an evaluation of W-2 takes place in a limited number of communities rather than statewide, so too should the process evaluation. A statewide outcome evaluation must be accompanied by statewide process data, and process evaluations for selected-site outcome evaluations should take advantage of this more intensive data collection design. Some mix of methods may be valuable, with a few intensive case studies conducted early in the program in order to discover issues that may be particularly important to highlight in broader survey efforts.

Any process evaluation must begin early in program implementation. A process evaluation of W-2, one of the most radical of the new generation of welfare-to-work programs, would be one of the first such efforts for a large-scale welfare and employment program. A prompt and early beginning would increase the chance that it will provide the information to administrators in Wisconsin and in other states that is necessary if programs are to be modified and outcomes explained.■

---

[1]The paper upon which this article is based appears in full in "Evaluating Comprehensive State Welfare Reforms: A Conference," IRP Special Report no. 69, University of Wisconsin–Madison, March 1997.

[2]Much previous research has indicated that there may be substantial deviations between the intended and actual implementation of particular programs. See A. B. Blalock, ed., *Evaluating Social Programs at the State and Local Level: The JTPA Evaluation Design Project* (Kalamazoo, MI: W. E. Upjohn Institute, 1990); C. H. Weiss, "Evaluating Social Programs: What Have We Learned?" *Society* 25, no. 1 (November/December 1987): 40–45.

[3]Head Start Project, *The Impact of Head Start: An Evaluation of the Effects of Head Start on Children's Cognitive and Affective Development* (Westinghouse Learning Corp. and Ohio University, 1969); see also P. Wu, "Structural Equation Models in the Analysis of Data from a Nonequivalent Group Design: A Reanalysis of the Westinghouse Head Start Evaluation," Department of Social Relations, Lehigh University, 1991.

[4]C. Doernberger and E. Zigler, "America's Title I/Chapter I Programs: Why the Promise Has Not Been Met," in *Head Start and Beyond*, ed. E. Zigler and S. J. Styfco (New Haven, CT: Yale University Press, 1993).

[5]For example, the Title I programs; see Doernberger and Zigler, "America's Title I/Chapter I Programs."

[6]A possibility to consider in investigating the unintended consequences of programs such as W-2 is that participants will likely have different rates of compliance. Program success, for example, may depend on an individual's psychological readiness for transition activities and the responsiveness of case workers to the needs of participants.

[7]L. M. Mead, "Welfare Policy: The Administrative Frontier," *Journal of Policy Analysis and Management* 15, no. 4 (1996): 587.

[8]There is no widely accepted theoretical basis for choosing program components to be evaluated and standard methods for acquiring process data, but a recent attempt by M. A. Scheirer is described in *A User's Guide to Program Templates: A New Tool for Evaluation Program Content*, New Directions for Evaluation, No. 72 (San Francisco, CA: Jossey-Bass, 1996).

[9]See J. S. Wholey, "Evaluability Assessment: Developing Program Theory," in *Using Program Theory in Evaluation*, ed. L. Bickman, New Directions for Program Evaluation, No. 33 (San Francisco, CA: Jossey-Bass, 1987).

A management information system for Wisconsin Works

Michael Wiseman

Michael Wiseman is Professor of Public Affairs and Urban and Regional Planning, University of Wisconsin–Madison, and an IRP Affiliate.

An important building block for welfare program administration is the management information system, or MIS, which encompasses procedures for collecting, storing, and retrieving information essential for operating and improving the program. Successful delivery of the new state welfare systems will depend heavily upon MIS adequacy.[1] In addition, the Personal Responsibility and Work Opportunity Reconciliation Act of 1996 includes extensive requirements for data on use of public assistance. This article explores the key information needs of the new intervention programs, identifies strategic problems in designing and implementing management information systems to support such programs, and makes a case for a particular strategy for MIS construction.

Management information and public assistance ideology

Government-sponsored programs to aid people in need require methods of determining need and of calibrating and delivering aid. The program's ideology and mission determine the character of the information required for successful management. When ideology and mission change, so should the MIS. Many management problems arise from disjunction between the requirements of the mission and the information that is available.

Public assistance ideology can be thought of in terms of two models: "passive" public assistance and "active" public assistance. *Passive* assistance systems respond to *circumstances* by alleviating them. If the circumstances can be readily assessed, program operation is a matter of delivering the goods or the money to recipients from whom no specific action is required. "Hunger," for example, is a "circumstance" and "food" the obvious prescription for relief. In contrast, *active* systems attempt to alter the *situation* of recipients and imply some movement or action by those seeking aid.[2]

These two models translate into different styles of delivering help, and differences in the management information systems required. A passive public assistance system emphasizes current transactions. In Aid to Families with Dependent Children (AFDC) as we knew it, eligibility and treatment were conducted on a monthly basis, and any month's assessment was in principle independent of what had gone on before. Active systems, in contrast, are history oriented and emphasize case management: situations are diagnosed, treatments are prescribed, and outcomes, that is, changes in situation, are observed. What happens next is very much a function of what came before. Management information must include a history of transactions and must provide access not only to the current action but also to those which preceded it. These differences have concrete implications. A passive, current-transaction system has no memory. At each point at which eligibility is reassessed or payments are redetermined, the necessary data fields may be overwritten to preserve storage. The file for each case can have a fixed format. In contrast, an action-oriented MIS must store history. Since one life differs from another, the dimensions of information associated with each case will vary. The file will be event oriented, and the more events that have occurred over the history of the case, the bigger the file will be. Most existing management information systems are not history oriented, and as a result they are inadequate for the new generation of active policies.

Management information for active intervention

A practical approach to determining "what managers need" under the new systems is to begin with what managers ideally would do, work backward to the information needed to do this job, and then to ask why such information is or is not on the manager's desk. Because required actions are quite program specific, I discuss the issues with reference to a particular program, Wisconsin Works (W-2).[3] My approach would be the same for any other state.

The Thompson Administration is committed to rapid and complete implementation of W-2. The critical importance of management in its plans is dramatized by a shift in responsibility for operation from what was the Department of Health and Social Services to the state's employment service agency, the Department of Industry, Labor, and Human Relations, now reorganized as the Department of Workforce Development (DWD). DWD will, in turn, be responsible for competitive subcontracting with various public and private organizations for W-2 operation. Dramatic in scope and ambition, W-2 offers an unparalleled administrative challenge.

The W-2 management hierarchy

W-2 MIS requirements may be derived from a review of what the "agent," or responsible entity, at each level of the W-2 organization hierarchy is expected to do. W-2 presents four levels of information demand: (1) the case

manager, (2) the site manager, (3) the system manager or state agency, and (4) the federal agency.

The front line of an active system is the *case manager*, or, in W-2, the "financial and employment planner" (FEP).[4] The FEP is the person ultimately responsible for allocating individuals to services and program tiers. The *site manager* (the local agency) is responsible for allocating cases among FEPs, coordinating resources for FEP use, contracting with providing agencies, coordinating activities with community agencies, and reporting to the state. The *state manager* or *state agency* (in Wisconsin, DWD) is responsible for designating agencies for site operation, evaluating agency performance, and financial management. The *federal agency* is in principle responsible for assuring that federal funds are used in a manner consistent with the enabling legislation.[5] These last three agencies may all play a role in disseminating information on methods.

It is likely that similar structures will exist in other state programs. An active system must involve a casework function that includes direct interaction with participants and a serial process of assessment, prescription, and evaluation. Since recommendations (prescriptions) for strategies to move participants to self-support are conditioned by local labor markets and services, the casework function must be managed locally. Resources for public assistance are generally collected from statewide sources and are used with state administrative and legislative oversight. This requires a state agency. And finally, over half of all public assistance costs are borne by federal taxpayers. Stewardship over these funds requires federal agency oversight.

Information requirements: The case manager

In considering the information required for these four agents, I will pay the greatest attention to the case manager. If the information is not collected here, nothing worthwhile will rise to the upper reaches of the administrative hierarchy.

The case file

At the front line, the case manager needs an information system that maintains a transactions history of each case for which the case manager is responsible. This is the electronic version of the standard welfare case file. It is here that assessment results and program prescriptions are recorded. "Time clocks" begin running as each adult participant enters levels of W-2 employment or starts receiving services, so the case record must be *relational*, linking all transactions for a family and children to the history of actions by and for the adults.

The case file should allow aggregation to four summary FEP reports, one of which is essential to individual case management and three of which provide summary information on the manager's entire case portfolio. One covers allocation by activity, another case flows, and the remaining two reports provide measures of time in status. All are described and illustrated in the full conference article (see note 1). Here I discuss only the first in detail.

The Participant Activity Report

The W-2 operations plan envisions a progression of cases from intake upward through the hierarchy to unsubsidized placement. The *Participant Activity Report* for a caseload will thus present cases cross-classified by status and stage.

A sample activity report is sketched in Figure 1. In the activity report, the categories identified on the horizontal axis are steps in the system process. The first step is intake, and thereafter come various assignments (in some rare instances the sequence may extend beyound four). The vertical axis of the activity report identifies the statuses created by the program plus additional classifications that are common to most welfare-to-work programs. I have included categories for missing information regarding participants' status (status unknown) and point in process.

The first three rows in Figure 1 cover the three W-2 statuses that involve active oversight: subsidized placement, community service employment, and W-2 Transitions. (Unsubsidized placement is an outcome.) When made operational, the chart would be full of numbers. The number in cell A, for instance, tells how many people in this FEP's case portfolio began the current month in a second W-2 assignment to a subsidized job placement. The number in cell B shows how many of the caseworker's entire portfolio of clients began the month in the subsidized employment status. The chart includes a category for persons who have passed through intake but are, at the beginning of the month, unassigned or between assignments; this is "Hold," and C is the total number of persons in this group.

The activity report yields many useful numbers. Cell H is the total number of persons for which the case manager has responsibility. The ratio E/H, that is, the proportion of these people for whom the case manager knows neither status nor assignment, is surely an indicator of loss of control. The complement of the ratio F/H, that is, of unassigned participants to total participants, is a type of participation rate. The larger the share of the caseload that falls in the upper rows, that is, in the most work-ready categories, the greater the likely turnover in the subsequent month, since these people are actively involved in welfare-to-work activity.

Readers with experience in public assistance administration will at this point be begging for a reality check. The activity report calls for a great deal of information that

	Activity						
Participant Status	Intake	Assignment One	Assignment Two	Assignment Three	Assignment Four	Assignment unknown	Status Totals
Subsidized Placement			A				B
Community Service Job							
W2 Transitions							
Hold							C
Intake, in process	D						
Exemption							
Status Unknown						E	F
Totals						G	H
	Totals By Point in Process						

Figure 1. Participant Activity Report for W-2 (point in time; probably beginning of month).

Note: Shaded blocks identify impossible participant status/activity combinations. Letters identify cells discussed in the text.

may be cumbersome to collect and time-consuming to record. Looking at the chart performs a useful function, however, in that it assists in identifying minimum information requirements. It would be difficult, for instance, to claim that cases were really being managed if the number in cell F, that is, cases for which status is unknown, is large.

The activity report also helps in thinking about an "information expansion path," that is, the order of importance of numbers. The choice here is in part a matter of policy. I believe that it is most important to know the number of cases in intake process (cell D), since processing intake is essential to delivering aid to people in need. Beyond this information, in order of importance, are: total case count (H), since knowing this signals an environment of well-defined case management responsibility; a count of cases for which status is unknown (see cell F), since this is a measure of oversight detail; and a count of cases for which information on stage in the process is unknown (see cell G). Note that full information does not an active program make. Our MIS may record the number of persons in "Hold" with remarkable precision, but if this number is large relative to the total number of cases the worker is managing, the program fails a key test of "active" policy.

In brief, the three remaining status reports necessary for the frontline case manager consist of:

1. The *Participant Transitions Report*, which covers an interval of time rather than a point in time and summarizes movement of individuals through the W-2 case management progression.

2. *The Time in Status Report*. This covers the elapsed time for which the participant has been receiving assistance; in W-2, this clock is set for each adult on first entry. The time in status report summarizes the duration of participants in each status and in the program overall.

3. The *Hours in Activity Report* records hours required for assistance-to-work activities. It is a second clock. In welfare initiatives such as Wisconsin's "Pay for Performance," the sanction applied for nonparticipation is proportional to hours of participation missed. Effective application of such a sanction requires that actual hours of participation be systematically recorded. And many programs also devote attention to the number of hours of activity that each status involves—the aim often being to raise this activity level so that time spent in activities approximates full-time work.

In sum, the focus of activist public assistance policy is case management, and management information begins at the ground level. In principle, an activist assistance-to-work system requires case managers who keep track of participants by (a) retaining history, (b) knowing current activity, (c) following changes, and (d) watching the clock. The number of things that might be recorded about the assistance process is potentially very large, and information is not costless to assemble. Therefore it is important to develop a clear "information expansion path" to assist in focusing effort on the information likely to be of greatest value in assessing performance.

Moving up the ladder

The ideas developed for thinking about the ground level can be applied upward in the administrative hierarchy. The question to be asked before developing the information system is "What does this person need?"

Site managers are responsible for supervising caseworkers, contracting for services, managing funds, responding to the community, and maintaining relations with the *state agency*. The key is knowing what is going on. For sure, part of this information comes from the fabled managerial activity of "walking around," but numbers count as well. Sitewide aggregations of the activity, case flows, time in status, and hours in status reports, at least in their minimalist versions, allow the achievements of individual case managers to be compared to agency norms. This comparison is meaningful only if participants are assigned to case managers at random. Otherwise, participants must somehow be differentiated, for example, by distinguishing those without work experience from those who report it. The choice of categories for such differentiation is a management problem for which state assistance and coordination would be useful.

The site manager must also monitor training, community service employment, and other program activities. The site manager has the best vantage point for assembling information on service providers and evaluating their productivity, information that must then be shared with line staff. The site manager is also responsible for budget. Each activity has an associated cost, so at the site level the various case reports are complemented with corresponding data on outlays. Finally, the site manager may be responsible for some review of post-program experience, including the duration of placements and the extent of recidivism.

Under W-2, the *state manager* or *state agency* is responsible for site oversight. In principle, the state agency is unconcerned about the performance of the contractors used by each site in operating W-2 (auditor concerns are another matter). Achievement is its concern. Moreover, the site aggregations of the participant activity, participant transitions, time in status, and hours in status reports are by definition of interest to the state if they are of interest to the site manager. In both cases, comparisons of agency performance must be "normalized" on the basis of participants' characteristics, just as caseworkers' achievements are.

The *federal agency* will need data to monitor the use of federal funds, assess overall program achievement, and support dissemination of information on methods. The federal role is further discussed below.

Challenges

Wisconsin's current management information system (for AFDC) is named Client Assistance for Reemployment and Economic Support, or CARES. Like systems in most other states, CARES is primarily oriented to supporting transactions linked to check writing and federal reimbursement. CARES is inadequate to the task of operating W-2, and major effort is being devoted to modifying the system to support the new program. The substantial challenges in designing and implementing such a management information system are what I term the issues of peaceful transition, graceful degradation, policy feedback, planned flexibility, system linkage, and meeting federal mandates.

Peaceful Transition. An operation as complex as public assistance does not change from one system to another overnight. Rather, a path of adjustment must be planned and carried out. The logic of W-2 suggests that the transition should involve gradually closing the AFDC payments system while simultaneously expanding existing work-for-welfare programs. This is what the state's "Pay for Performance" initiative (see this issue, p. 2) is about.

Graceful Degradation. It is important not only to enhance the accumulation of and access to information, but also to assure that all is not lost should it prove impossible to sustain the MIS at the levels intended. Over the long run, the capability for orderly retreat may be as important as the capacity for innovation.

Policy Feedback. Data systems like Wisconsin's CARES do not automatically produce data pertinent to policy analysis. Developing such information almost always takes major programming effort; staff time may be consumed by data extraction and programming issues at the expense of analysis. In Wisconsin, there exists nothing like the participant activity, transitions, time in status, or hours in activity reports for either case workers or site managers, even in high-profile demonstration sites such as the Work-Not-Welfare operations in Pierce and Fond du Lac Counties.[6] One immediate consequence is that these demonstrations have failed to provide management with as much information for W-2 planning as they might have done. These failings also diminish the utility of the state's administrative data for scholarly research.

Planned Flexibility. The state encountered problems with implementing the CARES system, and that in the context of a well-established transfer system. Neither AFDC nor the JOBS program were changing significantly at the time; it was the operating system that was being modified. Nevertheless, the change proved disruptive and has yet to be fully accomplished.

In contrast, the W-2 MIS will be implemented simultaneously with a new program. Past experience makes it highly likely that W-2 will change, and change significantly, as a result of problems encountered or opportunities discovered in the context of operation. The MIS must be planned with the expectation that the underlying program will change. The system of reports described in this paper is deliberately generic, in order to allow for such variation.

System Linkage. In the present system, both Medicaid and certain types of child care assistance are closely linked to AFDC. With W-2, access to child care, health insurance, and child support assistance will, in principle, be disconnected from participation in direct employment assistance. In practice, the systems will have to be linked in order to meet other program needs. For example, the current plan calls for direct deduction from grants to make the copayment for health insurance for persons in W-2 Transitions and community service jobs. Analysis of many policy questions will require that data be linked both across the services network and to other systems, including the social security and tax withholding systems.

The new federal mandate

The federal welfare reform legislation is schizophrenic in its treatment of states. On the one hand, it establishes block grants for "temporary assistance" to "increase the flexibility of states." On the other, the law creates remarkably rigid requirements for information collection, requirements that cannot be met at the present time by the management information systems available in any state. An elaborate set of work requirements is established for Temporary Assistance for Needy Families (TANF), the program that replaces AFDC.[7] To make sure that states are meeting these requirements, the law mandates quarterly reports that include the information needed to assess TANF participation rates and provide other data. The reports must be delivered quickly, and penalties are specified for laggards.

How should states respond? History suggests respectful skepticism. The federal reporting requirements are yet another example of a long congressional history of dealing with nagging concerns by requiring data that nominally address the problem. Little, if any, consideration seems to be given to analysis strategies or the consistency of data requirements with other administrative or legislative goals.[8] A major part of the information required by the new law could be derived from the case management system used to develop the four basic case manager reports and the site manager outcomes report. If employment is the objective of the state's own system, and if development of an effective MIS for support of the operating agencies is part of the state plan, federal requirements may, for the most part, take care of themselves. Those data items important for federal reports that are not generated by the MIS could best be obtained, at least cost, through special surveys.

Developing an MIS for Wisconsin Works: Research concerns

Apart from the fundamental issues of system design, other problems that require attention in developing a management information system for W-2 include: (1) calibrating the triggers, (2) building the incentives, and (3) managing intake.

Calibrating the triggers. Costs vary substantially across the various W-2 service tiers. Current forecasts of W-2 costs are based on estimates of the characteristics of incoming clients and transition flows from tier to tier that have little empirical support. Once W-2 is under way, experience will accumulate, but to date there has been no study that identifies developments that should trigger management concern or, for that matter, elation. At some point, some number will be pleasing. At some point, some number will be alarming. What are the numbers? What are the critical values that start bells ringing?

Building the incentives. An information system can produce nothing if fed nothing. The system developed here is based on what I believe to be the information essential to certain agents, in particular the case manager (FEP in W-2) and the site manager. We need to consider the actual operations of these agents and the extent to which incentives are adequate to ensure that the information they need as input is properly recorded. Perhaps the biggest problem is figuring out how to get things started, since for good agents the best incentive for proper input is useful output; yet no output can be received until information is input.

Managing intake. A common theme in both state and federal welfare reform is the "end of entitlement." As an issue for MIS design, loss of entitlement has significant implications at intake. Under AFDC as operated in Wisconsin, families in need had the right to receive benefits within 30 days of application, and failure to deliver gave grounds for legal action. Thus the system was to a significant extent self-policing. Without entitlement, it is possible that contractors will be tempted to discourage entry, especially by problem cases. The state's commit-

ment to timely aid would be confirmed by steps to measure the elapsed time of persons in intake for W-2, for instance, through the time in status report described earlier.

In sum, within every state, the Management Information System developed for TANF is important as an indicator of the nature of the program being implemented, as a source of information on operations and consequences for families, and as an object for study. The design of the W-2 MIS is an excellent example of the intersection of concerns of public management, policy analysis, and social science research. Errors in design and failures of implementation will complicate management, impede analysis, and diminish the usefulness of administrative data for research.■

---

[1]The paper upon which this article is based appears in full in "Evaluating Comprehensive State Welfare Reforms: A Conference," IRP Special Report no. 69, University of Wisconsin–Madison, March 1997. I have benefitted from discussions with members of the Steering Committee of the Wisconsin Works Management and Evaluation Project, Paul Saeman of the Wisconsin Department of Workforce Development, and Thomas Corbett of IRP. Opinions expressed here are mine.

[2]The distinction between "circumstance" and "situation" is narrow, but significant. Circumstances are generally cast in terms of surroundings or external factors, "situation" implies an actor. Individuals get into situations, not into circumstances.

[3]For a summary of the W-2 program, see this issue, p. 2.

[4]The primary administrative roles of the FEP under W-2 are described in this issue by E. Boehnen in "The Actors, Decisions, and Complexities of Welfare Reform."

[5]Under current law, responsibility for public assistance remains distributed across a number of agencies, but here federal interests will be treated as if they were unified.

[6]E. Boehnen and T. Corbett, "Work-Not-Welfare: Time Limits in Fond du Lac County, Wisconsin," *Focus* 18, no. 1 (1996): 77–81.

[7]These work requirements are summarized in Table 2 of the unabridged discussion (see note 1).

[8]The participation rate requirements imposed by the Family Support Act of 1988 are a case in point; see M. Wiseman, P. L. Szanton, E. B. Baum, and others, "Research and Policy: A Symposium on the Family Support Act of 1988," *Journal of Policy Analysis and Management* 10, no. 4 (1991): 588–666. Available as IRP Reprint no. 656.

Access to IRP information via computer: the World Wide Web site

IRP has a World Wide Web site that offers easy access to Institute publications. The Institute site includes publications indexes, updated semiannually, information on IRP publications, and ordering information. It provides basic information about the Institute's staff, research interests, and activities such as working groups, conferences, workshops, and seminars. The Web site also includes an annotated list of affiliates, with their particular areas of expertise. It offers an extensive set of links to poverty-related sites and data elsewhere.

Publications available on the Web site include files of formatted text of *Focus* articles, and selected Discussion Papers and Special Reports in both unformatted (ASCII) versions and formatted (Adobe Acrobat or Postscript) files. From the Web site, charts and graphs are available for immediate viewing and for downloading and printing.

IRP's home page on the Web can be found at: **http://www.ssc.wisc.edu/irp/**

Invited comment: Reshaping state information systems

Welfare has been reformed and will never be what is was. But far from being an end for welfare data and research, the reforms have increased the need for a more refined capacity to understand what is happening and for evaluating new programs.

In California, we are just going through the process of defining and negotiating how the Personal Responsibility Act will be implemented. There are four proposals: one by Governor Wilson's Administration; another by the counties, who are the administrative entities of welfare in California; a third by the Office of the Legislative Analyst; and a fourth developed by the advocates and prepared by the Western Center for Law and Poverty. All of these proposals differ in specifics, but there is substantial common ground. All emphasize the importance of limited time on aid, rapid and extensive transitions from welfare to work, and a limited state role accompanied by extensive county flexibility. All of the proposals speak to welfare reform as more than just a transition from Aid to Families with Dependent Children (AFDC) to Temporary Assistance for Needy Families (TANF). All acknowledge the comprehensive nature of welfare reform by including such areas as General Assistance, a safety net for those who have received five years of aid, and child support enhancements; by acknowledging the fundamental importance of child care and other supportive services; and by stressing that child well-being must not be lost as we emphasize the primary goal of personal responsibility for adults.

As researchers, it is our responsibility to define what are the important dimensions, how to measure them, and how to evaluate the impacts. All this, while the system is in a state of continual flux. To do this, we must first understand that we will need to have data at various levels: case workers and support staff must have current, reliable information on participants; county management must have tracking and analysis systems that allow them to understand what is working and what is not working; and state administrators will want to know the overall effects of program change and be in a position to support "best practices." How we will accomplish this is still unclear. What is clear is that the federal data collection and reporting efforts will not serve these purposes except to a limited degree.

What do we have, then? Well, we have some "knowns." For example, we know we will not have a "grand controlled experiment" with some randomly assigned people getting the old AFDC system and some getting the new TANF-derived program. We will not have one new system in which "one size fits all," but will have many new systems with different treatments for different people. It will not be a "pipeline" where people go in one end and leave at the other, but will be an open process, with people entering at various points and service needs determined dynamically, depending on need and availability. We know we will want to measure "process," such as how many people get up-front diversion services, and of what type. We will also want to know about "outcomes"—how many people are working and what their earnings are; whether children are living in safe environments and developing in healthy ways. We will want to know what happens to people who leave the system—those who are successful and those who are not.

What are we doing in California? First, we have set about trying to figure out how to meet the federal reporting requirements. Our current sense is we will do this using a modified form of the prior AFDC quality control (QC) reporting process and bundle TANF tracking with the remaining food stamps quality control tracking. Although this seems to meet the federal needs, we recognize that it will not meet other needs. First, the data are point-in-time, cross-sectional data with no capacity to follow cases and people over time. Indeed, the QC process provides for "cleaning" cases, so that tracking the "cleaned" cases would not provide a good picture of what is really happening. Additionally, data on a small sample of participants for a large, diverse state, with highly flexible, county-operated programs will not give us detailed understanding. So we are doing "first things first," recognizing that we will need to do more.

To determine what else we will need to do, we have formed an interdepartmental data and evaluation workgroup with representatives from various technical organizations in the California Department of Social Services, the Department of Health, the Employment Development Department, the Department of Education, and other state departments with supportive service roles. At the same time, the California Legislature has charged the University of California to make an assessment of what data and analyses will be needed under welfare reform. Many individuals participate in both groups. These efforts are only now getting under way; we anticipate a fairly clear set of concepts by late Spring of 1997.

The forum on data needs sponsored by the Institute for Research on Poverty captured most if not all of the dimensions. Indeed, we cannot accomplish all that has been proposed, and a related task will be to determine what is realistic and can be accomplished within available resources.

Werner Schink
Chief, Research Branch
California Department of Social Services

Designs for evaluating devolution

Burt S. Barnow and Robert A. Moffitt

Burt S. Barnow is Principal Research Scientist, Institute for Policy Studies, and Adjunct Professor of Economics, Johns Hopkins University, and Robert A. Moffitt is Professor of Economics at Johns Hopkins University and an IRP Affiliate.

In designing evaluation proposals for social welfare programs, an evaluator must first determine the questions to which answers are needed. Designs that are appropriate for answering some questions are often inappropriate for answering others, and multiple data-gathering approaches may be needed. In addressing any particular set of research hypotheses, there are five common questions to be considered:

What is the intervention/activity of interest?

What is the counterfactual to the intervention/activity of interest?

What is the population of interest?

What are the time frames of interest?

What are the outcomes of interest?

In previous evaluations of welfare programs, the *intervention* studied has usually been a new job search, training, or employment intervention; the *counterfactual* has been an existing employment and training program or no program at all; the *population of interest* has consisted of volunteers or mandatory assignees to the treatment—almost always individuals on the welfare rolls at the time of intervention (the general nonrecipient population has been excluded, by and large); the *time frame of interest* has typically run several years beyond random assignment; and the *outcomes of interest* have been earnings, wage rates, employment, and receipt and amount of welfare payments. If the consequences of the devolution of responsibility to the states are to be adequately evaluated, many of these specifications are likely to change.

The research questions

Activity/intervention of interest

The scope of the intervention will be vastly greater than anything tested in other evaluations to date. Devolution has already brought major changes in Aid to Families with Dependent Children (AFDC), the Food Stamp program, child care provision, and social services. Even if it were desirable or possible to focus solely on, say, the AFDC successor program, Temporary Assistance to Needy Families, the interventions already under way in most of the states are multifaceted and involve, most commonly, some combination of increased work requirements, financial work incentives, heightened sanctions for failure to comply, time limits, relaxation of AFDC-Unemployed Parent rules, family caps, and changes in asset requirements.

Disentangling the effects of interventions that have multiple components creates what we term the "bundling" problem: the intervention is, in actuality, a "bundle" of different interventions, all of which are implemented more or less simultaneously. This will make it extremely difficult to estimate the separate effects of individual interventions. It is not clear, however, that one should be particularly interested in the effect of adding an individual component to the current program environment or to the new environment after devolution. To the extent that the effect of the sum of the interventions may be more than the sum of the effects of each individually, it is arguable that the first order of business is to estimate the effect of the bundle, and only secondarily to estimate the effects of the components.

In an experimental environment, it might be possible to test many of the major interventions if it were desired to do so, although the design matrix required would be very large. But if nonexperimental evaluations are the only options available, then an evaluation must rely on "natural" variation—that is, on the variation in the choices that states and localities make. If this variation is very large relative to the number of states and localities, it may be difficult to infer the effects of individual components, even if it is possible to estimate the effects of the entire bundle.

The counterfactual

The definition of a counterfactual is also problematic, because states and localities have already begun implementing major changes in their programs. Thus it is, in some sense, already too late for the "before" half of a before-and-after study. However, there are two ameliorating factors. First, to the extent that information on past behavior can be gathered from historical administrative data, it may be possible to establish a baseline prior to the current changes. Second, it is quite likely that the implementation of devolution will take many years. Changes in programs from, say, 1997 to 1999 may dwarf anything that had happened by the end of 1996. If so, there is still time to measure major effects—a possibility that also has implications for an evaluation design, for it implies that a good evaluation must be capable of flexibility in an evolving policy environment.

Population of interest

As we mentioned previously, most previous welfare evaluations have had as their population of interest individuals on the welfare rolls or some subset of them (e.g., those eligible for employment and training programs). The new reforms are likely to affect not only more recipients but also the nonrecipient population, through well-known "entry" effects.[1] Most observers expect these effects to be negative, that is, to *deter* entry. Such is often explicitly the intent of the policy makers implementing the programs. The expansion of the population of interest calls for a much broader population base for the evaluation than has been the case previously—limited, perhaps, to the poor or near-poor population or, for some purposes, to particular subgroups.[2] It might, for instance, be possible to draw a special sample of those who are near to exhausting their benefits.

Time frames of interest

How long should evaluators track the sample(s) of participants? If time limits are included in the sites, it will be very desirable to do so for at least one year beyond the point at which participants are terminated from the rolls. If five-year limits are imposed, this may be quite a long intervention. Rather different in spirit is the question of how far the time frame should extend prior to the intervention examined. Nonexperimental evaluations require, by and large, more historical data than do experimental evaluations because there is more need to control for the "histories" of the individuals and localities involved.

Outcomes of interest

At first blush, the outcomes of interest should be similar to the outcomes in previous studies: earnings, wage rates, employment, receipt of welfare, and amount of welfare received. The broadening of the population, particularly to nonrecipients, requires that exit and entry rates be included also, and there may be interest in more aggregate outcomes such as caseloads and dollars spent on welfare.

The trade-offs between different objectives may be more stark in the new environment than in the past. By all appearances, the goal of reducing the caseload is likely to assume more prominence vis-à-vis increasing recipients' well-being than it has in the past. Indeed, policy makers have expressed willingness to accept reductions in recipients' incomes and increases in poverty rates in exchange for caseload and cost reduction, particularly if it is achieved by altering the rules of the program in a way that, it is argued, society prefers. The notion of a time limit, for example, is partly based on the simple notion that only a certain amount of support should be given, regardless of the consequences.

Another implication of the current state-level interest in new norms and "expectations of behavior" is that implementation becomes more important and is, to some extent, itself an outcome of interest. If the major goals of state policy makers are to require people to work while on the welfare rolls and to set time limits to the state's commitment, and if these policies are aimed, not at changing behavior, but simply at enforcing what are asserted to be societal norms, then the goal of the program is merely to implement and enforce those rules, regardless of the consequences. That by itself may be difficult to do on a large scale, and it is not yet clear how successful states will be. But an evaluation design that ignores this point of view runs the risk of failing to answer the questions that some policy makers want answered.

A classification scheme for nonexperimental evaluations

Table 1 summarizes the types of nonexperimental evaluations. In such studies, the estimation of a program effect (or "treatment" effect) is necessarily based on a comparison of individuals or groups who have been exposed to different programs. We term these "quasi-experiments," in contrast to true controlled experiments.[3] Each of the types of quasi-experiment in Table 1 makes a different type of comparison. Table 1 considers four different "generic" types of evaluation, and subcases within them: pure before-and-after designs, pure cross-section designs, designs which combine before-and-after with cross-sectional elements, and cohort designs. Each design faces different threats, and each may, therefore, generate different estimates of impact for the same program.

Pure before-and-after designs simply follow individuals or groups over a time period within which a program change has occurred. The change in their outcomes is attributed to the change in the program. The threats to this design are of two distinct types: aging (sometimes called maturation or life-cycle) effects, and systematic external changes in the environment. Aging effects might be ignored for short periods, but over longer periods the change in outcomes may be affected by natural life-cycle patterns. Also important are other changes—for example, in the local labor market or in the neighborhood environment—which occur simultaneously with the program change and which therefore confound the measurement of its effects.

If aggregate data are used, at either a state or national level, this approach is usually termed "time-series modeling." Aggregation has very little advantage per se, but aggregate data are often available for longer time periods and for more cross-section units (see the next generic type) than are individual, micro data. Particularly in a before-and-after evaluation, where reliance on the stability of the local economic environment is so important, a longer time series can be invaluable in separating the

Table 1
Classification Scheme for Nonexperimental Evaluations

Generic Type of Study	Specific Type of Study	Description
Pure Before-After		Units examined over time and outcomes measured; program has changed over time; attribute change in outcomes to change in program; can have multiple "before" and multiple "after" time periods
	Individual units	Recipients or nonrecipients; in one area, most commonly; usually do not have a long time series; sometimes have subannual data and sometimes not
	Aggregates	Fixed geographic unit, usually a state; also called time-series modeling; usually have relatively long time series and often have subannual data
Pure Cross-Section		Comparison of different units at a point in time (e.g., week, month, or year); program differs across units; attribute difference in outcomes across units to program differences
	Individuals within areas	Usually recipients only since recipient–nonrecipient comparisons usually not reliable; different individuals are treated differently; danger of selection bias
	Across areas	Individuals or aggregates; danger of site effects
Cross-Section and Before-After		Combination of two; have units that are treated differently; measure outcomes of all observations over time
	Individual units within areas	Different recipients or nonrecipients are treated differently; treatment changes over time; permits "fixed effects" (or "differences in differences") as well as "autoregressive" models that use lagged variables ("history") to control for "heterogeneity"
	Individual units across areas	Different recipients or nonrecipients in different areas with different programs, and program changes over time
	Aggregates across areas	Aggregate "time-series" modeling but using multiple areas, usually states
Cohort Design		Multiple birth or program entry cohorts who are each followed over time; program is changing over time; changes in cohort experiences are attributed to program change
	Individual units within areas	Individual data on multiple cohorts within a single area
	Individual units across areas	Multiple cohorts in multiple areas; can have cohort and area "fixed effects" by comparing cohort differences across areas
	Aggregates across areas	Possible if aggregate data can be disaggregated by age

Note: Each type can utilize administrative data, survey data, or both.

influence of general economic events from the program change in question. Aggregate data are also often significantly less expensive to obtain than micro data, and they can be used to detect entry effects.

These evaluation types can be further distinguished by whether administrative data, household survey data, or both, are used. Aggregate data on caseloads or on wages can be obtained from administrative data, or individual data may be drawn from welfare records or wage records, for example. Again, individual data are generally preferred, but they are also usually available for fewer time periods and fewer areas.

Pure cross-section quasi-experiments are rare, even though they correspond most closely to controlled experiments. Comparing recipients to nonrecipients, different types of recipients, or different areas at a single point in time so obviously runs the risk of confounding program effects with other differences among individuals or areas that the method is almost never used. Instead, these types of comparisons are conducted when multiple periods of data are available; thus they fall under our third generic category of *combined cross-section and before-and-after data*.

This category is, in fact, the most common type of nonexperimental evaluation and covers a large number of different subtypes. One classic method of evaluation is a comparison of participants to nonparticipants over time, using the history of their behavior to control for heterogeneity between the groups. This method has a fairly long history in the evaluation of job training programs, but it has almost never been used for the evaluation of welfare programs because welfare recipients and nonrecipients are generally thought to be sufficiently

different as to be noncomparable, even if subgroups with the same histories are compared. Somewhat more common are comparisons among different types of recipients who are given different treatments (e.g., different employment and training programs), who are on waiting lists, or who are otherwise treated differently by the program. Fixed effects, autoregressive, matching, and other techniques are often used to control for differences in histories between the groups compared.[4] The main danger with these methods is that the different groups being compared differ along dimensions that we cannot observe and that make them noncomparable. With a few exceptions, therefore, most of the possible types run a serious risk of selection bias in the ways in which people are assigned to categories.

More credible are combination designs that compare recipients, nonrecipients, or both together, in different areas over time, thereby making use of the variation in program type between areas to measure program impacts. Other than poor matching of areas on the basis of their initial conditions, the chief threats to this design are, again, uncontrolled differences in growth rates or other time-related changes in the outcome variables across areas. This problem is the counterpart to the unobserved site-effect problem in the pure cross-sectional comparison across areas. Comparison-site or matched-site designs aim to reduce these threats by choosing areas that are similar in a few measurable dimensions, but such designs are not successful if the areas differ in too many other ways. In addition, this method requires that programs in the different areas can be compared along some common measure or scale; that may be difficult for the diverse and complex array of current program changes.

Different areas can also be compared over time using aggregate data on caseloads, earnings, or other variables. The chief advantage of such an approach lies, as noted, in the greater number of time periods and areas available with aggregate data. However, as with the use of micro data, this method requires that programs in different areas and over time can be ordered along a single dimension or a few dimensions.

The final generic method presented in Table 1 is a cohort design that measures the effects of programs by comparing the experiences of different cohorts who come into contact with the system at different calendar times. A well-known example of this method is the evaluation of the 1981 OBRA legislation by the Research Triangle Institute. This evaluation compared the welfare exit and employment outcomes of a cohort of AFDC recipients before 1981 and a cohort after 1981; the latter experienced the OBRA legislation in full. The differences in outcomes between the two cohorts were attributed to OBRA.[5]

This method can be extended in many ways. Multiple cohorts over time, if they are measured prior to the intervention, can be used for a comparison which permits the evaluator to incorporate changes in the economic and social environment. Cohorts in different areas can be compared, and the treatment measured as the across-area difference in cohort differences. Historical data on the individuals within each cohort can be collected and used as controls for heterogeneity. For any of these designs, administrative data, household survey data, or some combination of the two could be used.

Should recipients or nonrecipients be treated as separate groups in any of these quasi-experiments? Although use of recipiency as a defining characteristic is possible and often desirable, recipiency itself is self-selected and may consequently pose a threat to designs which stratify on recipiency. For example, comparisons of the outcomes for recipients in different areas or at different times rely for their validity on the presumption that these populations are the same in unobservable as well as observable dimensions. If they are not, differences in response may be the result of their underlying differences in characteristics, not of the differences in treatment.

Data sources and issues

There are three general sources of data that can be used for the evaluation: aggregate administrative data, individual administrative data, and individual survey data. These forms of data are not simply substitutes, but can also serve as complements—certain types of questions can only be answered with specific types of data.

Aggregate administrative data, collected by state and local governments for administrative purposes, are generally available on a monthly basis and include total cases in the program, entries to the program, exits from the program (possibly including the reason for exit), average benefit levels, and activity levels. Such data can be linked to demographic and economic data for the same area to estimate time series and determine the impact of program changes on caseload entries, exits, average number on the rolls, and benefit levels. Aggregate data are relatively cheap to obtain, and often go back for many years—in some states, to the early 1970s.[6] The low cost implies that time-series analyses can be used to supplement other, more expensive approaches. The long time frame means that evaluations can capture the effects of a wider range of economic conditions than if only a few years are available.

Individual administrative data include information maintained at the individual level on program participation and individual characteristics. Examples include benefits received in programs such as AFDC, Food Stamps, unemployment insurance, and Supplemental Security Income (SSI). Because participation in these programs is generally conditioned on income, we can also obtain information on earnings, wage rates, and hours

worked. The time period over which such data are available will vary from state to state and program to program. Some states, for example, routinely discard unemployment insurance wage records after three years. Confidentiality statutes may pose access problems in certain states and programs.

Although administrative data are less expensive to obtain than household survey data (see below), they are not cheap. There are practical difficulties: records may be disorganized, with many errors, or not available in machine-readable form. By definition, a study using administrative data is limited to the variables available in the data—there is no flexibility to add or modify data elements. Administrative data tend to be rather weak on demographic information and can only track individuals while they are in the system. To know anything about the earnings of families who leave the welfare rolls, for example, IRS records or unemployment insurance wage records must be obtained. These two problems affect the ability to gather preprogram information on individuals in the evaluation, some of whom were off welfare.

Survey data. Surveys of individuals of interest have advantages over the two other data sources, but they also have limitations. Perhaps the most important advantage of surveys is that data can be gathered on any topic amenable to survey questioning. Besides the usual questions on outcomes and demographic variables, surveys can seek to cover topics not covered or inadequately covered in administrative data: motivation, mental health, intelligence, education, detailed work histories, and so on. In addition, a survey can be designed to cover whatever population is of interest, generating information on a comparison group that would not be available from administrative data.

Perhaps the most significant disadvantage of survey data is their expense, particularly when the focus is a low-income population. Major costs may be incurred in developing a sampling frame, for example. Screening costs can be extremely large when only a small proportion of the population is sampled in an area. Another major disadvantage is that surveys can only gather data prospectively, because of recall problems, making it difficult to address the problem of the "before."

Conclusions

The most attractive designs are those which combine cross-section before-and-after variation and which examine cross-cohort variation. It is important that the threats to these designs be examined and that the data collection plan take this into account. Supplementing individual administrative data with survey data, for example, can permit checks for self-selection into recipiency; and supplementing an individual-level analysis with a time-series analysis using aggregate administrative data can be used to check whether program effects are confounded with general trends. Although resource and time constraints may obviously limit such an ambitious data collection strategy, it should be regarded as a goal to be aimed for.■

---

[1]R. A. Moffitt, "The Effect of Employment and Training Programs on Entry and Exit from the Welfare Caseload," *Journal of Policy Analysis and Management* 15 (Winter 1996): 32–50.

[2]For some approaches, such as caseload modeling, it is not necessary to establish a population base for drawing the population of interest.

[3]D. T. Campbell and J. C. Stanley, *Experimental and Quasi-Experimental Designs for Research* (Chicago, Ill.: Rand-McNally, 1969).

[4]The method of "selection bias modeling," associated with one econometric tradition, is sometime associated with these methods also. However, it is in fact not a separate method from any of those presented in the table, each of which could be formulated as a "selection bias model."

[5]See Research Triangle Institute, *Final Report: Evaluation of the 1981 Amendments* (Research Triangle Park, NC, 1983).

[6]The period that can be analyzed depends on more than just the availability of data. For example, New Jersey officials warned us that data prior to 1978 were of uncertain quality. Also, if the program structure changes in major ways, it may not be advisable to assume that the same model applies before and after the change. For example, the changes instituted by OBRA in 1981 may make it inadvisable to use pre-OBRA data in some states.

The Midwest Welfare Peer Assistance Network (WELPAN): A model

Elisabeth Boehnen, Thomas Corbett, and Theodora Ooms

Elisabeth Boehnen is Database Administrator and Thomas Corbett is Acting Director of IRP; Theodora Ooms is Executive Director of the Family Impact Seminar.

On October 25, 1996, welfare officials from seven Midwestern states met in Chicago to explore the creation of a regional network that would meet regularly to discuss state welfare initiatives. This meeting was remarkable not so much for the agenda nor the substantive discussion as for the sense of connection made among the participants. By the end of that session, state officials who had initially been cautious about such a forum were enthusiastic about its prospects. The meetings were perceived as an opportunity to escape the daily pressures of management and think more deeply about basic issues and concerns, to discuss successes and failures, commiserate about anxieties, and find stimulation and challenge in what others were doing. With remarkable swiftness, participants reached consensus on an agenda for the future, agreed on the type of environment they wanted to establish, and set ground rules regarding participants and a framework for communication.[1]

Over the following months there emerged a regional information and support network, the Midwest Welfare Peer Assistance Network (WELPAN). Funded by the Joyce Foundation, this network is composed of state welfare officials and policy researchers from the states within which the Foundation's activities are concentrated—Illinois, Indiana, Iowa, Michigan, Minnesota, Ohio, and Wisconsin. The network model builds upon commonalities among geographically contiguous states that are somewhat alike in demographics, labor markets, and political cultures. In addition to their demographic and economic likenesses, the Midwestern states are also among the more aggressive innovators in welfare policy, making them ideal candidates for exploring the merits of a regional network.

The rationale for regional networking is simple. Traditionally, welfare policy making and management were *vertical* in direction, in that Congress, the federal bureaucracy, or the courts dictated most of the rules that states and local governments carried out. With the passage of federal welfare legislation in August 1996, that clear vertical flow of communication no longer exists, at least in the same form. It is logical, therefore, for the states to look toward one another as sources of innovation and technical assistance. But if they are to succeed, mechanisms to facilitate the *horizontal* flow of information will be helpful. If states act independently as they assume responsibility for tasks formerly shared with or largely determined by the federal government, inefficiency and redundancy are inevitable. By sharing knowledge and experience they can more effectively address difficult problems and reduce the uncertainties unavoidable in this era of change.

The regional network model offers a basis for the creative pooling of intellectual resources, insights, and expertise among states. It is a way to respond the new conceptual, technical, and financial challenges posed by the current revolution in welfare policy management.[2] Under the right conditions, such a model might also develop closer ties between government and academia. Academics who ordinarily shy away from the time investment necessary for working with individual states might make a personal time commitment if several states were involved. Such a collaboration not only broadens the knowledge base available to government officials but also assists them in sorting through the information they have available and in applying it to their program needs. It also increases the probability that state planning, monitoring, and evaluation efforts are most efficiently carried out. The exposure of academics to real policy questions is likely to stimulate new and productive research ideas and practical applications.

In formulating a proposal for a regional network in the summer of 1996, IRP enlisted the cooperation of the Family Impact Seminar (FIS), a Washington-based, nonpartisan policy institute that has two decades of experience in organizing seminars and roundtables for congressional staffers and government officials in Washington. If the model was successful, FIS would be well positioned to collaborate with other national organizations, such as the American Public Welfare Association, to help establish similar networks in other regions. From the start, it was intended that WELPAN members would decide the membership, format, and topics for the agenda. The role of FIS and IRP was to organize and coordinate the meeting logistics and agenda, prepare background materials, and facilitate the discussion. Over time state officials would assume primary responsibility for planning and coordinating the network, while IRP and FIS would provide technical assistance and broaden the network's knowledge base by drawing on research, ideas, and innovations from other regions.

...a chance to brainstorm, to be reflective, to problem-solve. . . . That's of great practical use, to hear about particular strategies, operational issues that you can leverage into your thinking as you're developing your program.

— Sally Titus Cunningham, Iowa Department of Human Services

State policy makers have few institutions that function as knowledge "brokers"—that synthesize the accumulation of research findings to provide them with brief, nonpartisan overviews on particular topics. IRP and FIS hoped to encourage a neutral, nonpartisan forum in which participants could review their state's experiences and freely discuss the design, implementation, and evaluation of welfare reform initiatives; identify the type of information and technical assistance they would like to receive from each other, from research centers, from other states, or from national sources; discuss networking activities that would assist their state welfare reform initiatives; and establish a structure for a supportive network.

When FIS and IRP began calling potential state participants in August 1996, officials expressed interest but noted that they were extremely busy. There were also pragmatic issues that required attention: what types of state officials would make ideal participants (level of authority and substantive expertise were in question here); how to nurture a trusting atmosphere so that real communication would take place; and how the project sponsors should balance the need to give the fledgling network a structure while working toward the goal of state assumption of responsibility. Given the uncertainties about the nature of the undertaking, the pressures upon state officials now facing the reality of welfare reform, and the short-term funding for the project, a real possibility existed that the network might be very short-lived.

In the productive first meeting described above, however, participants were very quickly able to agree on a preliminary list of six primary issues for future discussion:

1. *Defining success.* How can states move quickly to create outcome indicators that will be responsive to the demands being placed upon them by other state and federal officials, the public, and the media?

2. *Integrating workforce development and welfare reform*. Replacing income support with employment will challenge interagency coordination and partnering with the private sector. Participants also wanted to examine job creation, training, and retention.

3. *Changing the culture of welfare agencies*. Agency focus will shift from issuing checks to work and behavioral modification, necessitating change in the operation of frontline workers and in office culture. The issue of privatizing some welfare functions was included.

4. *Data systems and management information issues*. States want the ability to track clients over time using complex indicators. Data systems need to be upgraded in their ability to perform these activities and to exchange information with other existing systems.

5. *Federal Medicaid regulations and how to meld them with the regulatory flexibility in Temporary Assistance to Needy Families (TANF).* (TANF is the new federal program that replaces Aid to Families with Dependent Children.) Officials were interested in advocating increased flexibility in Medicaid.

6. *Family formation and parental responsibility.* Given that improvement in this area is a primary goal of TANF, there is particular interest in these issues, especially child support enforcement.

Because participants represented significantly different agency perspectives—some were senior researchers, others career managers, others political appointees—they did not necessarily share a common language. They very early reached the conclusion that ongoing discussions about these complex topics required them to develop one. And they agreed that an important goal of the WELPAN group should be to produce documents that would be useful not only to share with other state agency staff (particularly data experts), but also as background information during discussions with legislatures crafting new state welfare bills. The first such working paper attempts to establish a common framework of evaluation concepts and terminology and pulls together outcomes, indicators, and goals that the group identified as important within welfare reform.[3]

In two subsequent meetings, intense discussion identified what participants believed to be the *core concerns* of the states' welfare plans—those outcomes for which state officials firmly believe they will be held accountable by their governors, legislatures, and the public. These include welfare dependency, labor supply, program costs, parental responsibility, family formation and stability, and economic well-being (making work pay).[4] The participants wanted to develop more effective strategies for meeting the core concerns; to obtain feedback throughout the implementation process to make it possible to correct and adjust the welfare reform plan; to be prepared to respond constructively to questions from the media and the public; and to educate themselves, legislators, and others about the unintended positive or negative consequences of welfare reform.

The primary agenda item for the fourth meeting was the changing culture of welfare offices, which will continue to be a topic as agencies move from an entitlement focus to a self-sufficiency focus and from a process-based system to an outcome-based system.

Participants have also discussed perhaps developing working groups that might address specific issues and draft recommendations for the core group to consider. Such working groups might explore, for example, problems with management information systems and the integration of administrative data; workforce development issues, such as the communication between state officials and the business community; interstate child support enforcement; and successful approaches to paternity establishment.

The first WELPAN meetings suggest that:

1. A regional network can successfully address problems and share information and ideas in the areas of program and service development, implementation, monitoring, and evaluation.

2. Although representatives come to the table with a different set of perspectives, responsibilities, and experiences, they can reach consensus regarding the most important issues in the new world of welfare devolution. And they can also work toward consensus on the critical indicators and measures of a successful welfare reform program.

3. Closer ties between state officials are likely to result in increased information sharing, cross-state problem solving, and consensus on particular issues (such as child support enforcement).

Many challenges remain. These include ways for members to assume increasing ownership of and responsibility for the network, to facilitate communication among themselves outside of the meetings; to disseminate information concisely and efficiently; and to maintain the balance and integrity of the group during membership changes, the addition of other members, or the introduction of outside experts. Not least is the question of how the network will become self-sustaining once the Joyce Foundation grant ends.

Both the agenda and the format of WELPAN continue to evolve. The meetings remain small and informal—from one to two officials attend from each state. The participants have agreed to extend sessions from one to two days and have received additional funding from the Joyce Foundation to keep the network going. The interest in WELPAN already expressed by other states and the U.S. Department of Health and Human Services suggests that it constitutes a promising model for states wishing to explore a regional effort and that its products can provide insight to other states. There is reason to believe that the network offers a useful demonstration of the kind of horizontal communication that may be essential to states struggling to respond to the unfamiliar world of devolution.■

---

[1]The authors thank the members of the WELPAN network for their review of this draft and their helpful comments.

[2]Concern about the new challenges facing states predated the passage of welfare reform in August of 1996. In December 1995, for example, IRP sponsored a one-day workshop on State-Level Indicators of Children's Well-Being at the National Academy of Sciences in Washington. The workshop brought together a number of state officials and was not considered an end in itself so much as an impetus toward later activities. Several participants subsequently discussed the possibility of a regional approach to at least some of the technical and developmental needs arising from welfare policy devolution. Unmi Song, a program officer from the Joyce Foundation who attended the December workshop, noted the number of participating Midwestern states, and began discussing the idea of bringing them together to share ideas and problems. *Focus* 18, no. 1 (special issue 1996): 42–48 summarizes the workshop; pp. 47–48 discuss the regional model as it was envisioned at that time.

[3]The paper, "Defining Success in Welfare Reform," was prepared by Thomas Corbett and Elisabeth Boehnen on the basis of discussions and presentations at WELPAN meetings. Its distribution is currently restricted to WELPAN members.

[4]This is a very abbreviated version of the dialogue on measures of success by WELPAN participants. A full treatment of the topic will appear in "Measures of Success," a paper in preparation by Elisabeth Boehnen and Thomas Corbett.

The equity implications of the National Health Service reforms in the United Kingdom

Carol Propper

Carol Propper is Professor of Economics in the Department of Economics and School for Policy Studies, University of Bristol, United Kingdom.

In many countries—the United States, the Netherlands, Scandinavia, New Zealand, and Russia, for example—reforms are being implemented to make the health care sector more competitive. In 1991, the Conservative government of the United Kingdom introduced reforms which moved the National Health Service from a "public monopolistic integrated" model to a "public contract" model of health care, by introducing price competition between hospital care providers. These reforms alter the incentives faced by users, buyers, and sellers of health care. The stated intention of the U.K. government was that the reforms would do so in a way that would increase the responsiveness of service providers to users without sacrificing the equity properties of the prereform system. Both the general incentives of this model of health care and the precise incentives of the U.K. system have been discussed at some length.[1] Here I focus on the potential distributional impact of the reforms, which are discussed under two broad headings: changes in financing and the impact of competition and contracting. Much of my discussion is necessarily in terms of the potential for changes which may adversely or positively affect different groups in society, for our knowledge of the impact of these reforms is still very patchy. In contrast to the United States, the British government does not believe strongly in program evaluation; the reforms were introduced without trials, and in the first few years of their operation access to data was extremely limited.

The NHS reforms

The reforms did not alter the financial basis of the NHS, but the level of its funding has grown quite rapidly since the inception of the reforms. Under the prereform system, financing was allocated directly to family doctors (the GPs) and hospitals from the Department of Health. GPs used their funds to provide primary care to the patients registered with them; they referred any patients requiring secondary care to hospitals or specialists, but did not pay for that care. Hospitals used their allocated budgets to provide hospital-based medical services to any patient referred to them by a local GP. Under the reformed system, the funds for hospital care are not allocated directly to the hospitals. Instead, they are given to public third-party purchasers (defined below) who are responsible for buying this care from whichever hospital providers they wish to choose. No changes were made in the way in which GPs are paid to supply primary care.

The NHS hospitals, the main suppliers of secondary care, have been progressively transformed into publicly owned, self-governing NHS trusts. Now that their budgets are determined by contracts won from third-party purchasers who are not restricted in their choice of hospi-

Established in 1948, the National Health Service (NHS) provides health care to all residents of the United Kingdom. The system is financed by the national government, which provides free care in publicly owned hospitals and in the offices of primary care providers who, although paid from public funds, are self-employed. These primary care providers (family doctors), known as General Practitioners (GPs), handle 90 percent of all episodes of patient care and act as gatekeepers to hospital-based and specialist care. GPs practice in groups; patients choose their GPs and are then registered with the practice group. Specialist physicians practice in the public hospitals on a salaried basis, but can also spend time in private practice (such practice, as a whole, accounts for less than 5 percent of health care expenditure). Hospital staff are public-sector employees.

The annual budget for the NHS in 1996 was £42bn, U.S. $64.97 billion (in $1996), which is just over 7 percent of GDP. The majority of NHS funding—96 percent—is from general taxation (a mix of income, payroll, and expenditure taxes), and the remaining 4 percent from patient copayments for prescription drugs. A small private sector operates alongside the NHS and specializes in treatments for which there are waiting lists for NHS care (primarily elective surgery). Private medical insurance is held by only 13 percent of the population and is used to pay for the costs of specialist and hospital treatment in the private sector.

In 1991 the way in which NHS hospital-based care was delivered and funded was reformed. This article discusses these reforms.

tal, the incentive of competition has been introduced. One goal of the reforms was to reduce political control of health care suppliers; politicians have stressed that the responsibility for local services now rests with local purchasers and providers and not with politicians. Hospital management has been given greater control over inputs, outputs, and pay. But ownership of the physical assets, such as the hospital buildings themselves, remains in the public sector, and managers are subject to central ministry scrutiny of investment decisions and control over access to capital markets. Control over cash flows from year to year is retained by the central government; trusts must break even every year and are also subject to price controls (though the form of these controls is such that they appear to be widely broken).[2]

The reforms as originally intended had only one type of purchaser. These are the District Health Authorities (DHAs), agencies of the Department of Health that are responsible for purchasing health care for all the population in a particular geographical area. But the reforms also allowed the development of another set of purchasers, known as GP Fund Holders (GPFHs). These consist of a subset of GPs, self-selected from the whole population of family doctors, who have chosen to hold a budget (the fund) which they use to buy for their patients a limited set of the health services sold by hospitals. The funds a GPFH receives are deducted from the government budget for the DHA which covers the geographical area in which it is located. GPFH practices tend to be larger than other primary care practices.[3]

There are important differences between the purchasers. First are their size and location. DHAs are considerably larger—their average population is around 350,000. There are currently 100 DHAs, which cover all the population of the United Kingdom that is not covered by GPFHs. Fund holders have a budget for secondary care for a considerably smaller number of people, since each GPFH practice has a budget only for patients registered with the practice. GPFHs are not evenly distributed, but tend to be more heavily concentrated in the more affluent suburban areas around large conurbations, though there are exceptions. Individuals cannot directly choose whether they are covered by a DHA or a fund holder. If their GP is a fund holder, they will be covered by the fund holder; if their GP is not a fund holder, they will be covered by the DHA.

Second, the number of GPFHs has been rising, and the number of DHAs falling. By April 1996 there were 3,735 Fund Holding Practices, covering 52 percent of the population of England. From 1995 onwards, the Department of Health initiated schemes to increase the range of services that GPFHs can buy. The long-run aim of the Conservative government has been for GPFHs to be the main purchasers. A residual purchasing role is envisaged for DHAs, though the nature of this role has not yet been clearly defined. The Labour party, which took office at the beginning of May 1997, has suggested that it will abolish the GPFHs. However, the arrangements envisaged to replace this set of purchasers are not yet fully articulated, and in practice they may not be very different from a number of schemes which have grown up around fund holding, all of which increase the role of the family doctor in the allocation of secondary care.[4]

Third, the two types of purchaser differ in the incentives they face. DHAs are public bodies, required to break even each year. GPFHs are self-employed individuals who fully own their practices; they are allowed to keep the savings from their budgets. Officially, such savings must be invested in the practice rather than spent, but because GPFH books are not heavily audited this requirement is not easily monitored.

Fourth, the reform model was one of monopolistic third-party purchasers, with the attendant issues of how to create the appropriate incentives for purchasers to compel providers to be efficient, innovative, and responsive to consumers' preferences. But there are elements of competition on the purchasing side. DHAs do not compete with each other. Nor, because competition between GPs for patients is very limited, do GPFHs directly compete with each other for patients. But GPFHs compete with DHAs, in so far as the DHAs lose funds to GPFHs for the patients the latter are responsible for.

Finally, GPFHs have not only the opportunity to select their own patients but also some incentives to do so, whereas the DHAs do not. Patient selection by health care providers—for example, by avoiding patients expected to incur the greatest use of medical resources—is notoriously difficult to monitor, and indeed the Department of Health has not tried in any active way to do so. Whether GPFHs do actually practice patient selection is discussed below.

The impact of funding

Overall financing of health care

The NHS, as noted above, is funded from general taxation, with limited user charges. Payment for the NHS does not depend on age or health status, and depends on employment status to a lesser extent than in systems funded more heavily from either payroll taxes or private insurance. General taxation tends to be a more progressive form of health care funding than either social insurance or private health insurance systems, and studies comparing the NHS with other European and with the American health care systems show the NHS to be among the more progressive.[5] Table 1 shows the progressivity of all financing and of public financing in a number of countries. The entry in the row headed "Public Finance" in the United Kingdom is basically the financing of the NHS (because 96 percent of NHS funding comes from

taxation). Thus both NHS financing and the overall financing of the UK health care sector are more progressive than those of the other countries in the table. In addition, because NHS payments are not based heavily on employment status, differences in levels of payment among individuals who have similar incomes but different employment status—which gives rise to horizontal inequity in a tax system—are less in the NHS than in several European systems (where payroll taxes play a larger part in health care financing), or in the American health care system (where much insurance is supplied though employers).

Because the reforms did not change the method of financing NHS care, any changes in the progressivity of financing of the NHS will depend on changes in the structure of general taxation. During the 1980s, the U.K. government increased the proportion of indirect to income taxation, reducing the progressivity of the tax system, but fewer such changes have taken place since the introduction of the NHS reforms. So Table 1 is probably a pretty good guide to the present progressivity of financing for the NHS.

Although the NHS dominates the health care market in the United Kingdom, a small private health care market does exist (see box on p. 67). In an attempt to encourage the growth of private insurance, the reforms included tax breaks for the purchase of private health insurance, which does not cover residential (home health) care. Purchase of such insurance does not allow individuals to reduce their tax contributions to the NHS and as a result, it is purchased predominately by the better off. The tax breaks do not seem to have stimulated much increase in take-up, but recent research suggests that the performance of the NHS—in particular longer waiting lists for medical procedures—affects the purchase of private health insurance.[6] Changes which would have a large effect on waiting lists could, therefore, have a significant impact on the progressivity of health care financing in the United Kingdom. NHS funding has grown considerably since the reforms, but projected growth is much lower.[7] If this leads to declines in public-sector care and longer waiting lists, with a consequent increase in the purchase of private insurance, then poorer individuals, less able to afford such insurance, will be the losers.

Table 1
Kakwani Indices of Progressivity of Health Care Payments in Various OECD Countries

Country	Public Finance	Total Payments
U.K. (1992)	0.122 (for the NHS)	0.06
U.S.A. (1987)	0.09	-0.13
Netherlands (1992)	-0.099	-0.067
Sweden (1990)	0.089	0.027

Source: E. van Doorslaer et al., "The Redistributive Effect of Health Care Finance in 12 OECD Countries," Working Paper no. 8, Equity Project, University of Sussex and ITMA, University of Rotterdam, 1996.

Note: The Kakwani index is a measure of departures from progressivity. The index is bounded between -2.00 and 1.00. A positive value indicates a progressive system.

On the financing side, the large equity changes are likely to be in the areas of payment for nursing homes and residential care, because both the Conservative and Labour parties are committed to general tax funding of the NHS. Reforms introduced alongside the NHS reforms both increased the role for private provision of nursing home and residential care and reduced the extent of tax financing for such care. Among those requiring such care, therefore, differences in access based upon income are likely to grow.

Allocations to purchasers

In a system of third-party payers with tax financing, the distribution of resources across consumers will depend on how the government allocates funds to the third-party payers. Funding to DHAs is by means of capitation payments—a fixed amount per area resident not registered with GPFH—adjusted for measures of need. The reforms have been accompanied by changes in the formula used to allocate money to DHAs. These changes have tended to redistribute financing towards poorer urban areas and away from rural and suburban areas.

But the far more contentious issue in funding is the allocation between the two types of third-party payer, the DHAs and the GPFHs. There are two issues to consider. The first is the repeated claim that GPFHs have been financed more generously than DHAs, and that this has led to a two-tier system in which patients of GPFHs get faster access to care or better care than individuals covered by DHAs. Establishing the basis for this claim or its validity has been difficult. Comparison of the allocations to the two types of buyer has been hindered by the poor quality of available data, and there is considerable regional variation. Evidence that GPFH patients have better access to hospital care is more widespread, but this may be for several reasons. First, the patients of some general practitioners may have always had preferential treatment because of informal networks between hospital doctors and family practitioners and because family doctors vary in quality. The reforms may simply have made this more visible (not unlikely, given that GPFHs were a self-selected group). Second, fund holders may have more income per patient. Third, fund holders are the marginal buyer for providers who have annual budget constraints and are operating in a market in which there is believed to be excess capacity. There is evidence that, in the face of competition, suppliers lower prices to GPFHs and may also increase quality.[8] This will allow GPFHs to get more and/or better treatment for their patients for a given budget. But this indicates greater mar-

ket power, not greater funding. Finally, if there are spillovers in production within hospitals, the actions of GPFHs may have improved services for all, not just for their own patients (who have, however, benefitted the most).

The second is the opportunity for GPFHs to "skim the cream"—to pick patients who are expected to incur the lowest net cost (that is, net of the price reimbursed per patient). The budgets of GPFHs have been based upon historic cost, which may reduce the incentives for skimming the cream, but some incentive remains because, as noted, GPFHs keep any savings from their budgets. Research has shown that the composition of the fund holder patients is such that many of the high-cost patients could be identified with relative ease by the fund holder.[9] To date, there is little evidence of systematic discrimination against patients who might need high-cost care, but this may be at least partly because the issue has not been properly researched.

The impact of competition and contracting

The intention of the reforms was to introduce competition on the supply side, through the mechanism of contracting. The evidence on efficiency gains at a global level is limited. It appears that competition has had some impact on pricing behavior. Prices appear to be lower where market competition is higher, and there is anecdotal evidence that hospitals are charging GPFHs less by cost shifting from their GPFH buyers to the more passive DHA buyers.[10] Although the DHAs are the larger buyers, their concern to maintain continuity in local supply appears to have made them very much the weaker party in negotiation with NHS trusts. The changes have had little discernible effect on productivity, but have considerably increased transactions costs (primarily the cost of contracting).[11] There is limited evidence that GPFHs have reduced the growth in prescription drug costs and have shifted care at the margin from hospital to physician's office/outpatient services, again accompanied by higher administration costs.[12]

The impact of these changes on the distribution of care across different groups is completely uncharted. Research on the performance of the NHS through the later 1970s and 1980s showed, using simple measures of morbidity, that the NHS appeared to meet its policy objective of allocating care on the basis of need.[13] As Table 2 indicates, there appear to be no systematic differences across income groups in the receipt of NHS care, once the higher health-care needs of lower-income individuals are taken into account. Compared with other European health care systems, the whole of the prereformed U.K. health care system (including the private sector) was more progressive than some and less progressive than others.[14] There are currently no comparable results for the postreform NHS.

Table 2
Percentage Shares of NHS Expenditure, Standardized for Need, 1974–1987

Income quintile	1974	1982	1985	1987
Bottom	24.6	22.5	22.7	22.7
2nd	21.6	20.3	22.7	21.2
3rd	19.3	21.1	19.7	19.9
4th	17.9	21.7	18.9	19.8
5th	16.6	14.5	16.1	16.3
Concentration index	-0.083	-0.092	-0.070	-0.062

Source: C. Propper and R. Upward, "Need, Equity and the NHS: The Distribution of Health Care Expenditure 1974–87," *Fiscal Studies* 13, no. 2 (1992): 1–21.

Note: The concentration index presented at the bottom of each column is a summary measure of the extent of departure from proportionality. It is bounded between -1 and 1, a positive value indicating a regressive distribution.

The finding that the NHS allocates care according to need has been challenged by some smaller-scale studies, particularly of the behavior of family practitioners, who are pivotal in the NHS because they both deliver primary care and act as gatekeepers to secondary care. These studies indicate that wealthier individuals get more, or better, care.[15] As noted above, it may be that the two-tier findings for GPFHs are simply a continuation of prereform behavior that is made more visible by the operation of the internal market.

The more passive behavior of the DHAs also has implications for equity. The challenge in health care systems with monopsonistic third-party payers is to create incentives for these buyers to get providers to be more efficient and responsive to consumer needs. To date, despite considerable monitoring of DHAs, the Department of Health has had mixed success, and the purchasing function in DHAs remains underdeveloped. This means that those populations covered by DHAs may get poorer-quality or less timely care. Again, there is no published research to support or refute this hypothesis.

Finally, it is well established that different forms of contracts give providers greater or lesser incentives for selecting patients whose medical care costs less than the reimbursement received for treating them. The contracts negotiated by DHAs with hospitals tend to give the hospital a lump sum, with some indication of the total amount of services to be provided, either at global level and/or at speciality level. Such payments are prospective, and, as in any prospective payment method, give the hospital some incentive to select lower-cost patients (generally the healthier ones). However, such contracts also allow a hospital to cross-subsidize between patients, that is, to use the surplus from the patients who are healthier to subsidize care of the sicker. Cross-subsidization is a goal that hospitals appear to pursue when competition is weak. Given that competition is weak, or rather that hospitals appear to have greater bargaining

power than their DHA buyers, we might expect hospitals to cross-subsidize rather than engage in patient selection.

But there is one area in which patient-selection does appear to be a more important issue—care for the elderly. Changes in the allocation of responsibility between the health service and the social services for this group, coupled with social service budget cuts, have meant reductions in payment for long-term care for the elderly, generating concern that poorer elderly persons may have more limited access to such care.

The need for more research

Much of the discussion in this article has been in terms of the potential for equity differences, because the knowledge base upon which to analyze the impact of changes in the NHS is very limited. However, there is a growing body of research on the NHS reforms and on their equity consequences. Some studies are examining whether the overall goal of allocation according to need has been maintained. Others are focusing more on equity in treatment for particular conditions and in treatment of different age groups, and still others are examining the behavior of GPFHs in making referral decisions and patient selection. When these are complete, we will know more about both the macro and the micro equity impacts of the reforms.[16]■

---

[1]W. P. M. M. van de Ven, F. T. Schut, and F. H. Rutten, "Forming and Reforming the Market for Third-Party Purchasing of Health Care," *Social Science and Medicine* 39, no. 10 (1994):1405–12; C. Propper, "Agency and Incentives in the NHS Internal Market," *Social Science and Medicine* 40, no. 12 (1995):1683–90; A. Maynard, "Can Competition Enhance Efficiency in Health Care? Lessons from the Reform of the UK National Health Service," *Social Science and Medicine* 39, no. 10 (1994):1433–46.

[2]C. Propper, "Market Structure and Prices: The Responses of Hospitals in the UK National Health Service to Competition," *Journal of Public Economics* 61 (1996): 307–35

[3]Under the original reform proposals, GPFHs had to have practices with at least 11,000 patients. This limit was lowered to 9,000 in 1991 and further reduced to 7,000 in 1992.

[4]See H. Glennerster, A. Cohen, and V. Bovell, "Alternatives to Fundholding," Welfare State Programme, London School of Economics Discussion Paper WSP 123, 1996.

[5]A. Wagstaff, E. van Doorslaer, et al., "Equity in the Finance of Health Care; Some International Comparisons," *Journal of Health Economics* 11 (1992): 361–87; E. van Doorslaer, et al., "The Redistributive Effect of Health Care Finance in 12 OECD Countries," Working Paper no. 8, Equity Project, University of Sussex and ITMA, University of Rotterdam, 1996.

[6]T. Besley, J. Hall, and I. Preston, "The Demand for Private Health Insurance: Do Waiting Lists Matter?" Institute for Fiscal Studies, London, Working Paper W96/7, 1996.

[7]A. Maynard and K. Bloor, "Introducing a Market to the United Kingdom's National Health Service," *New England Journal of Medicine* 334 (Feb. 29, 1996): 604–8.

[8]C. Propper, D. Wilson, and N. Soderlund, "The Effects of Regulation and Competition in the NHS Internal Market: The Case of GP Fund-Holder Prices," mimeo, Department of Economics, University of Bristol, 1996.

[9]M. Matsaganis and H. Glennerster, "The Threat of 'Cream-Skimming' in the Postreform NHS," *Journal of Health Economics* 13 (1994): 31–64.

[10]Propper, "Market Structure and Prices"; Propper et al., "The Effects of Regulation"; S. Ellwood, "Pricing of Services in the UK National Health Service," *Financial Accountability and Management* 12 (1996): 281–301.

[11]N. Soderlund, I. Csaba, A. Gray, R. Milne, and J. Raftery, "The Impact of the NHS Reforms on English Hospital Productivity—an Analysis of the First 3 Years," *British Medical Journal*, forthcoming; Maynard and Bloor, "Introducing a Market."

[12]J. Dixon, and H. Glennerster, "What Do We Know about Fundholding in General Practice?" *British Medical Journal* 311 (Sept. 16, 1995): 727–30.

[13]C. Propper and R. Upward, "Need, Equity and the NHS: The Distribution of Health Care Expenditure 1974–87," *Fiscal Studies* 13, no. 2 (1992): 1–21; O. O'Donnell, and C. Propper, "Equity and the Distribution of National Health Service Resources," *Journal of Health Economics* 10 (1991): 1–21.

[14]E. van Doorslaer, A. Wagstaff, et al., "Equity in the Delivery of Health Care: Some International Comparisons," *Journal of Health Economics* 11 (1992): 389–411.

[15]J. Le Grand, "Equity in the Delivery and Finance of Health Care: A Comment," *Journal of Health Economics* 10 (1991): 57–64.

[16]Inter alia, G. Davey-Smith and C. Propper, University of Bristol, are analyzing the receipt of NHS care for arthritis and other common conditions, the consortium headed by the Kings Fund, London, is examining the behavior of GPFHs, and a consortium led by E. van Doorslaer, University of Rotterdam, the Netherlands, and A. Wagstaff, University of Sussex, U.K., is examining equity in the financing and delivery of health care at the macro level.

Robert Lampman, Emeritus Professor of Economics at the University of Wisconsin–Madison, died on March 4, 1997. Lampman's career exemplified the Wisconsin Idea of academic abilities in service to the state. He could also be regarded as the intellectual father of the War on Poverty and of the Institute for Research on Poverty.

As a staff member of President Kennedy's Council of Economic Advisers and a key author of the historic chapter on poverty in the 1964 *Economic Report of the President*, Lampman played an instrumental role in calling the nation's attention to poverty in America. Largely because of him, the Institute for Research on Poverty was established at the University of Wisconsin–Madison in 1966. The Institute's first major project was the negative income tax (NIT) experiment, a pioneering research undertaking to test the effects of income maintenance programs. Lampman played a central role both in the policy debates about the NIT and in shaping the NIT experiment.

He continued to provide guidance to IRP and to generations of graduate students at the University of Wisconsin while producing innovative work, theoretical and quantitative. In *Ends and Means of Reducing Income Poverty* (Chicago: Markham Press, 1971) he presented the range of possibilities for the reduction of income poverty. By 1984 his analytic net had widened to incorporate in an accounting system all social welfare spending in the United States. Using the system that he devised, he was able to estimate what the expansion of social welfare has accomplished and what its costs have been (*Social Welfare Spending: Accounting for Changes from 1950 to 1978* [Orlando: Academic Press]). His system made it possible to compare social welfare spending in the United States with spending in other welfare states.

Lampman was born in Wisconsin and received his undergraduate degree at the University of Wisconsin in 1942. He served in the Navy from 1942 to 1946, then returned to the university, receiving his Ph.D. in 1950. After ten years on the faculty of the University of Washington, he joined the Wisconsin faculty in 1958 and served there until his retirement in 1987. In 1972 he was named William F. Vilas Research Professor. In May 1989, his work was honored by an IRP-sponsored conference at the University and by an accompanying special issue of *Focus* (vol. 12, no. 3).

Lampman's research and career were marked by an open, questioning approach admired by all who came into contact with him in the course of academic life. The following quotations from some of his writings illustrate his outstanding character and intellect.

On Poverty

It is paradoxical that in this time of great prosperity in the richest nation in the world there should still be a substantial part of our population with incomes far below what is thought of as the American standard.

In the period since World War II great advance has been made in raising the total national income and the income per family and per person. Has similar progress been made in reducing the numbers in low-income status? What are the socioeconomic characteristics of the group that remains in low-income status? In what respects does this group differ from the total population? To what extent do "handicapping" characteristics of old age, nonwhite color, loss of breadwinner, and low education seem to explain the persistence of low incomes? Is the low-income problem peculiarly associated with any region or occupation or family size; are any important number of our children afflicted by low family income? These are questions that relate to an appraisal of the present low-income problem. (p. 3)

. . . . . .

It is notable that reduction of the numbers in poverty has been accomplished with little change in the share of total income going to the lowest income groups. Government policy aimed at moderating economic inequality seems merely to have prevented a fall in the share of income of the relatively poor. A more aggressive government policy could hasten the elimination of poverty and bring about its virtual elimination in one generation. (p. 4)

From "The Low Income Population and Economic Growth," Study Paper No. 12, Joint Economic Committee, Congress of the United States, December 16, 1959.

The general theme is Poverty, and the problem of working in the United States today against a very ancient enemy of mankind. . . . universities are very much concerned with the development of not only knowledge for its own sake, but knowledge that will serve the needs of a democratic country.

From Robert Lampman's introduction to an IRP conference on "Poverty Research, Communications, and the Public," April 1966

The three theories about causes of poverty . . . show ways in which our system selects people to be poor. These have to do with risks, barriers, and personal differences. Some remedies are suggested by this three-point analysis.

It is consonant with the "risk" theory that poverty will be minimized to the extent that frequency of disability, premature death, family breakup, loss of savings, and unemployment can be reduced. To the extent that a basic risk cannot be done away with, individuals, private groups, and governments can take steps to insure against the loss of income associated with the risk.

Poverty is sometimes seen as the result of failure of successive lines of defense against it. The first line of defense is earnings. The second line of defense is property income and savings. The third line is insurance, assistance, and charity. Note that this phrasing of the problem seems to assume that the normal position is nonpoverty and that the problem is to prevent people from falling away from this norm. However, some may never have reached the norm in the first place. Another framework for consideration of risk is suggested by what might be called the life-cycle classification of causes of poverty according to phase of life. Some persons are born into poverty. Others enter it in childhood because of death or disability of a parent. Some enter it in adulthood because of a personal disaster or failure to insure against all risks. In this "risk theory" the emphasis is upon randomness and historical accident, as in a fable Carl Sandburg told of two cockroaches washed off a roof by a rainstorm. One fell in a rock pile and the other in a garbage pail. When they met again the first cockroach asked the other, "How does it happen that you are so fat while I am so lean?" The answer was, "It is because of my foresight, industry and thrift."

A second class of remedies, which are identified with the "social barriers" theory of poverty, includes such things as breaking down practices of racial discrimination in hiring, housing, and education: improving mobility of labor from rural to urban occupations; and bettering chances for women and elderly people to work in a wider range of occupations. These remedies also include improving the environment of the poor and integrating the poor with the rest of the community. William Penn alternated the wide and narrow streets in Philadelphia so that the rich and poor would know each other.

The "social barriers" theory says that if poor people are different from the nonpoor, it is because of the fact of poverty rather than because of innate traits. One hundred years ago the Irish drank because they were poor, rather than vice versa. According to this theory, poverty itself is what is transmitted. It is an inheritable disease. The observable personal differences which are asserted to be symptoms rather than causes will abate if the conditions of poverty are remedied. Here the analogy to public health matters is clear.

A third theory is that people are selected to be poor on the basis of personal differences (which may or may not be transmissible) of ability, of motivation, of moral character, of will and purpose. Some philosophers consider life a matter of survival of the fittest and a contest which rewards the morally as well as the financially elect, and appropriately visits the punishments and rewards unto the second or third generation. However, if we want to reduce poverty, we may strive to reduce personal differences of ability and motivation. Here again there is a wide range of steps that can be taken. (pp. 138–40)

From *Ends and Means of Reducing Income Poverty* (1971).

The Lampman Question

It is right to call the war on poverty—first enunciated in President Johnson's State of the Union message and promptly endorsed by Congress in the Economic Opportunity Act of 1964—a logical extension of Franklin D. Roosevelt's Social Security Act and Harry S. Truman's Employment Act. It is also correct to identify it as in the general pattern laid down by the more advanced welfare states of Western Europe. But no other President and no other nation had set out a performance goal so explicit with regard to "the poor." No one else had elevated the question, "What does it do for the poor?" to a test for judging government interventions and for orienting national policy.

This question served as a flag for the great onrush of social welfare legislation commencing in 1965 and the consequent expansion in the role of the federal government. When poverty became a matter of national interest, Washington moved into fields where state and local governments had held dominant if not exclusive sway up to that time. This movement was manifested by the enactment of such measures as Medicare and Medicaid, and aid to elementary and secondary education. It led to uniform national minimum guarantees in the food stamp program, in cash assistance to the aged, blind, and disabled (under the title of Supplemental Security Income), and in stipends for college students in the form of Basic Educational Opportunity Grants—all adopted in the first Administration of President Richard M. Nixon. Other interventions—notably equal opportunity legislation, the provision of legal services for and on behalf of the poor, and "community action"—made little impact on the budget, but reflected new efforts by the federal government to be an integrative force in national life. (pp. 66–67)

"What Does It Do for the Poor? A New Test for National Policy," *The Public Interest,* January 1974.

Recent IRP discussion papers

Mead, L. M. "Welfare policy: The administrative frontier." 1996. 23 pp. DP no. 1093-96.

Farber, N. B. and Iversen, R. R. "Transmitting values about education: A comparison of black teen mothers and their nonparent peers." 1996. 29 pp. DP no. 1094-96.

Reynolds, A. J. and Temple, J. A. "Extended early childhood intervention and school achievement: Age 13 findings from the Chicago Longitudinal Study." 1996. 38 pp. DP no. 1095-96.

Mead, L. M. "Are welfare employment programs effective?" 1996. 46 pp. DP no. 1096-96.

Bird, E. J. "Exploring the stigma of food stamps." 1996. 19 pp. DP no. 1097-96.

Levine, P. B. and Zimmerman, D. J. "An empirical analysis of the welfare magnet debate using the NLSY." 1996. 35 pp. DP no. 1098-96.

Klawitter, M., Plotnick, R. and Edwards, M. "Determinants of welfare entry and exit by young women." 1996. 43 pp. DP no. 1099-96.

Levine, R. and Zimmerman, D. "The intergenerational correlation in AFDC participation: Welfare trap or poverty trap?" 1996. 26 pp. DP no. 1100-96.

Meyer, D. R and Cancian, M. "Life after welfare: The economic well-being of women and children following an exit from AFDC." 1996. 34 pp. DP no. 1101-96.

Yelowitz, A. S. "Using the Medicare buy-in program to estimate the effect of Medicaid on SSI participation." 1996. 57 pp. DP no. 1102-96.

Hoynes, H. W. "Work, welfare, and family structure: A review of the evidence." 1996. 61 pp. DP no. 1103-96.

Hoynes, H. W. "Local labor markets and welfare spells: Do demand conditions matter?" 1996. 56 pp. DP no. 1104-96.

Dominitz, J. and Manski, C. F. "Perceptions of economic insecurity: Evidence from the Survey of Economic Expectations." 1996. 35 pp. DP no. 1105-96.

Plotnick, R. D. and Hoffman, S. D. "The effect of neighborhood characteristics on young adult outcomes: Alternative estimates." 1996. 22 pp. DP no. 1106-96.

Olson, C. M., Rauschenbach, B. S., Frongillo, E. A., Jr. and Kendall, A. "Factors contributing to household food insecurity in a rural upstate New York county." 1996. 28 pp. DP no. 1107-96.

Betts, J. R. "The impact of school resources on women's earnings and educational attainment: Findings from the National Longitudinal Survey of Young Women." 1996. 35 pp. DP no. 1108-96.

Yelowitz, A. S. "Did recent Medicaid reforms cause the caseload explosion in the Food Stamp Program?"1996. 40 pp. DP no. 1109-96.

Hauser, R. M. and Huang, M. H. "Trends in black-white test-score differentials." 1996. 57 pp. DP no. 1110-96.

Garasky, S. "Exploring the effects of childhood family structure on teenage and young adult labor force participation." 1996. 45 pp. DP no. 1111-96.

Kost, K. A. "'A man without a job is a dead man': The meaning of work and welfare in the lives of young men." 1996. 23 pp. DP no. 1112-96.

Holzer, H. and Neumark, D. "Are affirmative action hires less qualified? Evidence from employer-employee data on new hires." 1996. 47 pp. DP no. 1113-96.

Mauldon, J. "Predicting hunger and overcrowding: How much difference does income make?" 1996. 25 pp. DP no. 1114-96.

Holzer, H. J. "Employer demand, AFDC recipients, and labor market policy." 1996. 30 pp. DP no. 1115-96.

Sandefur, G. D. and Wells, T. "Trends in AFDC participation rates: The implications for welfare reform." 1996. 24 pp. DP no. 1116-96.

Smith, P. A. "The effect of the 1981 welfare reforms on AFDC participation and labor supply." 1997. 50 pp. DP no. 1117-97.

Yelowitz, A. S. "Will extending Medicaid to two-parent families encourage marriage?" 1997. 45 pp. DP no. 1118-97.

Holzer, H. J. "Why do small establishments hire fewer blacks than large ones?" 1997. 22 pp. DP no. 1119-97.

Frongillo, E. A., Jr., Olson, C. M., Rauschenbach, B. S. and Kendall, A. "Nutritional consequences of food insecurity in a rural New York State county." 1997. 21 pp. DP no. 1120-97.

Falcón, L. M., Tucker, K. and Bermudez, O. "Correlates of poverty and participation in food assistance programs among Hispanic elders in Massachusetts." 1997. 48 pp. DP no. 1121-97.

Holzer, H. J. and Ihlanfeldt, K. R. "Customer discrimination and employment outcomes for minority workers." 1997. 41 pp. DP no. 1122-97.

Bauman, K. "Shifting family definitions: The effect of cohabitation and other nonfamily household relationships on measures of poverty." 1997. 26 pp. DP no. 1123-97.

Wolaver, A. M., McBride, T. D. and Wolfe, B. L. "Decreasing opportunities for low-wage workers: The role of the nondiscrimination law for employer-provided health insurance." 1997. 63 pp. DP no. 1124-97.

Wu, L. L., Cherlin, A. J. and Bumpass, L. L. "Family structure, early sexual behavior, and premarital births." 1997. 35 pp. DP no. 1125-97.

Reynolds, A. J. "The Chicago Child-Parent Centers: A longitudinal study of extended early childhood intervention." 1997. 41 pp. DP no. 1126-97.

Geweke, J. and Keane, M. "An empirical analysis of income dynamics among men in the PSID: 1968–1989." 1997. 83 pp. DP no. 1127-97.

Brady, P. and Wiseman, M. "Welfare Reform and the Labor Market: Earnings Potential and Welfare Benefits in California, 1972-1994." 1997. 34 pp. DP no. 1128-97.

Sandefur, G. D. and Cook, S. T. "Duration of public assistance receipt: Is welfare a trap?" 1997. 34 pp. DP no. 1129-97.

Meyer, D. R. and Bartfeld, J. "Patterns of child support compliance in Wisconsin." 1997. 21 pp. DP no. 1130-97.

McBride, T. D. "Uninsured spells of the poor: Prevalence and duration." 1997. 35 pp. DP no. 1131-97.

Order form for FOCUS NEWSLETTER (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: 608-265-3119)

Name: _____

Address: _____

_____
City State Zip

(Multiple copies of any issue: $1.00 each)

Focus articles also appear on the IRP World Wide Web site, http://www.ssc.wisc.edu/irp/

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 1–June 30
❑ Discussion Papers and Reprints ($70.00)

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

❑ Please indicate here if this is a change of address.

Focus

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

UNIVERSITY OF WISCONSIN MADISON