3

Ongoing Major Research on Welfare Reform: What Will Be Learned

Peter H. Rossi

Maryland School of Public Affairs

Welfare Reform Academy

Committee to Review Welfare Reform Research

www.welfareacademy.org

Part of a forthcoming volume
Family Well-Being After Welfare Reform
Douglas J. Besharov, Editor

3

Ongoing Major Research on Welfare Reform: What Will Be Learned

Peter H. Rossi[*]

This paper assesses four major research projects currently underway, each designed to gather information on what is happening to low-income families subsequent to the welfare reforms instigated by the passage of the Personal Responsibility and Work Opportunity Reconciliation Act of 1996 (PRWORA). Each of the projects is planned to describe some aspects of the welfare system in some places in the United States and the conditions of low-income households before and, in some cases, after the provisions of PRWORA have gone into effect. Although data collection has begun and the broad outlines of the research designs have been laid out, all the projects are still in progress. Changes in plans are likely to be made, especially for data collection efforts, which have not yet begun. Some data have been collected in each project, but the data collection will not be completed for several years. And, of course, except for the waiver experiments, none of the data collected has been analyzed in detail, although some findings have been released.

This review is based on only a few published documents: Most of them are unpublished memoranda, proposals, drafts of papers, and questionnaires intended to be used in the field. *Accordingly, this review must be regarded as a description of work still in progress.*[1] Of course,

---

[*]Peter H. Rossi is S. A. Rice Professor Emeritus at the University of Massachusetts (Amherst).

[1]Author's note: The five research organizations involved—the Bureau of the Census, the Manpower Demonstration Research Corporation, the Urban Institute, Mathematica Policy Research, Inc., and Abt Associates, Inc.—generously made available the materials on which this review is based. I am especially grateful to Daniel Weinberg and Michael McMahon (Bureau of the Census), Gordon Berlin and Charles Michalopoulos (Manpower Demonstration Research Corporation), Fritz Scheuren, Genevieve Kenney and Anna Kondratas (Urban Institute) and Howard Rolston (Department of Health and Human Services), all of whom sent me materials and patiently answered my questions. I circulated earlier versions of this paper to them, and their comments helped correct errors in my descriptions of the projects. I did not

a final report reviewing findings can only be attempted some years hence, when data collection will be complete and analyses will have been written.

Summary

America's public welfare system "as we knew it" changed dramatically with the passage of the Personal Responsibility and Work Opportunity Reconciliation Act of 1996. The major change was that Aid to Families with Dependent Children (AFDC) was abolished and replaced by Temporary Assistance for Needy Families (TANF), which mandated a five-year time limit on federally funded assistance, new work requirements, and a cap on federal funds based on historical AFDC funding. Otherwise, it is a program to be designed in detail and run by the states, emphasizing employment, marriage, and reduced out-of-wedlock childbearing.

Although few tears were shed over the death of AFDC, both advocates and opponents of PRWORA became concerned about how the new law, especially the TANF provisions, would affect low-income families. PRWORA advocates worried whether TANF would reduce welfare dependency and move families off the welfare rolls as intended. Opponents worried that low-income families would be plunged into abject poverty and that the children of the poor would suffer greatly. Fueled by these concerns, several major research projects were started and ongoing projects modified in order to provide empirical data on the changes wrought by TANF and other PRWORA-mandated changes and to estimate the new program's impact.

This paper assesses the prospects of each of four major research projects to provide empirically based findings concerning how well low-income families fared before and after PRWORA, how those prospects varied from state to state, and whether those trends can be credibly attributed to PRWORA.

The four research programs are as follows:

- **The Survey of Program Dynamics.** The Bureau of the Census is extending the Survey of Income and Program Participation (SIPP) to create the Survey of Program Dynamics (SPD), a longitudinal survey consisting of repeated interviews of a national sample of households from several years before PRWORA to several years after enactment. SPD oversamples low-income households.

- **Assessing the New Federalism**. The Urban Institute's extensive study, Assessing the New Federalism, is examining the devolution shifting many welfare programs from the federal to the state level. Administrative studies will describe the

always agree with some of their judgments, but I did take them into account in this revision.

changes in health and welfare systems in each state and take an intensive look at thirteen states both pre- and post-PRWORA. Much of the Urban Institute project will center on a pre-PRWORA survey and several post-PRWORA national household sample surveys—the National Survey of America's Families (NSAF)—with subsamples large enough to support precise estimates for each state in each of the thirteen states in which intensive administrative studies are being undertaken. NSAF also will oversample households at 200 percent or less of the poverty level, supporting detailed findings for low-income families. Additional surveys are planned for 2001 and, tentatively, for 2003.

- **Project on Devolution and Urban Change.** The Manpower Demonstration Research Corporation (MDRC) is undertaking the Project on Devolution and Urban Change, which consists of intensive studies of poor households in four major cities—Cleveland, Miami, Philadelphia, and Los Angeles. Using administrative data, the program participation and earnings of cohorts of poor households entering the Food Stamp or AFDC/TANF programs at given points in time will be tracked before and after TANF went into effect. In addition, it will include surveys of a sample of AFDC/TANF households, intensive studies of the TANF programs in the four cities, institutional analyses, and ethnographic studies within selected poor neighborhoods.

- **Child Impact Waiver experiments.** The U.S. Department of Health and Human Services (HHS) is augmenting five ongoing randomized waiver experiments with measures of effects on children in projects referred to as the Child Impact Waiver experiments. The five waiver experiments began before PRWORA and compare randomly selected families who have been receiving welfare benefits under the old AFDC rules (control-group families) with randomly selected families who received benefits under the waivers (experimental-group families). Because these extensions are among the best of the waiver experiments and will provide information on effects on children, they are considered a fourth major research project for the purposes of this paper. Most important, the waiver provisions being studied resemble quite closely the TANF provisions enacted by the states being studied. The five states and the contractors involved are Florida (MDRC), Minnesota (MDRC), Connecticut (MDRC), Iowa (Mathematica Policy Research, Inc.), and Indiana (Abt Associates).

The question that most interests the policy community is: What have been the net effects of TANF (that is, effects uniquely attributable to TANF) on the employment and well-being of low-income households? The "gold standard" design for estimating net effects is the randomized experiment. Only the Child Impact Waiver studies are experiments; the other three research projects provide nonexperimental data that cannot support estimates of net effects that are as credible. Because they must rely on "before-and-after" research analyses, they will be unable to distinguish between the effects of welfare reform and those of other economic and policy changes occurring at the same time. The best use of the nonexperimental research projects is to provide descriptive information about what has happened to low-income households and families as they experience TANF.

Weaknesses in research design. Each research project suffers from inadequacies that diminish its usefulness to greater or lesser degrees, as follows:

- **Survey of Program Dynamics.** The SPD's most attractive feature is its promise to provide longitudinal data on national samples of households for five years before and after the enactment of PRWORA. However, the most serious problem facing the SPD is a low response rate, with just 50 percent of the original sample responding in 1998 and 1999. Especially worrisome is that higher percentages of low-income households stopped cooperating. Furthermore, it is likely that less than one-quarter of the sample will produce complete longitudinal data covering all or even most of the interviews over the ten-year period. The Bureau of the Census has begun efforts to reach and obtain the cooperation of the nonresponding households, hoping to raise the response rates to above 60 percent. Even if efforts to reduce future attrition are that successful, the rates will not be high enough to satisfy most researchers.

- **Assessing the New Federalism.** A central component of the Assessing the New Federalism project is the NSAF, a series of nationally representative surveys conducted in 1997 and 1999, with additional surveys planned for 2001 and, possibly, 2003, to permit comparisons between households before and after welfare reform. Response rates for the first NSAF survey were average for well-run national telephone surveys—65 percent for families with children and 62 percent for families without children, but "average" may not be good enough for surveys that are highly policy relevant.[2] NSAF staff have attempted to compensate for low response rates by weighting the data. The large samples for each of the thirteen states also make it possible to examine how families in states with different TANF plans have fared. However, the decline in welfare rolls has considerably reduced the number of welfare families in the survey taken after TANF went into effect. The small sample sizes within each state will restrict the ability of analysts to estimate the impact of welfare reform, especially subgroup differences at the level of individual states.

---

[2]Editor's note: In their response to Rossi, the Urban Institute researchers argue that comparisons to other surveys should be based on an alternate "weighted" response rate that is slightly higher (about 70 percent).

- **Project on Devolution and Urban Change.** MDRC's Project on Devolution and Urban Change focuses on welfare recipients in four major cities, combining administrative data from 1992–2002 with surveys of welfare mothers who were on the rolls in 1995 and a yet-to-be-determined year in the post-PRWORA era. Response rates for the first survey in 1997 were very good, averaging 79 percent across the four sites. The project should provide a rich—and textured—picture of the characteristics and circumstances of the most vulnerable families in the cities studied. However, the findings cannot be generalized to the broader welfare population nationally or even to other urban neighborhoods.

- **Child Impact Waiver experiments.** HHS is augmenting five ongoing waiver randomized experiments, comparing families who have been receiving welfare benefits under the old AFDC rules with families who received benefits under the state's welfare reform plan. The importance of these continuing experiments is considerable. Because the experiments resemble each state's TANF program and are evaluated using a rigorous evaluation, they represent perhaps the best attempt at measuring the impact of welfare reform in selected states. Although the projects all rely on the universally preferred randomized experiments, generalizing findings to other states is clearly hazardous. It also is not clear whether the experimental conditions have been maintained with fidelity. Given the widespread publicity accompanying welfare reform, it is possible that control-group families may not have been subjected to AFDC rules and that such families did not fully realize that their welfare benefits were not subject to state TANF rules.

What will and will not be learned from the PRWORA research projects? Assuming successful completion of data collection for the four projects, what will be learned from them? There can be little doubt that once released, the descriptive findings will be met with intense interest. Public and partisan interest in how the poor are faring under PRWORA is strong. These studies will provide that desired information in great detail.

- **National-level descriptive findings.** Both the SPD and Assessing the New Federalism will provide findings concerning how American households were doing socioeconomically before and after PRWORA. For the nation as a whole, we will know whether earnings and employment of low-income households improved or declined. We also will know whether in certain respects the children in low-income households are better or worse off.

- **State-level descriptive findings.** Assessing the New Federalism will provide detailed information on thirteen states, and the Child Impact Waiver experiments will provide information on five states.

- **Local-level findings.** The Project on Devolution and Urban Change will provide detailed descriptive data on the welfare population of four large cities. The five Child Impact Waiver experiments will provide similar information in selected localities.

The prospects for information on the changes in the condition of the poor that can be credibly attributed to PRWORA are not as rich. The five Child Impact Waiver experiments will provide good impact estimates for the five TANF-like programs, but the extent to which those findings will be able to be generalized to the nation as a whole is limited.

As for the nonexperimental studies—SPD, NSAF, and Urban Change—their before-and-after designs generally will not support very credible impact assessments. Nevertheless, some attempts to estimate the net effects of PRWORA generally, or TANF specifically, may turn out to be more than merely suggestive, especially when findings are strong, consistent, and robust. For example, if analysis using NSAF data were to find that states with more generous earnings disregards had relatively higher caseloads, holding other interstate differences constant, and that such effects were muted when combined with a strict time limit or work requirement, then the findings could be regarded as supporting a causal inference, if they held up under different specifications and were confirmed in comparable analyses in the SPD and UC data sets, as well as in the Child Impact Waiver experiments. Such a convergence of findings across data sets, however, is not likely to occur often.

Other problems hamper research on PRWORA. First, devolution means that each state can design a different version of TANF. As a result, PRWORA cannot be evaluated as a national program—only state programs can be evaluated. Second, state programs can be expected to change over time as states change them in reaction to experiences. Third, larger forces compete with the programs. If labor market conditions for entry-level jobs change over time, those changes may affect the employment of welfare clients more than TANF. Fourth, some of the effects of PRWORA can be expected to be manifest quickly but others may take a number of years to show.

Welfare reform research in the future. In the near term, it is important to support some of the ongoing research projects by providing additional resources to strengthen their contributions. In particular, SPD will not be very useful unless response rates can be materially improved. Accordingly, it is important that the funds the Bureau of the Census needs in order to raise the response rates to SPD be appropriated. It is heartening that some funds are already at hand, but the bulk of the funding is not yet forthcoming. If SPD can be materially improved, it will provide extremely useful information on the changes accompanying welfare reform.

The effort to raise response rates should be monitored carefully: If at some point it becomes clear that response rates will not be materially improved, then the effort ought to be discontinued and the unexpended funds put to some better purpose (see below). In addition, serious consideration ought to be given to actions ranging from releasing the data sets with strong warnings about their limitations to suppressing their release entirely.

The prospects for improving NSAF through investment of additional resources do not appear to be good. The Urban Institute and Westat have done as much as possible to compensate for NSAF's weaknesses. The low response rate to NSAF-I is troubling and is irremediable except through careful weighting. However successful the weighting scheme may be—and even if NSAF-II has achieved a more acceptable response rate—two surveys, before-and-after, are quite a weak design for estimating net effects. The major value of the Urban Institute studies will come from the detailed descriptive data for the thirteen states. It will be difficult to capture the diversity of state policies in a statistical model, given the many variations chosen by states and the frequency of their change.

MDRC's Project on Devolution and Urban Change is not far enough along to make any judgment concerning its prospects. Especially critical will be its ability to analyze cohort experiences using administrative data. Assuming successful statistical modeling of quality data, Urban Change will provide good estimates of effects within four important localities, supplemented by qualitative data on four local welfare systems. It is clearly too early to judge whether any steps can be taken in the short term to improve Urban Change or, indeed, whether improvements will be needed.

The ongoing waiver experiments could prove to be valuable if strong efforts are made to ensure the fidelity of control-group conditions in each experiment. The danger is great that the control groups will be treated inappropriately and that the members of the control groups will not understand that they are not subject to state TANF rules. Maintaining the integrity of the control groups means not only more effort on the part of the contractors but also the possibility that additional funds need to be given to support training of agency personnel and to provide for more frequent reminders to control-group members of the special rules governing their welfare benefits. Unfortunately, given that the experiments have been underway for nearly four years since the start of TANF, it may now be too late to bolster their integrity.

It is also quite clear that when the four projects have been completed and analyzed, their findings will leave many critical questions unanswered. Almost certainly, PRWORA will be shown to be successful in meeting some of its goals in some of the states and failing to meet other goals in others. Questions will be raised about the effectiveness of time limits, family caps, income disregards, and other elements of the reform bundle. To answer those questions, further research will be needed. What form should such research take?

Perhaps the best strategy over the next decade or so would be to authorize and fund randomized experiments testing variations on the administrative and policy bundles. Those studies could be accomplished through state initiatives, but they are not likely to happen without federal funding. For example, questions about the effectiveness of family caps in reducing fertility might best be answered by conducting randomized experiments in which the experimental group is subject to family caps and the control group is under a no-family-caps condition (or vice versa). Randomized experiments can also be designed to observe the effects of varying the generosity of income disregards. Factorial experiments might be started to test the effects of various combinations of provisions that make up administrative and policy bundles.

This research strategy should lead to the accumulation of knowledge about how best to design welfare to achieve the dual objective of providing a safety net for the poor and facilitating entry into employment and higher income. A more expanded version of the current strategy of HHS's Administration for Children and Families (ACF) is proposed here. The current ACF experiments are not designed to unbundle TANF as much as to test proposed additions to the bundle. To understand how the bundles work, it is necessary to design experiments that vary such critical elements as earnings disregards, family caps, and time limits.

Some progress has been made toward implementing this strategy: ACF has funded three experiments and has issued a request for proposals for a fourth to be funded in 2000. All the experiments are designed to test measures aimed at improving TANF. For example, an experiment in Virginia will test the effectiveness of postemployment services in helping TANF clients retain their newly obtained employment. An evaluation planned in the future will involve four to ten states in MDRC-run experiments on measures aimed at employment retention and advancement in employment. The new experiments are patterned after the waiver strategy followed in the last decade or so of AFDC.

Studies also are needed to provide detailed descriptions of how the poor will fare under the welfare policy changes instigated by PRWORA and under whatever other policy changes occur. We should be planning now how best to collect the data that will support an empirically based understanding of what is happening to the poor and what policy changes are likely to improve their condition.

When AFDC was a more or less uniform national program, national surveys such as the Current Population Survey (CPS) or the SIPP may have served the purpose of monitoring the well-being of the poor. However, as discussed earlier, devolution has meant that state-level rather than national-level data are needed. An obvious move would be to enlarge sample sizes of existing ongoing national surveys to provide adequate state sample sizes. NSAF provides a good example in its selection of a small sample of critical states, an approach that the national surveys, including SPD, might want to emulate.

It is likely that in the end, SPD will not be very useful. Hence, serious consideration ought to be given to bolstering other ongoing large-scale surveys. In particular, it would be useful to augment SIPP and CPS by enlarging their sample sizes, especially bolstering their coverage of poor families. Ideally, I would like to see the sample sizes in at least the largest states increased enough to support state estimates.

Up to this point, the CPS has provided good monitoring data on the condition of the poor for the nation as a whole. Expanding the CPS sample to provide detailed data on a sample of states—and expanding CPS variables to include more information on how families with children are faring would be extremely useful. Additional efforts also should be made to address the problem of underreporting of welfare receipt. A parallel expansion of SIPP to conduct annual panel studies in a sample of states, especially in the ten to fifteen states that contain most of the poor, would be able to provide information on post-PRWORA changes in some detail. I recommend that the National Research Council of the National Academy of Sciences or a similar body examine the suitability of using SIPP for this purpose, paying special attention to and attrition and nonresponse and their impact upon obtaining valid and reliable analyses.

Finally, a serious issue is how to promote responsible analyses of these data sets. Neither the research nor the policy communities will be content with only descriptive analyses. If Wisconsin poor families are better off (or worse off) in 2001 than they were in 1997 but California poor families show an opposite pattern, then some analysts certainly will try to discern whether the differences between the two states' versions of TANF are the source of the difference. To some extent, we can expect that competition among analysts will provide constructive criticism. In any event, those who release public data sets should warn potential users about the limitations of their data as well as provide full and detailed documentation about the data sets.

Research on PRWORA

After decades of relative stability in structure, public welfare began to change in radical ways in the 1980s. Under encouragement from the Bush and Clinton administrations, state welfare departments were urged to apply for waivers permitting them to depart from federal regulations to try out new ways of providing support to the poor under Aid to Families with Dependent Children. Most states applied for and were successful in receiving waivers. When the Personal Responsibility and Work Opportunity Reconciliation Act was enacted in August 1996, many of the states' changes—and additional modifications—were institutionalized.

In the widest sense, the welfare reform receiving so much attention today consists of the cumulative changes made over the past ten to fifteen years. "Welfare as we knew it" is really welfare as it was in the late 1980s. By the time PRWORA was enacted, the "old" welfare in most states had been transformed, in some instances radically so.

An important change in welfare also was underway by the time PRWORA was enacted. Average monthly enrollment in AFDC began to decline in 1994 and continued to decline after

PRWORA. Some of the decline can be attributed to economic prosperity and the accompanying high employment rates, but as the decline approached 50 percent by mid-1999 (compared with 1994), it also became apparent that fewer families were applying for welfare. It appears that in the early years, the decline went largely unnoticed in either policy or welfare research quarters. Perhaps PRWORA accelerated the downward trend in enrollment, but perhaps not. In any event, this long-term trend would become an important aspect of the social changes accompanying PRWORA and, as will be discussed later, presents a vexing problem for existing welfare reform research.

The major provisions enacted under PRWORA were:

- The abolition of AFDC as an entitlement program, to be replaced by TANF, a time-limited welfare program with federal support to each state limited by historical funding patterns.

- Strong emphasis on moving TANF recipients off the rolls and into employment.

- Discretionary powers given to the states to design their own versions of TANF. Significant but less drastic changes were made in other welfare programs such as the Food Stamp Program, Medicaid, and Supplemental Security Income.

Within broad limits the states may design their own versions of TANF. Some states have elected to continue policies they put in place under AFDC waivers received prior to the passage of PRWORA. Other states have adopted shorter time limits than required for TANF. All states have changed, although the changes in each state differ. We can safely anticipate that more changes will occur over the coming years as states try new policies, find some wanting, and move on to implement modifications.

Although federal regulations can be expected to produce some degree of uniformity across states, it is likely that no two states will develop exactly the same set of provisions in their welfare programs over the next few years. Furthermore, it also is likely that many states will modify and change their programs in response to implementation problems or pressures from interested constituencies. It is also possible that federal TANF program requirements might change: For example, if time limits were found to produce extreme hardships for some families, Congress might very well modify those provisions or allow the states to do so. In addition, some states may establish state-financed programs designed to mitigate perceived hardships for some groups of poor families. Indeed, some states assumed the costs of providing food stamps to legal immigrants when the initial provisions of PRWORA declared food stamp eligibility restricted to citizens.[3]

The evaluations of the effects of some of the waivers put in place over the past decade can provide at least some hints about the effects to be expected from some features of the state TANF programs. Among the more promising waiver experiences are those that were put in place accompanied by randomized experiments intended to evaluate their effects and which were continued after PRWORA took effect. More than a dozen states have continuing AFDC waiver experiments. Strong inferences about TANF effects can be made for states in which the waiver provisions being tested closely match provisions adopted under TANF. In other continuing waiver experiments, mismatches make convincing inferences about TANF effects more difficult. Although the best of the waiver experiments will provide useful information on TANF effects, the findings will be limited to a handful of states and to comparisons between TANF and AFDC. In any event, a need will remain for more information on TANF in all its various manifestations throughout the country.

At the point of enactment, both the nature and the consequences of the changes enacted under PRWORA were largely unknown (and remain so). Perhaps the changes would lead to improvements, but perhaps not; the changes might very well produce some unanticipated effects. New research efforts were needed to find out what shape welfare reform would take in each of the states, the effects of the new welfare systems on public and private agencies and, most of all, the effects on poor households and the children within them.[4] The four research projects reviewed in this paper were designed to provide the needed empirical answers. For reasons discussed in some detail in the next section, none of the projects were easy to design, and none will be easy to carry out and analyze. Furthermore, any research conducted on the scale necessary to be relevant to all states or even a few key states will be expensive and take years to complete.

For all the research projects, the evaluation of PRWORA is primarily the evaluation of the additional changes embodied in that legislation and in the welfare plans designed by the states. It is not the evaluation of welfare reform represented by all the changes that began taking place before 1996. The decline in welfare rolls beginning in 1994 illustrates that the pre-PRWORA welfare changes may have had effects that precede those of PRWORA. In addition,

---

[3]This provision was modified in 1998 to allow most naturalized citizens to qualify for the Food Stamp Program.

[4]Existing ongoing research efforts were judged to be deficient for these purposes. The Current Population Survey (CPS) collects no information on child well-being and was not designed to provide sufficiently large samples for more than a handful of states. Both the National Health Interview Survey and the Survey of Income and Program Participation have inadequate sample sizes, especially for individual states.

the research projects are concerned mainly with the TANF provisions of the PRWORA legislation. Accordingly, the studies reviewed here are mainly concerned with TANF changes imposed on top of an already changed welfare system.

Of the four major research projects that this paper reviews, three were designed specifically to study the effects of PRWORA/TANF—we will call them “prospective” studies. The fourth project consists of extensions to five waiver experiments that were underway before PRWORA and are being modified to be relevant to TANF—we will treat them as a group, calling them “ongoing” studies.

The four major prospective research efforts currently underway[5] are designed on scales large enough to promise to generate findings relevant to significantly important places or states:

- **Survey of Program Dynamics**. The Bureau of the Census is extending the Survey of Income and Program Participation (SIPP) to create the Survey of Program Dynamics (SPD), a longitudinal survey consisting of repeated interviews of a national sample of households from several years before PRWORA to several years after enactment. SPD oversamples low-income households.

- **Assessing the New Federalism.** The Urban Institute’s extensive study, Assessing the New Federalism, is examining the devolution shifting many welfare programs from the federal to the state level. Administrative studies will describe the changes in health and welfare systems in each state and take an intensive look at thirteen states both pre- and

[5]Several additional efforts not covered in this paper are underway:

- The Rockefeller Institute (SUNY-Albany) has started the states Capacity Study project, a study of the administration of the new welfare programs in several states. Because this effort is not directly concerned with measuring the effects of such changes on poor families, it is not reviewed in this paper.

- A group of prominent social scientists (Ronald Angel, Linda Burton, P. Lindsay Chase-Lansdale, Andrew Cherlin, Robert Moffit, and William J. Wilson) have begun a longitudinal study of poor families based on samples taken within Boston, Chicago, and San Antonio, supplemented by ethnographic studies in poor neighborhoods within each of those cities and child development studies of welfare children. This collaborative effort was funded in 1998, and fieldwork has been underway since early 1999. In addition, an inventory and description of ten major surveys relevant to family well-being post-PRWORA can be found in Pat Dean Brick, “A Descriptive Review of Ten Major Surveys,” draft memorandum, Urban Institute, Washington, D.C., 1999.

post-PRWORA.[6] Much of the Urban Institute project will center on a pre-PRWORA survey and several post-PRWORA national household sample surveys—the National Survey of America's Families (NSAF)—with subsamples large enough to support precise estimates for each state in each of the thirteen states in which intensive administrative studies are being undertaken. NSAF also will oversample households at 200 percent or less of the poverty level, supporting detailed findings for low-income families. Additional surveys are planned for 2001 and, tentatively, for 2003.

- **Project on Devolution and Urban Change**. The Manpower Demonstration Research Corporation (MDRC) is undertaking the Project on Devolution and Urban Change, which consists of intensive studies of poor households in four major cities: Cleveland, Miami, Philadelphia, and Los Angeles. Using administrative data, the program participation and earnings of cohorts of poor households entering the Food Stamp or AFDC/TANF programs at given points in time will be tracked before and after TANF goes into effect. In addition, the study will include surveys of a sample of AFDC/TANF households, intensive studies of the TANF programs in the four cities, and ethnographic studies within selected poor neighborhoods. The first survey was completed in 1999.

- **Child Impact Waiver Experiments.** The U.S. Department of Health and Human Services (HHS) is augmenting several ongoing waiver experiments, adding measures of effects on children.[7] The five studies, known as the Child Impact Waiver experiments, compare randomly selected control-group families who have been receiving welfare benefits under the old AFDC rules with randomly selected experimental-group families who received benefits under the waivers. Because these extensions are among the best of the waiver experiments and will provide information on effects on children, they are considered a fourth major research project for the purposes of this paper. Most important, the waiver provisions being studied resemble quite closely the TANF provisions enacted by the states in question. The five states and the contractors involved are Florida (MDRC), Minnesota (MDRC), Connecticut (MDRC), Iowa (Mathematica Policy Research, Inc.), and Indiana (Abt Associates).

---

[6]Although NSAF interviewing took place in 1997, the survey collected information about 1996, covering the period before most PRWORA provisions went into effect.

[7]In addition to the five ongoing waiver experiments discussed in this paper, HHS is funding the continuation of experiments in Vermont, Texas, and Arizona. All eight waiver experiments were chosen for continuation because waiver provisions closely resembled their states' subsequent TANF plans. The five ongoing waiver experiments chosen to be discussed in this paper were selected because they each added child impact measures, bringing them more into line with the prospective studies.

The central purpose of this paper is to assess the prospects of the four research projects for providing credible findings on how the TANF programs will affect low-income families. Because none of the prospective research projects have gone beyond the first steps in data collection, the paper focuses on the ability of their research designs to support credible findings concerning the state welfare programs and their effects. For research programs that are already engaged in data collection, the paper examines how well the designs have been implemented. Of course, none of the evaluations have proceeded far enough to produce any findings relevant to TANF, and the research designs of the prospective studies are not yet in their final forms.

The four research projects each have several objectives, some of which are unique to each study. As far as evaluating welfare reform is concerned, there are three central objectives:

- **Describing how public welfare changed under PRWORA.** The MDRC, the Urban Institute, and the waiver experiment studies share the objective of describing how public welfare changed in each of the sites studied, whereas the Census Bureau's SPD does not.

- **Describing how low-income families fared before and after welfare reform**. In this area, the studies will give close attention to levels of employment, earnings, TANF recipiency, family functioning, and child well-being. All four projects share this objective, although MDRC primarily will use program administrative data supplemented by two cross-sectional household surveys, SPD will be a panel study, and NSAF will rely on three (possibly four) cross-sectional household surveys. The Child Impact Waiver experiments deal primarily with families who have been on welfare and not the entire population of low-income families.

- **Estimating the net effects of welfare reform.** Only the Child Impact and MDRC projects currently intend to estimate the net effects of TANF or other changes enacted by PRWORA. The Bureau of the Census is not planning to analyze the SPD data set in detail but will make it available for public use. The Urban Institute is not quite settled about analysis plans but will make its data set available for public use. Because the three latter data sets will be made available for public use, all likely will be used by a variety of analysts to estimate net effects.

Findings bearing on all three topics undoubtedly will be met with great interest on the part of everyone interested in welfare reform. The greatest concern is likely to be with estimates of net impact. Surely changes will occur in how clients fare under the new welfare system, but the central question is whether those changes can be credibly attributed to the changed welfare system or to other forces or events.

In addition, the first and second descriptive tasks, although demanding great care and skill, can be accomplished in fairly straightforward ways. System changes can be discerned in documentary materials and supplemented by interviews with key welfare officials—and, perhaps, with rank-and-file welfare workers. The methodology lying behind sample surveys is well known; although household surveys are not easy to conduct, given sufficient resources, excellent surveys can be conducted. In contrast, estimating the net effects of a full-coverage social program is much more difficult, for reasons discussed in the next section.

Why Estimating the Net Effects of PRWORA Will Be Difficult

This section describes the difficulties that will beset attempts to estimate the effects of PRWORA. The principals of the four prospective projects are very much aware of the problems and discuss them at length in project documents. Because the problems facing the ongoing waiver experiments differ from those facing the prospective studies, they are discussed in a separate section.

Urban Change, Survey of Program Dynamics, and Assessing the New Federalism. The net effects of a program are estimated by measuring relevant outcomes among program participants and comparing those outcomes with what would have been measured had the subjects not experienced the program. In a literal sense, such comparisons are impossible. The best approximation is achieved by conducting randomized experiments, in which subjects are randomly assigned to different groups that either experience the program or serve as controls from which the program is withheld. The conditions to which the control groups are subjected are known as *counterfactuals.* Because the assignment to experimental and control groups is random, the two groups differ only by chance. Hence, measures of outcome that differ between the two groups to an extent greater than what can be expected from chance can be reasonably attributed to the effects of the program. Program net effects are always estimated comparatively. The comparison is between the program being evaluated and the "program" represented by the conditions prevailing in the control group.

In principle, randomized experiments can be used to estimate the effects of PRWORA or TANF, although it would be extremely difficult to carry them out for a variety of reasons. First, it is not clear what counterfactual conditions should apply to control groups. Full coverage (or saturated) programs, such as PRWORA or TANF, present special difficulties because such programs cover all persons or households who are eligible. A control group cannot be formed without subjecting households assigned to control conditions to some policy-relevant alternatives.

The quintessential evaluation question is whether a given program is effective compared with an alternative program. In the usual randomized experiment, outcomes in the experimental group who have experienced a program are compared with those in a control group who have experienced some other treatment, usually the status quo. For example, in the welfare-to-work experiments conducted under AFDC waivers, the experimental groups each experienced a new program designed to facilitate the movement of welfare clients into paid employment, whereas the control groups experienced the conventional state welfare program, usually AFDC. The policy issues were whether and how the changes embodied in the waiver provisions differed from AFDC in their effectiveness.

It is not clear how the control conditions should be defined in a randomized experiment involving TANF:[8] Should the control condition be a pre-1996 AFDC program? If so, then the experimental findings compare TANF with a condition that most observers thought was ineffective and to which few would want to return. A study showing that compared with AFDC, TANF did not produce a significantly greater movement into employment would be unlikely to lead to a clamor for a return to AFDC. Rather, it would lead to proposals for modifying TANF to strengthen its effects on employment, for which the research findings would provide little guidance. Strong arguments for using AFDC as the counterfactual control-group condition also can be made, however. Critics of TANF have claimed that poor families will become worse off, suggesting that AFDC is an appropriate comparison group for TANF studies.

The issue of how control conditions should be defined need not be settled entirely one way or the other. Comparisons with AFDC are needed to settle the issue of whether TANF has deleterious effects on the poor. We also need to test how TANF can be improved. The latter question suggests that a randomized experiment strategy would be to test the effectiveness of alternatives to TANF, defining the control condition as the TANF status quo.[9] For example, some consensus may exist among policymakers that time limits of some kind are going to be a feature of any alternative to TANF, but it may not be clear whether longer or shorter time limits will be more effective. A sensible approach might be to define the experiment as testing the relative effects of longer and shorter time limits than those currently specified. Note that in this approach, the experimental intervention is variation in the size of time limits, and the control condition is

---

[8]Robinson Hollister, “What Works, for Whom? Learning from the Experience of Welfare Reform,” *Poverty Research News* (Summer 1997): 3–6.

[9]A demonstration quasi-experiment called Jobs-Plus is an example of such research. See James A. Riccio, *A Research Framework for Evaluating Jobs-Plus, a Saturation and Place-Based Employment Initiative for Public Housing Residents* (New York: Manpower Demonstration Research Corporation, 1998). Jobs-Plus, conducted by MDRC and financed by the U.S. Department of Housing and Urban Development and the Rockefeller Foundation, is based in eight public housing sites throughout the country and is testing the ability of a combination of community organization, job training and job search activities, and enhanced incentives for working to move welfare families living in public housing into employment. The housing projects for the demonstration were selected randomly from among a set of three roughly equivalent projects in each of eight cities, with two projects designated as controls. The main variations on TANF consist of community organization efforts and enhanced incentives. Note that Jobs-Plus is primarily a variation on TANF for public housing communities and is not intended to be broadly applicable to TANF. Of course, if Jobs-Plus proves effective, that finding might suggest that community strategies should be considered in TANF.

the TANF status quo. The research question then becomes, "Can we improve on TANF?" No experiments using this strategy currently are underway or planned.[10]

In all three of the prospective research projects this paper reviews, the main contrasts that will be available will be between how households were faring before 1997 under AFDC and their condition after living under TANF. Of course, the studies also can be used as cross-sections. For example, NSAF data on the thirteen states can be used to estimate the effects of state TANF plan differences on the well-being of poor households. Or, children in TANF families can be compared with children in comparable non-TANF families. Such analyses can hardly ever be credible and are certainly far inferior to even before-and-after analyses. Accordingly, they are not discussed in detail in this paper. All can be analyzed with nonexperimental research designs, using as nonrandomized controls or as comparisons the experiences of poor households pre-TANF (essentially AFDC in 1996 or earlier).

These before-and-after research designs are perhaps the best designs possible under the conditions existing when the research programs were formulated—the projects had to get rapidly underway before PRWORA was enacted but when it was clear that some sort of welfare reform was inevitable. Consequently, the design options were limited. Nevertheless, the findings resulting from such designs are widely viewed by evaluators as subject to many credibility problems. The most serious problems are twofold.

First, the client population may change as a consequence of TANF. Indeed, the welfare rolls began to decline in 1994, several years before the enactment of PRWORA, hinting that the composition of welfare clientele under TANF might be different in important respects compared with AFDC clientele. At least some of the decline in caseloads appears to be because some families who might be eligible for benefits have decided not to apply for them. Such "entry effects" cannot easily be studied or even detected in the MDRC and Urban Institute projects, especially the conditions of those families who were "deterred" from applying for welfare.

Second, before-and-after designs are not able to distinguish convincingly the effects of interventions from other changes occurring at the same time. For example, the studies may detect a significant increase in employment of single mothers after TANF, but that increase may be a result of increased employment opportunities resulting from a tighter labor market. Changes in other programs, such the Earned Income Tax Credit, or changes in the federal minimum wage also may affect employment.

[10]It is my understanding that the Administration for Children and Families of HHS has provided grants to thirteen states to plan, design, and evaluate variations in TANF provisions for improving job retention and advancement for TANF recipients.

Some of the difficulties in estimating TANF effects lie in the nature of the reforms themselves. Within quite broad limits, each state can develop its own version of welfare reform. For example, some states have quite generous income disregards, whereas others are less generous; some have quite short time limits on eligibility, but other states permit longer periods of eligibility. Some have imposed family caps, which involve not increasing income maintenance payments when children are born after initial eligibility. And so on.[11]

In effect, TANF cannot be evaluated as a national program; only state TANF programs can be evaluated. In addition, one can expect within-state variation in implementation, especially in such states as California or New York, where welfare is administered by local political jurisdictions (for example, counties), with some local discretion allowed in design. In most other states, TANF plans permit considerable discretion to "street-level bureaucrats" in implementation, potentially leading to within-state heterogeneity.

Within each state, TANF is designed as a package of interventions, typically consisting of several provisions, some of which can be expected to have quite opposite effects. For example, generous earned income disregards may make employment quite attractive, leading to higher household incomes for some households, but sanctions imposed for noncompliance with work provisions may serve to reduce income for other households. Accordingly, if a research project judges a state's program to be effective, it will be quite impossible to attribute those effects to any specific provision in the "bundle" that constitutes that state's version of TANF because only the bundle can be evaluated.[12] Furthermore, the effects of some provisions may be counterbalanced by the opposite effects of other provisions to the extent that the bundle may appear to have had no effects. The consequence is that it will be difficult, if not impossible, to

---

[11]The diversity that TANF state plans have developed are described in detail in an Urban Institute report summarizing state plans as of October 1977. See L. Jerome Gallagher, Megan Gallagher, Kevin Perese, Susan Schreiber, and Keith Watson, "One Year after Federal Welfare Reforms: A Description of State Temporary Assistance for Needy Families (TANF) Decisions as of October 1997," Occasional Paper No. 6, Urban Institute, Washington, D.C., 1998.

[12]In principle, it is possible to study the separate effects of provisions making up a bundle. Factorial experiments using several experimental groups, each consisting of unique combinations of programs, would provide estimates of the relative effectiveness of the separate programs and combinations of programs that make up a bundle. Factorial experiments have been used in the two major income maintenance studies as well as in the housing allowance experiment. For the income maintenance studies, see Peter H. Rossi and Katherine Lyall, *Reforming Public Welfare* (New York: Russell Sage, 1976). For the housing allowance experiment, see Philip K. Robins, Robert G. Spiegelman, Samuel Weiner, and Joseph G. Bell, editors, *A Guaranteed Annual Income: Evidence From a Social Experiment* (New York: Academic Press, 1980); and Joseph Friedman and Daniel H. Weinberg, "Housing Consumption in an Experimental Housing Allowance Program: Issues of Self-Selection and Housing Requirements," unpublished paper, 1980. One of the waiver experiments, described below, had a factorial design.

discern which elements of each state's TANF bundle are contributing (and how much) to success or failure.

Not only the state bundles but also the names given them vary. Although "Temporary Assistance for Needy Families" is the title given in the federal legislation to the successor to AFDC, some of the states have given different titles to their versions. This diversity produces difficulties for national surveys. It will be hard to write questions about TANF program participation in national surveys so that they are understood properly by respondents.[13]

TANF will take some years to settle into a relatively steady state. HHS published final regulations in 1999; accordingly, each state's plans did not begin to take final form until that year.[14] Implementation failures and modifications undoubtedly will occur in some states' TANF programs. For example, several states have experienced difficulties in training front line personnel, who under AFDC primarily determined welfare eligibility, to take on the task of facilitating client movement to employment.[15] As a result, the versions of TANF experienced in the first few years will change after a few years of trial and error. All three research projects are designed to measure what happens in the first few years of PRWORA and may therefore be measuring early versions of those programs rather than what will appear after the programs have settled down. Of course, determining how long such a period of adjustment will take is problematic.

It is also possible that some or all of the state PRWORA plans may never reach a steady state but will change continually over time. In addition, Congress may alter provisions in PRWORA, forcing changes on the states. Although evaluation of a frequently changing program can be undertaken, interpreting findings will be quite difficult.

Even if TANF functions as its authors expected, its fate depends heavily on forces outside the influence of the program. In particular, moving TANF clients into jobs cannot succeed unless jobs are available in the economy for which clients can be considered and unless the child care industry provides enough slots to meet the needs for child care. To be sure, the states might provide public-sector positions if the labor market does not offer private-sector employment, but arranging for and financing public-sector jobs will be difficult and expensive. Much more problematic is whether the child care industry can provide slots for the increased demand at

---

[13]This difficulty may be one of the main reasons that the CPS's estimates of TANF participation increasingly has departed in recent years from counts obtained from administrative data.

[14]Nevertheless, many states wrote their versions without benefit of the federal regulations and had to change them to conform to the final federal regulations.

[15]General Accounting Office, *Welfare Reform: States Are Restructuring Programs to Reduce Welfare Dependency*, Report No. HEHS-98-109, 1998.

prices that are affordable under state subsidies as well as meet quality standards. Unfortunately, none of the research projects plan to collect data directly relevant to those issues. For example, the sample surveys will collect data on respondents' child care arrangements, but no studies of child care markets are planned.[16]

Finally, issues of what should be regarded as criteria of success and when measurement should be undertaken remain. TANF has multiple goals as well as the potential to produce unanticipated and unwanted effects. The key outcome of greatest interest concerns the extent to which welfare clients will move off the rolls, obtain employment, and thereby be better off. Because clients are expected to be better off when employed, researchers are interested in the earnings of those who move into employment. Because some analysts have predicted that children will be worse off under PRWORA, they also have considerable interest in measures of child well-being. Accordingly, it is appropriate to ask whether each of the projected studies has planned to collect data relevant to the important outcomes that PRWORA is intended to affect.

The outcome measures used are restricted to those that can be obtained by the data collection modes used. For example, the quality of day care used by the households studied cannot be measured directly, only through the reports of household respondents.[17] Nor is it possible to measure directly other important outcomes, such as the quality of parent–child relationships or the cognitive abilities of children.

Some effects can be expected to appear relatively quickly, but others can be expected to occur after several years or longer. If TANF increases employment and earnings, such effects should appear within the first two years or so, whereas effects on the developmental trajectories of children may take a much longer period of time to detect. Steadiness of employment also requires observations over longer periods of time, whereas determining whether clients have found employment might require only a single interview. No one can expect that initial earnings of new entrants in the labor force will be high: It will take time to observe what shape earnings will take over time.

Ongoing waiver experiments. Relating the five ongoing waiver experiments to TANF presents somewhat different problems. Because each is a randomized experiment, the studies avoid all the difficulties stemming from nonrandomized designs, as described above. The waiver studies can produce unbiased estimates of the net effects of the waiver provisions involved in each case compared with the prewaiver AFDC programs in place in that state. The families in the experimental and control groups experience the same historical events and trends, so results will be independent of, say, trends in the local economy and labor market.

---

[16]HHS is funding research to be conducted by Abt Associates on low-income child care markets; the planned survey may provide the needed information.

[17]Little empirically based knowledge exists on how to measure the quality of day care.

Using the waiver experiment findings to judge the effects of TANF, however, is problematic in several ways. First, although the waiver provisions being tested are fairly close to the TANF programs enacted in the five states involved, some differences exist, especially in time limits. Accordingly, ascertaining the implications of waiver experiment findings for TANF must be judgment calls, advanced through argumentation and clearly subject to dispute. (See the appendix for comparisons between TANF state plans and experimental treatments.)

Second, the waiver experiments are plagued by some of the same problems affecting the prospective studies. In particular, four of the plans are bundles of provisions that cannot be separately evaluated, the exception being the Minnesota experiment.[18] The experiments also are studies of welfare reform in its earliest years.

Third, the subjects enrolled in the waiver experiments will become increasingly unrepresentative of all welfare families over time. For example, the members of the households all will be five years older at the end of a five-year-long experiment—and likely to be older than current welfare clients. Any changes in the characteristics of welfare clients enrolled after PRWORA will not be detected. Accordingly, entry effects cannot be studied.

Fourth, maintaining the integrity of the waiver experiments will be difficult. In particular, it is essential that the control-group families be subject to AFDC rules throughout the time periods of the experiments. It also is essential that they know that they are under such rules. PRWORA and TANF have received so much general publicity in the mass media that control-group families may not know that time limits, family caps, and other TANF rules do not apply to them. In addition, control-group families are typically a small proportion of the total welfare caseload, and welfare personnel mistakenly may apply the wrong rules to such families. That this is a very real danger can be seen in the failure to maintain the fidelity of treatments for both the control group and the experimental group in the New Jersey waiver experiment evaluating that state's Family Development Program.[19] A central feature of the New Jersey program was a family cap, under which family benefits were not increased when children conceived after enrollment were born to mothers enrolled in the plan. Despite the fact that the control group was not under the family cap rule, it was found that the benefits of more than a score of control-group families had not been increased after the birth of children. In addition, a survey of control-group families found that few knew that the family cap rules did not apply to them. Given these

---

[18]In Minnesota, the waiver provisions concerning mandatory services and the financial incentive are tested separately and in combination, making it possible to estimate the effects of each provision separately and in combination.

[19]Michael J. Camasso, Carol Harvey, Radha Jagannathan, and Mark Killingworth, *A Final Report on the Impact of New Jersey's Family Development Program: Experimental–Control Group Analysis* (Brunswick, N.J.: Rutgers University, 1998).

findings, it was clear that the experimental conditions had been violated to the extent that the resulting data can hardly be regarded as valid.[20]

The research organizations running the Child Impact Waiver experiments are fully aware of the need to be vigilant in maintaining the integrity of the experimental and control groups. Some have placed safeguards in the management information systems software, which block inappropriate actions with regard to the participating families. All are determined to provide reminders to participating families about the relevant rules to which they are subject. Perhaps these measures will succeed in maintaining fidelity in treatments—but perhaps not.

Issues of Research Design

Clearly, identifying research design problems is easier than devising effective countermeasures. Each research project has been designed to address the difficulties each expects to encounter. Those efforts and their likely successes or failures are described in some detail in the next sections.

Urban Institute: Assessing the New Federalism. Of the three welfare reform projects, the Urban Institute's Assessing the New Federalism is the most extensive.[21] About $66 million in grants from various private foundations constitute the initial funding. Assessing the New Federalism consists of three major research activities:[22]

- **State database.** The Urban Institute has constructed and made available on its web site a large database containing hundreds of data items relevant to welfare reform for each of the fifty states and the District of Columbia. The data items range widely, from state demographic characteristics and vital statistics to AFDC and General Assistance caseload data. Many variables are disaggregated and tabulated separately for important population subgroups. Some of the series enable tracking changes over time, although usually they

---

[20]The Family Development Program also was assessed using administrative data on all welfare clients before and after the onset of the program. See Michael J. Camasso, Carol Harvey, Radha Jagannathan, and Mark Killingworth, *A Final Report on the Impact of New Jersey's Family Development Program: Results from a Pre-Post Analysis of AFDC Case Heads from 1990–1996* (New Brunswick, N.J.: Rutgers University, 1998).

[21] This section is based on *Methodology Reports* 1–8 (Urban Institute 1999) and on interviews and correspondence with Fritz Scheuren, Anna Kondratas and Genevieve Kenney. In addition, I attended, upon invitation, a February 1998 meeting of the project advisory group, at which the progress of the project was discussed.

[22]A fourth major activity consists of dissemination of findings to the national and state policy communities.

do not cover more than the previous ten years. The Urban Institute plans to add to the database as new measures become available. The state database is designed to serve as an information source to state and local policy analysts and welfare agencies as well as social scientists.

- **State administrative case studies.** Intensive case studies have been undertaken of the welfare and health systems of thirteen states (Alabama, California, Colorado, Florida, Massachusetts, Michigan, Minnesota, Mississippi, New Jersey, New York, Texas, Washington, and Wisconsin), which contain more than half of the U.S. population. The case studies will describe each state's welfare system before the full implementation of PRWORA and again in 1998 or 1999, after PRWORA changes were made. The case studies are based on state documents and legislation as well as interviews with key state administrative personnel. Reports describing each state's pre-PRWORA welfare and health systems on the basis of research conducted in 1996 and 1997 are being released as they are completed. The post-PRWORA case studies have been started and will serve as the bases for additional reports.[23]

- **The National Survey of America's Families**. NSAF is planned as a set of national surveys of American households. The first, NSAF-I, was conducted in 1997 after the enactment of PRWORA, but before full implementation. Many of the questions in the interview asked about conditions and events experienced in 1996, the year before the enactment of PRWORA. It was followed in 1999 by a second survey, NSAF-II. The two NSAF surveys will constitute a before-and-after research design enabling comparisons between household measures from two cross-sectional surveys. NSAF-III is currently being planned for 2001, and NSAF-IV is tentatively planned for 2003. NSAF-III and NSAF-IV (if undertaken) will provide data on conditions experienced as PRWORA and TANF mature.

NSAF-I and NSAF-II are complex household surveys. Although their findings can be projected to the entire United States, they also provide large enough samples within each of the thirteen states chosen for the administrative case studies to make possible findings concerning each of those states. (More details on the surveys are given below.)

The three subprojects are intended to provide a comprehensive description of the changes in health and welfare policies resulting from devolution. The resulting data sets will support analytic studies of such important issues as whether the states have engaged in a "race to the bottom" by lowering benefits to the poor. Both the state database and the state case studies will provide valuable descriptions of the fifty states and the state administrative changes

---

[23] A compilation of TANF provisions for all fifty states as of October 1997 has been issued. See Gallagher *et al.*

accompanying PRWORA. Although data on the central evaluation questions concerning the impact of PRWORA on low-income households can come from the two NSAF surveys, the administrative data sets will play important roles in explaining interstate differences in the NSAF findings. This review will concentrate primarily on describing and assessing NSAF.

NSAF-I and NSAF-II are quite similar in design. Each consists of two complementary surveys: (1) a sample of telephone-owning households reached by random digit dialing and (2) an area probability sample of nontelephone-owning ("nontelephone") households reached by interviewers screening dwelling units for such households in a sample of tracts shown in the 1990 Census to have high proportions of nontelephone households.[24] Both samples include only households with at least one person under age sixty-five. Considered together, the two samples are designed to constitute a probability sample of all age-eligible households in the United States.[25] The Urban Institute contracted with Westat, a well-known survey research firm, to design the samples, conduct the interviews, and prepare the resulting data sets for analysis. All told, about 50,000 respondents within about 45,000 households were to be in the sample.

Within each of the main samples, the following categories of households of special interest were oversampled:

- Households within each of the thirteen states designated for intensive administrative data collection (about 3,000 households in each state).

- Households with incomes 200 percent or less of the poverty line.

- Households with children age eighteen or under.

Within households, respondents were selected according to the following rules:

- For households with children age eighteen or under, the adult who was the primary caretaker of a randomly selected child under age six was chosen; in households with children between ages six and eighteen, the primary caretaker of a randomly selected child was chosen as respondent. When households contained children in both age ranges, separate adult caretakers were chosen when that was appropriate. The main intent of this

---

[24]To make the interview situation as comparable as possible in the two samples, NASF-I interviewers carried with them a cellular phone. When an eligible respondent gave permission to be interviewed, the interviewer used the phone to contact Westat's telephone interviewing center, which then conducted the interview over the phone. Unfortunately, the sampling strategy for the nontelephone households turned out to be inefficient: Tracts that had high proportions of such households in 1990 no longer had that characteristic when screened in 1997.

[25]People living in institutions, in shelters, on the streets, or on military bases were excluded.

rule was to obtain respondents who knew the selected children well enough to be able to answer questions about them in a knowledgeable way.

- In households containing no children age eighteen or younger, an adult was randomly selected as respondent.

- In households with children and adults who were not parents of the children present, one of the nonparents was selected as a respondent.

- In households with large numbers of adults, more than one adult was selected.

In NSAF-I, the interviewing proceeded as follows: A short screening interview identified the eligibility of the household in terms of composition and income and identified which members would be subjects for a longer interview. All the longer interviews contained sections on employment, income, program participation, and health insurance and obtained information on the health status of the respondent as well as other adults in the household. Those who were selected as primary caretakers were asked additional questions about the selected children.

Interviewing for NSAF-I began in early 1997 and was completed in November 1997. The interviewing period took longer than initially planned because extra effort (and funds) had to be expended in order to achieve a high response rate. The final weighted response rates[26] were computed as shown in table 1.

In table 1, the weighted response rates are shown separately for households with children and all other households. Rates are shown for the screening interview, for the extended interview conditional on completing the screening interview, and for the combined screening and extended interview. Because critical data are obtained only when screening interviews are followed by completed extended interviews, the substantively important response rates are 64.5 percent for households with selected children and 61.7 percent for households without selected children. Those response rates mean that slightly more than one-third of sampled households with children

[26]Because the survey is complex, with different elements of the population being chosen with various probabilities, the response rate was weighted to bring each segment sampled into line with those probabilities and with the sizes of the segments involved. (Sampling probabilities are varied to produce desired sample sizes of critical categories. Hence they are part of the design, being determined by the researchers' interests in subgroup sizes.) For example, the nonresponses in a segment that was oversampled would count less in the final weighted response than a segment that was undersampled. The weighting objective is to arrive at a rate that reflects what the NASF-I fieldwork would have achieved had it been a simple random sample. The weighting scheme produces rates that are one to two percentage points higher than the response rates obtained from unweighted data. See Urban Institute, *1997 NSAF Response Rates and Methods Evaluation,* NSAF Methodology Report No. 8 (Washington, D.C.: Urban Institute, 1999).

and almost two-fifths of sampled households without children were not included in the resulting data set.

Response rates varied by geographic locations of the respondents. For example, the highest response rate for households with selected children (74.5 percent) was obtained in Minnesota, whereas the lowest (55.5 percent) was obtained in New Jersey. A higher response rate (83.3 percent) was obtained in the area sample of nontelephone households, and a lower rate (64.6 percent) for households contacted in the telephone sample.[27]

Table 1. Weighted Response Rates for NSAF-I

Types of Respondents	Percentage
Response rates for households with children	
Completed screening interviews	77.8
Completed extended child interview after screening	84.1
Completed both screening and extended interview	65.4
Response rates for households without children	
Completed screening interviews	76.6
Completed extended adult interview after screening	79.9
Completed both screening and extended interview	61.7

Note: Data combine telephone and area samples.
Source: Brick and Flores-Cervantes (Washington, D.C.: Urban Institute, 1998).

Although the response rates of NSAF-I are not high, they are typical of well-run national telephone surveys.[28] For example, a 1995 national telephone survey of adults ages twenty-five to seventy-five, which screened households for eligibility and followed up with a mailed questionnaire, achieved an overall response rate of 61 percent; that is, 61 percent of eligible respondents completed both the screening telephone interview and returned a long mailed questionnaire.[29] In an article summarizing response rates achieved in twenty-nine large national surveys conducted by nongovernment survey organizations in the 1990s, Massey *et al.* found that

[27]Response rates for telephone surveys are usually lower than for face-to-face surveys.

[28]Better response rates are achieved by sample surveys run by the Bureau of the Census, presumably because respondents believe that responding to government surveys is more important.

[29]Statistics are from the Survey of Midlife in the United States (MIDUS), conducted for the MacArthur Foundation Research Network on Successful Midlife. See Orville G. Brim and David L. Featherman, “Surveying Midlife Development in the United States,” unpublished paper, Life Trends, Vero Beach, Fla., 1997. It should be noted that the MIDUS survey did not oversample poor households, nor did it include adults under age twenty-five.

the average response rate was 62 percent, with about half of the response rates ranging between 60 and 70 percent.[30]

Much better response rates, however, are achieved in face-to-face interviewing of low-socioeconomic status (SES) populations. For example, significantly higher response rates have been achieved in face-to-face surveys of low-income populations by MDRC survey contractors: In eleven surveys of welfare participants, eight had response rates of 80 percent or greater, and the lowest response rate was 70 percent.[31] As will be shown below, MDRC achieved an overall response rate of 79 percent in the first round of interviewing in the Urban Change project. The nontelephone portion of NSAF-I also had notably higher response rates: NSAF-I response rates for completed interviews with adult-only households were 80.5 percent for nontelephone households, compared with 61.0 percent for comparable households that had telephones.

Response rates matter in any survey. For surveys whose subject matter is closely related to the probability of responding, however, response rates can be critical: “Average” may not be good enough. For almost all surveys, response rates are lower for minorities, the poor, the poorly educated, and young adults, all of whom are population elements of special interest in NSAF. It is almost a certainty that as a group, nonresponding households in NSAF have those characteristics to a greater extent than those who responded.

The Urban Institute funded Robert Groves and the Survey Research Center at the University of Michigan to interview nonresponding households with a shortened version of the NSAF interview.[32] Although that study had a low response rate,[33] the resulting data provide some critical information on the characteristics of the nonresponders. Apparently, compared with respondents, nonresponders to the screening interview tended to be of lower SES and nonrespondents to the extended interview tended to be of higher SES. Of course, these differences could be measured in the shortened version of the NSAF-I questionnaire, but important unmeasured differences may exist. This is of concern, because unmeasured differences

---

[30]James T. Massey, Dan O’Connor, and Karol Krotki, “Response Rates in Random Digit Dialing (RDD) Telephone Surveys,” *Proceedings of the Section on Survey Methods, (*Washington, D.C.: American Statistical Association, 1998).

[31]Memorandum from Greg Hoerz of MDRC dated April 23, 1998. The surveys were done under subcontract to several different survey organizations.

[32]Urban Institute, *Early Nonresponse Studies of the 1997 National Survey of America’s Families* (NSAF Methodology Report No. 7. Washington, D.C.: Urban Institute, 1999).

[33] The survey received a 48 percent response rate. However, recall that this is a survey of households who repeatedly were contacted in NSAF-I and either did not initially respond or refused, sometimes repeatedly.

might lead to quite different responses and hence bias. Groves' report[34] also analyzed differences among respondents who participated on the first contact compared with those who responded after repeated contacts, finding systematic differences according to how much effort it took to achieve cooperation. Clearly some biases in response exist. Although the biases are not large, they do indicate that NSAF-I underrepresents critical demographic groups. For some purposes, these differences are ignorable, but for other purposes, they may be critical. In any case, users of the NSAF data should be aware of the potentials for bias and qualify their findings accordingly.

The technical research undertaken by the Urban Institute on the NSAF-I response problem is commendable as highly responsible. The NSAF-I data set will be analyzed by many social scientists, who need to be aware of the biases introduced by nonresponse, and the Urban Institute has taken pains to describe those biases.

No widely accepted standards have been adopted for judging whether the response rate for a survey is acceptable. Clearly, the higher the response rate, the better, but whether a particular response rate is acceptable is a judgment call. Affecting such judgments are such factors as the policy importance of the survey, how the typical response-rate patterns are related to the central topics of the survey, and whether the survey mode and design adhere to standards concerning best survey practices. In the case of the NSAF, the response-rate issue is clearly important. The central survey topics are politically salient, and the response rates are lowest for population subgroups that are of central interest. Arguably the telephone survey mode used may have been less expensive, but face-to-face interviews typically have higher response rates. At best, these considerations make NSAF response rates worrisome; at worst, they support a judgment that the survey may not be valid.

Consistent with best practices, NSAF will attempt to compensate for potential nonresponse bias by devising weighting schemes for the data set[35] that reflect the probabilities of nonresponse within population subgroups.[36] The rationale underlying such weighting schemes is that people or households interviewed from among highly nonresponding groups can serve as proxies for those who did not respond from those groups. For a simplified example, if 25 percent of below-poverty-level inner-city households with children responded to the survey, the assumption is made that those who did not respond in that subgroup can be represented by those

---

[34]Urban Institute, *Early Nonresponse Studies*.

[35] Weights also will be constructed reflecting the different probabilities with which elements of the sample were selected. Those sampling weights are based on the sample design used. Westat also plans to construct "poststratification" weights, which are designed to bring the sample into line with known characteristics of the U.S. population as determined in the Census and updated by the CPS.

[36]Urban Institute, *1997 NSAF Snapshot Survey Weights,* NSAF Methodology Report No. 3 (Washington, D.C.: Urban Institute, 1999).

who did. Accordingly, if the average household income for responders in that population subgroup is $6,000, that average is also attributed to the nonrespondents. Hence, responding households are given a weight of four in any calculations that include that population subgroup.

How well such weighting schemes compensate for nonresponse biases depends on how good the assumptions concerning proxies are.[37] Those assumptions, in turn, are affected by how much is known about how nonresponse varies by respondent features that can be identified in the survey. In most surveys, those assumptions cannot be tested. NSAF-I is somewhat better off than the usual survey because of the information contained in Groves' research.[38] However, even if the patterns of nonresponse within subgroups are well understood, it can never be known how well the responding households used as proxies represent the nonresponding households. Indeed, about the best that can be said about weighting schemes designed to compensate for nonresponse is that they are usually better than having no scheme at all, the latter being a strategy that is based on the clearly false assumption that nonresponders collectively are identical to the overall average responder.

In addition, the Urban Institute staff have contrasted NSAF-I's weighted findings[39] with those of the Current Population Survey (CPS) on comparable data items.[40] Comparisons of family composition, work experience, earnings, income, and poverty status by key demographic characteristics show a great deal of agreement between NSAF-I and the CPS. Some minor disagreements were found: Compared with CPS data, NSAF-I had larger proportions reporting full-time, full-year work; weekly earnings in the NSAF-I were slightly higher for women and slightly lower for men; and poverty rates were higher in NSAF-I (14.8 percent, compared with 14.1 percent in the CPS). Given that the instruments used were not exactly comparable, the discrepancies found are certainly minor. Although the comparability found is heartening, it is not to be regarded as definitive. The CPS is arguably the best comparison standard available for some outcome measures, but it also has defects. It has consistently underestimated welfare and Food Stamp Program participation levels when compared with administrative data, and the gap

---

[37] The weighting scheme used in NSAF-I was not assessed for this paper.

[38] Urban Institute, *1997 NSAF Benchmarking Measures of Child and Family Well-Being,* NSAF Methodology Report No. 6 (Washington, D.C.: Urban Institute, 1999).

[39] Poststratification weights were applied to NSAF-I using home ownership and educational attainment variables in the CPS. Sensitivity analysis indicated that the poststratification weights are a minor factor in the CPS's comparisons on other demographic and socioeconomic variables.

[40] Urban Institute, *Survey Methods and Data Reliability,* NSAF Methodology Report No. 1 (Washington, D.C.: Urban Institute, 1999).

has been growing in recent years.[41] For example, in 1997 the CPS estimates of welfare participation were only two-thirds of what was reported in administrative records. Comparisons of Survey of Program Dynamics data for welfare recipients against data available in administrative benchmarks would be helpful in assessing the quality of the data. In addition, the comparisons are limited: Important response biases on respondent characteristics may not be captured by the surveys.

The Urban Institute staff has thoroughly investigated the psychometric properties of scales measuring adult and child functioning,[42] showing them to have generally adequate to good internal reliability. The scales also correlate in "expected ways" with family structure and socioeconomic characteristics. That is, adults and children who are poor and in female-headed households tended to be worse off compared with those having opposite characteristics. Because many of the instruments were adapted from existing scales, attempts were made to compare NSAF-I findings with other studies. Unfortunately, differences in the populations sampled and in the scales used made such comparisons difficult to evaluate.

Urban Institute and Westat staffs have done considerably more than the usual sample survey staffs in conducting technical research to best counteract and compensate for potential nonresponse bias in the first NSAF survey. Nevertheless, whether those efforts are sufficient is likely to be an open issue. Consequently, analysts using the data set should be aware that those compensatory efforts may not have been successful and interpret findings accordingly. Although some strong confidence may be given to NSAF-I sociodemographic findings, more caution should surround the use of measures of child well-being, for which tests of potential response bias may not exist. Accordingly, it is important that the documentation accompanying the NSAF-I data set distributed for public use contain detailed discussions of how the data are to be used responsibly. The number of cases receiving welfare is relatively small in each state, however, making it unlikely that the effects of welfare reform can be discerned for all but major differences.

Recall that the main purpose of the NSAF-I survey was to provide good descriptions of how the populations of the thirteen selected states and the nation were faring before PRWORA went into effect. As far as the nation as a whole is concerned, some of the information that can be obtained from NSAF-I is already available from other national surveys, such as the CPS and the SIPP, both conducted by the Bureau of the Census. The unique contributions of NSAF-I are its large samples for each of the thirteen selected states, the existence of administrative case studies

[41] Richard Bavier, "An Early Look at the Effects of Welfare Reform," draft memorandum, Office of Management and Budget, Washington, D.C., 1998.

[42] Urban Institute, *NSAF Benchmarking Measures*, 1999.

for each of those states, and its attention to important substantive topics, especially measures of the well-being of children, which are not collected in either SIPP[43] or CPS.

Considered by itself, the NSAF-I study can support a variety of interesting analyses. It can be expected that much attention will be given to interstate comparisons. To an extent limited by the fact that only thirteen states are being studied, NSAF-I can be the vehicle for analyses of interstate differences in pre-PRWORA welfare systems and the conditions of the low-income populations in those states that are related to those systemic differences. However, it should be recalled that in 1996, AFDC was already in transition in many states: The AFDC waivers obtained in the 1990s had changed AFDC in radical ways in states such as Wisconsin, Michigan, Massachusetts, and Florida. The measures that are unique to NSAF-I likely will support interesting analyses, many centering on the issue of how children living in poverty households differ from those living in more affluent households.

Of course, the NSAF-I data set is a cross-sectional survey. Causal inferences can be made, but they cannot be definitive or highly credible. To find that children in AFDC families in Mississippi are worse off than comparable children living in Minnesota can scarcely be attributed with great confidence to the less generous Mississippi welfare system because the two states differ in so many other ways.

Data collection for NSAF-II was completed in 1999. It was designed using the same sampling strategy and much the same interview schedule as the NSAF-I, with the addition of questions concerning experiences under PRWORA, use of the Earned Income Tax Credit, and changes to questions arising out of NSAF-I experiences. About two-thirds of the telephone numbers used in NSAF-I have been contacted again along with a new sample of telephone numbers. A similar strategy was followed for the nontelephone households. It is estimated that about two-thirds of the resulting sample was obtained from the NSAF-I telephone numbers and addresses. Although no special measures were taken to reach the same respondents as in NSAF-I, many of the respondents reached in recontacted households were the same people interviewed in NSAF-I. The NSAF-II sample design has some of the statistical advantages of a panel study, especially greater precision in detecting changes between NSAF-I and -II. Final NSAF-II response rates have not yet been calculated, although there are indications that they will be somewhat higher than those for NSAF-I.

Considered as a separate survey, NSAF-II certainly will be of interest in its own right. However, the major interest will be in the search for differences between the findings of NSAF-I and NSAF-II and relating the differences found to the workings of PRWORA. Despite the disclaimers from Urban Institute and Westat staffs that the two surveys together cannot be considered as properly supporting such causal inquiries, many analysts no doubt will try to draw

[43]Note, however, that the Survey of Program Dynamics will include questions on child well-being.

such inferences. In 1999, NSAF-I public use data sets began to be released, and NSAF-II public-use data sets began to be released in 2000.

As is the case with any public-use data set, no way exists to prevent any given use of the data sets, no matter how flawed the analytic approach may be. The main corrective to misuses of the NSAF data set doubtless will be the intensive scrutiny and criticism that will be given to all analyses by the community of policy analysts and evaluation experts.

The essential feature of a before-and-after design is that it relies on differences between measures taken before and after a program is put into place as the basis for inferences about the effects of the program in question. The major problem with that design is that the differences detected may be a result of the program as well as anything else that transpired in the time between the two sets of measures. For measures that may be affected by changes occurring independently of the program, it is not possible to sort out definitively the amount of change caused by the program from the change resulting from other causes. It is possible that the employment levels and earnings of low-income households may show an increase between NSAF-I and NSAF-II, but that increase may stem from, say, changes in the labor market. Even if employment levels and earnings decline between NSAF-I and -II, that may not mean that PRWORA was ineffective, because that program may have prevented a more precipitous decline. The obverse also may be true: The increase in earnings and employment might have been greater without PRWORA.

The limitations of NSAF's before-and-after design mean that plausible causal inference from the resulting data sets must draw heavily upon information gathered from other sources. For example, any inferences from NSAF about the effects of PRWORA on employment must take into consideration trends in employment generally in the thirteen states, especially for low-wage jobs. Inferences about the effects of time limits on family well-being need to be bolstered by information on how the time limits were administered in each of the states. (Of course, other factors that might affect those topics need to be considered as well, such as local labor market conditions.) Although all such analyses will be fragile, the best (and most credible) will be those that test and rule out alternative explanations for the outcomes under discussion. These considerations enhance the importance of the thirteen administrative case studies and the state database, two other potential sources of information about changes accompanying the implementation of PRWORA.

Bureau of the Census: Survey of Program Dynamics. For about fifteen years, the Bureau of the Census has been conducting a national longitudinal household survey known as the Survey of Income and Program Participation (SIPP).[44] The survey concentrates on issues

---

[44]This section is based on Daniel H. Weinberg, Vicki J. Huggins, Robert A. Kominski, and Charles T. Nelson, *A Survey of Program Dynamics for Evaluating Welfare Reform,* paper presented at the Statistics

involving household income, employment, and participation in a wide variety of welfare programs as well as associated topics, such as physical and mental health and household use of child care services. SIPP is based on an area probability sample, with oversampling of households living in low-income areas. About 37,000 households are interviewed in every four-month period. The sample consists of four equivalent panels, each of which is interviewed repeatedly over a period of four years. Interviews every four months query each household about employment, income, and program participation for each month of the preceding quarter and collect data on changes in household composition. Special supplemental interviews are administered from time to time; they cover topics such as child care arrangements or physical disabilities. SIPP is the major source of information about household patterns over time in the relationships among employment, poverty, and welfare receipt and the events that change them.

The PRWORA legislation contained a provision that directed and funded the Bureau of the Census to extend data collection from the SIPP households recruited in 1992 and 1993 through 2001, appropriating $70 million for that purpose.[45] This extension of SIPP, now called the Survey of Program Dynamics (SPD), will provide longitudinal data on about 19,000 of the approximately 30,000 panel households for four to five years before and for five years after the enactment of PRWORA.

The original plans called for SPD to start in 1996. Because PRWORA was not passed until August 1996, the survey could not be started until 1997, causing a significant gap in data collection for the 1992 SIPP panel, which had last been interviewed in January 1995. To fill the gap, a "Bridge Survey" was designed to gather data on income and program participation in the missing years. The Bridge Survey was essentially a modified version of the 1997 March supplement of the CPS and collected information for the missing period. As a result, data for 1995 and 1996 for the 1992 panel and for 1996 for the 1993 panel are not as detailed as in previous years for SPD households and are based on a longer period of recall. Specially prepared questionnaires also will be used with subgroups of particular interest, such as adolescents and married adults.

The Bridge Survey was followed in 1998 by interviews using questionnaires tailored for SPD purposes. SPD interviews will proceed on an annual basis until 2002. Each annual interview will cover the previous year (for example, information on 2000 will be collected in the 2001

Canada Symposium XIV, 1997, as well as questionnaires to be used in SPD furnished by the Bureau of the Census.

[45]The initiative for this provision stemmed from long-range planning activities in anticipation of welfare reform funded by HHS and the U.S. Department of Agriculture, which supported the Bureau of the Census in preparing plans for extending SIPP some years before the enactment of PRWORA.

interview). Note that SPD data for the post-PRWORA years will not be as detailed as for the pre-PRWORA years. New topics, however, especially measures of child well-being,[46] were added beginning with the 1999 SPD.

Unfortunately, Congress did not appropriate enough funds to interview all the households in the 1992 and 1993 panels. The Census Bureau thus reduced the 1998 SPD sample size to about 18,500, retaining all the households residing in low-income areas and sampling a smaller share of those living in more affluent areas.[47]

Despite the budget and data problems, SPD will provide data for a national panel study of households repeatedly interviewed over about a decade, the years more or less evenly divided into pre- and post-PRWORA years. The long pre-PRWORA period is intended to serve as a firm baseline measurement against which to contrast the changes that might be wrought by PRWORA. The long post-PRWORA interviewing period will track households through the early implementation stage of PRWORA and into the first few years of more settled forms of PRWORA.

The Bureau of the Census does not intend to undertake any sophisticated analyses of the effects of PRWORA. The SPD data set will be released for public use, and effectiveness analyses will be undertaken by social scientists in research universities and in the major research institutes.[48]

Maintaining the cooperation of households throughout an extended period of repeated interviewing is a difficult task. In both the 1992 and 1993 panels, about one-fourth of the households who were initial participants were no longer cooperating by the end of their initial four years of participation. An additional 18 percent attrition rate was experienced in the Bridge Surveys described above. Attrition rose again in the 1998 SPD survey, to 50 percent. The

---

[46]The measures closely resemble those being collected in the NSAF. Child Trends, Inc., is responsible for item development for both the Bureau of the Census and the Urban Institute. Because the items were added only beginning with the 1999 SPD wave, no pre- and post-PRWORA comparisons will be possible.

[47]The following subsamples will be included: (1) all households at 150 percent or less of the poverty line, (2) all households with children between 150 percent and 200 percent of the poverty line, (3) 90 percent of households with children at 200 percent of the poverty line or above, (4) 82 percent of childless households between 150 percent and 200 percent of the poverty line, and (5) 27 percent of childless households at 200 percent of the poverty line or above.

[48]Such efforts are already underway. Financed by grants from several private foundations, Thomas MaCurdy of Stanford University's Hoover Institute is planning interrelated analyses of SIPP, SPD, NSAF, and the Panel Study of Income Dynamics that will focus on the impact assessments of PRWORA and use sophisticated econometric modeling.

response rate did not go below 50 percent in 1999, likely because the Census Bureau made special efforts, including incentive payments, to bring some nonrespondents back into the sample. Given unchanging survey strategies, additional attrition is likely in future SPD waves. The final SPD data set, incorporating all the data collected, can be safely expected to include less than 50 percent of the households who gave interviews in the 1992 and 1993 panels. The result is that analyses requiring uninterrupted measures across the full decade of SPD interviews will be based on data sets that have experienced perhaps as much as 75 percent attrition.

Although cutting back on the sample size for SPD does not introduce any potential bias, the fairly sizeable attrition likely does. Because something is known about each of the households who stopped cooperating, the nature of the bias has been investigated. It appears that nonresponse was greatest among the poorest respondents, precisely the group about whose condition under PRWORA generates the most interest. We can gain some insight into the nature of the biases that the correlates of attrition introduce in the later interviews by examining the responses of households of similar characteristics who cooperated. Although weights can be calculated based on the data on nonresponding households in an attempt to offset attrition biases, the success of the weighting scheme in dealing with selection biases will be unknown.

Concerned about the high nonresponse rates, the Bureau of the Census conducted the Exploratory Attrition Study in early 1999 to ascertain the feasibility and cost of obtaining cooperation from nonrespondents. A sample of 406 households, composed partly of nonrespondents to the original 1992 and 1993 SIPP along with those who did not respond to the Bridge Survey, were traced to their current addresses; when contacted, interviews were attempted. Three levels of monetary incentives were used: $0, $50, and $100. Although the total response rate was 37 percent, those offered $100 as an incentive had a 44 percent response rate, compared with 29 percent for those offered no payment.

Encouraged by these results, the Bureau of the Census has asked Congress to appropriate funds to attempt to reach and convert to cooperation all the nonrespondents who were at or below 200 percent of the poverty line when they had been cooperating. Using the $100 incentive payment and a $40 incentive for each subsequent completed interview, the Bureau of the Census expects to achieve a response rate of close to 63 percent by the 2003 SPD interview. The current budget contains about $1 million to start on the project, with prospects for an additional $18 million to complete the process. Of course, it is not known whether the projected 63 percent response rate can be achieved. Certainly, there would be some improvement, perhaps even greater than expected. The value of the SPD data set would certainly be enhanced.

The Bureau of the Census also is investigating the feasibility of linking the 1992 and 1993 panels to Social Security Administration records of employment and earnings. Because social security numbers are available both for those who consistently responded and for those who did not, it will be possible to have that data for all who initially enrolled and for the entire period pre-and post-PRWORA. Assuming that the linking procedures can be reasonably successful, employment and earnings data would be available for 80 percent or more of the SPD panels. The effort would be a significant enrichment of SPD.

The information obtained from households beyond 1995 for the 1992 panel and 1996 for the 1993 panel is not of the same character as that obtained prior to those dates. In SIPP, households are interviewed every four months; in SPD, data are obtained annually. Information recalled over a four-month period is almost certainly more accurate than similar information obtained by recall over a full year. In short, the post-PRWORA information is likely to be more subject to recall errors than is the pre-PRWORA information.

Attrition and data differences thus lower the value of SPD data sets for analyzing the impact of PRWORA. Without corrective action to raise response rates, the percentage of cases with full sets of interviews may well be around 25 percent, too low by any standards. Note also that the planned efforts to improve the response rate will not improve the numbers for full panel information. The newly cooperating respondents all will have gaps consisting of the interviews missed while not cooperating.

Response rate problems are even more serious for some specialized surveys conducted periodically as part of the SPD. For example, the 1998 Adolescent Self-Administered Questionnaire achieved a 60 percent response rate, but because the rate is calculated based on an overall SPD response rate of 50 percent, the findings are based on adolescents from just 30 percent of the original households.

Putting possible data quality problems aside, what can be learned from SPD about the impact of PRWORA? Even under the conditions that SPD data are excellent, the study is still a before-and-after design, differing from NSAF primarily by having more information on the life trajectories of households for both periods. That information advantage is considerable, but it does not counteract the basic deficiency of such designs, namely an inability to distinguish between the effects of PRWORA and the effects of other trends occurring at the same time.

MDRC: The Project on Devolution and Urban Change. The Manpower Demonstration Research Corporation has conducted many of the better-known state welfare waiver experiments over the past decade or so, accumulating an impressive record of research accomplishments.[49] MDRC's Project on Devolution and Urban Change (UC) combines many of

---

[49]This section is based on MDRC project design papers and various communications with Gordon Berlin and Charles Michalopoulos. See Charles Michalopoulos, Hans Bos, Robert Lalonde, and Nandita Verman, *Assessing the Impact of Welfare Reform on Urban Communities: The Urban Change Project and Methodological Considerations*, MDRC project design paper (New York: Manpower Demonstration Research Corporation, 1997); Janet Quint, *Urban Change: Implementation Study Design*, MDRC Project Design Paper (New York: Manpower Demonstration Research Corporation, 1997); Claudia Coulton and Nandita Verma, *Using Social Indicators to Study Change in Low-Income Neighborhoods Before and*

the features of SPD and NSAF, although it focuses on cities rather than states or the nation as a whole. It also differs from the other studies in its greater reliance on administrative data for measuring employment and earnings and in having an ethnographic component. Of all the research reported in this paper, UC is more sharply focused on TANF, being concerned more with recipients of welfare than with poor households generally.

> The Urban Change project currently has a projected budget of about $13 million and is funded through private foundation grants. It is anticipated that some of the project components will turn out to be more expensive than projected: The final budget could be closer to $16 million. In keeping with MDRC policy, the resulting data sets eventually will be released for public use.

UC is planned as a set of community studies involving four cities: Cleveland, Philadelphia, Miami, and Los Angeles. The choice of those cities was heavily influenced by the openness of the political jurisdictions to MDRC access to administrative data. In each of the four cities, several research operations are planned (some of which are underway):

- Ethnographic studies in three poor neighborhoods in each site, in which about a dozen single mothers will be interviewed about their experiences with the welfare system and their work experiences.

- Implementation studies of TANF in each site, which will be based on documents, interviews with agency officials, focus groups with welfare workers, messages given to clients by agencies, and media stories about local TANF issues and activities.

- Institutional studies of the impact of TANF on for-profit and nonprofit institutions, such as schools, police departments, banks, local businesses, and private social service agencies.

- Neighborhood indicator studies using existing social indicators (for example, fertility, mortality, and crime). The indicators will be used to identify and track changes in poor neighborhoods over time.

- Sample surveys of single mothers, which consist of a 1998 survey of 1,000 mothers who were on AFDC in May 1995 and who were living in high-welfare or high-poverty

---

After the Implementation of Welfare Reform, MDRC Project Design Paper (New York: Manpower Demonstration Research Corporation, 1997); Kathryn Edin, *Urban Change Project Ethnographic Component,* MDRC Project Design Paper (New York: Manpower Demonstration Research Corporation, 1997).

neighborhoods, and a second survey (of the same size and composition) of mothers enrolled in TANF at a yet-to-be-determined post-PRWORA point. The interviews covered employment, program participation, income, fertility, and measures of family and child well-being.

- Longitudinal administrative data covering successive cohorts of nonelderly AFDC/TANF and Food Stamp Program enrollees from 1992 through 2002. Earnings records from employers' unemployment insurance reports will measure employment and earnings in covered employment. AFDC/TANF and Food Stamp Program records will be used to measure participation in those programs.

The six research projects that make up UC are designed to complement each other to produce a rich description of the systemic changes in welfare sparked by PRWORA and accompanying changes in the circumstances of poor households. Several are unique to UC. For example, none of the other major welfare reform studies contain an ethnographic component, and UC is the only study that will provide information on how the members of the "street bureaucracy" of welfare agencies interpret their roles and behave toward welfare clients.

Especially characteristic of MDRC is the important role given to administrative data. That data will be used in a multiple-cohort time-series design for estimating effects of PRWORA. That design is critical to UC estimates of the effects of PRWORA. Accordingly, the focus of this review is on how UC plans to use administrative data.

Although the details are still being worked out, the plan for using the multiple-cohort time-series design is as follows:

AFDC/TANF and Food Stamp Program administrative data will be used to define cohorts of people enrolled in those programs. The UC project will restrict its concerns to the nonelderly people on the Food Stamp rolls. A cohort is a group of people or households defined as being in some identifiable state at some point in time; hence, a 1992 AFDC/Food Stamp cohort can be defined as everyone on AFDC and/or Food Stamp rolls in that year.[50] They remain in that cohort until the end of the research period, and data for that period constitute the cohort data file. The research project runs from 1992 to 2002.

Using the employer files of state employment security agencies, each cohort's employment and earnings will be obtained for each quarter in each year from the start of the cohort to the end of the research period. Administrative records from AFDC/TANF will be used

[50]The research team also is considering defining each cohort as those who enrolled in AFDC/TANF or the Food Stamp Program over a given period of time. In addition, no decision has been made on how long the initial period should be (for example, a year, a half year, or some shorter period).

to track enrollment in AFDC/TANF and receipt of income maintenance. The earliest cohorts will provide data on pre-PRWORA earnings, employment, and welfare recipiency as well as post-PRWORA measures. Cohorts defined after 1997, of course, will only reflect experience under TANF. AFDC/TANF administrative data also contain information on some of the demographic characteristics of members of enrollees, including household composition, race, and marital status. The separate administrative data sets can be linked through such common unique identifiers as social security numbers, birth dates, and names.

The multiple-cohort design has several desirable features. First, the resulting data will provide much detail on both pre- and post-TANF participation and labor force behavior and thus a firm basis for characterizing those periods. Second, the data are not based on recall but are recorded close to the time of occurrence; for example, the data on earnings are based on payroll records and reports that employers turn in every quarter with their employment tax returns. Third, the data can be used to show how PRWORA may have changed the composition of TANF clients compared with the AFDC population. For instance, it may turn out that people with earnings potential may prefer to seek employment rather than enroll in TANF, lowering the average educational attainment of new enrollees in TANF compared with new enrollees for AFDC. Conversely, the more generous earnings disregard under TANF might encourage the working poor to enroll.

The multiple-cohort time-series design also has problems. The data set is confined to households enrolled in AFDC/TANF, Food Stamp Program, or both. Although this set includes a large percentage— perhaps the majority—of the poor, many eligible households are not enrolled in either program. The more inclusive set consists of those enrolled in the Food Stamp Program, but the size and composition of this group can be affected by changes in food stamp eligibility criteria, which were changed in PRWORA and may change again in the future.

The earnings and employment data set reflects only employment and earnings that are covered by the unemployment system. As several researchers (notably, Edin and Lein) have shown, many AFDC recipients participate in the “underground economy” in employment that is not recorded in those records as well as receive income in the form of gifts from others.[51] It is problematic whether the more generous income disregards put in place under most state TANF plans will reduce participation in the underground economy or increase the reporting of such income. A related problem is that the administrative data sets will not cover people for any

[51]The two surveys and the ethnographic studies may provide estimates of the extent of such income sources, but it is unlikely that those estimates will be able to supplement the administrative data except for those who are included in the survey samples. Underreporting of “under-the-table” earnings and gifts is a problem in most surveys. Edin and Lein report that they were able to obtain such information only after establishing strong rapport with respondents. See Kathryn Edin and Laura Lein, *Making Ends Meet: How Single Mothers Survive Welfare and Low-Wage Work* (New York: Russell Sage Foundation, 1997).

portion of the study period in which they lived out of state. Of course, in-migrants and people leaving the state will constitute relatively small groups. MDRC hopes to reach by telephone those who have left the study communities.

Although administrative data have only minimal response problems, they are not entirely without quality problems. Linking several data sets rarely can be done without error. Some data elements may be missing. That said, administrative data typically suffer from considerably fewer data quality problems than survey data.

The multiple-cohort time-series design has only limited ability to separate TANF effects from those of other historical events or trends. For example, the several cohorts will come up against time limits at different time points. If the effects of time limits are approximately equal across the different cohorts, one can be confident that such effects are independent of historical trends and events. However, trends whose influences cannot be identified may occur post-TANF. For example, the effects of a trend change (for better or worse) in the employment prospects for people seeking low-skill jobs likely cannot be clearly distinguished from the effects of TANF.

The two sample surveys of welfare participants are another important project component. The surveys are planned as before-and-after studies and are intended to measure outcomes that are not available in administrative data, such as measures of family functioning, health status, child well-being, and so on. The first survey, completed in 1998, consisted of samples in each of the four cities of 1,000 single mothers who were enrolled in AFDC in May 1995 and who lived in selected neighborhoods high in poverty and welfare participation. Much of the interview was focused on respondents' experiences over the year prior to the interview. The overall response rate to the 1998 survey was 79 percent, clearly a high value. The second survey will be of 1,000 single mothers who were enrolled in TANF after PRWORA and who live in the same neighborhoods. The second survey focuses on experiences under TANF as well as contemporary issues. Both surveys were designed as face-to-face interviews.

An important feature of the UC surveys is that their samples are drawn from the AFDC/TANF administrative data files and therefore can be linked to those files and employment and earnings files. That feature considerably enriches the resulting data set by adding employment, earnings, and welfare participation data from the administrative files for periods before and after the time of interview.

Comparisons across the two UC surveys may present problems. Each of the UC surveys is drawn from different populations. It is quite possible that the May 1995 AFDC clients differ in important ways from the post-PRWORA TANF clients; TANF applicants self-select themselves, may face different eligibility rules, and hence may be quite different from AFDC clients. Comparisons across the two surveys will require statistical adjustments.

HHS ongoing studies: The Child Impact Waiver experiments. Some of the Department of Health and Human Service's waiver experiments put in place during the 1980s had been completed by the time PRWORA was enacted, but many of those started in the 1990s were still underway.[52] The states were given the option of continuing the latter, and more than a dozen states elected to continue their experiments.

Although policy concerns about the impact of TANF centered on whether welfare clients would be weaned from dependency on benefits and moved into the labor force, strong interest also was generated about TANF's effects on children. As single mothers became employed, young children would move into out-of-the-home child care with unknown effects on their well-being. Furthermore, the new time limits and sanctions for noncompliance raised questions about what would happen to children when benefits were reduced or ended. Unfortunately, not much could be learned about effects on children from the waiver experiments. Almost all the waiver experiments concentrated on outcomes such as adult employment and earnings. Data collection instruments typically contained few variables concerned with outcomes for children.

HHS decided to make some of the best continuing waiver experiments more relevant to TANF concerns by augmenting them with special data collection efforts adding child outcome variables. To make the augmented data collection efforts comparable across experiments, HHS issued invitations for state participation in a planning phase, the product of which would be a standard data collection instrument covering outcomes for children. About twelve states agreed to participate. After the planning phase was over, the participating states were asked to submit applications for support for augmenting their waiver experiments. Out of the eleven applications submitted, five were awarded support: Connecticut, Florida, Indiana, Iowa, and Minnesota. The five states agreed to use survey instruments, mostly comparable across states, designed to capture the differences in child outcomes between experimental-group and control-group families. The Child Impact studies will concentrate on families with children between ages five and twelve, although some studies also will study younger children or adolescents. (Because the surveys take place three to four years after random assignment, the studies will capture the effects of the welfare reform experience on children who entered the program between the ages of one and

[52]This section is based on proposals submitted by the states to HHS for funding to augment the ongoing waiver experiments. See Connecticut Department of Social Services, *Child Impact Studies Project* (Hartford: Connecticut Department of Social Services, 1997); Florida Department of Children and Families, *Child Impact Studies* (Tallahassee: Florida Department of Children and Families, 1997); Indiana Division of Family and Children, *Child Impact Studies: Operational Phase* (Indianapolis: Indiana Division of Family and Children, 1997); Iowa Department of Human Services, *An Application for Continuing Funding for the Operational Phase: Iowa Child Impact Study,* (Des Moines: Iowa Department of Human Services, 1997); Minnesota Department of Human Services, *Child Impact Studies: Operational Phase* (St. Paul: Minnesota Department of Human Services, 1997). Additional information was obtained from comments on an earlier draft of this paper from Howard Rolston of the Administration for Children and Families, HHS.

nine.) About $12 million in federal funding has been allocated to the five waiver experiments through 1999, augmented by some state and private foundation funding.

Although the five waiver experiments are concerned with employment and earnings effects, a new emphasis, child outcomes, has been added. The child outcome measures common to all five waiver experiments are shown in the last two columns of table 2; the first two columns show, respectively, the expected direct effects of the waivers and the intermediate effects that are expected to lead to the anticipated child outcomes. Of course, child outcomes cannot be measured retrospectively. Each of the five states will measure those outcomes[53] for several years beyond the start of the experiments. The Child Impact surveys initially were planned to take place between 1997 and 2000. Table 3 shows programmatic features of the Child Impact Waiver experiments.

[53]The absence of baseline measures of child outcomes means that the estimates of effects will not be as precise as for employment and earnings, for which baseline measurements can be used as covariates.

Table 2. Common Construct Measures for HHS Child Impact Studies

Target of Welfare Policies	Other Variables Likely to Be Affected by State Policies	Aspect of Child's Environment Likely to Be Affected by Previous Columns	Child Outcomes
Income • Total income • Sources of income (mother's earnings, father's earnings, child support, AFDC, Food Stamp Program, SSI, foster care/adoption) • Stability of Income • Financial strain/material hardship *Employment* • Any vs. none • Health benefits through employment • Wages (hourly) • Hours of employment • Stability of employment • Education/licenses • Job skills (hard) • Multiple jobs concurrently • Barriers to employment (harassment, violence) *Family Formation* • Nonmarital birth/marital birth • Child/family living arrangements • Marital status, whether married to biological or nonbiological father	*Psychological Well-Being* • Depression • *Stability and Turbulence* • Foster care • Stability in child care • Stability in income • Number of moves of residence • Change in marital status or cohabitation • Why child not living with family *Absent Parent Involvement* • Whether child support provided Paternity establishment • Frequency of contact with child *Use of Health and Human Services* • Food stamps • Medicaid (awareness, use, eligibility) • Child care subsidy (awareness, use, eligibility) • Access to medical care *Consumption* • % of income spent on child care and rent	*Child Care* • Type • Extent • Quality (group size, ratio, licensing, parent perception) • Stability • Child care calendar for past several years *Home Environment and Parenting Practices* • Child abuse/neglect (administrative data) • Domestic violence/abusive relationships • Family routines • Aggravation/stress in parenting • HOME (Emotional Support and Cognitive Stimulation Scales)	*Education* • Engagement in school (ages 6–12) • School attendance (all children) • School Performance (all children) • Suspended/expelled (all children) • Grades (ages 6–12) *Health and Safety* • Hunger/nutrition (ages 5–12) • Rating of child's health (ages 5–12) • Regular source of care (ages 5–12) • Teen childbearing (ages 14–17) • Accidents and injuries (all children) *Social and Emotional Adjustment* • Behavior problems Index (ages 5–12) • Arrests (all children) • Positive Behaviors/Social Competence Scale (ages 5–12)

Source: Author.

Table 3. Features of Child Impact Waiver Experiments

	Connecticut	Florida	Indiana	Iowa	Minnesota
Program name	Jobs First	Family Transition Program	Indiana Welfare Reform	Family Investment Plan (FIP)	Minnesota Family Investment Plan
Contractor	Manpower Demonstration Research Corporation	Manpower Demonstration Research Corporation	Abt Associates	Mathematica Policy Research	Manpower Demonstration Research Corporation
Year started	1996	1994	1995	1993	1994
Sample	Manchester and New Haven: 1,200 families in both experimental and control groups	Escambia County: 600 families in both experimental and control groups	Clustered state sample: 1,300 families in experimental group and 1,300 in control group	Nine counties: 2,150 families in experimental group and 1,050 in control group	Eight counties: 1,977 families in experimental group and 1,323 in control group.
Focal child(ren)	Randomly selected child age 5–12	All children ages 5–12	All children ages 5–12	All children ages 5–12	Randomly selected child age 5–12
Time limits	21 months	24 months out of each 60 month period	24 months for adults	None	None
Earnings disregard	100% disregard until family reaches poverty level	First $200 and 50% of any additional earnings	No change from AFDC	100% for first four months and generous work allowances and disregard thereafter	Generous income disregards with payments continued until family reaches 140% of poverty level
Family cap	Yes: reduced additional payments	No	Yes	No	No
Child care eligibility	Extended	24 months after leaving welfare	Extended	Transitional child care extended to 24 months	No enhancement over AFDC
Sanctions	Graduated sanctions for noncompliance increasing with each violation	For noncompliance	For noncompliance	For noncompliance	For noncompliance
Child Impact survey date	2000	1998–1999	2000	1997–1998	1997–1998
Special features	Emphasis on maintaining fidelity	Emphasis on maintaining fidelity	PRWORA provisions applied to experimental group in May 1997	Controls switched to FIP in March 1997	Two experimental treatments. Emphasis on maintaining fidelity

Source: Author

Although the five augmented experiments concentrate on the same child outcomes, they differ considerably in their waiver provisions, as shown in table 3. Three of the five have time limits. Four of the five have more generous income disregards (compared with AFDC), but even those four vary considerably, from the most generous 100 percent earnings disregard (up to the poverty level) in Connecticut to the less generous Florida provisions. One of the experiments, Indiana's, is statewide, whereas all the others are conducted in one or more subareas within the states in question. Two states have family cap provisions, and three do not. All the states apply sanctions for noncompliance with waiver provisions, but only Iowa and Connecticut cut off benefits after repeated violations.

The relevance of the ongoing experiments to TANF depends heavily on the extent to which the experimental conditions match those adopted by the states as their TANF provisions. By and large, the matches are quite close: TANF provisions apparently followed quite closely the treatments used in the experiments.

In all five states, changes were made in the experimental conditions to bring them more closely into line with TANF. (The appendix to this paper describes the changes made.) Many of the changes appear to be quite minor in that they affect small numbers of participants or are small changes. For example, the only change made in Minnesota's experiment was the imposition of a five-year time limit on the experimental group, a change that makes the experimental group subject to all the provisions of Minnesota's TANF. The greatest number of changes was made in Indiana's experimental group, but most of the changes appear to be quite minor.

In Iowa the changes are major ones: The experiment essentially was discontinued in March 1997, and both control and experimental families were subjected thereafter to TANF provisions. However, because TANF provisions are almost identical to the experimental conditions, the 3.5-year pre-TANF experiment is relevant to that state's TANF conditions. Of course, measurements taken after March 1997, including child outcomes, cannot be analyzed as stemming from a randomized experiment.

Whether each of the five experiments tests conditions that are close enough to the relevant state TANF provisions is clearly a judgment call. My view is that the experiments in Minnesota, Florida, Indiana, and Connecticut are close enough[54] but that the data collected in Iowa after March 1997 will not be useful.

The five waiver experiments began and continue in the context of turbulent times for public welfare. The national as well as the state and local mass media have given much space and

[54]Three additional waiver experiments—in Vermont, Texas, and Arizona—are being continued by HHS, although they are not participating in the Child Impact Waiver experiment. The waiver conditions in the three states are quite close to the state's TANF provisions.

time to public announcements and discussions of PRWORA and TANF. For state and local welfare agencies, the times have meant changes both in professional duties and in agency regulations and procedures. Under these conditions, maintaining the fidelity of experimental conditions is a challenge to the success of the experiments. Ideally, welfare personnel should treat families in the experimental group according to the conditions assigned to them, and participating families also should know the critical features of those conditions. These strictures apply not only to the families who are active cases but also to those who are not. Families who have dropped off the rolls should know that if they were to apply again, the experimental conditions would still govern. Similarly, those in the control group who have dropped off the rolls should know that AFDC provisions would apply should they return to the rolls.

In four of the experiments—Connecticut, Florida, Indiana, and Minnesota—the operational plans include special efforts to maintain experimental conditions, including periodic communications to participating families, separate welfare staff assigned to participating families, and subroutines added to MIS systems to prevent inappropriate actions. Of course, how successful those efforts will be is unknown.

What Will Be Learned?

This section describes what can be learned from the studies reviewed. It first considers the three prospective research projects, then discusses the Child Impact Waiver experiments.

The prospective research projects. The three prospective PRWORA research projects have already spent more than $150 million, and millions more will be spent over the next five years. Assuming successful completion of data collection for the three projects, what will be learned from them? When released, the descriptive findings undoubtedly will be met with intense public and partisan interest in how the poor are faring under PRWORA. These studies will provide that desired information in great detail.

The studies will be able to provide descriptive findings on the following levels of aggregation:

- **National level.** The Urban Institute's Assessing the New Federalism project and the Census Bureau's SPD will provide findings concerning how American households were doing socioeconomically before and after PRWORA. For the nation as a whole, we will know whether earnings and employment of low-income households have improved or declined. We also will know whether children in low-income families are better or worse off in certain respects. The Urban Institute's state database also will provide detailed data on how the fifty states used the opportunities offered by devolution to design their particular versions of PRWORA. We will know which states offered generous income disregards, which adopted stricter time limits than those called for in TANF, and which put caps on payments when mothers on TANF had additional children.

- **State level.** Assessing the New Federalism will provide detailed information on each of thirteen states. The administrative case studies promise to provide rich information on how the states ran AFDC and what changes were made under TANF. Implementation issues also will be studied. NSAF's sampling design is intended to provide large enough state samples (about 3,000 households in each of the thirteen states) to describe how the circumstances of the poor changed after PRWORA. Comparisons of AFDC and TANF will be able to be made among the states and within each state. The Urban Institute will provide detailed descriptions of PRWORA and TANF plans for each of the thirteen states. The Census Bureau's SPD is not based on a large enough sample to provide data on more than the largest states—perhaps only New York and California—and even then, the samples are quite small.

- **Local level.** MDRC's Urban Change project is concerned with four cities and promises to provide detailed longitudinal descriptions of how several cohorts of poor households changed in employment and program participation over the historical periods pre- and post-TANF. UC will provide descriptions of how TANF was implemented in the cities as well as changes in several poor neighborhoods in all the cities. UC's findings will be especially valuable for learning about how the changes making up PRWORA are being interpreted in practice, with the obvious limitation that such findings are limited to the four urban places under study.

Taken together, the descriptive information resulting from the three research projects surely will be valuable. It will be of great interest to find out how various subgroups among low-income households are faring under PRWORA. For example, when TANF participants are cut off from support, either through sanctions or from coming up against time limits, what happens to them? Will job-seeking success be more difficult to achieve among teenage single mothers than among older mothers? Will TANF participants in urban neighborhoods characterized by extreme poverty have different experiences compared with participants living in small towns? Whether these analyses can be carried out, however, depends on whether the data sets contain enough households in the situations in question. For example, the NSAF surveys do not contain enough TANF participants in many of the thirteen states to sustain subgroup analyses in those states, although the UC study will likely support such analyses in each of the four cities being studied.

In general terms, the three prospective research projects will produce information bearing on the following topics:

- Employment pre- and post-PRWORA (whether adults are employed and the time patterns of employment, such as steadiness of employment).

- Welfare benefits pre- and post-PRWORA (participation patterns and amounts of cash and in-kind benefits received, including food stamps, Supplemental Security Income, housing subsidies, Earned Income Tax Credit, and other programs directed at the poor).

- Earnings from employment pre- and post-PRWORA.

- Sanctions pre- and post-PRWORA (benefit reductions through noncompliance and benefit loss resulting from time limits).

- Household composition pre- and post-PRWORA (marriage formations and dissolutions and births to welfare clients).

- Child care pre- and post-PRWORA (child care arrangements, some quality measures, turnover in arrangements).

- Health and health care pre- and post-PRWORA (self-ratings of health status, depression, medical care accessibility).

- Child well-being measures pre- and post PRWORA (school attendance, parenting quality, child behavior problems, food insecurity).

Conspicuously missing from the list above are findings concerning the impact of PRWORA and TANF. They are omitted because the before-and-after designs cannot support credible impact analyses. Nevertheless, it can be safely predicted that attempts will be made to estimate the effects of PRWORA and TANF because of widespread interest in policy circles and in the general public.

An obvious approach to impact analysis would be to exploit the longitudinal features of the data sets, contrasting individuals, cohorts, or political jurisdictions before and after PRWORA and TANF were implemented. Because there is reason to believe that many changes affected households during the research period in addition to welfare policy, any analysis must try to take such changes into account. Especially important are macro-level changes in the economy and changes in other programs, such as the Earned Income Tax Credit program. The changes occurring at the micro level may be more difficult to model and take into account. For example, increased earnings of partners and in-kind transfers may make it possible for TANF recipients to leave the program and join other households. Whether such nonwelfare changes can be adequately taken into account in the before-and-after analyses will be problematic. Indeed, it is likely that analysts may come to quite different impact assessments depending on how their statistical models are constructed.

The data sets produced by the three projects also can be analyzed as cross-sections—snapshots taken at particular points in time. A likely strategy will be to contrast households who face different TANF plans or to compare households who are TANF participants with comparable households who are not. In such analyses, the major obstacle to credibility is whether analysts have properly managed to model and adjust for selection biases.

A related approach might be to contrast the clients of different welfare plans on the state, county, or city level. The point of such analyses would be to discern how the features of different plans affect outcomes for households subject to them. The cross-sectional approach is especially fraught with potential for error. As discussed earlier, the TANF plans are bundles that are impossible to "unbundle." It is unlikely that any pairs of jurisdictions exist whose TANF plans differ in just one or even two features. It will be difficult to make a strong case for any impact estimate based on such comparisons.

Although statistical modeling cannot make up fully for the inherent weaknesses of before-and-after designs, it appears to be the best strategy for the impact analyses of the prospective data sets. Furthermore, such analyses can yield valuable information. When practiced skillfully and sensitively, such analyses can yield quite credible causal statements.[55] Estimates that are consistent across several data sets and are robust under alternative model specifications will be especially credible.

The MDRC and SPD data sets are the most promising for yielding credible findings. Both contain many before- and after-PRWORA observations, allowing for firmer estimates of pre-PRWORA trends to extrapolate into the post-PRWORA period and better estimates of the important outcomes of employment and earnings. MDRC's data sets are particularly promising because the nonsurvey information provides for more detailed understanding of the four PRWORA local programs.

The Child Impact Waiver experiments. The descriptive information resulting from the waiver experiments likely will not be of much more than local interest. Whatever benefits they may yield will come from impact estimates. To the extent that the changes in the experiments made the experimental conditions equivalent to the TANF plans in those states and control-group conditions are maintained with reasonable fidelity, the experiments will provide the only estimates of TANF effects based on randomized designs that will be available in the near future. The waiver experiments' findings therefore will be of considerable importance.

[55]Over the past decade, considerable thought has gone into how best to analyze nonexperimental data. See James Heckman and Richard Robb, "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes," in *Drawing Inferences from Self-Selected Samples,* edited by H. Wainer (Berlin: Springer-Verlag, 1986); Paul R. Rosenbaum, *Observational Studies* (New York: Springer-Verlag, 1995).

What Will Not Be Learned from the Four Studies

As emphasized in this paper, a major limitation to before-and-after designs is their inability to distinguish between the effects of historical events or trends and the effects of the program being developed. Accordingly, differences between pre- and post-PRWORA conditions cannot be uniquely attributed to PRWORA. If TANF rolls are much smaller than AFDC rolls, it could be a result of welfare reform or the consequence of other trends occurring in the post-PRWORA period. For example, in the several years before PRWORA, welfare rolls were declining; that trend simply might persist into the post-PRWORA period. Similarly, inequality in the distribution of household income has been increasing over the past decade or so. An increase in income inequality post-PRWORA may simply be a continuation of that trend. In short, the findings of the three projects will not reveal much about the overall effects of PRWORA. Of course, this limitation does not apply to the augmented waiver experiments because they are based on randomized designs.

The decline in welfare rolls also has the effect of considerably reducing the number of welfare families in each of the surveys taken after TANF went into effect. NSAF-I completed interviews with about 2,700 families receiving welfare assistance; the numbers in individual states ranged from 83 in Alabama to 379 in Wisconsin and averaged 188 in the thirteen states.[56] Because the rolls have further declined since 1997, NSAF-II will have considerably smaller sample sizes for welfare families, perhaps as much as 25 percent smaller. The small sample sizes will restrict the ability of analysts to estimate the impact of welfare reform, especially subgroup differences at the level of individual states.

The four projects also have another important limitation: As planned now, none of the research extends much beyond the first few years of PRWORA. For many states, the period covered will be one in which each state will be developing its version of PRWORA—writing regulations, training staff, and disseminating knowledge of the new system to its clients. For many states, this process will take several years to complete. The “permanent” version of PRWORA may not appear, at the earliest, until beyond the research period. The NSAF and UC second surveys may reflect a stage in the evolution of PRWORA before the transition to the permanent PRWORA versions will have occurred. The waiver experiments end in the period 1998 through 2000. Accordingly, the versions of PRWORA being studied in the four projects may not look like the PRWORA that exists later in the first decade of the twenty-first century. This timing, however, may have some strengths. For example, if PRWORA has some important unwanted effects—say, a sharp rise in the number of children in abject poverty—it would be useful to know about that as quickly as possible so that countermeasures can be undertaken.

It also is important to keep in mind that research on the effects of PRWORA will not

---

[56]By design, the Wisconsin sample was larger than in other states.

come to an end when the four research projects are completed. Indeed, the projects may be extended: An NSAF-III for 2001 is now in the planning stage; SIPP samples may be enlarged to provide more detailed information; and the waiver experiments can be extended. New research projects may be funded to carry on when these three projects have ended. It is almost certain that interest in learning more about poverty, unemployment, and PRWORA will continue indefinitely into the future.

Although the research projects cover many important outcomes that might be affected by PRWORA, they will not be able to track how the changing conditions of the poor affect child abuse, substance abuse, or housing adequacy because none of the studies plan to measure them. Missing also are measures of social approval or disapproval as experienced from peers, kin, or the larger society. Of course, limitations exist on how many outcomes can be measured with the kinds of survey instruments used, and well-tested, reliable, measurement instruments do not exist for some alternative outcome measures. Furthermore, the research designers may have thought that such areas of behavior were unlikely to be affected by PRWORA. In any event, even if all the studies successfully overcome data collection problems, not all questions concerning the changes accompanying welfare reform will be answered.

Finally, an important limitation arises because a large part of the 1994–1999 decline in enrollment may be a result of "entry effects," wherein fewer eligible families apply for benefits and, perhaps, fewer are able to successfully complete the enrollment process. Anecdotal evidence indicates that some state welfare agencies have made the application process difficult to complete, thereby discouraging prospective applicants. Whatever the cause for the decline, the entry effects preceding and accompanying PRWORA are producing important changes in welfare.

Entry effects can take several forms. In some states, welfare workers are authorized to provide emergency cash payments to people who apply for TANF benefits in exchange for becoming ineligible for TANF benefits for a specified period of time (for example, six months). The payments often are offered when it appears that the applicant may need immediate financial help, such as paying rent arrears or repairing a car, rather than long-term support. Some welfare departments have made applying for TANF a tedious and lengthy process, with the result that some applicants never complete the process. Of course, some eligible families simply may have come to believe that the new work requirements of TANF may be more trouble than the payments are worth and have not applied at all.

The research projects will have various difficulties providing data that directly bear on entry effects. Because SPD is a longitudinal study, however, it is better designed to study those effects. SPD can identify eligible families and determine the socioeconomic and employment characteristics of those who do not apply at various points in time in the pre-and post-PRWORA periods.[57] Urban Change and Assessing the New Federalism may be able to study entry effects indirectly by contrasting the composition of enrollees before and after PRWORA. Least able to provide information on entry effects are the five Child Impact Waiver experiments. The experimental and control groups were formed by randomly allocating people already enrolled at specific points in time and cannot provide information on how the composition of new enrollees subsequently may have changed.

Balancing the Books

The three prospective welfare reform research programs reviewed in this paper were driven by a concern for collecting the best possible information on how PRWORA would affect poor households. The constraints that had to be met in designing the studies were formidable. Everyone agreed that to launch randomized experiments after PRWORA started was simply out of the question. Of course, the waiver experiments could be used, although they could provide information on just a few states. Except for administrative data and extending SIPP, there was no way to collect extensive series of pre-PRWORA observations covering many states. Administrative data did not contain measures of such critically important outcomes as child well-being, and the task of pooling data from states with different data systems was formidable. SIPP had some child well-being measures, but they were not as extensive as desired. And so on.

In their design phases, each of the new projects countered its constraints in different ways. Taken as a group, it can be argued that the studies were about the best bundle of studies that could have been put together. To be sure, arguments can be made that each project could be improved, but the improvements would likely result in marginal benefits.

As the saying goes, hindsight is twenty/twenty. Taken together, the three prospective projects will spend about $200 million, and the extension of the five waiver experiments will cost more than $12 million more. Given this level of funding, would other strategies have been more productive? For example, would it have been more productive to strengthen SPD, possibly by making intensive efforts initially to raise response rates, provide for interviews every four months (rather than annual post-PRWORA interviews), and allow for following all the households in the two cohorts? Alternatively, might it have been better to strengthen NSAF-I and -II by designing them as in-person surveys, which typically experience better response rates than telephone surveys? Would it have been better to provide funds to MDRC to expand its community studies to statewide studies? Perhaps any one of those alternatives would have led to richer data sets, but none of them would have made estimating the impact of PRWORA any easier. On balance, the current investment in three separate approaches, each with significant faults, may have been the optimal strategy, resulting in complementary data sets.

[57] In addition, SPD plans to ask about reasons for applying for welfare, reasons for not applying among eligible respondents, and reasons for denied applications.

Whether extending the five waiver experiments was worthwhile requires considering different issues. In all five instances, extending the experiments took advantage of the already sunk costs of carrying the experiments through 1997. The investments in extensions are minor compared with the total costs of the experiments, and the potential returns are quite high in comparison. That said, it is doubtful that extending the waiver experiment in Iowa made any sense because the experiment essentially ended around the time that the state PRWORA plans were implemented. The waiver experiments in Connecticut, Florida, Indiana, and Minnesota have much greater potential to provide useful information.

Welfare Reform Research in the Future

In the near term, it is important to support some of the ongoing research projects by providing additional resources to strengthen their contributions. In particular, SPD will not be very useful unless response rates can be materially improved. Accordingly, it is important that the funds the Bureau of the Census needs in order to raise the response rates to SPD be appropriated. It is heartening that some funds are already at hand, but the bulk of the funding is not yet forthcoming. If SPD can be materially improved, it will provide extremely useful information on the changes accompanying welfare reform.

The effort to raise response rates should be monitored carefully: If at some point it becomes clear that response rates will not be materially improved, then the effort ought to be discontinued and the unexpended funds put to some better purpose (see below). In addition, serious consideration ought to be given to actions ranging from releasing the data sets with strong warnings about their limitations to suppressing their release entirely.

The prospects for improving NSAF through investment of additional resources do not appear to be good. The Urban Institute and Westat have done as much as possible to compensate for NSAF’s weaknesses. The low response rate to NSAF-I is troubling and is irremediable except through careful weighting. However successful the weighting scheme may be—and even if NSAF-II has achieved a more acceptable response rate—two surveys, before-and-after, are quite a weak design for estimating net effects. The major value of the Urban Institute studies will come from the detailed descriptive data for the thirteen states. It will be difficult to capture the diversity of state policies in a statistical model, given the many variations chosen by states and the frequency of their change.

MDRC’s Urban Change project is not far enough along to make any judgment concerning its prospects. Especially critical will be its ability to analyze cohort experiences using administrative data. Assuming successful statistical modeling of quality data, UC will provide good estimates of effects within four important localities, supplemented by qualitative data on four local welfare systems. It is clearly too early to judge whether any steps can be taken in the short term to improve UC or, indeed, whether improvements will be needed.

The ongoing waiver experiments could prove to be valuable if strong efforts are made to ensure the fidelity of control-group conditions in each experiment. The danger is great that the control groups will be treated inappropriately and that the members of the control groups will not understand that they are not subject to state TANF rules. Maintaining the integrity of the control groups means not only more effort on the part of the contractors but also the possibility that additional funds need to be given to support training of agency personnel and to provide for more frequent reminders to control group members of the special rules governing their welfare benefits. Unfortunately, given that the experiments have been underway for nearly four years since the start of TANF, it may now be too late to bolster their integrity.

It is also quite clear that when the four projects have been completed and analyzed, their findings will leave many critical questions unanswered. Almost certainly, PRWORA will be shown to be successful in meeting some of its goals in some of the states and failing to meet other goals in others. Questions will be raised about the effectiveness of time limits, family caps, income disregards, and other elements of the reform bundle. To answer those questions, further research will be needed. What form should such research take?

Perhaps the best strategy over the next decade or so would be to authorize and fund randomized experiments testing variations on the administrative and policy bundles. Those studies could be accomplished through state initiatives, but they are not likely to happen without federal funding. For example, questions about the effectiveness of family caps in reducing fertility might best be answered by conducting randomized experiments in which the experimental group is subject to family caps and the control group is under a no-family-caps condition (or vice versa). Randomized experiments can also be designed to observe the effects of varying the generosity of income disregards. Factorial experiments might be started to test the effects of various combinations of provisions that make up administrative and policy bundles.

This research strategy should lead to the accumulation of knowledge about how best to design welfare to achieve the dual objective of providing a safety net for the poor and facilitating entry into employment and higher income. A more expanded version of the current ACF strategy is proposed here. The current ACF experiments are not designed to unbundle TANF as much as to test proposed additions to the bundle. To understand how the bundles work, it is necessary to design experiments that vary such critical elements as earnings disregards, family caps, and time limits.

Some progress has been made toward implementing this strategy: ACF has funded three experiments and has issued a request for proposals for a fourth to be funded in 2000. All the experiments are designed to test measures aimed at improving TANF. For example, an experiment in Virginia will test the effectiveness of postemployment services in helping TANF clients retain their newly obtained employment. An evaluation planned in the future will involve four to ten states in MDRC-run experiments on measures aimed at employment retention and advancement in employment. The new experiments are patterned after the waiver strategy followed in the past decade or so of AFDC. They should lead to the accumulation of knowledge about how best to design welfare to achieve the dual objective of providing a safety net for the poor and facilitating entry into employment and higher income.

Studies also are needed to provide detailed descriptions of how the poor will fare under the welfare policy changes instigated by PRWORA and whatever other policy changes occur. We should be planning now how best to collect the data that will support an empirically based understanding of what is happening to the poor and what policy changes are likely to improve their condition.

When AFDC was a more or less uniform national program, national surveys such as the CPS or the SIPP may have served the purpose of monitoring the well-being of the poor. However, as discussed earlier, devolution has meant that state-level rather than national-level data are needed. An obvious move would be to enlarge sample sizes of existing ongoing national surveys to provide adequate state sample sizes. NSAF provides a good example in its selection of a small sample of critical states, an approach that the national surveys, including SPD, might want to emulate.

It is likely that in the end, SPD will not be very useful. Hence, serious consideration ought to be given to bolstering other ongoing large-scale surveys. In particular, it would be useful to augment SIPP and CPS by enlarging their sample sizes, especially bolstering their coverage of poor families. Ideally, I would like to see the sample sizes in at least the largest states increased enough to support state estimates.

Up to this point, the CPS has provided good monitoring data on the condition of the poor for the nation as a whole. Expanding the CPS sample to provide detailed data on a sample of states—and expanding CPS variables to include more information on how families with children are faring would be extremely useful. Additional efforts also should be made to address the problem of underreporting of welfare receipt. A parallel expansion of SIPP to conduct annual panel studies in a sample of states, especially in the ten to fifteen states that contain most of the poor, would be able to provide information on post-PRWORA changes in some detail. I recommend that the National Research Council of the National Academy of Sciences or a similar body examine the suitability of using SIPP for this purpose, paying special attention to attrition and nonresponse and their impact upon obtaining valid and reliable analyses.

Finally, a serious issue is how to promote responsible analyses of these data sets. Neither the research nor the policy communities will be content with only descriptive analyses. If Wisconsin poor families are better off (or worse off) in 2001 than they were in 1997 but California poor families show an opposite pattern, then some analysts certainly will try to discern whether the differences between the two states' versions of TANF are the source of the difference. To some extent, we can expect that competition among analysts will provide constructive criticism. In any event, those who release public data sets should warn potential users about the limitations of their data as well as provide full and detailed documentation about the data sets.

Appendix 3–A

Post-PRWORA Changes Made to HHS' Ongoing Experiments

Note: This appendix is freely adapted from a summary of the changes supplied by Howard Ralston.

The specific changes made to the conditions governing the experimental and control groups in the five states involved in the Child Impact Waiver experiments are described below:

Connecticut

Changes affecting only experimental-group cases.

- **Participation allowances.** Connecticut originally provided fixed participation allowances for work-related child care and transportation expenses. If actual expenses exceeded the allowance, the state would pay a supplement for the difference. Under the changes, the transportation supplement was retained, but the supplement for child care expenses has now been eliminated for families with only one child "in care." In addition, a flat-rate special allowance has been put in place for participation in short-term activities, such as workshops or training sessions, in place of paying for child care and transportation costs.

- **Work participation exemption criteria.** Exemption for caring for a child under age one is no longer available to teen parents.

- **Time-limit extensions**. A prohibition against time-limit extensions was removed for when recipients were penalized for failure to comply with work requirements. In its place, the state required "individual performance contracts" establishing conditions recipients must meet to qualify for benefit extensions.

- **Penalties.** In place of progressive reductions in grant amounts for repetitive failure to cooperate with effort to verify eligibility or child support enforcement, a full-family sanction is now imposed for the first violation. However, the sanction can be lifted immediately by demonstrating cooperation. Sanctions for employment services violations or unjustified quits/fires remain progressive.

Changes affecting both experimental and control cases.

- **Expansions in eligibility.** Women are now eligible throughout pregnancy instead of just the last trimester.

- **Reduction in penalties**. There is no longer a penalty for transferring assets.

- **New TANF ineligibility criteria**. The following are no longer eligible for benefits:

  - Individuals convicted of drug-related felonies
  - Individuals fleeing felony prosecution or confinement
  - Individuals failing to participate in digital imaging identification procedure
  - Individuals convicted of fraudulently collecting public assistance
  - Noncitizens who have not applied for citizenship
  - Noncitizens who have not been state residents for at least six months.

Florida

Changes affecting both experimental and control cases.

- **Sanctions for noncompliance with work requirements.** The first instance of noncompliance is met with full-family sanction, but benefits are reinstated immediately upon compliance. The second instance is met with full-family sanction, and full benefits are restored after thirty days of compliance, although benefits may be continued for children under age seventeen. The third instance of noncompliance is met with full-family sanction, which can be restored after three months of compliance. After six consecutive months of compliance, the sanction counter is set to zero.

TANF differences from experimental-group conditions.

- **Time limits.** TANF includes a five-year lifetime time limit in addition to the time limits applying to experimental cases. The experimental time limits restrict benefits to twenty-four months in any sixty-month period, with some exceptions.

- **Asset limits.** The TANF asset limit is $2,000, whereas the limit for the experimental group is $5,000. TANF motor vehicle exclusion is $8,500, but it is $8,560 for experimental-group cases.

- **Work participation exemption.** TANF exempts adults caring for a child under 3 months old, whereas the experiment exempts those caring for a child under 6 months old.

Indiana

In May 1977 the experimental-group conditions were changed to match those in the state’s new TANF program. The control-group conditions were not changed.

The changes made to the experimental group were as follows:

- **Time limits.** Months of benefit receipt were calculated cumulatively rather than consecutively for a total of twenty-four months. Months of ineligibility caused by program sanctions are considered as months of receipt. In addition, a recipient may earn an additional month of assistance for each consecutive six months of full-time employment.

- **Minor parents.** Minor parents and their children are required to live with specified relatives with the income and resources of the host relative considered in determining eligibility.

- **Fraud penalty.** A recipient convicted of fraud will become ineligible for assistance for twelve months following the first and second offense and become permanently ineligible upon conviction of a third offense.

- **School attendance requirement.** Excessive unexcused absences by a dependent child results in referral to a case manager for diagnosis and the development of a plan for remediation. Remediation failure may result in fiscal sanction affecting both caretaker and child.

- **Family benefit cap**. A monthly voucher equal to the cash increment for another child is available for children subject to the family cap.

- **Child care payment**. In place of receiving assistance equal to the family's benefit payment level before employment, employed recipients may opt to receive only a payment to cover child expenses.

- **Child support enforcement.** Failure to cooperate with paternity establishment will result in denial of benefits.

- **Personal responsibility agreement**. Each parent or caretaker must sign a personal responsibility agreement.

- **Penalties for illegal drug use.** Drug abusers will be referred for assessment and treatment. Failure to comply with treatment will result in a sanction of $90 per month.

- **JOBS volunteers.** Exempt individuals who volunteer to participate in the JOBS program will be sanctioned if they do not attend or participate regularly.

- **Eligibility.** Recipient eligibility is determined using the Federal Poverty Guideline for family size, instead of the 185 percent-of-need standard.

- **Voluntary quits.** The needs of someone who voluntarily quits a job or reduces hours of work will be disregarded in determining a family's eligibility and assistance payment for six months.

- **Transitional child care.** Transitional child care is limited to twelve months during the period of the demonstration.

After the implementation of PRWORA, Indiana also established a new randomized experiment. The new experimental group experienced PRWORA TANF conditions, whereas the controls the pre-PRWORA AFDC conditions.

Iowa

Iowa terminated its experiment shortly after the implementation of its TANF plan. Because the experiment had been in place for three to three and a half years and the experimental conditions closely match the state's TANF plan, the experiments findings are arguably relevant to TANF. The differences between the experimental conditions and TANF are as follows:

- **Work transition period.** In the experiment, earnings of new workers are disregarded in the initial four months of employment. Under TANF this disregard was abolished.

- **Time limit.** No time limit existed in the experiment. TANF adopted a five-year time limit.

Minnesota

The experimental and control conditions in Minnesota's experiment were maintained. The only difference between the experimental conditions and TANF is that time limits are imposed on TANF participants.

References

Bavier, Richard. 1998. "An Early Look at the Effects of Welfare Reform." Draft memorandum. Washington, D.C.: Office of Management and Budget.

Brick, Pat Dean. 1999. "A Descriptive Review of Ten Major Surveys." Draft memorandum. Washington, D.C.: Urban Institute.

Brim, Orville G., and David L. Featherman. 1997. “Surveying Midlife Development in the United States.” Unpublished paper, Life Trends, Vero Beach, Fla.

Camasso, Michael J., Carol Harvey, Radha Jagannathan, and Mark Killingworth. 1998a. *A Final Report on the Impact of New Jersey’s Family Development Program: Experimental–Control Group Analysis.* New Brunswick, N.J.: Rutgers University.

Camasso, Michael J., Carol Harvey, Radha Jagannathan, and Mark Killingworth. 1998b. *A Final Report on the Impact of New Jersey’s Family Development Program: Results from a Pre-Post Analysis of AFDC Case Heads from 1990–1996.* New Brunswick, N.J.: Rutgers University.

Connecticut Department of Social Services. 1997. *Child Impact Studies Project.* Hartford: Connecticut Department of Social Services.

Coulton, Claudia, and Nandita Verma. 1997. *Using Social Indicators to Study Change in Low-Income Neighborhoods Before and After the Implementation of Welfare Reform.* MDRC Project Design Paper. New York: Manpower Demonstration Research Corporation.

Edin, Kathryn. 1997. *Urban Change Project Ethnographic Component*. MDRC Project Design Paper. New York: Manpower Demonstration Research Corporation.

Edin, Kathryn, and Laura Lein. 1997. *Making Ends Meet: How Single Mothers Survive Welfare and Low-Wage Work.* New York: Russell Sage Foundation.

Florida Department of Children and Families. 1997. *Child Impact Studies.* Tallahassee: Florida Department of Children and Families.

Friedman, Joseph, and Daniel H. Weinberg. 1980. “Housing Consumption in an Experimental Housing Allowance Program: Issues of Self-Selection and Housing Requirements.” Unpublished paper.

Gallagher, L. Jerome, Megan Gallagher, Kevin Perese, Susan Schreiber, and Keith Watson. 1998. “One Year after Federal Welfare Reforms: A Description of State Temporary Assistance for Needy Families (TANF) Decisions as of October 1997.” Occasional Paper No. 6. Washington, D.C.: Urban Institute.

Groves *et al.* 1998. *Preliminary Report on Non-Response Studies of NSAF*

Heckman, James, and Richard Robb. 1986. “Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes.” In *Drawing Inferences from Self-Selected Samples*. H. Wainer, ed.. Berlin: Springer-Verlag.

Hollister, Robinson. 1997. “What Works, for Whom? Learning from the Experience of Welfare Reform.” *Poverty Research News* (Summer): 3–6.

Indiana Division of Family and Children. 1997. *Child Impact Studies: Operational Phase.* Indianapolis: Indiana Division of Family and Children.

Iowa Department of Human Services. 1997. *An Application for Continuing Funding for the Operational Phase: Iowa Child Impact Study*. Des Moines: Iowa Department of Human Services.

Manpower Demonstration Research Corporation. 1997. “The Institutional Study Design: Urban Change Project.” MDRC Project Design Paper. New York: Manpower Demonstration Research Corporation.

Massey, James T., Dan O’Connor, and Karol Krotki, 1997. “Response Rates in Random Digit Dialing (RDD) Telephone Surveys,” *Proceedings of the Section on Survey Methods.* Washington, D.C.: American Statistical Association.

Michalopoulos, Charles, Hans Bos, Robert Lalonde, and Nandita Verman. 1997. *Assessing the Impact of Welfare Reform on Urban Communities: The Urban Change Project and Methodological Considerations.* MDRC Project Design Paper. New York: Manpower Demonstration Research Corporation.

Minnesota Department of Human Services. 1997. *Child Impact Studies: Operational Phase.* St. Paul: Minnesota Department of Human Services.

Quint, Janet. 1997. *Urban Change: Implementation Study Design.* MDRC Project Design Paper. New York: Manpower Demonstration Research Corporation.

Riccio, James A. 1998. *A Research Framework for Evaluating Jobs-Plus, a Saturation and Place-based Employment Initiative for Public Housing Residents.* New York: Manpower Demonstration Research Corporation.

Robins, Philip K., Robert G. Spiegelman, Samuel Weiner, and Joseph G. Bell, eds. 1980. *A Guaranteed Annual Income: Evidence From a Social Experiment.* New York: Academic Press.

Rosenbaum, Paul R. 1995. *Observational Studies.* New York: Springer-Verlag.

Rossi, Peter H., and Katherine Lyall. 1976. *Reforming Public Welfare.* New York: Russell Sage.

Urban Institute. 1997. *National Survey of America’s Families: Questionnaire.* Washington, D.C.: Urban Institute.

——. 1999. *Survey Methods and Data Reliability.* NSAF Methodology Report No. 1. Washington, D.C.: Urban Institute.

——. 1999. *1997 NSAF Sample Design.* NSAF Methodology Report No. 2. Washington, D.C.: Urban Institute.

——. 1999. *1997 NSAF Snapshot Survey Weights.* NSAF Methodology Report No. 3. Washington, D.C.: Urban Institute.

——. 1999. *1997 NSAF Benchmarking Measures of Child and Family Well-Being.* NSAF Methodology Report No. 6. Washington, D.C.: Urban Institute.

——. 1999. *Early Nonresponse Studies of the 1997 National Survey of America's Families.* NSAF Methodology Report No. 7. Washington, D.C.: Urban Institute.

——. 1999. *1997 NSAF Response Rates and Methods Evaluation.* NSAF Methodology Report No. 8. Washington, D.C.: Urban Institute.

Urban Institute Policy and Benefits Staff. 1998. "Comparing the NSAF Results to the Current Population Survey." Internal memorandum. Washington, D.C.: Urban Institute Policy and Benefits Staff.

U.S. General Accounting Office. 1998. *Welfare Reform: HHS' Progress in Implementing Its Responsibilities.* Report No. HEHS-98-44. Washington, D.C.: General Accounting Office.

———. 1998. *Welfare Reform: States Are Restructuring Programs to Reduce Welfare Dependency*. Report No. HEHS-98-109.Washington, D.C. General Accounting Office.

Weinberg, Daniel H., Vicki J. Huggins, Robert A. Kominski, and Charles T. Nelson. 1997. *A Survey of Program Dynamics for Evaluating Welfare Reform.* Paper presented at the Statistics Canada Symposium XIV.

Comments

The Survey of Program Dynamics

Daniel H. Weinberg and Stephanie S. Shipp[*] [**]

The Survey of Program Dynamics (SPD) is a ten-year longitudinal survey designed to provide data about families before and after the 1996 nationwide welfare reform. The SPD's value derives from three characteristics: (1) It was designed to focus on welfare, (2) its sample is representative of the 1992 and 1993 civilian noninstitutionalized population, and (3) its response rates are comparable to those of other longitudinal household surveys. Even so, the problem of attrition of respondents has necessitated the use of incentives and special efforts to return nonrespondents to the survey.

Because of respondent attrition, researchers have questioned the usefulness of data from the SPD; thus, the purpose of this paper is to evaluate the quality of the SPD data. The conclusions drawn from the analysis that follows are that (1) the SPD data are representative of the population when compared with the Current Population Survey (CPS) and (2) the SPD response rates are comparable to those of two other major longitudinal household surveys—the Panel Study of Income Dynamics (PSID), conducted by the Survey Research Center at the University of Michigan, and the National Longitudinal Survey of Youth (NLSY), conducted by the National Opinion Research Center for the Center for Human Resource Research at Ohio State University. Attrition, however, is still a problem for the SPD. An experimental study that the Census Bureau conducted in 1998 concluded that monetary incentives were successful in gaining

---

[*]This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications and is released to inform interested parties and to encourage discussion. The authors would like to thank the participants in an American Enterprise Institute seminar for the opportunity to develop and present this information. They particularly thank Donald Hernandez for key work analyzing attrition; Karen King and Michael McMahon for information about the current status of the SPD; Stephen Campbell, Charita Castro, and Arthur Jones for assistance in computations; Jenny Hess, Larry Long, and Kenton Kilgore for their excellent comments; and Roberta Payne for assistance in computations; and Roberta Payne for preparing numerous drafts of this manuscript.

[**]Daniel H. Weinberg is chief of the Housing and Household Economic Statistics Division at the U.S. Census Bureau; Stephanie S. Shipp is assistant division chief for Labor Force Statistics and Outreach in the Housing and Household Economic Statistics Division at the U.S. Census Bureau and chair of the Survey of Program Dynamics steering committee.

cooperation from panel nonrespondents, a finding suggesting that SPD should adopt the use of monetary incentives to reduce attrition.

This paper addresses the following questions:

- What role does the SPD play in measuring the effects of welfare reform?
- How do the SPD's response rates compare with those of the 1968 PSID and 1979 NLSY?
- What affected SPD attrition?
- How do data from the SPD compare with data from the CPS March Demographic Supplement?
- What was learned from the SPD Exploratory Attrition Study[1] and the use of incentives?
- What response rates can be expected if the Census Bureau receives funding to regain the participation of nonrespondents to the 1997 SPD and the 1992 and 1993 Survey of Income and Program Participation (SIPP)?

Research Questions

What role does the SPD play in measuring the effects of welfare reform? The SPD is a national longitudinal survey that follows the same families for up to ten years, from 1992 through 2001. In 1996, Congress mandated that the Census Bureau continue to collect data from households who participated in the 1992 and 1993 panels of the SIPP, households that had already completed survey participation by January 1995 or 1996, respectively (see box 1). This additional data collection allows the Census Bureau to obtain information on changes in program participation, employment, and earnings as well as measures of adult and child well-being in the post-1996 time period. The data collected from the 1992 and 1993 SIPP panels provided us with three years of longitudinal baseline data prior to major welfare reform. Data collected in those panels included information on the factors that determine program eligibility, program access and participation, transfer income and in-kind benefits, detailed economic and demographic data on employment and job transitions, earnings and other types of income, and family composition. The SIPP data (1992–1995 for half the sample, 1993–1995 for the other half), combined with the SPD data (1996–2001), will provide ten years of annual panel data capturing both pre- and post-welfare reform data. As do most longitudinal surveys, the SPD follows people in the original sample who move or form new households.

---

[1]As will be explained later, this study was conducted to test the feasibility and costs of finding and interviewing nonrespondents from the SIPP sample—the sample of households used for the SPD survey.

Several other surveys also will contribute to our understanding of the changes that result from welfare reform. The 1996 SIPP will provide nearly four years of longitudinal data—from April 1996 through March 2000—for almost 37,000 households. A special welfare reform module was collected from August to November 1998.[2]

The 1997 SPD data (collected in 1996) were released in February 1999. The 1998 SPD data (collected in 1997) were released in February 2000. Although we released calendar year files, our main focus is to develop a longitudinal processing system to create a unified data file with common formats. It is only with such a file, and an appropriate longitudinal weight, that sophisticated before-and-after analysis of the effects of welfare reform can take place. Developing such a processing system is a major undertaking.

The Census Bureau's CPS and the Urban Institute's National Survey of America's Families (NSAF) are cross-sectional surveys that will be used to study the effects of welfare reform. The CPS already has been used to study other nonexperimental welfare changes, such as those made in 1981 to the Aid to Families with Dependent Children (AFDC) program.[3] The NSAF data are being collected specifically to evaluate the 1996 changes.[4]

Researchers also hope to learn about different components of the 1996 changes by looking at preexisting, continuing experimental studies, such as welfare waiver demonstration projects.[5] Other useful approaches include ethnographic studies, such as the Manpower

---

[2]One-quarter of the SIPP sample is interviewed each month about the previous four months. See Charles Nelson and Pat Doyle, "Recommendations for Measuring Income and Program Participation in the Post Welfare Reform Era," in *1999 Proceedings of the Section on Government Statistics and Section on Social Statistics, American Statistical Association* (Alexandria, Va.: American Statistical Association: 1999).

[3]See, for example, Robert A. Moffit, "Incentive Effects of the U.S. Welfare System," *Journal of Economic Literature* 30 (March 1992): 1–61.

[4]See the Urban Institute's website, which discusses the NSAF survey, available from: http://newfederalism.urban.org/nsaf/index.htm, accessed November 1, 2001.

[5]See, for example, Johannes M. Bos and Veronica Fellerath, *LEAP: Final Report on Ohio's Welfare Initiative to Improve School Attendance among Teenage Parents* (New York: Manpower Demonstration Research Corporation, 1997); Thomas M. Fraker, Lucia A. Nixon, Jan L. Losby, Carol S. Prindle, and John F. Else, *Iowa's Limited Benefit Plan: Summary Report* (Washington, D.C.: Mathematica Policy Research and the Institute for Social and Economic Development, 1997); Evan Weissman, *Changing to a Work First Strategy: Lessons from Los Angeles County's GAIN Program for Welfare Recipients* (New York: Manpower Demonstration Research Corporation, 1997); Nancy Wemmerus and Bruce Gottlieb, *Relaxing the FSP Vehicle Asset Test: Findings from the North Carolina Demonstration* (Washington,

Demonstration Research Corporation's (MDRC) Urban Change Study and the General Accounting Office's studies of welfare reform in selected states.[6] Each survey and study will provide insight into some aspects of welfare reform and should be considered part of the portfolio needed to understand that major program change.

The SPD is a unique tool for evaluating welfare reform because of its welfare reform–specific content (see boxes 1 and 2) and its ability to analyze the economic and social well-being of families both at two points in time and longitudinally over a ten-year period. The importance of the SPD is that it will provide a national longitudinal picture of welfare before, during, and after the enactment of welfare reform. Because it is a national survey, it will serve as a benchmark to the numerous state and city studies.

Rossi notes that "the question that most interests the policy community is, What have been the net effects of TANF (uniquely attributable to TANF) on the employment and well-being of low-income households?" He states that the "gold standard" for estimating the net effects of welfare reform is the randomized experiment. The SPD cannot measure the effects directly, but it can, through modeling, decompose the impact of economic changes and welfare reform. Several studies have done this using pooled time-series, cross-sectional data.[7] The same models could be tested using longitudinal data from the SPD for the 1992–2001 period. Furthermore, even gold-standard random-assignment impact studies use modeling to account for differential attrition from the treatment and control groups.

Rossi also expresses concern that attrition may compromise the amount and representativeness of data from the SPD, a problem made worse by the possible issue of incomplete longitudinal data for some households. Although we also are concerned with attrition

---

D.C.: Mathematica Policy Research, 1999).

[6]See, for example, U.S. General Accounting Office, *Welfare Reform: States Are Restructuring Programs to Reduce Welfare Dependence*, Report No. GAO/HEHS-98-109 (Washington, D.C.: U.S. General Accounting Office, 1998); Manpower Demonstration Research Corporation, *Biannual Report 1996/97* (New York: Manpower Demonstration Research Corporation, 1998).

[7]Rebecca Blank, "What Causes Public Assistance Caseloads to Grow?" Working Paper No. 6343 National Bureau of Economic Research, Cambridge, Mass.; Council of Economic Advisers, *The Effects of Welfare Policy and Economic Expansion on Welfare Caseloads: An Update* (Washington, D.C.: Council of Economic Advisers, 1999); Lewin Group, *Determinants of AFDC Caseload Growth, Final Report Prepared for the U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation* (Fairfax, Va.: Lewin Group, 1997); James P. Ziliak, David N. Figlio, Elizabeth E. Davis, and Laura S. Connolly, "Accounting for the Decline in AFDC Caseloads: Welfare Reform or Economic Growth?," Discussion Paper No. 1551–97 (Madison: Institute for Research on Poverty, University of Wisconsin, 1997).

and report later in the paper both on the representativeness of the remaining sample and on efforts to bring back nonrespondents, we must note that SPD analysts can compare initial and final conditions without necessarily using intermediate longitudinal data.

How do the SPD's response rates compare with those of the 1968 PSID and the 1979 NLSY? The usefulness of data from any study that interviews the same respondents over a period of years depends on whether the data represent the relevant populations.[8] Nonresponse by members of the original sample is a potential source of bias that can undermine the quality of estimates derived from longitudinal data. This section compares response rates between the initial interview and the most recent interview for three major national longitudinal household surveys: the SPD, the PSID, and the NLSY. See table 1 for a brief description of each survey and its universe. This section also discusses how the surveys have tried to minimize attrition.

Table 2 presents the current response rates for specified survey periods.[9] The current mortality-adjusted cumulative response rate for the entire survey period between initial sample selection and the most recent interview is 50 percent for the SPD, 64 percent for the NLSY, and 35 to 41 percent for the PSID (see the bottom line of table). The rates indicate the proportion of people designated for interview during sample selection who were successfully interviewed during each round of interviews.[10] Note also that the PSID-SRC (Survey Research Center) sample was intended to be a representative sample of the U.S. population, whereas the PSID-SEO (Survey of Economic Opportunity) subsample and the entire NLSY were representative only of selected portions of the population: low-income households in 1968 and people age fourteen to twenty-two in 1979, respectively.[11]

---

[8]Donald J. Hernandez, "Comparing Response Rates for the Survey of Program Dynamics (SPD), Panel Study of Income Dynamics (PSID), and the National Longitudinal Survey of Youth (NLSY)," internal memorandum, U.S. Bureau of the Census, January 28, 1999.

[9]These were the most recent, final estimates published or available for each survey at the time the source document was written.

[10]SPD respondents could have missed an intervening SIPP interview. They were eligible for the SPD sample if they participated in the first and last SIPP interviews for their panel.

[11]The SEO sample of the PSID was selected from low-income respondents to the 1967 SEO conducted for the Office of Economic Opportunity by the Census Bureau. Thus, these households had an extra opportunity for nonrandom attrition (nonresponse to the SEO). See Appendix A.

Box 1. Description of the Survey of Program Dynamics (SPD)

The Census Bureau conducts the SPD under the authority of the Personal Responsibility and Work Opportunity Reconciliation Act of 1996 (Public Law 104-193). P.L. 104-193 requires (and funds) the Census Bureau 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 law] 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 the causes of repeat welfare spells, and shall obtain information about the status of children participating in such panels.

The 1997 SPD “Bridge Survey” attempted to interview all sample participants in the 38,000 households that completed all waves of the 1992 and 1993 SIPP panels (73 percent of the original sample). The field staff interviewed 82 percent of those households (approximately 30,000) using a modified version of the March 1997 Current Population Survey (CPS) in May and June of 1997. This survey provides a bridge between the 1992–1993 SIPP data and the 1998–2001 SPD data.

A new core SPD questionnaire was developed for 1998 with the assistance of Child Trends and funding from the U.S. Departments of Agriculture and Health and Human Services. The 1998 survey also included a self-administered adolescent questionnaire (SAT). The SPD core instrument includes retrospective questions for all people age fifteen and over on jobs, income, and program participation as well as detailed questions about children under age fifteen. Because of budget constraints, the sample for the 1998 SPD was approximately 18,500 households, including all sample households with children in or near the poverty threshold and an overrepresentation of other households with children below or near the poverty threshold. The field staff obtained interviews from 89 percent of households eligible for the 1998 SPD. The 1999 SPD included extended measures of children’s well-being, the 2000 SPD includes a retrospective children’s residential history, the 2001 SPD will repeat the 1998 SAT, and the 2002 SPD will repeat the 1999 extended measures of children’s well-being.

Box 2. Welfare Reform-Specific Content in the SPD

Basic information: basic demographic characteristics, household composition, educational enrollment, work training, functional limitations and disability.

Economic information: employment and earnings; income sources and amounts; assets, liabilities, and program participation and eligibility information (including reasons for leaving programs and reasons not accepted into programs); health care use; health insurance coverage; and food adequacy.

Child well-being: school enrollment and enrichment activities, disability and health care use, contact with absent parent, child care arrangements, payment of child support on children's behalf, and residential history.

The SPD also includes two self-administered questionnaires:

(1) a series of questions for adults about marital relationship and conflict that includes a depression scale, and

(2) a questionnaire for adolescents ages twelve to seventeen on issues such as household routines and chores, parental monitoring, identification with parents, contact with nonresidential parent, delinquent behaviors, knowledge of welfare rules, crime-related violence, substance use, dating, sexual activity, and contraceptive use.

Response rates for some longitudinal surveys often appear higher than for the SIPP in the literature because they report their response rates on the basis of the number of households actually interviewed in Wave 1 rather than on the basis of the original sample selected for interview. Table 3 compares response rates for SPD, PSID, and NLSY at Interview 1 and at Interviews 11, 12, and 13 (the SPD's most recent interviews). Interview 1 response rates that are based on the sample selected are 91 percent, 76 percent, and 89 percent for the SPD (originally SIPP), PSID, and NLSY, respectively. A comparison of response rates from sample selection to the same number of interviews (11, 12, and 13) shows SPD rates comparable to those of PSID and somewhat lower than NLSY's. The Census Bureau conducted SPD interviews during the 1990s, when response rates for household surveys were generally somewhat lower than in the 1978–1980 period for the PSID and 1989–1991 period for the NLSY. Both the PSID and the NLSY used incentives throughout their field period to encourage participation, whereas the SIPP used no incentives.

Table 1: Summary of Three Longitudinal Surveys

	Survey of Program Dynamics Panel	Panel Study of Income Dynamics	National Longitudinal Survey of Youth
Purpose of survey	To provide panel data to evaluate the 1996 welfare reform legislation	To provide panel data to study demographic, social, and economic changes over an extended period of time	To gather information at multiple points in time on the labor market experiences of people who were age 14 to 22 in 1979
Universe	Civilian noninstitutionalized population in 1992–1993	Civilian noninstitutionalized population, (SRC sample); low-income households with householder under age 60 in 1968 (SEO sample)	Individuals age 14 to 22 in 1979
Original sample size	50,000 households	4,802 families (SRC); 23,430 people (SEO)	14,574 people
Time frame	1992-2001	1968-present	1979-present
Survey organization	U.S. Census Bureau	University of Michigan SRC	National Opinion Research Center

SEO=Survey of Economic Opportunity; SRC=Survey Research Center.
SOURCE: Author.

The SPD response rate held steady at 50 percent between the twelfth and thirteenth interviews because the Census Bureau made additional efforts to bring Wave 12 nonrespondents back into the sample and to encourage Wave 13 nonrespondents to respond. A $40 incentive was mailed to Wave 12 nonrespondents, and during Wave 13 field representatives were allowed to give $40 incentives to encourage nonrespondents to be interviewed.[12]

Examining the 1992 SPD data shows that some differential attrition occurred by income group, program participation characteristics, and family characteristics.[13] Table 4 compares the attrition rates for SPD sample cases, arranged according to the income-to-poverty ratio in the first interview month. Attrition rates are calculated for three periods: (1) between the first and last SIPP interview months for the 1992 and 1993 SIPP panels that provided the sample for the

[12]Field representatives requested that the regional office send the incentive with a letter requesting cooperation.

[13]Donald J. Hernandez, “The Survey of Program Dynamics: Sample Retention, the Accuracy of Change Estimates, and Implications for Assessing Welfare Reform, Research Brief,” internal memorandum, U.S. Bureau of the Census, November 4, 1998.

subsequent SPD Bridge Survey interview;[14] (2) between the last SIPP interview month in October 1994 and January 1995 and the SPD Bridge Survey interview in 1997; and (3) between the first SIPP interview month and the SPD bridge interview. Attrition rates are calculated as the percentage of the sample at the first time point who are not successfully followed and interviewed at the second point in time.[15]

Table 2. Response Rates for SPD, PSID, and NLSY (%)

	SPD	PSID-SRC	PSID-SEO	PSID: Total	NLSY: Always	NLSY: Currently
Sample-selection to Interview 1	90.9	77.0	50.8	66.5	89.0	89.2
Interview 1 to most recent Interview (see note below):						
All deceased included in base	51.6	45.2	45.2	45.2	69.6	86.7
Known deceased removed from base	53.6	53.0	53.0	53.0	71.5	NA
Sample selection to most recent interview (see note below):						
All deceased included in base	46.9	34.8	23.0	30.1	62.1	77.3
Known deceased removed from base	50.0	40.8	26.9	35.2	63.8	NA

Notes: Data collection year and wave (interview) number for most recent survey at the time this paper was prepared: SPD=1998 (Wave 12); PSID=1993 (Wave 26); NLSY=1996 (Wave 17). The label “always” means that a respondent never missed an interview; “currently” means that a respondent may have missed one or more interviews but is currently in the survey. See Appendix 1 for a comparison of SPD, PSID, and NLSY response rates.
NA=not available.
SEO=Survey of Economic Opportunity; SRC=Survey Research Center.
Source: Author calculations from survey documents and personal conversations; see Appendix A.

People with lower incomes have higher attrition rates than do people with higher incomes. But the differences are not enormous, and about 50 percent of the people of most direct interest to researchers for evaluating welfare reform were interviewed both in the first SIPP interview month and in the SPD Bridge Survey interview.

---

[14] The SPD Bridge Survey was a modified March 1997 CPS interview, designed to bridge the gap between the last SIPP interview (1994 or 1995) and the first SPD interview (1998). See box 1.

[15] See SIPP and SPD documentation for the specific rules used to ascertain whether or not a sample person is designated as interviewed at a particular point in time. More information available from: http://www.sipp.census.gov/sipp/ and http://www.sipp.census.gov/spd/, accessed October 5, 2001.

Table 3. Cumulative Survey of SPD, PSID, and NLSY Response Rates from Sample Selection to the First, Eleventh, Twelfth, and Thirteenth Interviews (%)

Survey	1st Interview	11th Interview	12th Interview	13th Interview
SPD	91	59	50	50
PSID	76	54	52	51
NLSY	89	79	78	77

Notes: SPD: 1st Interview=1992 or 1993; 11th–13th interviews=1997 to 1999. PSID: 1st interview= 1968; 11th–13th interviews=1978 to 1980. NLSY: 1st interview=1979; 11th–13th interview=1989 to 1991.
Source: Author calculations.

What affected SPD attrition? Budgetary and other reasons may have exacerbated attrition. First, to capture the pre–welfare reform situation of households (including prewaiver behavior), the 1992 and 1993 panels of the SIPP were used as the sampling frame for the SPD; thus, the SPD sample inherited a 27 percent attrition rate from the 1992 and 1993 SIPP panels.[16] Second, the budget was insufficient to interview all households in both the 1992 and 1993 SIPP panels for the length of the SPD; therefore, households that participated in both Wave 1 and Wave 9 or 10 interviews[17] were selected for the SPD sample. Third, because of budget constraints, the Census Bureau subsampled the 1997 SPD Bridge Survey sample for the 1998–2002 SPD. The low-income population and households with children were oversampled with certainty or near certainty to maximize the sample population most likely to receive welfare. If these types of households are also most likely to become nonrespondents, then measured attrition would be biased upward compared with a nonstratified sample.

How do data from the SPD compare with data from the CPS March demographic supplement? Tables 5 (for all respondents) and 6 (for young women) compare selected measures from the SPD with the CPS March Income Supplement.[18] The tabulations of both the SPD and CPS data presented here use normalized weights (the individual weight divided by the average sample weight); the results of the normalized weighting procedure resemble unweighted counts.

---

[16]For more details on the origins and sampling scheme of the SPD, see Daniel Weinberg, Pat Doyle, Arthur Jones Jr., and Stephanie Shipp, "Measuring the Impact of Welfare Reform with the Survey of Program Dynamics," in *Social Statistics Proceedings* (Alexandria, Va.: American Statistical Association, 1998).

[17]Some households were assigned to 9 waves and others to 10 waves of interviews; for households to be eligible for SPD, they must have completed the last assigned interview.

[18]Stephen Campbell, Charita Castro, and Arthur Jones in the Census Bureau's Housing and Household Economic Statistics Division defined the variables and wrote the SAS programs to produce these data.

The normalized weights, however, preserve the weighted relationship between variables. That is, the proportional distribution is the same whether normalized or cross-sectional weights are used. The results are not national estimates.

Table 4. Three Measures of Sample Attrition in the SPD, by Income Level

Income-to-Poverty Ratio	Attrition Rates (%)		
	Interview 1 to Interview 9 or 10	Interview 9 or 10 to SPD Bridge Survey	Interview 1 to SPD Bridge Survey
0.0 to <0.5	36	26	53
0.5 to <1.0	27	24	45
1.0 to <1.5	26	23	43
1.5 to <2.0	23	21	39
2.0+	18	21	35

Source: Author calculations.

Statistical differences between the SPD and CPS at the 90 percent significance level are asterisked.[19] SPD–CPS comparisons for women ages twenty to twenty-six in 1997 (ages twenty-one to twenty-seven in 1998)[20] are a proxy for potential young mothers. This group is useful in evaluating the potential of SPD data for examining the effects of welfare reform on young mothers.

In the comparisons for all survey respondents, some statistical differences are apparent, more so for the 1998 data than for 1997. For example, the percentage of people participating in programs for the SPD and CPS are quite comparable for 1997. A greater number of significant differences are found in the 1998 data, with the SPD showing a slightly higher percentage participating in programs; however, the percentages for the two years are still reasonably close. In the comparisons for women ages twenty to twenty-six in 1997 and twenty-one to twenty-seven in 1998, only a few differences are statistically significant, possibly indicating that the data compare quite well with the CPS; but more likely, the differences are significant because of the relatively small sample size of this age group.

---

[19]When multiple comparisons are made at the 90 percent confidence level, 10 percent of differences will appear to be statistically significant just as a result of chance. In table 3–8, about twelve of the seventeen SPD-CPS comparisons (one is dependent) are significant and are thus suggestive of sample differences. In contrast, table 3–9 shows that for young women, the much smaller number of significant differences suggests few sample differences.

[20]Because we have not yet constructed family and household variables for the 1998 SPD, tabulations are shown at the individual level. For household-level comparisons for 1997 (1996 data), see Appendix C.

Table 5. Comparison of Selected Variables Collected in the SPD and in the Current CPS March Income Supplement for All Individuals (Normalized Weights)

	Program Participation (%)			
Program	**SPD 1997**	**CPS, March 1997**	**SPD 1998**	**CPS, March 1998**
Temporary Assistance for Needy Families	1.3	1.4	1.5	1.1
Supplemental Security Income	2.4	2.0	3.0*	1.9
Food stamps	8.6*	10.0	7.6*	8.7
Public housing and rent subsidies	4.1	4.5	5.3*	4.2
Energy assistance	3.2*	2.6	3.5*	2.5
Free/reduced school lunch	13.4	14.3	13.8*	12.4
Type of income	**Distribution by type of income (%)**			
Wage and salary earnings	53.8*	50.5	48.8	49.9
Retirement income	9.1*	6.8	7.7	6.8
Income from at least one asset	47.8*	40.2	45.3*	39.8
Dividends	13.1*	11.5	18.6*	12.2
Work characteristic	**Distribution by work characteristic (%)**			
Worked at all during 1997-1998	67.8*	69.0	67.9	69.1
Worked 50+ weeks	80.7*	72.6	80.7*	73.7
Worked for one employer	85.9*	84.6	84.4	84.7
Had health insurance	87.6*	84.4	89.7*	83.9
Education	**Distribution by education (%)**			
No high school diploma	35.0*	40.2	37.8*	39.7
High school diploma	26.2*	24.6	24.9	24.6
Some college	21.1*	19.3	20.6*	19.4
Bachelor's degree or higher	17.7*	15.9	16.7	16.3

Note: Except for educational status, which is as of the interview date, the data are for the previous year.
*The SPD estimate is significantly different from the CPS estimate at the 90 percent confidence level.
Source: Author.

Table 6. Comparison of Selected Variables Collected in the SPD and in the CPS March Income Supplement for Women Ages 20 to 26 in 1997 and 21 to 27 in 1998 (Normalized Weights)

	Program Participation (%)			
Program	**SPD 1997**	**CPS 1997**	**SPD 1998**	**CPS 1998**
Temporary Assistance for Needy Families	7.5	7.4	6.1	6.7
Supplemental Security Income	2.2	2.1	2.9	1.9
Food stamps	14.0	13.8	10.9	12.8
Public housing and rent subsidies	6.8	5.8	7.2	6.3
Energy assistance	3.7	2.4	3.0	2.8
Free/reduced school lunch	11.0	10.5	11.1	10.3
Type of income	**Distribution by type of income (%)**			
Wage and salary earnings	79.8	78.5	76.7	79.1
Retirement income	0.8	0.3	0.2*	3.2
Income from at least one asset	35.0	34.7	39.1	36.0
Dividends	4.0	5.2	7.7	6.2
Work characteristic	**Distribution by work characteristic (%)**			
Worked at all during 1997-1998	80.7	80.0	79.2	81.0
Worked 50+ weeks	58.8	56.9	70.8*	63.4
Worked for one employer	73.2	71.7	71.2	73.4
Had health insurance	76.6	74.9	86.8*	74.3
Education	**Distribution by education (%)**			
No high school diploma	11.1	12.2	8.5	10.6
High school diploma	27.5	30.0	27.6	29.7
Some college	41.3	40.8	42.3	38.3
Bachelor's degree or higher	20.0	17.0	21.6	21.4

Note: Except for educational status, which is as of the interview date, the data are for the previous year.
*The SPD estimate is significantly different from the CPS estimate at the 90 percent confidence level.
Source: Author.

What was learned from the SPD Exploratory Attrition Study and the use of incentives? Addressing concern about the SPD response rates, the Census Bureau conducted the Exploratory Attrition Study[21] to assess the extent to which nonrespondents could be brought back into the sample. Other longitudinal surveys (such as the PSID and NLSY) have contacted early panel nonrespondents and successfully brought them back into the sample. A key element of this experiment was to test the effectiveness of monetary incentives in encouraging people who had not responded as much as five or six years earlier to re-enter the sample and respond to a current questionnaire. The project focused on people at or below 200 percent of the poverty threshold, because they are of interest in studies of welfare reform and their attrition is much greater than that of the higher income population.

Possible reasons why attrition is a problem for the SPD include the fact that SPD households will be followed much longer than SIPP households—ten years for SPD versus four years for the 1996 SIPP panel and three years for other SIPP panels. Furthermore, additional interviews were required because of the clear legislative mandate for SPD, but respondents from the 1992–1993 SIPP panels had been told that they had completed their eligibility for the survey. Moreover, some potential participants had refused the Census Bureau many times before, thus making them “hard-core” nonrespondents.

Because of the complexities and cost of programming a computer-assisted personal interview instrument for a small sample, a revised paper questionnaire and control card from Wave 9 of the 1993 SIPP panel were used to interview households in the experiment. Three incentive amounts were tested ($0, $50, and $100) to see if the size of the incentive affected response rates. Table 7 shows the results.

The Exploratory Attrition Study sample consisted of 358 randomly selected low-income (below 200 percent of poverty) cases that became nonrespondents in the 1992 and 1993 SIPP panels and 48 cases spawned[22] after SIPP but before the SPD Exploratory Study interview, for a total of 406 cases. Of the 406 cases, 373 were eligible to be interviewed. The eligible cases included those who were interviewed, those who refused, cases in which no one was home, those who were temporarily absent, and those who moved and could not be found at the time. The remaining 33 cases included vacant units, units under construction, units occupied by people

---

[21]Karen King, “Preliminary Evaluation of the Survey of Program Dynamics (SPD) Exploratory Attrition Study,” internal memorandum to the SPD Steering Committee, U.S. Bureau of the Census, June 17, 1999.

[22]New cases are spawned when new households are formed out of original sample unit households. For example, a child who marries and establishes a separate household is a spawned case.

whose usual residence is elsewhere, demolished units, units converted to a business, and units that had been moved (for example, a mobile home).

Table 7. SPD Exploratory Attrition Study Response Rates by Income and Incentive Amount

			Response Rate with Monetary Incentive (%)		
	Sample Size	Total Response Rate (%)	$0	$50	$100
Eligible cases	373	37	29	37	44*
0–99% of poverty threshold	132	39	34	42	42
100–200% of poverty threshold	241	35	26	33	44*
0–149% of Poverty Threshold	191	39	33	42	42
150–200% of poverty threshold	182	35	25	31	46[a]

[a]Response to the $100 incentive is significantly different from the $0 rate at the 90 percent level, but not significantly different from the $50 rate.

All the comparisons were tested at the 90 percent significance level. The response rate for all eligible cases was 37 percent. The response rate for those who received the $100 incentive (44 percent) was higher than the response rate for those who received no monetary incentive (29 percent). The response rate for the $50 group (37 percent) was not significantly different from the $0 group. The total response rate for those below the poverty threshold was 39 percent, not significantly different from the response rate of 35 percent for those above the poverty threshold. Incentives have a larger effect on inducing cooperation from those who refused their last SIPP or Bridge Survey interview (or where there was no one home or all occupants were temporarily absent) than on those who moved and could not be found. For those offered $100, the response rate for the former group was 54 percent, compared with 35 percent for unlocated movers. Obviously, incentives could be offered only if the cases were located.

What response rates can be expected if the Census Bureau receives funding to regain the participation of nonrespondents to the 1997 SPD and the 1992 and 1993 SIPP? If funding becomes available, the Census Bureau plans to interview a targeted sample of SIPP and SPD Bridge Survey nonrespondents. During the period 2000–2002, we will interview a targeted sample of SPD Bridge Survey (1997) nonrespondents, and in the period 2001–2002 we will interview a targeted sample of SIPP (1992–1995) nonrespondents.[23] The targeting will follow rules parallel to those used to subsample the 1998 SPD from the 1997 Bridge Survey sample. The proposal involves paying nonrespondents an initial $100 incentive in the first year and $40 maintenance incentives in subsequent years. If this plan to reinterview nonrespondents is implemented, the Census Bureau projects response rates to increase from the current (that is, 1999) 50 percent rate to 55 to 57 percent in 2000, to 62 to 64 percent in 2001, and to 60-63 percent in 2002.[24]

There is a risk that those who return to the study will differ from those who do not return. The sample of "turned-around" nonrespondents may not be fully representative, but it is likely to be more similar to nonrespondents in general than to people who have consistently responded during the past seven to eight years.

Conclusion

The SPD is only one of many tools for evaluating welfare reform, yet it has the potential to be particularly valuable. On the basis of comparisons with the CPS March Income Supplement, SPD data are representative of the national population.[25] SPD response rates are comparable to NLSY and PSID response rates, although attrition remains a problem. Despite these positive signs, the experimental evidence suggests that it will be worthwhile to pay incentives to current nonrespondents in order to bring them back and improve SPD response rates.

[23]SPD (1998 and later) nonrespondents are always approached for later interviews.

[24]Early results from the 2000 SPD interviewing show a 56 percent response rate, consistent with these projections.

[25]To fully use the SPD data, researchers must understand that complex modeling is needed to adjust for nonresponse and to incorporate other data sources (for example, state-specific variables that describe state welfare programs).

Appendix A

Comparison of SPD, PSID, and NLSY Response Rates[26]

Survey of Program Dynamics

Household response rates between sample selection and the first interview were calculated for the SIPP 1992 and 1993 panels combined, on the basis of results presented in McMahon and in Eargle.[27] People interviewed in the first and last waves of the 1992 and 1993 SIPP samples became the SPD Bridge Survey sample. Individual response rates between the first SIPP 1992 and 1993 panel interviews and the SPD Bridge Survey interview (1997) were derived by Donald J. Hernandez using the SIPP 1992 Panel Waves 1–10 Longitudinal File, the SIPP 1993 Panel Waves 1–7 Longitudinal File, and the U.S. Census Bureau internal SPD 1997 file available on the Housing and Household Economic Statistics Division server on December 4, 1998. Deceased are identified from the SIPP data for the period between Interview 1 and the final SIPP interview prior to the SPD interview. The response rate between SPD 1997 and SPD 1998 is preliminary, and both deceased and newly institutionalized populations were removed from the base.

Panel Study of Income Dynamics

The PSID User's Guide[28] notes that the original PSID sample actually consisted of two independent samples, one drawn by the Survey Research Center (the "SRC sample"), and the other selected from the Survey of Economic Opportunity (SEO), which was conducted in 1966 and 1967 by the Census Bureau for the Office of Economic Opportunity (OEO). The initial response rate for the SEO sample was calculated to be 74 percent and was based on the sample of households provided to SRC by the Census Bureau and the OEO (SRC 1972). This result does not include the effects of (1) attrition between sample selection and the first interview of respondents by the Census Bureau in 1967, which led to a response rate of 91.6 percent (OEO 1970); (2) sample loss through subsequent refusals to remain in the sample that became the SEO

---

[26]This appendix is based on material prepared by Donald Hernandez (1999).

[27]Michael F. McMahon, "Final 1992 Panel Type A and Type D Sample Loss Report," internal memorandum, U.S. Bureau of the Census, November 27, 1995; Judy Eargle, "Final 1993 Panel Type A and Type D Sample Loss Report," internal memorandum, U.S. Bureau of the Census, January 27, 2000.

[28]Martha S. Hill, *The Panel Study of Income Dynamics: A User's Guide* (Newbury Park, Calif.: Sage Publications, 1992).

component of the PSID, because about 25 percent of respondents refused to allow their names to be passed to SRC (see Hill); and (3) the failure of some sampled addresses to be transmitted from OEO to SRC (Hill). To calculate the PSID–SEO sample selection-to-interview response rate of 50.8 percent, the initial response rate of 91.6 percent was multiplied by the 75 percent rate of “willingness” to have names transmitted from the Census Bureau to SRC, and then by the 74 percent response rate obtained by SRC in seeking to interview households provided by the Census Bureau and the OEO. This formula does not take into account the fact that address information for some willing participants was not transmitted from the Census Bureau and OEO to SRC. Introducing this source of sample loss into the calculations reduces the current estimate to 50.8 percent. Of course, as in all the surveys discussed here, weighting procedures were designed to take into account various factors, including sample attrition. The response rate for the SRC sample was 76 percent. The SRC sample constituted about 60 percent of the initial PSID sample, whereas the SEO sample constituted about 40 percent of the initial PSID sample.

The response rates for “Interview 1 to the most recent interview” were obtained from the SRC (1972) and from table 2a of the documentation provided by PSID.[29] The first and most recent interview years were 1968 and 1993, respectively. Tecla Loup of the PSID staff was very helpful in identifying needed estimates and confirming the interpretation of specific estimates. Deceased are identified by PSID staff from the PSID data between the first and most recent interviews. Sandra Hofferth provided the estimated response rate for 1994, where the base was adjusted for mortality.

National Longitudinal Survey of Youth

The source for these estimates is *NLSY-79 Users’ Guide, A Guide to the 1979–1996 National Longitudinal Survey of Youth Data*.[30] The response rate of 89.2 percent between sample selection and first interview is obtained from table 3.3.1 of the document and is based on the cross-sectional and supplemental subsamples. The response rate of 69.6 percent between the initial interview (1979) and the most recent interview (1996) is obtained from table 3.7.1. Deceased, who numbered 224 by 1994 according to table 3.6.1, were removed from the base. An additional 39 deaths for years 1995 and 1996 also were removed from the base.[31] The “always” interviewed column of table 2 in this paper includes in the numerator those people who were

---

[29]U.S. Bureau of the Census, *A Panel Study of Income Dynamics: Procedures, Wave XXVI, a Supplement,1998*, available from: http://www.isr.umich.edu/src/psid/doc.html, accessed October 3, 2001.

[30]Ohio State University, *Center for Human Resource Research, NLS-Y Users’ Guide: A Guide to the 1979–1996 National Longitudinal Survey of Youth Data, Report Prepared for the U.S. Department of Labor* (Columbus: Ohio State University, Center for Human Resource Research, 1997).

[31]Reported in personal communication from Randall J. Olsen (1998).

interviewed in each of the seventeen interviews. The “currently” interviewed column of table 2 includes in the numerator people who were interviewed in at least the first interview and the current interview. If the base is limited to those not dropped from the survey or deceased, the proportion of those interviewed in the first interview who missed no more than one interview out of seventeen was 83.0 percent; when combined with the 4.6 percent who missed only two out of seventeen interviews, the response rate was 87.6 percent.

Appendix B

Appendix B Table 1. Household Income Distribution: The 1997 SPD Bridge Survey and the 1997 CPS March Income Supplement (%)

Income	SPD	CPS
Less than $5,000	3.1	3.4
$5,000 to 9,999	7.8	8.4
$10,000 to 14,999	8.3	8.6
$15,000 to 24,999	14.9	15.4
$25,000 to 34,999	14.0	13.7
$35,000 to 49,999	16.7	16.3
$50,000 to 74,999	18.5	18.0
$75,000 and over	16.8	16.4

Note: The distribution is similar for both the SPD Bridge Survey and CPS March Income Supplement household income. The one distinction between the two distributions is that the 1997 CPS March Income Supplement has a higher percentage of households with total income below $25,000 compared with the SPD—35.7 percent versus 34.1 percent. This difference is statistically significant.
Source: U.S. Bureau of the Census, March CPS 1997, and 1997 SPD Bridge Survey.

Appendix B Table 2. Selected Household Data from the 1997 SPD Bridge Survey and the 1997 CPS March Income Supplement

Variable	1997 SPD Bridge Survey	1997 March CPS
Average household income	$47,381.0	$47,123.0
Average age of householder	50.0[a]	48.4
Average number of children per household	0.7	0.7
Households with children under age 18 (%)	36.7[a]	37.6
Households receiving means-tested government transfers (%)		
Total	16.2	16.6
Temporary Assistance for Needy Families (TANF)	2.1[a]	2.5
Supplemental Security Income (SSI)	4.7[a]	4.4
Food stamps	7.6[a]	8.2
Energy assistance	3.3*	2.6
Housing assistance	4.7	4.9
Free lunch program	8.7	8.8
Households receiving selected means-tested benefits (%)		
Average household income	$20,110.0[a]	$19,119.0
Average age of householder	46.8[a]	44.2
Average number of children per householder	1.5	1.5
Households with children under age 18 (%)	65.7[a]	67.9

[a]The SPD estimate is significantly different from the CPS estimate at the 90 percent confidence level.
SOURCE: U.S. Bureau of the Census, March CPS 1997, and 1997 SPD Bridge Survey.

References

Blank, Rebecca. 1997. “What Causes Public Assistance Caseloads to Grow?” Working Paper No. 6343. Cambridge, Mass.: National Bureau of Economic Research.

Bos, Johannes M., and Veronica Fellerath. 1997. *LEAP: Final Report on Ohio’s Welfare Initiative to Improve School Attendance among Teenage Parents.* New York: Manpower Demonstration Research Corporation.

Council of Economic Advisors. 1999. *The Effects of Welfare Policy and Economic Expansion on Welfare Caseloads: An Update.* Washington, D.C.: Council of Economic Advisors.

Eargle, Judy. 2000. “Final 1993 Panel Type A and Type D Sample Loss Report.” Internal memorandum, U.S. Bureau of the Census, Washington, D.C. January 27**.**

Fraker, Thomas M., Lucia A. Nixon, Jan L. Losby, Carol S. Prindle, and John F. Else. 1997. *Iowa’s Limited Benefit Plan: Summary Report*. Washington, D.C.: Mathematica Policy Research and the Institute for Social and Economic Development.

Hernandez, Donald J. 1998. “The Survey of Program Dynamics: Sample Retention, the Accuracy of Change Estimates, and Implications for Assessing Welfare Reform, Research Brief.” Internal memorandum, U.S. Bureau of the Census, Washington, D.C. November 4.

———. 28 January 1999. “Comparing Response Rates for the Survey of Program Dynamics (SPD), Panel Study of Income Dynamics (PSID), and the National Longitudinal Survey of Youth (NLSY).” Internal memorandum, U.S. Bureau of the Census, Washington, D.C.

Hill, Martha S. 1992. *The Panel Study of Income Dynamics: A User’s Guide*. Newbury Park, Calif.: Sage Publications.

King, Karen. 17 June 1999. “Preliminary Evaluation of the Survey of Program Dynamics (SPD) Exploratory Attrition Study.” Internal memorandum to the SPD Steering Committee, U.S. Bureau of the Census, Washington, D.C.

Jones, Arthur, Barbara Downs, and Stephanie Shipp. 1999. “Measuring Welfare Reform: Preliminary results from the Survey of Program Dynamics.” *1999 Proceedings of the Section on Government Statistics and Section on Social Statistics, American Statistical Association.* Alexandria, Va.: American Statistical Association.

Lewin Group. 1997. *Determinants of AFDC Caseload Growth.* Final report prepared for the U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation. Fairfax, Va.

Manpower Demonstration Research Corporation (MDRC). 1998. *Biannual Report 1996/97.* New York: Manpower Demonstration Research Corporation.

McMahon, Michael F. 27 November 1995. "Final 1992 Panel Type A and Type D Sample Loss Report." Internal memorandum, U.S. Bureau of the Census, Washington, D.C.

Moffit, Robert A. "Incentive Effects of the U.S. Welfare System." *Journal of Economic Literature* 30 (March):1–61.

Ohio State University. Center for Human Resource Research. 1997. *NLS-Y Users' Guide: A Guide to the 1979–1996 National Longitudinal Survey of Youth Data*. Report prepared for the U.S. Department of Labor. Columbus.

Nelson, Charles, and Pat Doyle. 1999. "Recommendations for Measuring Income and Program Participation in the Post Welfare Reform Era." *1999 Proceedings of the Section on Government Statistics and Section on Social Statistics, American Statistical Association.* Alexandria, Va.: American Statistical Association.

U.S. Bureau of the Census. 1998. *A Panel Study of Income Dynamics: Procedures, Wave XXVI, a Supplement.* Available from: http://www.isr.umich.edu/src/psid/doc.html. Accessed 10/3/01.

U.S. General Accounting Office. 1998. *Welfare Reform: States Are Restructuring Programs to Reduce Welfare Dependence*. Report No. GAO/HEHS-98-109. Washington, D.C.: U.S. General Accounting Office.

Weinberg, Daniel, Pat Doyle, Arthur Jones Jr., and Stephanie Shipp. 1998. "Measuring the Impact of Welfare Reform with the Survey of Program Dynamics." In *Social Statistics Proceedings.* Alexandria, Va.: American Statistical Association.

Weissman, Evan. 1997. *Changing to a Work First Strategy: Lessons from Los Angeles County's GAIN Program for Welfare Recipients.* New York: Manpower Demonstration Research Corporation.

Wemmerus, Nancy, and Bruce Gottlieb. 1999. *Relaxing the FSP Vehicle Asset Test: Findings from the North Carolina Demonstration.* Washington, D.C.: Mathematica Policy Research.

Ziliak, James P., David N. Figlio, Elizabeth E. Davis, and Laura S. Connolly. 1997. "Accounting for the Decline in AFDC Caseloads: Welfare Reform or Economic Growth?" Discussion Paper No. 1551–97. Madison: Institute for Research on Poverty, University of Wisconsin.

Comments

The National Survey of America's Families

Kenneth Finegold and Fritz Scheuren[*] [**]

The Personal Responsibility and Work Opportunity Reconciliation Act of 1996 can be considered the centerpiece of welfare reform. Peter H. Rossi's paper examines the contributions of current research projects to PRWORA evaluation, including the Urban Institute's Assessing the New Federalism (ANF) project and ANF's National Survey of America's Families (NSAF). Rossi seems to prefer social experiments to survey analysis as ways of researching PRWORA. Among surveys, he seems to like the Census Bureau's Survey of Income and Program Participation (SIPP) and its offspring, the Survey of Program Dynamics (SPD), better than NSAF.

We welcome Rossi's efforts, which can only result in clearer thinking about the problems of analyzing the recent transformations of American social policies, but we disagree with some of his comments about ANF and NSAF. In this response, we suggest reasons for greater enthusiasm about the contributions of survey-based research, and we show that several of the potential problems that lead Rossi to be skeptical about NSAF have minimal effects or have been effectively addressed through methods such as poststratification reweighting. We also argue for greater caution about the limitations of social experiments such as the Child Impact Waiver studies and for greater attention to the possibilities of researching PRWORA through microsimulation, an approach Rossi overlooks that can integrate and build on what is learned from experimental and survey research.

ANF and NSAF Are Broader Than PRWORA, and PRWORA Is More Than TANF

The goals of ANF and NSAF are much broader than evaluating PRWORA. ANF is a multiyear, multidisciplinary project aimed at understanding the devolution of responsibility for health care, income security, job training, social services, and other policies and the effects of

---

[*]The authors thank Linda Giannarelli, David Merriman, Elaine K. Swift, Alan Weil, and Michael Wiseman for their helpful comments.

[**]Kenneth Finegold is associate professor of government at Eastern Washington University and a visiting scholar with the Urban Institute's Assessing the New Federalism Project; Fritz Scheuren is senior fellow at the Urban Institute and director of the institute's National Survey of America's Families.

this devolution on the well-being of children and their families. PRWORA is the most important piece of devolutionary legislation, but it is not the only one; indeed, the process of devolution was underway before PRWORA's passage in August 1996.

PRWORA, moreover, did more than just replace the Aid to Families with Dependent Children entitlement with the Temporary Assistance for Needy Families (TANF) block grant, which is Rossi's focus. PRWORA also made major changes to Medicaid and to the Food Stamp Program, two low-income programs with more participants and higher costs than AFDC or TANF.

The importance of PRWORA provisions other than TANF and of devolutionary policies other than PRWORA is particularly evident in health care, which has been a major focus for NSAF and ANF. A recent ANF brief, for example, used NSAF data to estimate the number of adults who could potentially receive Medicaid under the new parental eligibility rules included in PRWORA.[1] Another recent ANF brief used NSAF data to identify patterns of access to dental care for low-income children, a topic of concern for parents and policymakers that is only indirectly related to the replacement of AFDC by TANF.[2]

An NSAF designed solely to assess TANF might have been very different. It could have had, for example, more narrowly focused questions and screening, which would in turn have yielded larger samples of TANF recipients. Even higher response rates might have been achieved as well.

Nonetheless, we think that the survey as designed and conducted is more important for research on PRWORA and TANF than Rossi's evaluation might suggest. Researchers inside and outside the Urban Institute are already using ANF and NSAF data to analyze the effects of PRWORA. Douglas J. Besharov and Peter Germanis, for example, say, "The best source of data about the families that have left welfare are surveys of former welfare recipients ('leaver studies') that have been conducted by various states and by the Urban Institute."[3] They suggest

---

[1]Lisa Dubay, Genevieve Kenney, and Stephen Zuckerman, *Extending Medicaid to Parents: An Incremental Strategy for Reducing the Number of Uninsured*, Assessing the New Federalism Policy Brief No. B-20 (Washington, D.C.: Urban Institute, 2000), available from: http://newfederalism.urban.org/html/series_b/b20/b20.html, accessed October 3, 2001.

[2]Genevieve M. Kenney, Grace Ko, and Barbara A. Ormond, *Gaps in Prevention and Treatment: Dental Care for Low-Income Children,* Assessing the New Federalism Policy Brief No. B-15 (Washington, D.C.: The Urban Institute, 2000), available from: http://newfederalism.urban.org/html/series_b/b15/b15.html, accessed October 3, 2001.

[3]Douglas J. Besharov and Peter Germanis, "Welfare Reform—Four Years Later," *Public Interest* 140 (Summer 2000): 17–35.

that one surprising finding from this data—that almost half the leavers are not working regularly—"has profound implications for the economic and social condition of low-income families." Such analyses demonstrate the value of these new information sources better than any theoretical or methodological arguments we can make here.

NSAF and Survey Analysis

Because state discretion, arguably, is the central idea of devolution, national samples, such as the Current Population Survey (CPS), the SIPP, and the SPD, may not adequately capture state-by-state effects. And those effects can be considerable: States have new responsibilities for designing programs that provide cash assistance for families with children, child support, food stamps, health insurance coverage for children, child care, and education and training programs for low-income adults.

NSAF provides larger samples of poor and near-poor households, at both the national and state levels, than do CPS, SIPP, or SPD. Table 1 shows the achieved 1997 NSAF sample sizes for all households and for households under 200 percent of the poverty threshold. Comparisons to the March 1997 CPS are also provided, illustrating one of the gaps the NSAF fills: The CPS sample sizes, especially for households under 200 percent of the poverty threshold, were for the most part too small to allow for reliable estimates of this group by state. SIPP and SPD sample sizes for the low-income population were even smaller on a state basis.

Table 1. Comparison of 1997 NSAF and 1997 CPS Sample Sizes

Site	All Households		Households Below 200 Percent of the Poverty Threshold	
	NSAF	CPS	NSAF	CPS
Alabama	2,553	561	1,276	191
California	2,543	3,904	1,224	1,480
Colorado	3,175	678	1,249	195
Florida	2,368	2,018	1,158	724
Massachusetts	3,238	979	1,114	265
Michigan	2,776	1,392	1,061	362
Minnesota	3,285	573	1,182	134
Mississippi	2,390	518	1,293	229
New Jersey	3,567	1,249	1,193	296
New York	2,632	2,825	1,222	1,008
Texas	2,452	2,350	1,295	913
Washington	3,393	566	1,337	177
Wisconsin	5,355	607	2,111	145
Balance of United States	4,716	23,687	2,086	7,428
Total	44,461	41,907	18,801	13,547

Source: Fritz Scheuren and Kevin Wang, "The National Survey of America's Families: a Mid-Course Review," paper presented at the American Enterprise Institute Conference on Child Well-Being Under Welfare Reform, Washington, D.C., December 1999, available from: http://newfederalism.urban.org/nsaf/methodology_rpts/Methodology_16.pdf, accessed October 3, 2001.

The 1997 NSAF is a large national sample with sizable oversamples in thirteen states (collectively covering more than half the population of the United States). As table 1 displays, interviews for the 1997 NSAF were conducted in more than 44,000 households, yielding information on more than 100,000 people. Wisconsin was targeted for particularly intensive study, with separate large samples for Milwaukee and the balance of the state.

For the first round of the NSAF, data were obtained from February to November 1997. The survey asked an extensive battery of questions on the economic, health, and social characteristics of children, adults under age sixty-five, and their families. By design, households under 200 percent of the federal poverty threshold were oversampled. Westat conducted the data collection for the Urban Institute and Child Trends, Inc. The 1999 round of the survey has about the same sample sizes by state as the 1997 round; later rounds are expected to be similar in scope.

The NSAF is a dual-frame survey with two separate components. One is a random digit dialing (RDD) survey of households with telephones. The RDD approach is a cost-effective way to collect the desired data. However, because households without telephones ("nontelephone households") contain a significant proportion of low-income children, a supplementary area sample was conducted in person for those households. In the area sample, households within sampled blocks were screened, and all nontelephone households with someone under age sixty-five were interviewed. The dual-frame procedures pioneered for the 1997 NSAF were so successful that they have been repeated, with few modifications, in the 1999 survey.

All telephone interviewers worked in central interviewing facilities using computer-assisted telephone interviewing (CATI) technology. In-person interviewers provided cellular telephones to respondents in nontelephone households to connect the respondents to the interviewing centers for the CATI interview. Nontelephone household interviews were conducted in essentially the same way as for RDD households. At least 10 percent of each telephone interviewer's work was silently monitored for quality control purposes. For a more complete summary description of NSAF.[4] Full details on the survey are to be found in the NSAF Methodology Series.[5]

---

[4]See Fritz Scheuren and Kevin Wang, "The National Survey of America's Families: a Mid-Course Review," paper presented at the American Enterprise Institute Conference on Child Well-Being Under Welfare Reform, Washington, D.C., December 1999, available from: http://newfederalism.urban.org/nsaf/methodology_rpts/Methodology_16.pdf, accessed October 3, 2001.

[5]Available from: http://newfederalism.urban.org/nsaf/methodology.html, accessed October 3, 2001.

Even if examined only from the point of evaluating PRWORA, NSAF does reasonably well in achieving appropriate sample sizes for the thirteen targeted states. The design of the NSAF sample, for example, means that it has many more families receiving AFDC/TANF than in the comparable (that is, March 1997) CPS (table 2). Despite the praise that Rossi has for SIPP/SPD, those surveys have many fewer AFDC/TANF recipients than either CPS or NSAF.

Table 2. Families Receiving Public Assistance in the 1997 NSAF and 1997 CPS

	Number of Families Receiving Public Assistance	
Site	**NSAF**	**CPS**
Alabama	103	20
California	271	274
Colorado	146	15
Florida	198	87
Massachusetts	269	43
Michigan	208	67
Minnesota	236	23
Mississippi	147	25
New Jersey	189	38
New York	215	255
Texas	169	67
Washington	279	29
Wisconsin	401	20
Balance of United States	331	924
Total	3,162	1,887

Source: Fritz Scheuren and Kevin Wang, "The National Survey of America's Families: a Mid-Course Review," paper presented at the American Enterprise Institute Conference on Child Well-Being Under Welfare Reform, Washington, D.C., December 1999.

Criticisms of NSAF response rates. Rossi's discussion of NSAF response rates is, at best, incomplete. NSAF response rates both are higher than he reports and compare favorably with those of other household surveys.

Two NSAF response rates[6] have been used, although others (for example, Scheuren[7]) may have value as vehicles for quantifying that survey's limitations. Rossi cites only the lowest of these, "65 percent for families with children and 62 percent for families without children," without drawing out the distinctions made elsewhere in the extensive NSAF Methodology Series. For comparisons with other surveys, we have been using a different weighted response rate, about 70 percent, for the 1997 NSAF, and we indicate in our findings that we expect the weighted rate to be only slightly less for the 1999 NSAF. The discrepancy arises because in the

[6]Scheuren and Wang.

[7]Fritz Scheuren, "Quality Assessment of Quality Assessments: New Tools for a New Focus?" *Proceedings of the Section on Survey Research Methods, American Statistical Association* (Alexandria, Va.: American Statistical Association, forthcoming).

1997 NSAF documentation, we give two response rate calculations. The operational rates (at 62 percent and 65 percent) originally were used and are documented in Brick, Flores-Cervantes, and Cantor.[8] The second rate calculated (at about 70 percent) was chosen to be more nearly comparable to that used in other surveys; it is described in Brick, Kenney, *et al.*[9]

How do NSAF response rates at about 70 percent compare with other household surveys similar in scale? Such a comparison indicates that the NSAF response rates must be considered to be at the high end of the spectrum—certainly not average. The most recent published review of large national RDD surveys is by Massey *et al.*, who show that the median response rate for RDD surveys is below 60 percent.[10] Less than 20 percent of the surveys they reviewed had response rates above 64 percent.

Household response rates for the CPS, perhaps the best-known non-RDD survey, are considerably better than in NSAF, at around 93 percent. One reason for the high rate is that the CPS starts out as a face-to-face survey and calculates its response rates differently. The NSAF area component offers the closest parallel; for this part of the NSAF, we achieved response rates of about 80 percent. Another important reason for the difference is that the CPS allows proxies: Any responsible adult may answer questions for the household. In the NSAF, designated respondents generally were required. The differences in respondent rules have the effect of trading off lower response rates (in the NSAF) for potentially larger measurement errors (in the CPS). This effect is probably not small; in fact, allowing proxies might have increased the NSAF response rate by up to 10 percent.

Concerns about nonresponse bias. No matter what the response rate is, survey estimates will be unbiased only when no differences exist between respondents and nonrespondents on survey items of interest. Thus, although nonresponse bias can increase as the response rate decreases, the response rate is not in itself a direct indicator of the magnitude of nonresponse bias.

---

[8]J. Michael Brick, Ismael Flores-Cervantes, and David Cantor, *1997 NSAF Response Rates and Methods Evaluation,* NSAF Methodology Report No. 8 (Washington, D.C.: Urban Institute, 1999), available from: http://newfederalism.urban.org/nsaf/methodology_rpts/Methodology_8.pdf, accessed October 3, 2001.

[9]Pat Dean Brick, Genevieve Kenney, Robin McCullough-Harlin, Shruti Rajan, Fritz Scheuren, Kevin Wang, J. Michael Brick, and Pat Cunningham, *Survey Methods and Data Reliability,* NSAF Methodology Report No. 1 (Washington, D.C.: Urban Institute, 1999), available from: http://newfederalism.urban.org/nsaf/methodology_rpts/Methodology_1.pdf, accessed October 3, 2001.

[10]James Massey, Dan O'Conner, and Karol Krotki, "Response Rates in Random Digit Dialing (RDD) Telephone Surveys," *Proceedings of the Section on Survey Research Methods* (Alexandria, Va.: American Statistical Association, 1997): 707–712.

Because of its importance in understanding devolution, great efforts were undertaken in NSAF to examine differences between the characteristics of NSAF respondents and nonrespondents; the work is documented in detail in Groves and Wissoker.[11] This was not the only effort made. An extensive set of comparisons was carried out between NSAF and several other studies with overlapping content (for example, Brick 1999), notably the CPS, the National Health Interview Survey, and the SIPP, among others. A summary of those comparisons is forthcoming in Brennan *et al.*[12]

The 1997 NSAF results just mentioned and the continuing work done on nonresponse for the 1999 NSAF have found no evidence of large or systematic nonresponse errors in the NSAF statistics examined.[13] Even before poststratification adjustments for nonresponse, the special study we did for 1997 shows little, if any, bias. After adjustment, the remaining differences between NSAF and the results from comparable surveys (like CPS, which has better response rates) are minimal. In summary, the major results from NSAF, after adjustment, are robust against nonresponse bias.

Analysis of NSAF. Other sources of error in NSAF are similar to the problems of other large-scale household surveys, and NSAF arguably may be better off than most. In addition to nonresponse, measurement errors (for example, those caused by respondents misunderstanding certain questions) and, of course, sampling error (despite NSAF's large size) can affect research uses of the data. Researchers will need to bear in mind ordinary survey limitations when doing analyses. To this end, our releases of NSAF public use files, as Rossi recommends, "warn potential users about the limitations of their data as well as provide full and detailed documentation about the data sets."

The NSAF Methodology Series (now well along for both 1997 and 1999) is our attempt to do exactly what Rossi wants (see also Scheuren 2000, for an example). When finished, the 1997 series will have twenty-two volumes. We have already come quite far in developing a

---

[11]Robert Groves and Douglas Wissoker, *Early Nonresponse Studies of the 1997 National Survey of America's Families*, NSAF Methodology Report No. 7 (Washington, D.C.: Urban Institute, 1999), available from: http://newfederalism.urban.org/nsaf/methodology_rpts/Methodology_7.pdf, accessed October 3, 2001.

[12]Naill Brennan, Leticia Fernandez, Joyce Morton, Jeffrey Passel, Adam Safir, Fritz Scheuren, Kevin Wang, Sheila Zedlewski, and Stephen Zuckerman, *1997 NSAF National Benchmarking Measures,* NSAF Methodology Report No. 15 (Washington, D.C.: Urban Institute, 2000).

[13]Fritz Scheuren, forthcoming.

complete metadata system (for example, Dippo and Sundgren)[14] around the 1997 NSAF (with nineteen volumes completed to date). More than 2,000 researchers have registered as users of the existing public use files. Widespread availability will make for the needed “competition among analysts” (as Rossi puts it)—access that promises to speed up the search for new knowledge. The data are indeed complex, but the documentation and research reports that the Urban Institute and Child Trends are producing do help in warning about the data’s limitations and how to work around them.

Social Experimentation and the Child Impact Waiver Studies

Researchers trying to assess PRWORA should be at least as cautious about using data from the Child Impact studies or future social experiments as they are about using data from NSAF or other surveys. Randomized experiments, in our experience, generally cannot be readily scaled up to look at overall effects, can have serious nonsampling problems (for example, inability to track affected individuals), and can be badly out of date when completed—testing policy options that are not actually those implemented.

Rossi describes randomized experimentation as the “‘gold standard’ design for estimating net effects” of policy change and expects the Child Impact studies of waiver experiments in Connecticut, Florida, Indiana, Iowa, and Minnesota to produce more “credible” causal findings than the other, nonexperimental research projects. He further endorses randomized experimentation as the best strategy for future PRWORA research. Yet Rossi also identifies problems with the experimental approach and its application to the five Child Impact Waiver study states. In the following section, we discuss the serious issues he raises, along with several other issues. Some of the problems are intrinsic to or common with the experimental approach. A federal system creates additional problems in the conduct of social experiments.

“Bundling.” As Rossi notes, the experimental treatments in each of the waiver states combined several distinct welfare reform provisions. Only Minnesota followed a research design that permits estimation of their separate effects.

Tracking participants. It is difficult to track participants over time, particularly those who leave the location of the experiment. This attrition becomes especially important when the potential outcomes of the experiment include changes in the probability of outmigration, as the controversial “welfare magnet” hypothesis would suggest.

---

[14]Cathryn Dippo and Bo Sundgren, “The Role of Metadata in Statistics.” paper presented at the American Statistical Association’s International Conference on Establishment Surveys—II, Buffalo, N.Y., June 2000.

"Saturation" effects. Experiments affecting a small number of randomly selected participants do not allow assessment of the "community" or "saturation" effects that could follow from universal implementation.[15] Many observers, for example, have tied the success of welfare reform to changes in agency culture that lead caseworkers to view their tasks in terms of job placement rather than benefit calculation. Widespread implementation of welfare reforms might affect recipients' wages or rents in ways that experiments on a small group of recipients, whose identities are presumably unknown to employers or landlords, would not. These issues are sometimes addressed with quasi-experimental designs, in which the experimental treatment is applied to all participants at a particular site and experimental outcomes are compared with those at a matched control site, but then the closeness of the site match becomes a new source of concern.

Differences between experimental treatments and actual policies. Because it takes time to conduct a randomized social experiment, the treatments tested often are not those that turn out to be central to policy debates, or the key provisions of legislation are approved and implemented before the experimental data have been collected and analyzed. Rossi acknowledges differences between the welfare reforms tested under the Child Impact study waiver states and those contained in PRWORA, which lead him to question the usefulness of the Iowa data collected after that state implemented TANF. He concludes, however, that in the other four states, the correspondence between the welfare waivers and TANF is "close enough." This judgment is too sanguine. The Appendix lists thirteen changes that were made to bring the Indiana experiment under PRWORA; although Rossi finds most of those changes "quite minor," surely thirteen minor changes can add up to major discontinuities.

The only change Rossi lists for Minnesota is the implementation of the five-year lifetime limit on TANF benefits, but that may be the single most important provision in PRWORA. Early findings suggest that even before recipients become subject to loss of benefits, they take time limits into account and "bank" months of eligibility for use later, when recipients might have greater need for assistance or qualify for higher benefits.[16] Of the eleven states that implemented some form of time limit before passage of PRWORA, only Delaware had a lifetime limit, and its provisions included more generous exemptions than PRWORA and had not yet been implemented statewide.[17] PRWORA thus contained stricter time limits than those implemented

---

[15]David Greenberg and Mark Shroder, *The Digest of Social Experiments,* second ed. (Washington, D.C.: Urban Institute Press, 1997).

[16]Jeff Grogger and Charles Michalopoulos, "Welfare Dynamics under Term Limits," NBER Working Paper No. 7353 (Cambridge, Mass.: National Bureau of Economic Research, 1999).

[17]L. Jerome Gallagher, Megan Gallagher, Kevin Perese, Susan Schreiber, and Keith Watson, "One Year after Federal Welfare Reform: A Description of State Temporary Assistance for Needy Families (TANF)

in any state under waivers; the Minnesota welfare waiver, in contrast, contained no time limits at all. This difference may help explain why caseloads have dropped under PRWORA in Minnesota, as in every other state, whereas caseloads actually increased under the Minnesota waiver experiment.[18]

Unpredictable costs. The Minnesota waiver experiment also illustrates another problem of social experiments: Unless benefits are rationed (which would introduce new issues of selection bias and implementation procedures), it is not possible to fix costs in advance. At the beginning of an experiment, its effects on welfare receipt and work activity are unknown. Therefore, as long as members of the experimental group who are eligible for benefits can receive them, the costs of the experiment are also unknown and cannot be held equal to the costs with the control group. In the Minnesota experiment, outcomes generally were better than in the control group, but costs were higher, too. Thus, comparisons between the experimental and control populations do not directly indicate whether the more positive outcomes resulted from the experimental treatment per se or from the input of additional resources. Post hoc statistical adjustments or cost-benefit analysis can be applied to the data, but at best, either approach will approximate what the results would have been if costs had been constant.

Federal diversity. The diversity of the American federal system makes it difficult to know whether the experimental treatments would have similar effects in other states. President George Bush, arguing for federal approval of state waiver requests, said, “These states aren’t all the same. Welfare problems in Milwaukee are quite different than those in Juneau, Alaska, for example, or in California someplace” (see Teles).[19] The obvious diversity that might justify policies that differ in Wisconsin, Alaska, and California, however, also might limit the “generalizability” of waiver data across those states.

Nonrandom site selection. At each site, participants in an experimental treatment are selected randomly to avoid selection bias, but the sites themselves are not chosen by random selection. Rather, sites are selected through a politicized waiver process in which state officials request and implement reforms, subject to approval by federal officials. To the extent that

---

Decisions as of October 1997,” Assessing the New Federalism Occasional Paper No. 6, Urban Institute, Washington, D.C.:, 1998), available from: http://newfederalism.urban.org/pdf/oneyr.pdf, accessed October 3, 2001; Steve Savner and Mark Greenberg, *The CLASP Guide to Welfare Waivers: 1992–1995* (Washington, D.C.: Center for Law and Social Policy, 1995).

[18]Virginia Knox, Cynthia Miller, and Lisa A. Gennetian, *Reforming Welfare and Rewarding Work: A Summary of the Final Report on the Minnesota Family Investment Program.* (New York: Manpower Demonstration Research Corporation, 2000).

[19]Stephen M. Teles, *Whose Welfare? AFDC and Elite Politics* (Lawrence: University Press of Kansas, 1998), 136.

variables affecting waiver outcomes also determine which states seek waivers, how federal agencies respond to the requests, or how waiver provisions are implemented, waiver outcomes will yield biased predictions about outcomes from implementation in all states.

Compared with the average state, for example, Minnesota has a more equal income distribution, lower unemployment, a more competitive two-party (and now, perhaps, three- party) system, a moralistic political culture, below-average proportions of racial and ethnic minorities and of recent immigrants, higher spending, and a more progressive tax system.[20] It has also ranked consistently as the healthiest state in the nation.[21] At least some of those variables are plausibly related both to Minnesota's decision to test a relatively liberal welfare reform package, which emphasized support and incentives over sanctions and did not include time limits, and to the positive outcomes of the experiment.

Microsimulation as a Third Approach

None of the above concerns should be taken to mean that experimental data cannot be valuable or that future experiments along the lines proposed by Rossi should not be pursued. Of course, they can be carried out only to the extent permitted under PRWORA, which, for example, prohibits states, even with waivers, from eliminating or weakening mandatory work requirements. Clearly, assessments of PRWORA based on experimental data will be most credible when they can be corroborated with data from surveys or microsimulation.

Microsimulation, which Rossi does not discuss as an approach to researching PRWORA, addresses some of the problems of social experiments or surveys. The Urban Institute's TRIM3 and other microsimulation models calculate the effects of complex, large-scale governmental tax, transfer, and health programs at the individual, family, state, and national levels.[22]

---

[20]Martha Burt, Rob Geen, and Amy-Ellen Duke, *Income Support and Social Services for Low-Income People in Minnesota,* Assessing the New Federalism State Report (Washington, D.C.: Urban Institute, 1997), available from: http://newfederalism.urban.org/html/MNincome1.html, accessed October 3, 2001; Virginia Gray, Russell L. Hanson, and Herbert Jacob, *Politics in the American States: A Comparative Analysis*, 7th ed. (Washington, D.C.: CQ Press, 1999).

[21]Teresa A. Coughlin, Shruti Rajan, Stephen Zuckerman, and Jill A. Marsteller, *Health Policy for Low-Income People in Minnesota,* Assessing the New Federalism State Report (Washington, D.C.: Urban Institute, 1997), available from: http://newfederalism.urban.org/html/MNhealth1.html, accessed October 3, 2001.

[22]Linda Giannarelli, *An Analyst's Guide to TRIM2: The Transfer Income Model, Version 2* (Washington, D.C.: Urban Institute Press,1992); Constance F. Citro and Eric A. Hanushek, eds., *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling,* 2 vols. (Washington, D.C.: National Academy Press, 1991); Gordon H. Lewis and Richard C. Michel, eds., Microsimulation Techniques for

Microsimulation models are based on data from surveys such as the CPS, but the survey data are transformed by adjustments to match administrative aggregates and imputations of missing information. The resulting data can be used to explore patterns of program eligibility and participation, interactions in the effects of different programs, and what-if experiments that alter program parameters such as benefit levels or eligibility criteria.

One advantage of microsimulation is that its estimates of the number of program participants and the amount of benefits they receive are aligned to administrative totals. This process adjusts for the underreporting of transfer payments, which, as Rossi notes, has been a consistent problem in the CPS and other surveys. By applying the rules of each complicated and interrelated program to each unit, microsimulation also generates consistent eligibility data, which is usually not available directly from survey or experimental sources. Only with microsimulation, therefore, is it possible to directly estimate the relative contributions of participation and eligibility trends to caseload changes such as the decline in TANF and food stamp recipients since 1996.

Another advantage of microsimulation is that data for a long time series is already available. Assessments of PRWORA based on experimental or survey data often rely on the convenient fiction that control group data from waiver experiments or survey data that predate implementation of PRWORA measure conditions “before” welfare reform. In truth, welfare reform has been underway since the 1980s, as the Reagan, Bush, and Clinton administrations each encouraged states to request waivers. As the latest in a series of microsimulation models developed since the 1960s with support from the U.S. Department of Health and Human Services, TRIM3 can be used to track changes in participation and eligibility throughout that period. Microsimulation also makes it possible to estimate the impact of, say, 1997 TANF rules on 1995 CPS data, or of 1995 AFDC rules on 1997 data, thus controlling for some of the demographic changes that confound other efforts to estimate the effects of changes in program rules.

Yet another advantage of microsimulation is that it can be used to estimate the effects of potential future changes in program rules more quickly and more cheaply than can be done through social experiments. Microsimulation also avoids what social experiments cannot: the ethical issues and social consequences arising from the possibility that real human beings will be harmed as their TANF benefits, food stamps, or access to health insurance are altered. For these reasons, HHS frequently has used TRIM3 to test possible changes in AFDC and TANF, and the U.S. Department of Agriculture has used Mathematica’s microsimulation model of the Food Stamp Program for similar purposes.

Tax and Transfer Analysis, (Washington, D.C.: Urban Institute Press, 1990).

One final advantage to mention is the flexibility that microsimulation models have in absorbing disparate data sources and behavioral insights. Many microsimulation datasets are amalgams of several data sources. Numerous data-handling issues arise around the creation of such amalgams,[23] but when done carefully, microsimulation can represent an excellent heuristic for combining complex information.[24]

Microsimulation, of course, has limitations. TRIM3 and other "static" models apply new or experimental program rules to existing data, thus assuming that program changes do not generate such behavioral changes as increased or decreased employment or marriage rates. In contexts such as welfare reform, where behavioral changes are feasible or are themselves of great interest, microsimulation data is best viewed as indicating the limits to expected effects. To the extent that PRWORA encouraged more welfare recipients to work or change their living arrangements, for example, microsimulation of 1995 CPS data and 1997 TANF rules would underestimate the caseload decline. This effect occurs because some of the recipients who worked more or married would lose eligibility and others would be eligible for lower benefits, which would in turn reduce their probability of participating. Another issue is that adjusting survey data to align them to administrative totals is a tricky statistical proposition. It is also necessary to build algorithms that capture new policy provisions and to test and calibrate model results.

The Inescapable Problem of Causality

If we are to draw any conclusions about what PRWORA has wrought, what it might do in the future, or what effects changes in its provisions might have, we must infer causality. We know from administrative data that welfare caseloads have declined, and we can obtain additional descriptive information from surveys such as NSAF or from the ethnographic studies that are part of the MDRC Devolution and Urban Change project discussed by Rossi. But those sources cannot directly tell us why changes have occurred. For example, caseload trends do not by themselves indicate the extent to which they were caused by welfare reform or by the unusually strong macroeconomic conditions of the past few years.

Rossi's evident hope is that data from the social experiments will make these inferences clear cut. The Child Impact Waiver studies, however, demonstrate that inferences from social

---

[23]Chris Moriarity and Fritz Scheuren, "Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure," *Journal of Official Statistics,* 17 (2001): 407–422; Chris Moriarity and Fritz Scheuren, "A Note on Rubin's 'Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations,'" *Journal of Business and Economic Statistics (*under review).

[24]National Research Council, *Combining Information: Statistical Issues and Opportunities for Research* (Washington, D.C.: National Academy Press, 1992).

experiments to public policy are complicated by problems both in the construction of the experiments and in the correspondence between the experimental treatments and the policy changes that are eventually implemented.

Survey analysis and microsimulation offer alternative methods of inferring causality. Questions of causality can be explored by comparing survey data from before and after a policy intervention, using appropriate statistical techniques to control for other independent variables. Conclusions derived from survey analysis become even more credible when time-series data is also cross sectional, as is true for NSAF; pooled designs then can be used in sophisticated tests of alternative causal hypotheses. Microsimulation can be used to explore causality by running preintervention rules on postintervention data (or vice versa) and by conducting sensitivity analyses that test the impact of changes in programs on outcomes of substantive interest. Data from the first (1997) and second (1999) rounds of NSAF are already publicly available. TRIM3 data for 1997 will be publicly available soon, enhancing the ability of researchers to grapple with the problems of causality that surround PRWORA.

Summary

In this response, we have contrasted what can be learned from social experiments with what can be learned from survey-based research and from microsimulation, and we have argued for the need to use all three approaches to evaluate PRWORA. We believe that this view is more balanced than the conclusions presented by Rossi. What remains to be said is that all three approaches should be used together whenever possible: real experiments (for example, the Child Impact Waiver studies) and thought experiments (à la TRIM3 microsimulation) applied to representative populations of potential and actual recipients obtained through surveys (such as NSAF). The strengths and weaknesses differ for each approach.[25] Optimally, all three methods should be used to evaluate PRWORA. This is exactly what the Urban Institute and Child Trends are doing as part of the ANF project.

In a world in which public officials tested reforms one at a time and waited until the tests were completed and their results fully analyzed before approving any new legislation, randomized social experiments might give us all the information we needed to assess major

---

[25]One example of how microsimulation modeling can borrow strength from a survey can be found in Dedun Ingram, John O'Hare, Fritz Scheuren, and Joan Turek, "Exploiting Coincidences in Statistical Matching," *Proceedings of the American Statistical Association, Section on Survey Research Methods* (Alexandria, Va.: American Statistical Association, forthcoming). The authors use NSAF data to test a crucial data-handling issue that can arise in building a microsimulation model that employs two or more survey data sets that have been statistically matched together. NSAF itself was introduced into the TRIM3 modeling system this year.

policy changes. In the very different world of welfare reform, however, social experiments, surveys, and microsimulation all are necessary, and so is careful inference from the imperfect measures that each approach to policy research can provide.

References

Besharov, Douglas J., and Peter Germanis. 2000. "Welfare Reform—Four Years Later." *Public Interest* 140 (Summer):17–35.

Brennan, Niall, Leticia Fernandez, Joyce Morton, Jeffrey Passel, Adam Safir, Fritz Scheuren, Kevin Wang, Sheila Zedlewski, and Stephen Zuckerman. 2000. *1997 NSAF National Benchmarking Measures*. NSAF Methodology Report No. 15. Washington, D.C.: Urban Institute.

Brick, J. Michael, Ismael Flores-Cervantes, and David Cantor. 1999. *1997 NSAF Response Rates and Methods Evaluation*. NSAF Methodology Report No. 8. Washington, D.C.: Urban Institute. Available from: http://newfederalism.urban.org/nsaf/methodology_rpts/Methodology_8.pdf. Accessed October 3, 2001.

Brick, Pat Dean. 1999. "A Descriptive Review of Ten Major Surveys." NSAF Working Paper. Washington, D.C.: Urban Institute.

Brick, Pat Dean, Genevieve Kenney, Robin McCullough-Harlin, Shruti Rajan, Fritz Scheuren, Kevin Wang, J. Michael Brick, and Pat Cunningham. 1999. *Survey Methods and Data Reliability*. NSAF Methodology Report No. 1. Washington, D.C.: Urban Institute. Available from: http://newfederalism.urban.org/nsaf/methodology_rpts/Methodology_1.pdf. Accessed October 3, 2001.

Burt, Martha R., Rob Geen, and Amy-Ellen Duke. 1997. *Income Support and Social Services for Low-Income People in Minnesota.* Assessing the New Federalism State Report. Washington, D.C.: Urban Institute. Available from: http://newfederalism.urban.org/html/MNincome1.html. Accessed October 3, 2001.

Citro, Constance F., and Eric A. Hanushek, eds. 1991. *Improving Information for Social Policy Decisions: the Uses of Microsimulation Modeling*. 2 vols. Washington, D.C.: National Academy Press.

Coughlin, Teresa A., Shruti Rajan, Stephen Zuckerman, and Jill A. Marsteller. 1997. *Health Policy for Low-Income People in Minnesota.* Assessing the New Federalism State Report.

Washington, D.C.: Urban Institute. Available from: http://newfederalism.urban.org/html/MNhealth1.html. Accessed October 3, 2001.

Dippo, Cathryn, and Bo Sundgren. 2000. “The Role of Metadata in Statistics.” Paper presented at the American Statistical Association’s International Conference on Establishment Surveys—II, Buffalo, N.Y., June.

Dubay, Lisa, Genevieve Kenney, and Stephen Zuckerman. 2000. *Extending Medicaid to Parents: An Incremental Strategy for Reducing the Number of Uninsured.* Assessing the New Federalism Policy Brief No. B-20. Washington, D.C.: Urban Institute. Available from: http://newfederalism.urban.org/html/series_b/b20/b20.html. Accessed October 3, 2001.

Gallagher, L. Jerome, Megan Gallagher, Kevin Perese, Susan Schreiber, and Keith Watson. 1998. “One Year after Federal Welfare Reform: A Description of State Temporary Assistance for Needy Families (TANF) Decisions as of October 1997.” Assessing the New Federalism Occasional Paper No. 6. Washington, D.C.: Urban Institute. Available from: http://newfederalism.urban.org/pdf/oneyr.pdf. Accessed October 3, 2001.

Giannarelli, Linda. 1992. *An Analyst’s Guide to TRIM2: The Transfer Income Model, Version 2*. Washington, D.C.: Urban Institute Press.

Gray, Virginia, Russell L. Hanson, and Herbert Jacob. 1999. *Politics in the American States: A Comparative Analysis*. 7th ed. Washington, D.C.: CQ Press.

Greenberg, David and Marvin Shroder. 1997. *The Digest of Social Experiments.* Second edition. Washington, D.C.: Urban Institute Press.

Grogger, Jeff, and Charles Michalopoulos. 1999. “Welfare Dynamics under Term Limits.” NBER Working Paper No. 7353. Cambridge, Mass.: National Bureau of Economic Research.

Groves, Robert, and Douglas Wissoker. 1999. *Early Nonresponse Studies of the 1997 National Survey of America’s Families.* NSAF Methodology Report No. 7. Washington, D.C.: Urban Institute. Available from: http://newfederalism.urban.org/nsaf/methodology_rpts/Methodology_7.pdf. Accessed October 3, 2001.

Ingram, Dedun, John O’Hare, Fritz Scheuren, and Joan Turek. Forthcoming. “Exploiting Coincidences in Statistical Matching.” *Proceedings of the American Statistical Association, Section on Survey Research Methods.* Alexandria, Va.: American Statistical Association.

Kenney, Genevieve M., Grace Ko, and Barbara A. Ormond. 2000. *Gaps in Prevention and Treatment: Dental Care for Low-Income Children.* Assessing the New Federalism Policy Brief No. B-15.Washington, D.C.: The Urban Institute. Available from: http://newfederalism.urban.org/html/series_b/b15/b15.html. Accessed October 3, 2001.

Knox, Virginia, Cynthia Miller, and Lisa A. Gennetian. 2000. *Reforming Welfare and Rewarding Work: A Summary of the Final Report on the Minnesota Family Investment Program*. New York: Manpower Demonstration Research Corporation.

Lewis, Gordon H., and Richard C. Michel, eds. 1990. *Microsimulation Techniques for Tax and Transfer Analysis*. Washington, D.C.: Urban Institute Press.

Massey, James, Dan O'Conner, and Karol Krotki. 1997. "Response Rates in Random Digit Dialing (RDD) Telephone Surveys." *Proceedings of the American Statistical Association, Section on Survey Research Methods.* Alexandria, Va.: American Statistical Association.

Moriarity, Chris, and Fritz Scheuren. 2001. "Statistical Matching: a Paradigm for Assessing the Uncertainty in the Procedure." *Journal of Official Statistics* (17): 407–422.

Moriarity, Chris, and Fritz Scheuren. Under review. A Note on Rubin's "Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations." *Journal of Business and Economic Statistics.*

National Research Council. 1992. *Combining Information: Statistical Issues and Opportunities for Research*. Washington, D.C.: National Academy Press.

Savner, Steve, and Mark Greenberg. 1995. *The CLASP Guide to Welfare Waivers: 1992–1995*. Washington, D.C.: Center for Law and Social Policy.

Scheuren, Fritz. 2000. "Quality Assessment of Quality Assessments: New Tools for a New Focus?" Forthcoming. *Proceedings of the American Statistical Association, Section on Survey Research Methods.* Alexandria, Va.: American Statistical Association.

Scheuren, Fritz, and Kevin Wang. 1999. "The National Survey of America's Families: A Mid-Course Review." Paper presented at the American Enterprise Institute Conference on Child Well-Being Under Welfare Reform, Washington, D.C., December. Available from: http://newfederalism.urban.org/nsaf/methodology_rpts/Methodology_16.pdf. Accessed October 3, 2001.

Teles, Steven M. 1998. *Whose Welfare? AFDC and Elite Politics*. Lawrence: University Press of Kansas.

Comments

The Project on Devolution and Urban Change

Charles Michalopoulos[*]

We are pleased that Peter Rossi believes that Urban Change may provide the best hope for understanding the effects of Temporary Assistance for Needy Families (TANF) reforms among currently funded research efforts. We agree in general with most of Rossi's points, but we disagree with some of the details.

We agree that more experiments should be run and funded to understand the effects of potential policies. In fact, Manpower Demonstration Research Corporation, with funding from the U.S. Department of Health and Human Services, is currently conducting the Employment Retention and Advancement project, which will use random assignment in a number of states to study policies designed to help welfare recipients stay employed and advance into better jobs.

Rossi says, "The Personal Responsibility and Work Opportunity Reconciliation Act "cannot be evaluated as a national program; only state TANF programs can be evaluated. In addition, one can expect within-state variation in implementation, especially in states such as California and New York, where welfare is administered by local political jurisdictions . . . ." That is why Urban Change focuses on local areas and includes implementation research, to understand how policies have been implemented locally; ethnographic interviews, to study how welfare recipients understand welfare reform; neighborhood indicators data, to place welfare reform into a larger economic context; and institutional research, to investigate the effect of reform on local institutions.

We agree when Rossi worries that "a serious issue is how to promote responsible analyses of these data sets. Neither the research nor the policy communities will be content with only descriptive analyses." That is why Urban Change is more than an effort to collect descriptive data; it is a project that uses a range of complementary methods to collect, integrate, *and* analyze data, with the objective of understanding how welfare reform has affected the lives of families in the four counties under study. In the new world of welfare, where understanding the welfare system means understanding state, county, and even local implementation, Urban

---

[*]Charles Michalopoulos is a senior research associate at the Manpower Demonstration Research Corporation.

Change is uniquely positioned because it is the only study that examines implementation from several perspectives—including those of welfare agencies and staff; welfare recipients; and local institutions, including churches and other nonprofit service deliverers. Although the Survey of Program Dynamics and the National Survey of America's Families hold out the promise of more generalizable findings because they cover more areas, it will be extremely difficult to bring together the knowledge of local systems needed to analyze them responsibly. Urban Change has that expertise in place already.

We also agree when Rossi notes that "none of the planned research extends in time much beyond the first few years of PRWORA." In the four Urban Change sites, time limits first ended welfare benefits in October 1998 in Miami/Dade County, and they will begin ending benefits in Cuyahoga County, Ohio, in October 2000. Philadelphia's work-trigger time limit went into effect in March 1999. Termination time limits, however, will not begin until 2002 in Philadelphia and 2003 in Los Angeles County. In three of the sites, then, the follow-up period is long enough to capture the initial effects of some form of time limit. We agree that longer follow-up would be better still.

We agree, finally, when Rossi says, "Assuming successful statistical modeling of quality data, [Urban Change] will provide good estimates of effects within four important localities, supplemented by qualitative data on four local welfare systems."

We disagree, however, with some of Rossi's other comments.

Rossi asserts that Urban Change's "findings cannot be generalized to a broader welfare population nationally or even to other urban neighborhoods." Although it is true that Urban Change is studying only four counties, in 1997 more than 10 percent of all welfare recipients nationally were in those counties, so they may be as representative as any four large urban counties could be. In addition, Rossi's concern about generalizability is a standard criticism of most social experiments, the "'gold standard' design for estimating net effects," since most experiments are conducted in only a few sites.

Rossi also asserts that Urban Change cannot separate effects of welfare reform from effects of other policy and economic changes. In this regard, Rossi has missed the strength of our studying *four* counties and having longitudinal information on several million people starting in 1992. To take one example, the Earned Income Tax Credit was expanded simultaneously nationwide. TANF, however, was implemented in Florida in October 1996, in Pennsylvania in March 1997, in Ohio in October 1997, and in Los Angeles County in January 1998. If welfare reform is responsible for sudden changes in the pattern of welfare use or movement from welfare to work, those changes should correspond to some degree to the changes in welfare policies and implementation. If they occur at the same time in all four sites, they are more likely a result of federal policy changes or changes in the national economy.

We think Rossi has missed the strength of the six interwoven components in Urban Change: impact analysis, survey data, implementation research, ethnographic interviews, neighborhood indicators, and research on local institutions. The project has already benefitted from its multiple components. For example, its first two reports have used more than one component: A report on how reforms have been implemented by welfare administrators and understood by welfare recipients combined implementation research with ethnographic interviews, and another on the health status of current and former welfare recipients combined the client survey with ethnographic interviews.

The impact component would be as weak as Rossi portrays it without the other components. Adding them substantially strengthens the project's design. Consider some relatively simple examples:

- TANF reforms went into effect nationally in October 1996. But only one of the four Urban Change counties implemented TANF reforms that early, and Los Angeles County did not implement reforms until 1998. Without the benefit of the project's implementation research, an analyst might mistakenly look for effects of reform in 1996 in all sites.

- Time limits were scheduled to go into effect in Miami/Dade County in October 1998, and work-trigger time limits were scheduled to go into effect in Philadelphia in March 1999. In both cases, the counties realized they were not prepared to implement time limits and throw people off the rolls, and so they gave extensions to recipients who would have reached the time limits. Without implementation research, we might have looked for effects of time limits at the wrong time in the two places.

- All four counties implemented "enhanced earnings disregards," which allow welfare recipients to keep more of their benefits when they go to work. The disregards are usually expected to increase the number of people receiving welfare. Ethnographic interviews with clients in the four counties indicated that, except in Philadelphia, the clients did not know about or understand the enhanced disregards. An analyst working alone would look for the same patterns of welfare use in all four counties, but we know that the pattern might look different in Philadelphia and that in three of the counties, it might look different from the pattern predicted by economic theory.

Is Urban Change perfect? Not by any means. No research effort is. Rossi himself is somewhat ambivalent about how best to evaluate welfare reform. On the one hand, he argues that nonexperimental research is inadequate compared with random assignment experiments. On the other, he notes that "it would be extremely difficult to carry [randomized experiments] out for a variety of reasons." But, like Rossi, we think that "comparisons with [Aid to Families with Dependent Children] are needed to settle the issue of whether TANF has deleterious effects on the poor." We believe Urban Change promises to contribute to this endeavor.

Comments

State Welfare Reform Waiver Experiments

Howland Rolston[*][**]

The U.S. Department of Health and Human Services, along with five states and several other private and public funders, augmented five ongoing welfare reform waiver experiments to include more systematic and uniform measures of family and child well-being. The waiver programs being evaluated represent a range of policies that could be implemented under the Personal Responsibility and Work Opportunity Reconciliation Act and, in fact, have strong similarities to each state's Temporary Assistance to Needy Families program. Although limited to five states (one of which has only one site), their strong experimental research designs will provide credible information on the effects of alternative welfare reform policies on children. This will make the Child Impact Waiver experiments a vital complement to the various nonexperimental studies that will assess the effects of welfare reform on children.

By and large, I agree with Peter Rossi's analysis of the five state projects. My comments, therefore, relate to the few instances in which I would characterize the usefulness of the waiver experiments somewhat differently from Rossi. I will also comment on a few recent developments.

Rossi rightly points out that the most critical factor related to the utility of the experiments is the integrity of the policies applied to the experimental and control groups and their knowledge of those policies. If control groups receive experimental policies, or vice versa, then at best, effects will be understated and at worst, the results will be meaningless. As Rossi explains, however, the evaluation firms have put in place precautions to prevent crossover treatment, and no evidence shows that it has occurred.

Rossi's second point also needs to be addressed, namely, that the participants in the experimental and control groups correctly understand their respective policy regimes. Here recent

[*]Expressed here are the author's personal views; they do not represent the views of HHS or the administration.

[**]Howard Rolston is director of Planning, Research and Evaluation in the Administration for Children and Families of the U.S. Department of Health and Human Services.

evidence suggests that they generally do. For example, in Connecticut 89 percent of participants in the experimental group correctly identified that they were (or had been) subject to a time limit, whereas only 23 percent of control participants incorrectly answered that they were subject to such limits. In Florida the results were similar: 88 percent of those in the experimental group correctly answered, and 29 percent of control participants incorrectly answered. Given that even in the absence of welfare reform or the experiments, one would not expect welfare recipients (or former recipients) to have perfect knowledge of the rules of welfare, this wide disparity between experimental and control groups' view of their situation suggests that although the findings may understate the effects of time limits, they do not invalidate the experiments. More generally, it is consistent with the notion that given the broad public discussion of welfare reform, it is inconceivable that the behavior of controls has been totally unaffected by its message. Thus, we ought to regard the findings of the experiments as conservative estimates of how particular policies compare with Aid to Families with Dependent Children and as more accurate estimates of their relative effects in an environment of national welfare reform.

In keeping with his concern about the integrity of the experiments, Rossi concludes that "it is doubtful that extending the waiver experiment in Iowa made any sense because the experiment essentially ended around the time that the state PRWORA plans were implemented." I think that the judgment is less clear than Rossi asserts. In fact, the treatment–control difference *was* maintained for a period of 3.5 years for 57 percent of the sample and 2.5 years for 77 percent of it. Although Rossi is correct that knowing what rules would apply were a recipient to return to welfare is integral to the experiment, it is also true that when the control embargo was lifted, only 28 percent of the control group remained on welfare, and no attempts were made to contact people who were off welfare to inform them of their changed status. Furthermore, it is reasonable to believe that the effects of having been subject to different rules for up to 3.5 years would not immediately disappear when the rules were changed. Thus, although Rossi is right to identify Iowa as the experiment whose integrity was most threatened, in my judgment it is reasonable to assume that any effects the Iowa experiment might produce on children and families would still be discernible for a period of time after the control embargo was broken.

Rossi raises the question of whether an experiment based on a pre-1996 AFDC control group has utility, given that few people would want a return to AFDC even if welfare reform interventions proved ineffective; he presents arguments on both sides of the question. Rossi then rightly goes on to say that the question of what the control group should experience "need not be settled entirely one way or the other." To me this is correct, although I would place the emphasis a little differently, given the subject of the paper. As Rossi's paper illustrates, a consensus on understanding the effects of PRWORA is something of a long shot; it depends on a convergence of findings among the various nonexperimental analyses that will be attempted on the data sets that Rossi discusses in detail as well as on other data sets and the findings of the limited set of experiments. Previous experience suggests that this convergence may not occur. But because the welfare reform experiments represent the only set of projects with experimental designs, it seems clear to me that having some projects with a pre-1996 control group is highly desirable if we are to have a chance at understanding the effects of PRWORA.

I also strongly agree with Rossi's emphasis on experiments to test what he terms "how TANF can be improved." Rossi characterizes what should be tested as "improvements." Given the broad flexibility of TANF, however, it is more illuminating to characterize the experiments as tests of alternative ways in which states can use the flexibility that TANF provides to improve a range of adult, child, and family outcomes, including those that are described as the purposes of TANF [section 401(a) of the Social Security Act]. For example, the effects of most of the nascent strategies that states are using to improve employment retention and advancement are unknown. Furthermore, most of the approaches lend themselves to experimental evaluation of the incremental behavioral effects of these investment strategies. HHS and several states have already begun to use this opportunity to launch a series of experiments that can produce credible information for state and local governments about how to make those investments most effectively.

Also in line with this strategy, in several instances HHS is working with states to use factorial designs to isolate the effects of different policies. The value of this strategy, which Rossi advocates, is substantial. For example, the first final results from one of the five waiver experiments that Rossi discusses, the Minnesota Family Investment Plan, were particularly strong because they included multiple subgroups and three treatments, enabling the researchers to identify which effects derived from which part of the treatment and for which subgroups. The factorial design enabled researchers to conclude that a broad range of positive effects on families and children for long-term single recipients and two-parent families resulted from a generous financial incentive, whereas earnings effects and positive effects on full-time work resulted from work requirements.

Finally, two smaller points. First, Rossi asserts that no findings on child abuse will result from the studies. Although child abuse will not be fully examined, the survey instrument does ask questions about whether a child ever had to be removed from the home, and two of the states will be matching their samples to foster care records.

Second, the experimental analysis in Arizona will not be completed for several reasons, including a near identity of views among experimental and control participants regarding the policies to which they were subject. Although we are pleased that the five Child Impact Waiver experiments are not suffering from this fatal problem, it does serve as a reminder of Rossi's points regarding the need to maintain fidelity of experimental and control groups.

Social policy and program design related to improving the employment prospects of low-income parents, strengthening families and improving child outcomes is at a rare point where enormous opportunities are available not only to *do*, but to *learn* while doing. HHS, along with up to ten states, has launched a major effort to examine the effectiveness of state and local strategies to improve job retention and advancement. We also are in the early stages of projects that would establish similar efforts related to assessing supported work strategies for hard-to-employ welfare recipients and examining the effects of alternative strategies for implementing child care subsidies. A coherent strategy of field experiments can prove to be invaluable in improving state and local governments' efforts to support the advancement of low-income families.

Comments

Judith M. Gueron[*]

Peter Rossi's paper identifies a number of problems with the five-state studies. While he raises reasonable cautions, I would counter that those studies, even though they were launched before 1996, will provide some of the most important information on the causal effects of time limits, incentives, and other elements of TANF programs on children.

Rossi notes that the studies are not going to be that useful because they focus on people who have applied for or are receiving welfare and do not include data on people deterred from applying for welfare or children who were not born as a result of welfare reform. This is correct, but it is important to remember that those applying for or receiving welfare are a key group. If we do not see any effects on children in those families, we are unlikely to find effects on deterred entrants' children.

Rossi also points out that the studies could have a problem if the researchers are not able to maintain a distinction between people subject to the time limits or incentives and the control group that was supposed to continue in the old AFDC world. Again, this is an important issue, but some of the evaluations have done quite well in preserving this distinction.

Thus, overall, there is an enormous amount we can learn from these studies. They represent our best shot at determining causality because they employ powerful, random assignment research designs that allow analysts to avoid some of the problems they will confront using more general data sets.

Many of the most important policy choices that states have made and are going to make are at a more micro-level than one can get from looking at national data, despite Richard Bavier’s masterful job. We all know that TANF is not a program but a funding stream, and states will continue to need to make choices about how to spend those funds. Fortunately, HHS pushed to include measures on child outcomes in a range of studies so that we will be able to address questions about the results of different welfare policies.

So what are some of those questions and issues? First, over the years, MDRC's work suggests that the key story is going to be in the subgroups; the averages are not so important. At a recent Manpower Demonstration Research Corporation meeting, we were reviewing a study in which the authors had focused the presentation on major subgroups and wondered whether readers would think they should, instead, have emphasized the overall average effects. Bob

[*]Judith M. Gueron is president of the Manpower Demonstration Research Corporation.

Solow used a medical analogy to point to the legitimacy of the subgroup focus. If a pharmaceutical company had population-wide information on the results of a drug that proved effective only for the 5 percent of the population that had a specific illness, you would not want to emphasize the overall average impacts. The correct story would be the effectiveness with the sick 5 percent of the population. So subgroups are going to be important, and experimental data are going to be key to estimating impacts for them.

For adults on welfare, one key subgroup will be the most disadvantaged. We will want to look hard at the effects of time limits, incentives, different packages of services in different locations, different packages of treatments on people with mental health issues, domestic violence issues, homelessness, substance abuse, and language barriers. Is there evidence that these programs hurt or help the most disadvantaged groups on welfare? For example, do these programs increase stress or reduce income, leading to negative outcomes for children?

We all know that in welfare reform there are going to be winners and losers, yet this can be lost in the average. For the winners, are we going to see something like the reverse of Bill Wilson’s argument: more employment, a shift in values toward work, different habits for parents and children. Or will we see increased stress on families? We have to get down to this level to understand how welfare reform is playing out and the mechanisms through which it affects children.

A second point is that we will want to understand the effects of the building blocks of TANF programs, because at the state and local level, people are asking what choices they should make in using available TANF funds. For example, how make-work-pay incentive programs, time limits, or programs that are more or less mandatory affect family formation and child well-being? Those kinds of issues are being examined in these studies. The results will be intriguing and should be available in about six months. At this time, all I can say is that it won't be a "no news" story.

A third area is the effect of zero grants and sanctions on families. Are we seeing increases in the frequency of situations that might really affect children, such as homelessness, domestic abuse, or foster care? We are learning a bit from the leaver studies, but we can learn a lot more from these experiments.

Fourth, I was talking with a welfare administrator recently. He said that the next big issue is child welfare. This suggests that an important issue will be whether there a relationship between different welfare reform policies and foster care or the increase in child only cases. Actually, there is a decrease in child-only cases. There is an increase in the percentage of child-only cases, not the absolute number.

Fifth, child care outlays have increased substantially. A senior administrator in one state recently stated that he thought this increase would be the most important effect of welfare reform. When Lorraine Klerman asked whether we were going to know anything about Head Start, I think we will, because we will know about the number of people whose children are in different care arrangements. We have an opportunity to work on that issue.

Finally, *we all should be thinking hard about the lessons for the time in the future when we will not have the economy we have right now.* Right now, the system is awash in money and states can be generous in what they are spending to reach the hard-to-serve and on childcare, incentives, and services. But in the future, presumably, people will have to make choices. It would be marvelous if the situation did not arise, but it will.

Unfortunately, states are not using these flush times to build the database to help with these future choices. The studies underway will provide some of the answers, but new ones should be launched to get at the likely results of key policy alternatives. Doing this will require more microlevel data and analysis, where we can look at the relationship between environment, program features, and different outcomes, and in the process see whether there is a trade-off between results from different welfare programs.