FOR INTERNAL REVIEW ONLY

NOT FOR PUBLIC RELEASE

# Under-Reporting of Medicaid and Welfare in the Current Population Survey

Jacob Alex Klerman, Jeanne S. Ringel, and Beth Roth

Latest Revision: March 19, 2004

Address Correspondence to: Jacob Alex Klerman, RAND, 1700 Main St., Santa Monica, CA 90407, 1-310-393-0411, x289, Jacob_Klerman@rand.org.

This is a draft circulated for discussion purposes only. It has not yet received review by the California HealthCare Foundation (formerly the Medi-Cal Policy Institute) or RAND's internal quality control process. It is not to be publicly released or cited.

The research in this paper was conducted while the authors were Research Associates at the Center for Economic Studies, U.S. Census Bureau. Research results and conclusions expressed are those of the authors and do not necessarily indicate concurrence by the U.S. Census Bureau, the Center for Economic Studies, RAND, or its sponsors including U.S. Department of Health and Human Services, Administration for Children and Families and the Medi-Cal Policy Institute.

# PREFACE

Conventional estimates of the number of uninsured Californians are derived from the Current Population Survey (CPS). Unfortunately, CPS estimates of the number of people receiving Medi-Cal and welfare (AFDC/CalWORKs) are well below the numbers implied by official Medi-Cal records, suggesting that the conventional estimates of the number of uninsured Californians (and their characteristics) are seriously flawed.

To improve our understanding of these issues, the California HealthCare Foundation (through its then separate the Medi-Cal Policy Institute—MCPI) and the U.S. Department of Health and Human Services, Administration for Children and Families (DHHS-ACF) funded RAND to match CPS data to individual-level administrative data for the Medi-Cal program. With the cooperation of the California Department of Health Services (CDHS), the U.S. Bureau of the Census, and the California Census Research Data Center (CCRDC), that match was performed. This document describes the findings of the analysis of those matched data.

# TABLE OF CONTENTS

## LIST OF TABLES

# LIST OF FIGURES

## STATEMENT OF BENEFIT TO THE BUREAU

In what follows, the materials in italics are the "Statement of Benefit to the Bureau" from our original proposal. The nonitalicized materials are additional details and benefits based on our actual analysis and findings.

> *The proposed project relates directly to the Census Bureau's core mission to improve the quality of its data. These analyses will be informative about the quality of Census Bureau survey data on welfare participation and Medicaid. This study will identify factors associated with misreporting of program participation in the CPS. This information could be used in a variety of ways. Although the results will be based on California data, we believe that they will provide insights that will be applicable nationally. Using the estimates from California, national participation estimates could be adjusted based on the composition of the population (i.e., age, race/ethnicity) and characteristics of the programs (i.e., welfare program type—child-only, 1-parent, and 2-parent). Along the same lines, the results of our analyses for California will be used to improve the imputation procedures for non-response on program participation questions. We believe that this would be a very important contribution of the proposed work. Funding from DHHS-ACF, suggests that they also believe that this analysis is of considerable potential policy relevance.*

Increasing the benefit to the Bureau, we further note that concerns about the quality of the data are (correctly) pervasive. The Census Bureau's own technical notes, accompanying reports using these data (discussed in the body of this report) explicitly note that the survey data are at radical variance with aggregate counts from administrative data.

The tabulations included in this request for release address this concern directly. We first describe the procedures we used to match the two data sets (the Current Population Survey, CPS, and the Medi-Cal Eligibility Data System, MEDS, administrative data from California) and then the magnitude of the discrepancy in reported program participation between them. These tabulations concern Medi-Cal and its two components—welfare and Medi-Cal only. Unfortunately, sample sizes do not appear to be large enough to support additional stratification by Medicaid subprogram. However, we are able to present some tabulations by year.

In addition, the CPS currently includes questions on the number of months of participation in welfare and Medicaid. We explicitly compare those responses to the official data. We conclude that while there is some correlation between the CPS responses and the administrative data, the correlation is weak. People appear to be unable to respond correctly to this question. If we truly care about sub-annual estimates, we need a sub-annual survey. We have such a survey—the Survey of Income and Program Participation (SIPP). Therefore, in as much as interview time is a binding constraint (i.e., there are other questions that might be included), we believe that the Census Bureau should consider dropping these questions on the number of months of participation in welfare and Medicaid.

We are aware that the Census Bureau has revised the health insurance-related questions to improve the quality of the data. Unfortunately, despite our requests, the Census Bureau was unable to make available to our project either data for current years (2001, 2002, or 2003) or a consistent time series. (Some of our data are "validated" and some are "unvalidated.") As a result, our data set appears to be too small to be informative about the effect of those changes in the questionnaire.

> *In addition, the proposed analysis of the effects of the CPS reference period will provide valuable information about how people interpret the questions and whether refinements to the survey instruments are warranted. Not only could these findings be used to improve the CPS questions, but the findings could also be used to guide the design of future surveys that seek to elicit information on program participation.*

This request includes the tabulations on the reference period. As expected, they are informative on the reference period used by the CPS respondents. This is a crucial issue for interpreting CPS responses and for potentially redesigning the survey.

> *Finally, the CPS questionnaire and imputation procedures have been redesigned several times in an attempt to improve the accuracy of the survey-based estimates. In as much as our data cover the periods before and after these changes, we can and will explore the extent to which they appear to have improved the quality of the data. Of course, we would require more recent data to analyze the most recent questionnaire changes.*

We present limited results on this issue. They suggest that the CPS imputation procedures (known as "hot-decking") are biased and are biased in a plausible direction. However, imputation is sufficiently rare so that this is probably not an issue worthy of major attention by the Bureau.

*Overall, we believe that the analyses that we have proposed will be of great value to the Census Bureau and more broadly to the policymaking community. The estimates of program participation from the March CPS are widely cited and are used by many researchers to track changes and to evaluate the effectiveness of policy changes. It is important that decisions about future policy changes be made on solid evidence, and improving the quality of the estimates of program participation is an important step towards this important goal.*

# EXECUTIVE SUMMARY/POLICY BRIEF

High-quality survey data are crucial to our understanding of the effects of the Medi-Cal program in California, and the nation's social welfare system more broadly. We can tabulate the number of people enrolled in Medi-Cal from the official program records, the Medi-Cal Eligibility Data System (MEDS). However, beyond enrollment counts, understanding Medi-Cal's effects often requires survey data because information is needed on both enrollees and non-enrollees. For example, to assess take-up rates we need to know the number of people enrolled as well as the number of people who are eligible for the program. Similarly, if we want to look at take-up by sub-group, we need more detailed information about the characteristics (e.g., family structure, household income) of enrollees and non-enrollees. Along the same lines, if we are interested in assessing overall levels of health insurance coverage, we need information on the full population (enrollees and non-enrollees) and their private health insurance coverage. This type of information is not available in administrative data, which highlights the importance of high-quality survey data.

## THE CURRENT POPULATION SURVEY (CPS), UNDER-REPORTING, AND MATCHING

The U.S. Bureau of the Census's March Annual Demographic Survey (ADS) to the Current Population Survey (CPS) is the standard data source for analyses of the Medi-Cal program and the nation's social welfare system more broadly. The CPS is a large (about 50,000 households nationally, 6,000 households in California), household survey with information on program participation (including Medicaid/Medi-Cal and welfare), health insurance coverage, and other household characteristics. Two other features of the CPS data are crucial for policy analyses: (1) The ADS data are collected annually in a relatively consistent manner back to the late 1980s—allowing trend and time series analyses; and (2) The data are released promptly—results of the interviews conducted in March are publicly released in late-August or early-September of the same year—allowing nearly real-time tracking of changes.

Unfortunately, the CPS is known to under-report program participation, including Medi-Cal. The official CPS report notes the problem explicitly:

> The Current Population Survey (CPS) underreports medicare [stet] and medicaid [stet] coverage compared with enrollment and participation rates from the Centers for Medicare and Medicaid Services (CMS), formerly the Health Care Financing Administration. A major reason for the lower CPS estimates is that the CPS is not designed to collect health insurance data; instead, it is largely a labor force survey. Consequently, interviewers receive less training on health insurance concepts. Additionally, many people may not be aware that they or their children are covered by a health insurance program if they have not used covered services recently and therefore fail to report coverage. CMS data, on the other hand, represent the actual number of people (who) enrolled or participated in these programs and are a more accurate source of coverage levels.

Furthermore, some analyses suggest that the problem has gotten worse over time.

As we will show below, the under-reporting is substantial, but neither its causes, nor its effects, are well understood. Therefore, with funding from the Medi-Cal Policy Institute and the U.S. Department of Health and Human Services, Administration for Children and Families and the cooperation of the U.S. Bureau of the Census and the California Department of Health Services (CDHS), we matched individual-level CPS responses to their corresponding MEDS administrative data records. Specifically, as part of its interview, the CPS attempts to collect Social Security Numbers (SSNs) on all respondents age 15 and older. The MEDS data include SSNs for each enrollee. For this project, the Census Bureau supplied a version of the CPS data for 1990 to 2000 that included a scrambled version of the SSN, where available. In addition, the Census Bureau processed a version of the MEDS data for 1989 to 2001 replacing the original SSNs with the same scrambled SSNs. Where possible, we then matched the two files creating a single analysis file with both CPS and MEDS data. To preserve the confidentiality of CPS respondents and Medi-Cal enrollees, the data analysis took place at the UCLA site of the Secure Data Facility of the Census Bureau’s California Census Research Data Center. The authors had no access to identifiers (names or Social Security Numbers) and all research results were reviewed to assure that they did not indirectly reveal the identity of or information about CPS respondents or Medi-Cal enrollees.

**THE MAGNITUDE OF UNDER-REPORTING AND OUR IMPUTATION MODEL**

How serious is the problem of under-reporting? Previous analyses of this question using unmatched data have been limited by the inconsistencies between the two data sources. The CPS, administered in March, asks about program enrollment *at any time* in the last calendar year (i.e., the 2000 CPS asks about program participation between January and December 1999). Aggregate Medi-Cal data is usually reported in terms of persons covered per month. The extent to which discrepancies in aggregate counts based on unmatched data were real as opposed to being an artifact of different data concepts has therefore been unclear. Given the structure of our matched data, we can tabulate the individual level Medi-Cal data from MEDS to be consistent with the CPS questions and thus provide a better estimate of under-reporting in the CPS.

Figure ES.1 summarizes that analysis. It considers two age groups (adults—15-65 at the interview, and children—0-14 at the interview) and three program concepts: cash assistance/welfare (W—Welfare), and non-cash assistance Medi-Cal (MO—Medi-Cal Only), and all Medi-Cal (M)—the sum of the other two. Averaged over the entire period, CPS estimates of total Medi-Cal enrollment for adults are only 88 percent of the counts from the official MEDS administrative data, i.e., Medi-Cal is under-reported by about 12 percent. For children, reporting of Medi-Cal is even worse, about 70 percent. Unlike some national estimates, there is little evidence of a decline in reporting over time.

This overall pattern in Medi-Cal hides a strong divergence by Medi-Cal sub-program. Enrollment in welfare is severely under-reported, over the entire time period about 47 percent for adults and 45 percent for children. For welfare, there is clear evidence of a sharp drop in reporting rates over time. The timing of the drop (in the late 1990s) is nearly simultaneous with the implementation of welfare reform in California (i.e., CalWORKs), perhaps suggesting an increase in the stigma of welfare participation.

At the same time, reporting rates for Medi-Cal Only are much higher, 88 percent for adults and 113 percent for children (i.e., the CPS estimates for children are higher than the administrative data counts). Further analyses suggest that the results for welfare and Medi-Cal only are related. Many people with welfare report Medi-Cal, but not welfare. The net result is under-reporting of welfare and higher reporting rates (sometimes over-reporting) of Medi-Cal Only.

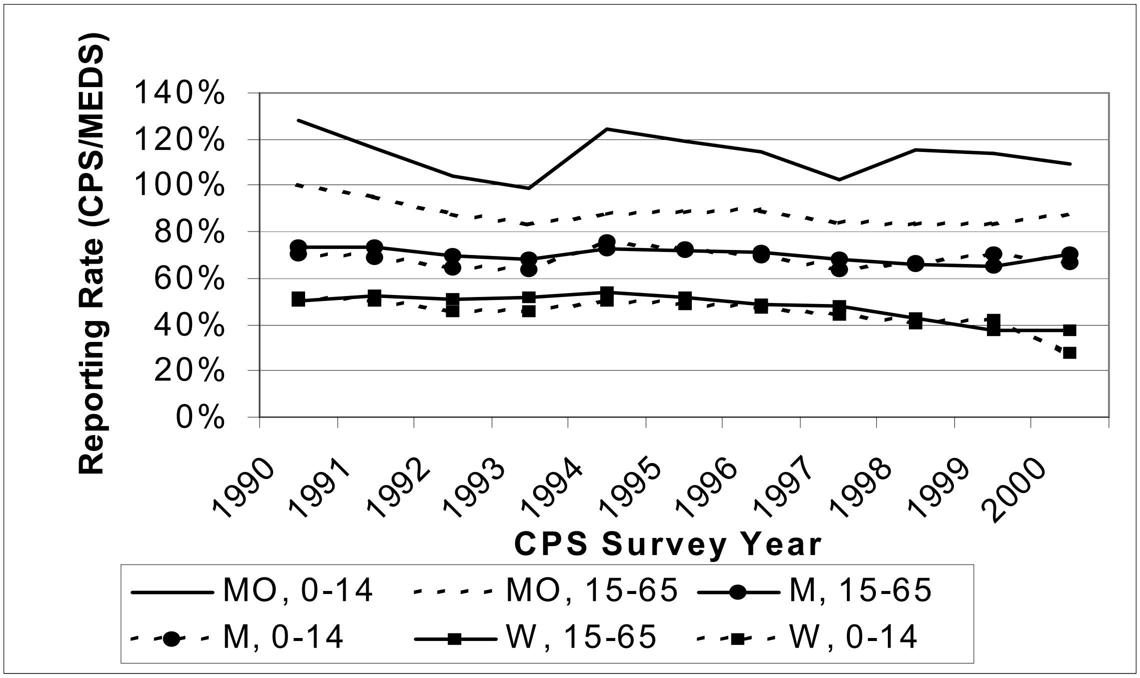


**Figure ES.1—Reporting Rates (CPS relative to MEDS) by Age and Program**
Source: Tabulations from RAND Merged MEDS File

This under-reporting is severe enough to have substantively important effects on our understanding of the effects of the Medi-Cal program. Here, we consider two effects. First, under-reporting will lead us to under-estimate take-up rates (the fraction of eligibles enrolled in the program) and thus to over-estimate the need for efforts to increase enrollment or new programs to provide additional coverage. Second, under-reporting will lead us to over-estimate the total number of uninsured people.

Our analysis proceeds as follows. For those providing a SSN, we overwrite the CPS Medi-Cal responses with the official information from the Medi-Cal administrative data (i.e., treating the MEDS information as the truth). However, our ability to match the survey and administrative data is constrained by the fact that only about 57 percent of CPS adults provide a SSN. Furthermore, children under 15 were never asked for a SSN. To address this problem, we build an imputation model to predict mis-reporting among those people without a SSN who we cannot match to the MEDS data. The response errors (i.e., reporting no Medi-Cal in the CPS given actually having Medi-Cal and reporting having Medi-Cal in the CPS given not actually having Medi-Cal) among those not providing a SSN are assumed to follow the general pattern in the sub-sample who do provide a SSN, with an adjustment to force the totals to align exactly (see the full report for details). The problem is more pronounced for children since SSNs are not collected in the CPS for people under age 15. To address this issue we use a combination of information from the head of household and our imputation model. Specifically, where the head of the household provides a SSN (as is true for about 61 percent of CPS children), we use the head's Medi-Cal status (from the MEDS or from our imputation) to impute Medi-Cal status to the child. Some Medi-Cal programs include children, but not adults. Therefore, in cases where the child has Medi-Cal, but the head of household does not, the child's data are not changed. Again, as with adults, the imputation includes an adjustment to force the CPS totals (after imputation) to align exactly to the MEDS counts (again, see the full report for details).

These imputations are performed for every observation in the CPS. The resulting individual level file allows us to construct improved estimates of take-up rates and uninsurance coverage. Using the individual-level imputation file, we can consider the effects of under-reporting by respondent characteristics (e.g., gender, age, income).

**THE EFFECTS OF UNDER-REPORTING ON ESTIMATES OF MEDI-CAL TAKE-UP**

If Medi-Cal enrollment is under-reported, then Medi-Cal take-up will also be under-reported. We have seen that the under-reporting of Medi-Cal is substantial, about 12 percent for adults and 30 percent for children. Our analyses of the matched file suggest that the under-reporting is also not uniform across sub-groups of the population, so the effects of under-reporting on take-up rates are also not uniform.

Table ES.1 summarizes our findings on Medi-Cal take-up. For our most recent data (calendar year 1999/survey year 2000), we estimate a raw (i.e., not adjusted for under-reporting) take-up rate (i.e., the fraction of the full population enrolled in any type of Medi-Cal) of 9 percent among adults and 24 percent for children. Accounting for under-reporting increases our estimated take-up rates by nearly half to 13 percent for adults and 35 percent for children.

**Table ES.1**
**Take-Up Rates: Unadjusted, Adjusted, Discrepancy**
**Medi-Cal, for Calendar Year 1999/Survey Year 2000**

| | Adults | | | Children | | |
|---|---|---|---|---|---|---|
| | **Raw** | **Imputed** | **Delta** | **Raw** | **Imputed** | **Delta** |
| All | 9% | 13% | 42% | 24% | 35% | 50% |
| Male | 7% | 9% | 37% | 24% | 35% | 49% |
| Female | 11% | 17% | 45% | 23% | 35% | 51% |
| SW w/kids | 25% | 36% | 46% | 23% | 35% | 51% |
| SW w/kids <50% FPL | 49% | 59% | 21% | 47% | 65% | 38% |
| SW w/kids 50%-100% FPL | 48% | 63% | 31% | 52% | 67% | 30% |
| SW w/kids 100%-150% FPL | 41% | 55% | 34% | 37% | 51% | 38% |
| SW w/kids 150%-200% FPL | 24% | 45% | 84% | 22% | 54% | 139% |
| SW w/kids >200% FPL | 8% | 16% | 104% | 7% | 13% | 72% |

NOTE: "Raw" is the unadjusted CPS estimate; "Imputed" is the adjusted CPS estimate, based on the multiply-imputed data set; and "Delta" is the percentage (not percentage point) increase in estimated take-up with imputation; "SW" is an abbreviation for single women.

As noted above, the effects of adjusting for under-reporting are not uniform across sub-groups of the population. They are relatively small for adult men, who are rarely eligible for Medi-Cal. They are larger for adult women as a group. Perhaps because of varying stigma of reporting Medi-Cal participation, the adjustments are smallest for female-headed families with children at or near the poverty line and larger for families with income just above the poverty line. This larger under-reporting for families with income just above the poverty line is particularly important because they have been the target of policy initiatives (e.g., the Section 1931(b) program and the proposed, but never implemented, expansion of Healthy Families to adults).

As expected, for children, there is little divergence in adjustment by gender. The patterns for children living in female-headed families by income are similar to those for adults, lower under-reporting for those below poverty, more under-reporting for those just above poverty. Again, differential stigma is a likely explanation. The divergence is particularly policy relevant because this group (near poor children) with the largest under-reporting has been the target of recent policy initiatives (e.g., the Medicaid Expansions, the Section 1931(b) program, Healthy Families).

**THE EFFECTS OF UNDER-REPORTING ON ESTIMATES OF UNINSURANCE**

Another consequence of under-reporting of Medi-Cal enrollment is that it will lead to over-estimates of the rates of uninsurance in the CPS. The magnitude of the over-estimate will depend on the extent to which those under-reporting have other sources of health insurance at some point during the year. If it were the case that everyone who under-reports Medi-Cal did not have any other source of insurance, then we could construct a better estimate of the number of uninsured by subtracting the estimate of under-reporting (i.e., the percent of people in the CPS who report no Medi-Cal, but who our imputation model, based on the matched data, suggests are enrolled) from the raw estimate of the percent of people who are uninsured in the CPS. Conversely, if it were the case that everyone who under-reports Medi-Cal also has private health insurance, then under-reporting would have no effect on the estimates of the uninsured. Our analyses suggest that the truth lies somewhere between these two extremes. Plausibly, we find that under-reporting is more common among those with private health insurance, but under-reporting also includes large numbers of people without private health insurance.

From our matched file, we tabulate rates of other health insurance among people who under-report Medi-Cal. Here we report adjusted estimates of uninsurance based on several different scenarios.

Table ES.2 summarizes our findings for rates of uninsurance. We estimate that under-reporting is about 4.1 percent for adults (i.e., 4.1 percent have Medi-Cal but do not report it to the CPS). Consistent with much higher rates of Medi-Cal coverage for children, the corresponding rate of under-reporting is much higher. We estimate that 11.8 percent of all children have Medi-Cal, but do not report it. This result is shown in the first row of the table. The second row of the table reports the unadjusted, or raw, estimate of uninsurance, 23.2 percent for adults, slightly lower, 21.9 percent, for children.

**Table ES.2**
**Take-Up Rates: Unadjusted, Adjusted, Discrepancy**
**Medi-Cal, for Calendar Year 1999/Survey Year 2000**

| Correction for Under-Reporting and Dual Coverage (DC) | Adults | | Children | |
|---|---|---|---|---|
| | **DC** | **UI** | **DC** | **UI** |
| A: Under-reported | | 4.1% | | 11.8% |
| B: Unadjusted, or raw | | 23.2% | | 21.9% |
| C: Assuming no dual coverage (B-A) | 0.0% | 19.1% | 0.0% | 10.1% |
| D: Dual coverage based on all Medi-Cal | 27.6% | 20.2% | 20.3% | 12.5% |
| E: Dual coverage based on False Negatives | 31.2% | 20.4% | 40.2% | 14.9% |
| F: Full imputation model | 34.4% | 20.5% | 45.0% | 15.4% |

Note: "DC"—Dual Coverage, UI—Percent Uninsured

Under the first scenario that we outlined above, in which none of the people who under-report Medi-Cal have any other source of insurance, we can subtract the under-reporting percentage from the raw estimate of uninsurance to obtain an adjusted estimate. This estimate is presented in the third row of the table. The results in row C show that if we assume no dual coverage, the rates of uninsurance drop dramatically: For adults from 23.2 percent to 19.1 percent, for children from 21.9 percent to 10.1 percent.

The next three rows of the table consider alternative methods for adjusting estimates for the magnitude of dual coverage. The results in row D use the information in the public use CPS file to estimate the prevalence of dual coverage (i.e., the percent of Medi-Cal enrollees who also report having private insurance) and use this information to break out the portion of under-reporters who have dual coverage. In the CPS public use file, about 27.6 percent of adults and 20.3 percent of children who report Medi-Cal also report private health insurance. The results in row E use that estimate of dual coverage to estimate uninsurance. As expected, accounting for some dual coverage increases the estimate of uninsurance above that when no dual coverage is assumed.

Because we have matched data, we can do a more sophisticated adjustment. Our population of interest is not everyone enrolled in Medi-Cal, but those enrolled in Medi-Cal who do not report this in the CPS (call them false negatives). We can identify those people in our matched sample and look at dual coverage among this group. Consistent with a stigma explanation, this group is much more likely to report having other health insurance (31.2 percent for adults; 20.3 percent for children) than everyone in the CPS who reports Medi-Cal enrollment. Therefore, when we use the matched sample to account for dual coverage based on the false negatives, rather than the full sample of Medi-Cal enrollees, we see that the estimated rate of uninsurance increases for both adults and children (relative to line C of Table ES.2).

The final row (row F) of the table presents our preferred estimates. They are based on our full imputation model and include a correction for those who spuriously report Medi-Cal, when the administrative data do not report Medi-Cal. Our analyses suggest that this group includes a large group of people who gained Medi-Cal between the end of the CPS reference year in December and the CPS interview in March. The full model suggests even higher rates of dual coverage (34.4 percent of adults and 45.0 percent for children).

Our preferred estimates are thus closer to the unadjusted estimate than to the simpler estimates. The adjustment is, however large. For calendar year 1999/interview year 2000, the simple estimate for adults is 23.2 percent. Our preferred estimate is 12 percent lower, 20.5 percent. For children, the unadjusted estimate is 21.9 percent; our preferred estimate is 15.4 percent, 30 percent lower.

**SUMMARY**

This Policy Brief has considered the quality of Medi-Cal information in the Current Population Survey, the standard data source for tabulations of Medi-Cal take-up and levels of uninsurance. The analyses are based on an imputation model derived from a match of individual-level survey data with individual-level administrative data for the Medi-Cal program. We find sizable under-reporting of Medi-Cal, leading to sizable under-estimates of Medi-Cal take-up and sizable over-estimates of the fraction of Californians who are uninsured. These results cover the period 1990 to 2000. The Census Bureau made some adjustments to the CPS interview towards the end of this period. Nevertheless, these results suggest caution in basing policy on unadjusted analyses of the CPS data. Analyses based on unadjusted data are likely to substantially overestimate the magnitude of the problem, especially for children.

## ACKNOWLEDGMENTS

Funding for this analysis was provided by the California HealthCare Foundation (CHCF; through its then separate Medi-Cal Policy Institute—MCPI) and the U.S. Department of Health and Human Services, Administration for Children and Families (DHHS-ACF). The project officers at CHCF/MCPI—Ingrid Aguirre Happoldt, and at DHHS-ACF—Audrey Mirsky-Ashby, Laura Chadwick, and Leonard Sternbach—have been waiting patiently for these results and we appreciate their interest and consideration.

This analysis is based on a unique dataset constructed by matching confidential Census Current Population Survey data to confidential administrative data on the Medi-Cal program. Doing so has required the cooperation of several groups. Gene Hiehle and the California Department of Health Services provided the Medi-Cal administrative data and have been supportive throughout this project. B.K. Atrostic and the U.S. Bureau of the Census's Center for Economic Studies provided the Current Population Survey data and handled the matching and deidentification tasks. Senior leadership of the California Census Research Data Center, especially, V. Joseph Hotz at UCLA and Andrew Hildreth at UC Berkeley, have provided crucial support during the negotiations. Center staff, especially Nelson Lim and Becky Acosta, have provided guidance and support with writing the proposal, using the Center, and doing the analyses. The contribution of each of these groups has been necessary in order to gain access to the data.

This research is an outgrowth of work begun under the RAND Statewide CalWORKs Evaluation. The constructive comments and guidance provided by CDSS employees during that effort has benefited this effort greatly (though they have not always agreed with our findings). They include Werner Schink, Lois van Beers, Nikki Baumrind, Wilistine Sayas, Aris St. James, and Paul Smilanick.

The data analysis for this project is based on the data preparation work of programmers in RAND's Research Programming Group under the direction of Jan Hanley, who led the effort and did much of the data preparation work herself. Beth Roth, an author on this report, is a member of that group. Other programmers involved in the effort included Christine DeMartini, Laurie McDonald, and Deborah Wesley.

Finally, at RAND this work proceeded within the Labor and Population Program's Center for the Study of Social Welfare Policy. Further information about RAND, the Labor and Population program, and the Center for the Study of Social Welfare Policy can be found at /www.rand.org, /www.rand.org/labor, and /www.rand.org/socialwelfare. The strong support of RAND, the Labor and Population Program, and its current and former Directors, Arie Kapteyn and Lynn Karoly (respectively), and Assistant Director Rebecca Kilburn, for this effort is gratefully acknowledged. Within RAND, this report has also benefited from the secretarial support of Christopher Dirks and Natasha Kostan. Finally <fill in the names> provided technical review that has improved the final product.

# GLOSSARY, LIST OF SYMBOLS, ETC.

| Symbol | Definition |
|---|---|
| ADS | CPS March Annual Demographic Survey |
| AFDC | Aid to Families with Dependent Children |
| CalWORKs | California Work Opportunities and Responsibility to Kids Act (1997) |
| CRDC | Census Research Data Center |
| CCRDC | California Census Research Data Center |
| CDHS | California Department of Health Services |
| CDSS | California Department of Social Services |
| CHCF | California HealthCare Foundation |
| CHIP | Child Health Insurance Programs (established by statute in 1997, operated by the states) |
| CMS | Centers for Medicare and Medicaid Services |
| CPS | Current Population Survey |
| ESHI | Employer Sponsored Health Insurance |
| HCFA | Health Care Financing Administration |
| MEDS | Medi-Cal Eligibility Data System |
| MCPI | Medi-Cal Policy Institute |
| NIPA | National Income and Product Accounts |
| OHI | Other Health Insurance |
| PIK | Person Identification Key |
| SIPP | Survey of Income and Program Participation |
| SSA | Social Security Administration |
| SSI | Supplemental Security Income |
| SSN | Social Security Number |
| TANF | Temporary Assistance to Needy Families |

# 1. INTRODUCTION

High-quality survey data are crucial to our understanding of the effects of the nation's social welfare system. If all one wants to know is the number of people participating in a program, then that information can be obtained from administrative data. However, very often, both researchers and policymakers want to know take-up rates (i.e., the fraction of people enrolled in the program) and the effects of the program on subsequent outcomes (e.g., probability of lacking any health insurance, probability of living in poverty, etc.). For these outcomes, we need richer data that can only be gleaned from surveys; in particular, we need: (1) information on the number and characteristics of nonparticipants; and (2) information on participating families not recorded in administrative data.

Unfortunately, there is considerable evidence that the quality of existing survey data on program participation is poor. There are indications that survey data significantly under-report participation in safety-net programs relative to aggregate administrative counts and also that the under-reporting has increased over time. However, most of the evidence to date is based on comparisons between aggregate administrative counts and estimates from survey data. It is our belief that a better understanding of the nature and scope of under-reporting can be obtained by comparing administrative and survey data at the individual level and that is what we seek to do in this report.

This document reports the results of a record-match study of individual-level administrative data for Medi-Cal—the Medicaid program in California, and the Current Population Survey (CPS). With funding from the California HealthCare Foundation (CHCF; through its then separate Medi-Cal Policy Institute—MCPI) and the U.S. Department of Health and Human Services, Administration for Children and Families (DHHS-ACF), and the cooperation of the U.S. Bureau of the Census, the California Department of Health Services (CDHS), and the California Census Research Data Center (CCRDC), we matched administrative data for Medi-Cal from the Medi-Cal Eligibility Data System, (MEDS) to March CPS data for 1990 to 2000. In California, everyone receiving cash assistance (sometimes referred to as welfare)——through Aid to Families with Dependent Children (AFDC), later Temporary Assistance to Needy families

(TANF)/California Work Opportunities and Responsibility to Kids (CalWORKs)—is automatically enrolled in Medi-Cal. Since the MEDS administrative data allow us to identify the “type” of Medi-Cal coverage (i.e., why the person is eligible for Medi-Cal), we are able to consider overall Medi-Cal coverage and its two components—welfare and Medi-Cal only (i.e., Medi-Cal, but not welfare)——in our analysis.

**PLAN OF THE REPORT**

This report proceeds as follows. The balance of this opening chapter reviews the existing literature on the quality of the CPS data on Medicaid and welfare. The second chapter provides background information on the Medicaid/Medi-Cal program, the MEDS (administrative) data, and the CPS (survey) data. It then characterizes the under-reporting problem, using separate tabulations from each data source. In the third chapter, we turn to the matched data file. For the subset of individuals who provide a valid Social Security Number (SSN), we describe the nature of reporting biases based on a one-to-one match of the survey and administrative data. Unfortunately, not all survey respondents provide a SSN. The fourth chapter provides a technical discussion of our methods for using information from the matched data to impute welfare and Medi-Cal for the entire California CPS sample. In the fifth chapter, we use the resulting multiply-imputed file to reconsider some of the substantive issues for which the CPS is used. In particular, we explore program take-up by (reported) household income and family structure and levels of uninsurance. The final chapter considers the implications of the results.

**PREVIOUS LITERATURE ON UNDER-REPORTING**

The conventional source for information on program take-up is the CPS, the largest annual, national survey. Beginning with the March 1995 CPS, the Census Bureau (Benenfield, 1996a), the Congressional Budget Office (Bilheimer, 1997), General Accounting Office (1997), and the Employee Benefits Research Institute (Fronstin, 1996) each publish annual CPS-based estimates of health insurance coverage and uninsurance. However, the CPS-based estimates of health insurance coverage are much lower and estimates of uninsurance much higher than tabulations from other surveys, such as Survey of Income and

Program Participation (SIPP) or the National Survey of America's Families (NSAF) (Bennefield, 1998; Lewis, Ellwood, Czajka, 1998; Fronstin, 2000).[1]

In addition and of particular relevance to this study, CPS estimates of Medicaid coverage (Medi-Cal in California) are much lower than corresponding tabulations from administrative data on Medicaid (and Medi-Cal in California), suggesting that survey respondents under-report Medicaid/Medi-Cal coverage. The Urban Institute's TRIM2 model (used by DHHS to simulate program costs) based estimates of the number of uninsured on administrative data from Health Care Financing Administration/ Centers for Medicare and Medicaid Services (HCFA/CMS) to partially correct for such under-reporting. For 1995, this correction for underreporting lowers the fraction of children (0-17) uninsured by 31 percent and the fraction of all non-elderly individuals (0-65) uninsured by 11 percent.[2]

As part of a discussion of the decline in Medicaid coverage, Ku and Bruen (1999) summarize the national issues and their effect on our understanding of policy.

1) "CPS data indicate that about 2.5 million fewer non-elderly people got Medicaid in 1997 than in 1995 (9.3 percent fewer), while administrative data indicate that 1.2 million (3.2 percent) lost Medicaid."

2) "CPS data indicate that more children lost coverage than adults from 1995 to 1997, while administrative data indicate [that] the declines were larger for adults."

3) "[T]he total number of nonelderly people who had Medicaid at any time in a given year was about 25 to 30 percent lower in the CPS than in administrative counts."

4) "[T]here appears to be a growing discrepancy between CPS and administrative data concerning the receipt of benefits like Medicaid, welfare, and food stamps in recent years. . . . Using measures of enrollment during the year, the CPS Medicaid participation estimates were 75 percent of administrative counts in 1995, but fell to 70 percent in 1997."[3]

---

[1] Other papers focusing on question wording for health insurance items include Rajan et al. (2000), and Nelson and Mills (2001).

[2] For similar comments about welfare, see http://www.census.gov/hhes/www/income/assess1.pdf.

[3] Furthermore, concern about the problem has increased. See, for example, Levit et al. (1992, pp. 45-46), reflecting minimal concern about undercounting. "CPS counts of people covered by Medicare and Medicaid programs are reasonably consistent with Health Care Financing Administration (HCFA) program data after allowing for the institutional component missing

5) "Some believe that respondents to the CPS may be reporting their current insurance status, rather than answering the actual question about insurance at any time in the prior year."

Such reporting biases would cause over-estimates of the number of uninsured Americans and, thus, of the demand for the programs being created by new policy initiatives. Lower than expected enrollment has in fact been a problem (Alpha Center, 2000). While simple reporting bias is unlikely to explain all the lower than expected enrollment, such reporting bias has explicitly been cited by some observers (e.g., Alpha Center, 2000).

The problem of under-reporting is perceived to be so severe that the official U.S. Bureau of the Census report on health Insurance (P60-220, 2002) notes it explicitly and at length in its "Technical Note":

> The Current Population Survey (CPS) underreports medicare [stet] and medicaid [stet] coverage compared with enrollment and participation rates from the Centers for Medicare and Medicaid Services (CMS), formerly the Health Care Financing Administration. A major reason for the lower CPS estimates is that the CPS is not designed to collect health insurance data; instead, it is largely a labor force survey. Consequently, interviewers receive less training on health insurance concepts. Additionally, many people may not be aware that they or their children are covered by a health insurance program if they have not used covered services recently and therefore fail to report coverage. CMS data, on the other hand, represent the actual number of people (who) enrolled or participated in these programs and are a more accurate source of coverage levels.

The problem of under-reporting appears to be particularly severe for welfare. Welfare recipients are categorically eligible for Medicaid. In fact, the CPS imputes Medicaid to anyone who reports receiving welfare. However, welfare reform appears to have worsened reporting of welfare in the CPS, perhaps because of confusion over program names, perhaps because of increased stigma of welfare receipt.

---

from CPS." They compare the 1991 CPS estimate (for 1990) of 24.3 million persons to the HCFA Medicaid program estimate of 25.3 persons. They attribute the difference (only about 4 percent) to "the institutionalized population not included in CPS and difficulties that surveys have capturing Medicaid recipients." They note that estimates of change over time (in particular 1980 to 1991) are quite similar across the CPS and HCFA data.

See Fronstin (1997), HCFA (1996), and Lewis, Ellwood, and Czjaka (1998) for claims that Medicaid under-reporting has increased.

## 2. MEDI-CAL, THE CPS, THE MEDS, AND UNDER-REPORTING

The core of this project is a data match between administrative data for California's Medicaid program—Medi-Cal (i.e., the MEDS data)——and CPS data. This section begins with a brief description of the Medi-Cal program. It then describes the administrative data (the MEDS) and the survey data (the CPS). Finally, we provide some simple tabulations using the *unmatched* data.

### THE MEDI-CAL PROGRAM

Since 1965, Medicaid——a joint federal-state program—has provided health insurance to current welfare recipients and some other qualifying families. During the 1980s and 1990s, coverage was significantly expanded, with particular attention to poor children (often referred to as "the percent programs") and families that are welfare-eligible, whether they are actually on welfare or not (the 1931(b) program).[4] California's Medicaid program, Medi-Cal, is a joint effort of the California Department of Health Services (CDHS), which administers the program and handles payments, and the California Department of Social Services (CDSS), which supervises county welfare departments that handle enrollment and re-enrollment.[5]

Figure 2.1 (and Table 2.1) shows Medi-Cal enrollment from the MEDS data (described below) according to the CPS concepts we will use in our main analysis. In particular, we tabulate the total number of individuals enrolled at any time in the calendar year. We distinguish welfare from other Medi-Cal (Medi-Cal only, or simply "MO"). Finally, we consider only the non-elderly, in two groups: those 0-14 as of the following March (who we refer to as "Children") and those 15-65 as of the following March (who we refer to as "Adults"; we discuss the reason for this child/adult break at 14/15 below).

[4] For more discussion of these eligibility changes and their effects, see Gruber (2000).

[5] For more information on Medi-Cal and its multiple programs see: http://www.medi-cal.org/ and its fact sheet: http://www.medi-cal.org/resources/view.cfm?section=Resources&itemID=1397. For more information on the administration of Medi-Cal, see Klerman and Cox (2003).

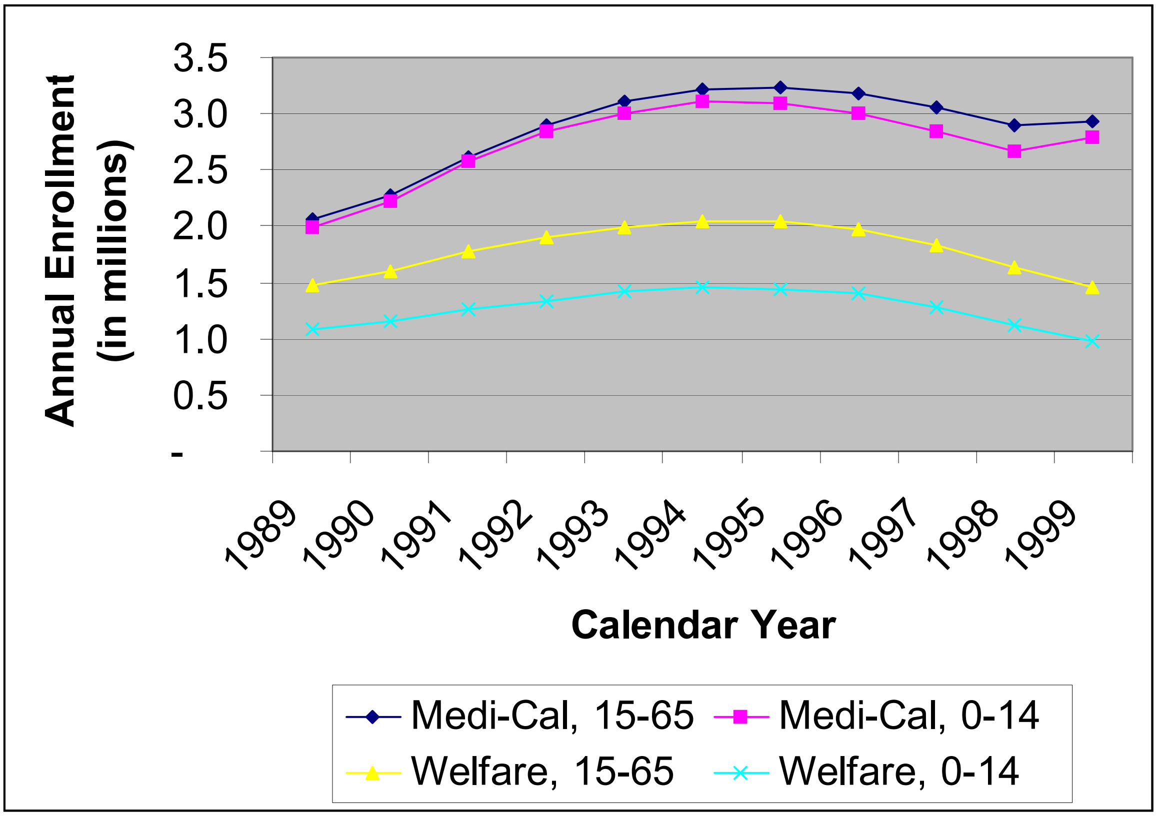


**Figure 2.1—Medi-Cal Enrollment, in California, by Age, Welfare and Total**

Source: Tabulations from RAND Merged MEDS File

**Table 2.1**

**Medi-Cal Enrollment in California (millions of persons)**

| | **Adults** | | | **Children** | | |
|---|---|---|---|---|---|---|
| **Year** | **M** | **W** | **MO** | **M** | **W** | **MO** |
| 1989 | 2.06 | 1.47 | 0.59 | 1.98 | 1.08 | 0.90 |
| 1990 | 2.28 | 1.60 | 0.68 | 2.22 | 1.15 | 1.07 |
| 1991 | 2.62 | 1.78 | 0.84 | 2.57 | 1.26 | 1.31 |
| 1992 | 2.89 | 1.90 | 0.99 | 2.84 | 1.33 | 1.51 |
| 1993 | 3.10 | 1.99 | 1.11 | 3.01 | 1.41 | 1.60 |
| 1994 | 3.21 | 2.05 | 1.16 | 3.10 | 1.46 | 1.65 |
| 1995 | 3.23 | 2.05 | 1.18 | 3.10 | 1.44 | 1.66 |
| 1996 | 3.18 | 1.98 | 1.20 | 3.00 | 1.40 | 1.60 |
| 1997 | 3.05 | 1.83 | 1.22 | 2.84 | 1.28 | 1.56 |
| 1998 | 2.89 | 1.63 | 1.26 | 2.67 | 1.13 | 1.54 |
| 1999 | 2.94 | 1.46 | 1.47 | 2.80 | 0.98 | 1.81 |

Source: RAND tabulations from merged MEDS file.

Note: M-Medi-Cal, W-Welfare, MO-Medi-Cal Only.

Tabulated according to CPS concepts: Enrollment is at any time in calendar year (not in each month); Adults are 15-65 in March of the next year; Children are 0-14 in March of the next year.

In the late 1980s, Medi-Cal had just over 4 million enrollees, approximately evenly divided between adults and children and between welfare and Medi-Cal Only (i.e., during the calendar year at least some Medi-Cal, but never welfare). During the early 1990s, the number of enrollees grew rapidly to over 6 million because of a combination of two factors. First, program eligibility was deliberately expanded. Second, California's deep recession made more people income-eligible, especially through rapid growth in welfare/cash assistance.

From the mid-1990s to the early 2000s, Medi-Cal enrollment has remained relatively stable, near 6 million. This stability is the result of offsetting trends in Medi-Cal sub-programs such as welfare and 1931(b). First, as in the rest of the nation, there has been a sharp drop in welfare/cash assistance since the early 1990s. Second, as was intended (but after a transition period), the 1931(b) program's growth has more than offset the shrinkage in cash assistance. Third, the other components of Medi-Cal such as Supplemental Security Income (SSI), Medically Needy, and "Other" are relatively stable.

**THE MEDI-CAL ELIGIBILITY DATA SYSTEM (MEDS)**

Real-time enrollment information in Medi-Cal is maintained in the MEDS. County welfare departments update this system as individuals are enrolled in or drop out from the program. Providers check the system to verify whether an individual is covered or not, and which services would be reimbursed by the Medi-Cal program. Individual-level extracts from the file provide a complete historical record of Medi-Cal eligibility for 1987 forward. Crucially for our purposes, the file includes linking information (name, Social Security Number—SSN), some basic demographics (gender, date of birth, race/ethnicity), and detailed Medi-Cal program information.

In our analysis below, we treat the MEDS records as "truth." This is a reasonable approximation given their use by providers in determining whether care will be reimbursed. However, the MEDS data are not always absolutely correct. Careful study of the MEDS data suggests some anomalies when counties had trouble updating the records (e.g., for two months in late 1990, there is a period of a few months when there appear to be no entries onto welfare for Los Angeles County). Furthermore, there is some retrospective eligibility that is recorded in the MEDS data but that might not be known to the respondent as of a survey interview. Card, Hildreth, and Shore-Sheppard note some seam bias (sharp increases in transition rates across versions of the MEDS file), which also suggests some reporting error in the MEDS.[6]

In addition, we note that some people may be enrolled in Medi-Cal but might not be aware of it. In particular, the Edwards v. Kizer decision requires California's counties to continue Medi-Cal eligibility for welfare leavers until their eligibility for continued Medi-Cal can be determined. Moreover, California's implementation of the Medicaid 1931(b) program and the provisions of California SB 87 have the effect of keeping many welfare leavers on Medi-Cal even without filing an application.[7] It is widely believed that many of these people do not realize they are covered.

**THE CURRENT POPULATION SURVEY (CPS)**

The CPS is a monthly survey of about 50,000 households conducted by the U.S. Bureau of the Census for the U.S. Department of Labor.[8] The CPS's primary purpose is to provide official monthly estimates of the unemployment rate, a key business cycle indicator. With its associated sampling weights, it represents the American non-institutional population.[9]

---

[6] See Card, Hildreth, and Shore-Sheppard (2001) for some further discussion of these issues. The seam bias problem should be less severe in the annual reference period of the CPS which we analyze than in the monthly reference period of the SIPP that Card, Hildreth, and Shore-Sheppard analyze. Note also that their biggest matching problems are with children, for whom we do not have SSNs and therefore do not match. Finally, note that below we limit our sample to the validated records which should increase the quality of the SSN data.

[7] Medicaid Section 1931(b) was a new program created by federal welfare reform (the Personal Responsibility and Work Opportunities Act of 1996) to guarantee Medicaid to any family that would have been eligible for welfare before welfare reform. Section 1931(b) also gave states the option of expanding 1931(b) eligibility to align it with eligibility for cash assistance. California did so with the net effect that welfare leavers with income up to about 165 percent of the poverty line remain indefinitely eligible for Medi-Cal. In practice, implementation of Section 1931(b) in California was delayed until early 1999, but indirect effects (the "Edwards Hold," see Klerman and Cox, 2004) were felt beginning in early 1998.

California SB 87 (chaptered September 30, 2000, effective July 1, 2001) streamlined continued enrollment in Medi-Cal for welfare leavers through adoption of an ex parte process and, in practice, a presumption of continued eligibility for Medi-Cal among welfare leavers. This implementation occurred after the period covered by our data.

[8] For more on the CPS, see http://www.bls.census.gov/cps/overmain.htm.

[9] The restriction of the CPS universe to the non-institutional population is potentially problematic for analyses of Medi-Cal. While most Medi-Cal enrollees are young, most Medi-Cal expenditures go to the elderly in nursing homes. That group is not in CPS's universe, which is the non-institutional

Since 1948, in its spring survey the CPS has included additional questions on annual income in the previous year.[10] Today, those additional questions are asked at the end of the March survey (corresponding to the arrival of W-2s and household preparation of federal income tax returns) and are referred to collectively as the Annual Demographic Survey (ADS). Over the years, the set of supplementary questions has grown.

Most important for our purposes, since 1980, the ADS has included detailed questions on health insurance coverage and welfare receipt in the previous calendar year (not as of the date of the March interview).[11] These questions began as an attempt to expand the definition of "income" to include employee benefits and noncash government benefits (Food Stamps, subsidized housing, medical assistance, etc.). Combining the questions on health insurance as an employee benefit with the questions on participation in government health insurance programs yielded a rough measure of total health insurance coverage; its complement provided an estimate of those without health insurance. Until the 2000 interview, the last included in our analysis file, there was no direct question in the CPS about being uninsured. Rather uninsured status is inferred from answers to questions about receipt of Medicaid and other types of health insurance (see below Table 2.2 for a summary of changes to the CPS questionnaire).

As this discussion suggests, the individual questions were not originally intended to generate an estimate of the size of the population without health insurance. With issues of uninsurance becoming more salient, in 1988, the Census Bureau refined the questions.[12] Questions about employer-based health insurance that previously had only been asked of employed individuals were asked of all individuals 15 or older, regardless of whether they worked. This

---

population. The MEDS data do not have a flag for institutional residence. As a partial correction, our analyses below exclude those age 65 and over.

[10] For more on the March Annual Demographic Supplement to the CPS, see: http://www.bls.census.gov/cps/ads/adsdes.htm.

[11] This discussion draws on Nelson and Mills (2001).

[12] In addition, in 1983, the Census Bureau began a second national survey, the SIPP). The SIPP is a moderate-sized panel survey with more detailed questions on income and program participation (as its name implies). The original vision appears to have been that the SIPP would replace the CPS-ADS for many purposes, including the measurement of health insurance. However, for a variety of reasons (including varying sample size, issues related to its panel structure, and slow data release; see Short, 2001), the CPS remains the primary data source for counts of the uninsured. For an analysis similar to this one for the SIPP, see Card et al. (2001).

change should have captured retiree coverage and COBRA benefits (i.e., benefits from a previous employer). In addition, for children, questions were added about health insurance coverage from individuals not residing in the household. This change should have captured coverage provided by non-coresident parents. Finally, the imputation methods for children's coverage were revised and additional questions on Medicaid were added (see Levit et al., 1992; Moyer, 1989; Swartz and Purcell, 1989; and EBRI, 2000).

Additional changes have been made since then. (See also Swartz, 1997; EBRI, 2000.) Table 2.2 presents a detailed chronology. Census analyses suggest that the changes in survey years 1996 and 2000 increased reported health insurance coverage by about 1 percentage point each.

**Table 2.2**
**Chronology of CPS-ADS Changes and Effect on Health Insurance Coverage**

| Year | Change | Effect on Health Insurance Coverage |
|---|---|---|
| 1981 | First health insurance questions (employer sponsored and government sponsored) on CPS-ADS | <Not applicable> |
| 1988 | Introduction of new CPS processing system | Minimal |
| 1989 | Addition of questions on child health insurance coverage (previously coverage of children was imputed based on adult responses | Moderate |
| 1993 | Switch to 1990 Census population controls | Minimal |
| 1994 | Switch from paper and pencil to Computer Aided Personal Interviews | Minimal |
| 1996 | Questions reordered and modified to improve information on Medicaid | Possibly moderate; see Swartz (1997). |
| 1998 | Indian Health Service no longer considered coverage | Minimal |
| 2000 | Verification questions added | Moderate (about 1 percentage point) |
| 2001 | Switch to 2000 Census population controls | Minimal (less than 1 percentage point) |
| 2001 | Addition of questions on state CHIP programs | |
| 2002 | Additional sample (78,000 rather than 50,000) to estimate state health insurance coverage rates | Minimal (less than 1 percentage point) |

Note: "Year" refers to the survey year. The CPS questions refer to the previous calendar year; i.e. "2001" refers to the survey conducted in March of 2001, collecting information about calendar year 2000. Effect is on the total national coverage rates. Purely because of sampling issues, effects are larger at the state level. Because of substantive issues, effects are often larger for components (e.g., Medicaid). "Minimal" is less than 1 percentage point; "Moderate" is more than one percentage point.

Corresponding to the fact that California has 12 percent of the national population, the annual March ADS to the CPS has about 6,000 California households, about 13,000 individuals (adults and children), and about 2,000 individuals on Medi-Cal.

**LEVELS AND TRENDS IN MIS-REPORTING**

Before turning to the more detailed results from the matched data, we conclude this section with an analysis comparing simple (unmatched) tabulations from the two data sources (i.e., a comparison of the aggregate administrative counts and the CPS estimates of the population enrolled). Figure 2.2 plots the ratio of CPS enrollment to MEDS enrollment. (Table 2.3 provides the underlying numbers.) These tabulations are made using the individual-level MEDS files. We have aligned the counts to match the CPS concepts so that enrollment is measured as being at any time in the past year. Age is as of March of the next year. The division between adults and children follows the CPS at 14/15 at the interview. California residence in the CPS is ascribed based on residence a year before the interview (not at the interview, as is conventionally done).

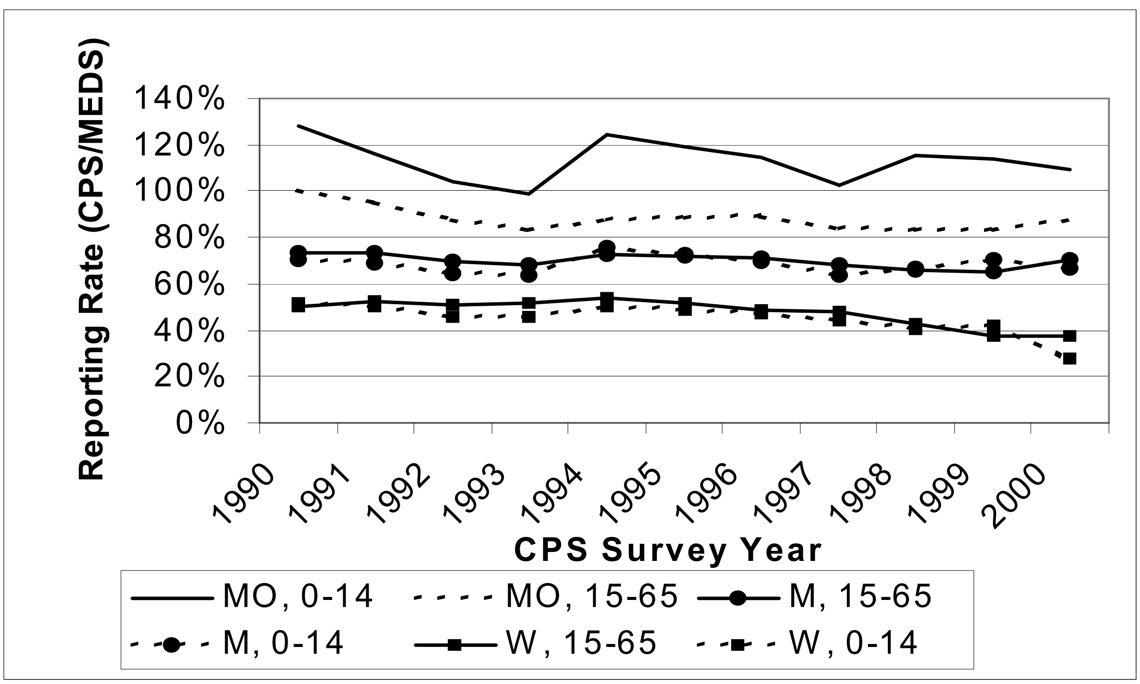


**Figure 2.2—Under-Reporting of Enrollment (CPS/MEDS)**

Source: Tabulations from RAND Merged CPS File and RAND MEDS file

**Table 2.3**
**Reporting Rates (CPS/MEDS)**

| Survey Year | Adults | | | Children | | |
|---|---|---|---|---|---|---|
| | M | W | MO | M | W | MO |
| 1990 | 74% | 50% | 100% | 71% | 51% | 128% |
| 1991 | 73% | 52% | 95% | 69% | 50% | 116% |
| 1992 | 70% | 51% | 88% | 64% | 46% | 104% |
| 1993 | 68% | 52% | 83% | 63% | 46% | 99% |
| 1994 | 72% | 54% | 88% | 76% | 50% | 124% |
| 1995 | 69% | 49% | 86% | 73% | 48% | 119% |
| 1996 | 71% | 49% | 89% | 70% | 47% | 114% |
| 1997 | 68% | 48% | 84% | 64% | 44% | 103% |
| 1998 | 66% | 43% | 83% | 67% | 40% | 115% |
| 1999 | 65% | 37% | 83% | 70% | 42% | 114% |
| 2000 | 71% | 37% | 87% | 67% | 28% | 109% |

Source: RAND tabulations from RAND CPS matched file and the RAND MEDS file.

Note: M-Medi-Cal, W-Welfare, MO-Medi-Cal Only. Tabulated according to CPS concepts: Data refers to calendar year preceding the survey year. Enrollment is at any time in calendar year (not in each month); Adults are 15-65 in March of the next year; Children are 0-14 in March of the next year.

As the figure and the table show, there clearly is under-reporting. CPS Medi-Cal counts are only about 70 percent of MEDS Medi-Cal counts. CPS welfare counts are only about 40 percent of MEDS welfare counts. Anomalously, CPS Medi-Cal Only counts are close to MEDS Medi-Cal Only counts. We return to this below.

Unlike the earlier characterization of the national data, there is little evidence of trend in California's Medi-Cal reporting rates. Perhaps there is some increase in Medi-Cal for adults in 1996 and 2000 with the latest changes. There is no evidence of an increase for children in 1996 or 2000; the estimates decline. The situation is very different for welfare. Welfare reporting rates for adults have fallen from about 50 percent to about 40 percent over these eleven years. For children, there is a large additional drop in 2000.

**DISCUSSION**

This section has provided basic background information. It described the Medi-Cal program and the two data sources. It then compared the two data sources in aggregate to provide a rough characterization of the mis-reporting. In the next section, we turn to the match: what it can contribute and what are the technical issues in using the information provided.

# 3. THE MATCHED DATA

In this chapter, we turn to the matched data so that we can investigate mis-reporting at the individual level. In order to understand mis-reporting, we begin by creating and analyzing a dataset containing only the highest-quality matches. In this chapter, we describe the congruence of the CPS responses to the MEDS information, where we treat the MEDS information as "truth."

What we are able to report is strongly constrained by Census disclosure rules. To preserve the confidentiality of CPS respondents, those rules limit our ability to report exact results for tables with small cell sizes in any of the cells. Thus, to preserve confidentiality, for most analyses, we pool observations across all of the survey years. In some cases, where the cell sizes are particularly small, we combine cells or report the predictions from simple regression models.

As we discuss here and in the next chapter, not everyone provides a SSN, so the matched data cannot provide a complete characterization of the quality of reports. We defer until the next chapter a discussion of how we use the information from the matched sample to make inferences about the entire population—both those who do provide a SSN and those who do not.

## MATCHING

Our first task in constructing the matched dataset is to match the administrative data (from the MEDS) to the survey data for California residents (from the CPS) where possible. Table 3.1 summarizes the sample selection rules for the analyses in this chapter. Appendix A describes the matching process, sample selection rules, and results in more detail. We note that we deliberately suppress the exact sample counts in each year to preserve our ability to present other more substantive results by year later in the analysis.

**Table 3.1**
**Sample for Matched Analyses**

| Sample | Validated | Unvalidated | Total |
|---|---|---|---|
| **Adults 15-65** | | | |
| Initial | 46,404 | 53,515 | 99,919 |
| No SSN | 18,665 | 18,677 | 37,342 |
| Movers | 1,717 | 1,169 | 2,886 |
| Imputed Data | 976 | 1,519 | 2,495 |
| Bad Match | 692 | 446 | 1,138 |
| Final | 24,354 | 31,704 | 56,058 |
| **Children, 0-14** | | | |
| Initial | 16,567 | 20,289 | 36,856 |
| No Parent | 205 | 251 | 456 |
| No Parental SSN | 6,773 | 7,595 | 14,368 |
| Final | 9,589 | 12,443 | 22,032 |

Our primary focus is on "adults" (defined as age 15 to 65 as of the March interview) residing in California. The reason for these two exclusions is straightforward. First, the CPS only collected SSNs for those aged 15 and older, so there is no possibility of matching "children" (defined as those under age 15). Second, the MEDS is administrative data from the state of California and, thus, only includes information on people enrolled in California's Medicaid program, Medi-Cal.

The basic sample begins with California adults. Consistent with the CPS questions on program participation that refer to participation in the previous year, we define "California adults" based on state of residence a year prior to the interview.[13] From this sample of California adults (as of a year prior to the interview), the "matched sample" drops the following groups:

- *No SSN*: Even for those age 15 and older, not everyone has a (scrambled) SSN on our internal CPS file.[14]

---

[13] Using state of residence a year before the survey is more consistent with the CPS reference period than using state of residence at the survey. We note, however, that it does not appear to be standard practice in analyses of program participation or health insurance coverage.

We also note that the CPS reference period induces some standard coverage issues. The sample is drawn in March of the following year and is, therefore, not completely representative of everyone alive in the reference year (or even as of the end of the reference year). The divergence will include births and deaths and changes in residence (in the United States at all, in an institution).

[14] In 1990, 1992, 1993, 1995, and 1996, our file includes "unvalidated SSNs," i.e., the SSNs are simply a scrambled version of the SSN provided by the CPS respondent.

- *Movers*: Our MEDS data only include welfare receipt in California. So for movers, responses about Medicaid and welfare might refer to enrollment in a different state, which would not be recorded in the MEDS data. For the analysis in this chapter, we therefore also drop movers (i.e., those who were not in California both a year before the survey and at the survey). This is an imperfect adjustment for movers. (See further discussion of this issue at the end of Appendix A.)
- *Imputed Data*: Imputed data on enrollment is not informative for the quality of unimputed responses. Imputed data on the matching variables might cause us to incorrectly accept or reject a SSN match. (See the next bullet.) We therefore drop from the analyses in this chapter anyone with imputed responses on program enrollment (Medicaid or welfare) and anyone with imputed responses on the basic demographics used in matching (gender and age; there are only a trivial number of such individuals).
- *Bad SSN*: To accept a match, we require a match on gender and on age plus or minus one year.

Note that even though we refer to this as our "matched sample," it is more properly the sample of people who *could have* matched to the MEDS (i.e., they provided a SSN). We do *not* require an actual match to the MEDS because we want to include people in our sample who did not receive Medi-Cal (who may or may not report their program participation correctly in the CPS). Our MEDS extract contains a record for each individual who has received Medi-Cal during 1989 to 2001. Individuals who did not receive Medi-Cal during this period

---

In 1991, 1994, and 1997-2000, our file includes a validated version of the SSN, not (a scrambled version of) the SSN provided by the CPS respondent. Exact details on the validation process are not available, but it appears that Census provided a list of SSNs and basic demographic information (name, gender, age) to the Social Security Administration (SSA). SSA then cross-checked the information against its SSN records. For some cases, a SSN was imputed onto [into?] the file based on name, place of birth, and birth date; for other cases, a provided SSN was deleted based on failure to match on these criteria.

Some of the tabulations below (including Table 3.1) tabulate results separately by validation status of the survey file (i.e., the entire year). To avoid ascribing lack of congruence between CPS and MEDS data to improper SSNs, other analyses below drop all unvalidated years.

should not and will not appear in the MEDS data. We leave them in the matched sample and infer that they never received Medi-Cal. We note, however, that for this group it is not possible to verify that the gender and age match across the two data sources.

Finally, the analyses in this chapter consider only the "validated" data in order to generate an analysis file with only the highest quality matches. When the CPS collects SSNs, it also asks permission to "validate" SSNs with the Social Security Administration (SSA). The details of this validation are unclear. It appears that for a respondent who does not provide a SSN, Census passes his/her name, gender, and age to SSA which attempts to impute a SSN from the official SSA SSN files (the NUMIDENT file). Conversely, it also appears that provided SSNs are checked and either replaced or deleted. The presumption is that this validation improves the quality of the matches. In some years (1991, 1994, and 1997-2000), Census provided us with the validated files; in other years (1990, 1992, 1993, 1995, 1996), Census provided us with only the unvalidated files. Presumably, validation improves the quality of the SSNs. Our results below support that presumption.

The CPS's failure to collect SSNs for children implies that we cannot directly apply similar matching methods to children. In the substantive analyses below, we make a rough imputation of the implications of our analysis of the matched data for children and, therefore, for the entire population. To do so, we assume that if the family's reference person (as identified by the CPS FAMREL variable) is actually enrolled in Medi-Cal/welfare, the children are as well. However, if the family's reference person is not enrolled in Medi-Cal/welfare, we do not use that information to impute enrollment for the child. Table 3.1 also gives the total number of California children and the number of them whose reference person is in the matched sample.

The net effect of these sample restrictions is that although the CPS has 99,919 California adults during the period 1990 to 2000, our narrow matched sample for data quality analysis is only a quarter of that, 24,354. Only six of the eleven years are validated, cutting our sample nearly in half. Only about 65 percent of the validated sample provides SSNs, about 2 percent of the sample in California a year before the survey is not in California at the survey, about 3 percent of the sample has imputed data, and about 1 percent is a "bad match."

For our analyses of CPS data quality, we will use these narrowly defined/highest quality matches. For our extrapolations to the full file, we will use all the matches, making multivariate corrections for the effects of validation (see the discussion of details in the next chapter). As expected, we note that validation both increases the fraction of adults with SSNs (from 60 percent to 65 percent). However, the bad match rate nearly doubles with validation (from 0.8 percent to 1.5 percent). Apparently, some of the SSNs added at validation are incorrect.

Based on our descriptive analysis, providing a SSN does not appear to be random (see Table 3.2). More advantaged people are more likely to provide a SSN and to appear in the final sample; less advantaged people are less likely to provide a SSN and be in the final sample. Combining the validated and unvalidated samples, we tabulate the fraction of people with given characteristics in the final (i.e., matched) sample. This is a rough proxy for presence of a SSN; about 85 percent of those in the full sample, but not in the final sample, are dropped because of a missing SSN. Overall, 58 percent of adults are in the final sample; i.e., provide a SSN, are not movers, do not have imputed data, and are not "bad matches," (These figures are weighted, unlike those in Table 3.1, which are unweighted). Minorities (black and Hispanic, 50 percent), high school drop-outs (51 percent), and those in poverty —less than half the poverty line (45 percent), half the poverty line to the poverty line (52 percent), the poverty line to one and a half times the poverty line (52 percent), and one and a half times the poverty line to twice the poverty line 54 percent)—and single females with children (54 percent) are less likely to be in the final sample. Those with at least some college (63 percent) and other health insurance (63 percent) are more likely to provide a SSN.

**Table 3.2**
**Percentage of People from Full Sample in Final Sample**

| Sample | % w/SSN |
|---|---|
| **Overall** | 58% |
| Male | 58% |
| Hispanic | 50% |
| Black | 50% |
| HS Drop Out | 51% |
| Some College | 63% |
| FPL<0.5 | 52% |
| 0.5<FPL<1.0 | 52% |
| 1.0<FPP<1.5 | 52% |
| 1.5<FPP<2.0 | 54% |
| Kids in Household | 57% |
| Single Female w/Kids | 54% |
| Other Health Insurance | 63% |
| Welfare | 65% |
| Medicaid | 55% |

Note: FPL—Federal Poverty Line.

Given that they are less socially advantaged, we might expect those enrolled in Medi-Cal to be less likely to provide a SSN. Offsetting this, Medi-Cal enrollees are required to supply a SSN each time they deal with the welfare office, so they are likely to know their SSN, and perhaps be less reluctant to give it out. In fact, those reporting to the CPS that they have welfare are more likely than others to provide a SSN (63 percent); those reporting to the CPS that they are enrolled in Medi-Cal are as likely as those not reporting Medi-Cal to provide a SSN (58 percent); and those with Medi-Cal only are less likely than others to provide a SSN (55 percent).

Combining both the validated and unvalidated years, our basic sample has 36,856 children. For almost all of them, we can locate a parent on the file. For about 61 percent of those parents, we have a SSN. This rate is similar to the 63 percent of adults that provide a SSN.

## SIMPLE REPORTING RATES

We have deliberately constructed this sample to maximize the congruence between the survey data and the administrative data. We have distinguished between validated and unvalidated SSNs; we have required that gender and age match across the survey data and the administrative data; we have dropped all imputed data; and we have dropped all movers.

Furthermore, we have used the administrative data to exactly mimic the CPS concept of program participation—Medi-Cal or welfare in any month in the previous calendar year (January to December). While the CPS data refer to the previous year, aggregate Medicaid and welfare tabulations from administrative data are usually published as monthly totals. With movement onto and off Medicaid/welfare, there is no direct relation between the counts for the individual months and CPS concept—the total number of individuals enrolled in the program at any point in a given calendar year.

This difference in concepts of participation is not an issue in our analysis. We have the monthly MEDS data and can tabulate it so that it exactly corresponds to the CPS question (i.e., any Medi-Cal in the previous calendar year, any welfare in the previous calendar year, and—for Medi-Cal Only—Medi-Cal, but no welfare in the previous calendar year).

From these ideal data, we want to compute "behavioral" mis-reporting rates (in contrast to the "imputational rates" that we define in the next chapter). Formally, the mis-reporting rates are:

$$\rho_{FP}^{b} = \frac{FP}{TN + FP} \qquad \rho_{FN}^{b} = \frac{FN}{TP + FN}$$

On the left we have the behavioral false positive rate. The denominator is all people who are not truly enrolled, the sum of true negatives and false positives. The numerator is the number of people who are not enrolled, but report that they are. Thus, the ratio is the fraction of people who are not enrolled, who report falsely that they are. On the right is the behavioral false negative rate. By analogy, it is the fraction of people who are enrolled (i.e., the sum of true positives and false negatives), who falsely report that they are not enrolled. The observed net under-reporting suggests that the false negatives are the more common group.

Table 3.3 reports these behavioral rates. The off-diagonals give the false reporting rates; i.e., the percentage of people truly in a given status, who give each possible response. Despite our efforts to develop a sample that should maximize the congruence between the survey and administrative data, the congruence of the two reports is distressingly poor, as can be seen in Table 3.3. We perform the analysis separately for each of the three programs of interest: Any Medi-Cal, welfare, and Medi-Cal only (i.e., Medi-Cal, but not welfare). There are eight possible outcomes for each individual, which are summarized in Table 3.3 below: Four outcomes for people for whom there is a corresponding MEDS record; two possible responses for people with a valid SSN, but for whom there is no corresponding MEDS record (CPS response is yes or no and their MEDS responses are implicitly "no"); and two possible CPS responses for those who do not provide a SSN (no information can be gleaned in terms of a MEDS response).

**Table 3.3**
**Congruence in the Matched Sample between MEDS and CPS Data**

| | | CPS | | | |
|---|---|---|---|---|---|
| | | **N** | **MO** | **W** | **T** |
| MEDS | N | 97.7% | 1.8% | 0.5% | 89.8% |
| | MO | 36.8% | 59.3% | 3.9% | 4.9% |
| | W | 20.4% | 32.0% | 47.6% | 5.3% |
| | T | 90.6% | 6.2% | 3.2% | 100.0% |

Notes: Entries: row percents (i.e., rows sum to 100%), except for last row and column ("T" for total) that give overall percentages. Computed from the "matched sample" (i.e., adults who provide a SSN).

N-No Medi-Cal, MO-Medi-Cal Only, W-Welfare, T-Total.

Then, the first column of Table 3.3 suggests that few people who do not have Medi-Cal report that they do have Medi-Cal (2.3 percent = 100.0 percent - 97.7 percent). The same is not true for those with Medi-Cal. Less than two-thirds (59.3 percent) of those with Medi-Cal Only actually report having Medi-Cal Only in the CPS (i.e., Medi-Cal, but not welfare); and less than half (47.6 percent) of those with welfare actually report having welfare in the CPS. These are distressingly low rates of congruence, particularly in a sample designed to only include the highest quality matches.

The overall congruence of reports of Medi-Cal is higher. Nearly three-quarters of those with Medi-Cal report that they have Medi-Cal (computed from the second and third rows of Table 3.2 by converting to the unconditional percentages, combining the Medi-Cal Only and Welfare groups and computed error rates). This divergence appears to be explained by those with welfare. While more than half of them report not having welfare (52.4 percent), most of them do report having Medi-Cal. Thus, the overall reporting of Medi-Cal is not as bad as considering welfare and Medi-Cal-Only separately would suggest. This also appears to explain why the raw reporting rates for Medi-Cal Only (i.e., the ratio of the CPS counts to the MEDS counts; see Table 2.3) are so much higher than the reporting rates for Medi-Cal or Welfare. It is not that reporting rates are higher for Medi-Cal Only; it is that many of those reporting Medi-Cal only in the CPS actually have welfare as well.

The implications of Table 3.2 for reporting rates are relatively subtle. The previous paragraph considered reports conditional on the truth (as recorded in the MEDS). Some of the errors are offsetting. If instead, we compare the gross rates (i.e., the ratio of the CPS total to the MEDS total), the reporting rates are higher. The gross reporting rate for Medi-Cal is 92 percent ((6.2%+3.2%)/(4.9%+5.3%)). The gross reporting rate for welfare is much lower at 60 percent (3.2%/5.3%). For Medi-Cal Only, the gross reporting rate implies over-reporting of Medi-Cal Only at 128 percent (6.2%/4.9%).

These gross reporting rates in the validated years are higher than for the unmatched comparisons that implicitly include both the matched and unmatched samples (about 70 percent of Medi-Cal, about 45 percent for welfare, and about 85 percent for Medi-Cal Only). Thus, while the matched data capture some of the under-reporting, it seems likely that there are additional considerations leading to under-reporting in the unmatched sample. The simplest explanation is higher false negative rates; people who do not provide a SSN are more likely to not report enrollment, even when they are enrolled. Below, we implement an algorithm consistent with that simple explanation.

Other explanations would consider not differential reporting, but differential coverage of the CPS. Even if the CPS survey process was perfect, the CPS only attempts to interview those in the non-institutional population. Anyone receiving Medi-Cal and in an institution would be in the MEDS count, but not in the CPS count. To address this concern, we delete everyone over 65 from both counts. This should eliminate most of the institutionalized population. The size of the remaining institutionalized population is unclear.

Similarly, while the CPS sampling frame is the non-institutionalized population, as with any survey, some people are missed. The CPS adjusts for such failure to interview using control totals derived from the Census that stratify on region, gender, and age. It seems likely, that within these cells, those with Medi-Cal are less likely to be interviewed. If this supposition is correct, then the CPS, even with adjustments, will under-report Medi-Cal enrollment. Again, the methods we propose below will correct for this weighting error, at least at the aggregate level.

**UNDERSTANDING THE REPORTING ERRORS**

Table 3.3 tabulates the MEDS data according to whether there was any program enrollment in the previous year. Standard cognitive approaches to survey response (Sudman and Bradburn, 1973; Groves, 1989) would suggest that a positive Medi-Cal response is more likely the more Medi-Cal/welfare receipt there was and also if there is Medi-Cal receipt in the survey month.

Our detailed MEDS data allow us to compute months of enrollment last year and enrollment in March of this year. The observation counts at each month are too small to allow reporting the raw rates. Instead for Medi-Cal, Figure 3.1 reports: (1) the raw rates for March of this year, and all of last year; and (2) not March of this year and none of last year. For the other combinations, we report the results of probit polynomial regressions.[15] Reporting the regression results also smoothes out some of the sampling variation.

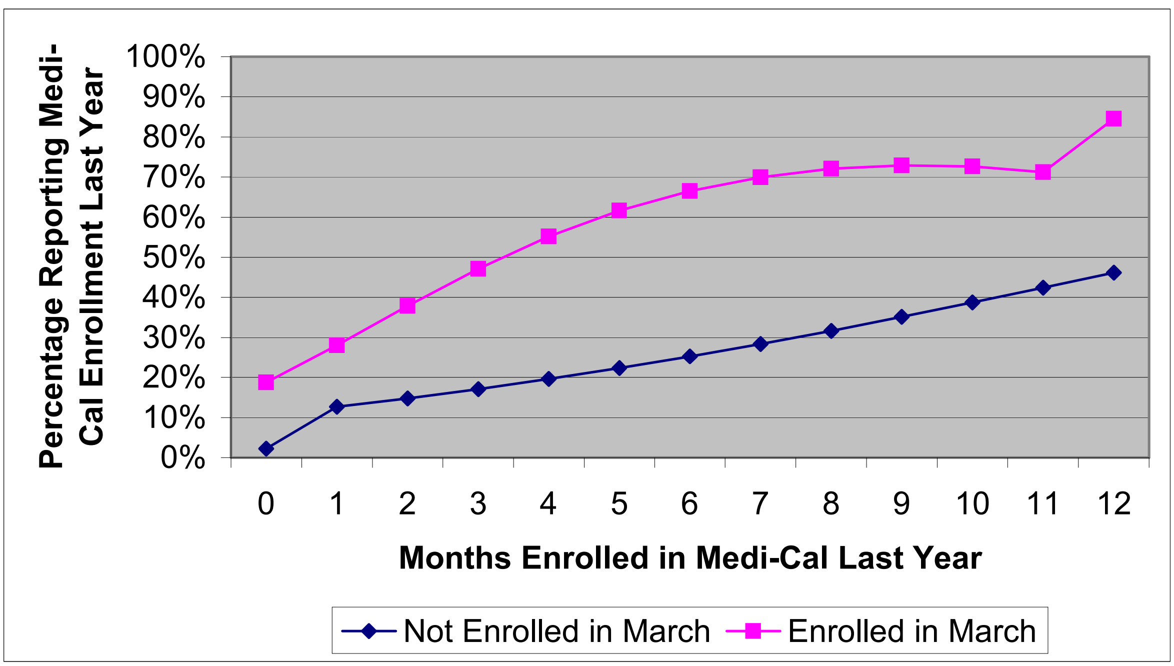


**Figure 3.1—CPS Reporting of Medi-Cal Given MEDS Pattern of Receipt**

---

[15] The probit analysis takes as its dependent variable the percentage of people reporting enrollment from the MEDS, given their actual status in March of the survey year (i.e., the vertical axis of Figure 3.1) and as its independent variables polynomials in months of actual enrollment from the MEDS (i.e., the vertical axis of Figure 3.1). The polymomial is quadratic for those on welfare this March and linear for those not on welfare this March. The analysis uses the CPS sample weights. We then plot the predicted probabilities from that model.

If reporting was perfect, the points for "0" would be zero and the other points would be 100 percent. Instead, we observe a clear dose-response relationship. The more months of enrollment last year recorded in the MEDS, the more likely a person is to report program enrollment in the CPS. People on Medi-Cal all of last year and in March of this year, report Medi-Cal at about 85 percent. People on in March, but not on all of the previous year, are less likely to report Medi-Cal, with reporting rates varying from 30 to 70 percent.

The CPS question only asks about enrollment last year, but enrollment in March clearly affects the probability of answering the CPS question about last year positively. The difference ranges from 15 to 30 percentage points. Finally, note that people enrolled in March but not at all last year—who should answer negatively—have a 20 percent probability of answering in the affirmative (i.e., they respond based on their current enrollment status rather than their enrollment status last year). These people appear to explain much of the false positives, people who report enrollment in the CPS but who are not actually enrolled (in the past year) according to the MEDS.

Figure 3.2 reports the same tabulations and probit regression predictions for welfare. The patterns are similar. People enrolled in March are more likely to respond positively. Even some (about 5 percent) of those enrolled in March, but not last year, respond positively, comprising many of the false positives. Both for those enrolled in March and for those not enrolled in March, a positive response is more likely the more months of enrollment in the past year. However, even for people enrolled in March and all of last year, only slightly more than half report their welfare participation in the CPS.

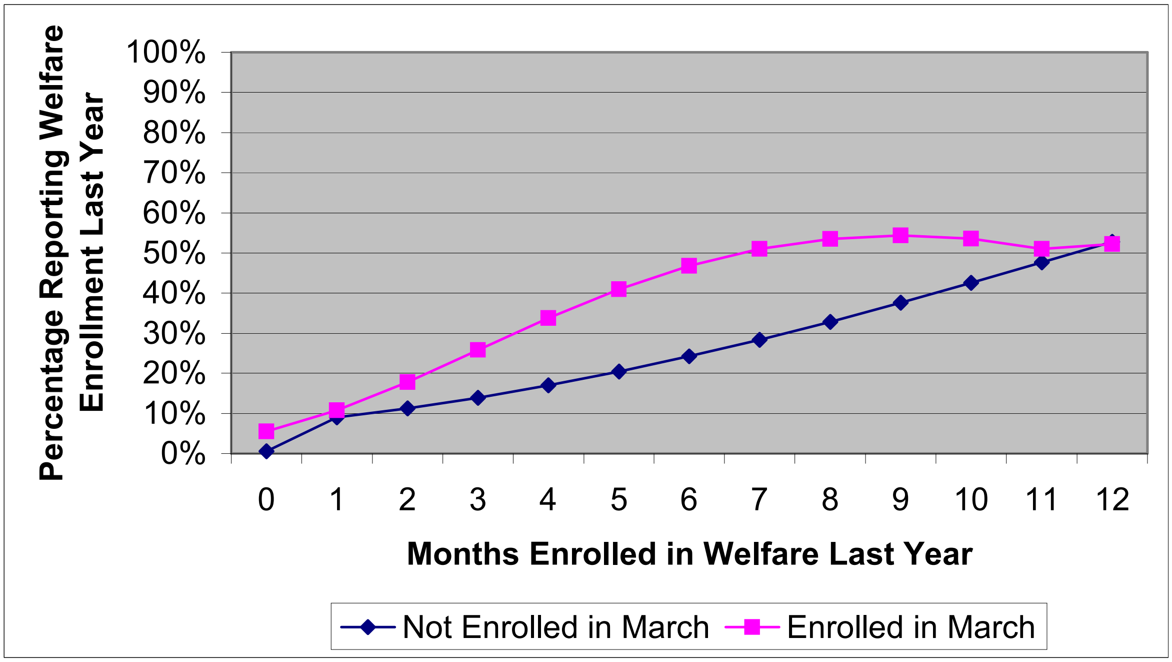


**Figure 3.2—CPS Reporting of Welfare Given MEDS Pattern of Receipt**

The CPS also includes a question about months of receipt of Medicaid and welfare last year; however, the individual cells are too sparse to allow reporting. We do find that there is some correlation between the MEDS and CPS data, but the correlation is only moderate (0.58 for Medi-Cal, 0.57 for welfare) and the mean numbers of months reported are very different. Consequently, we conclude that the monthly data are not very useful.

Some of the other welfare false positives, people who report to the CPS that they received welfare last year but who are not recorded as having welfare in the administrative data (i.e., the MEDS), are also understandable. The CPS distinguishes between AFDC/TANF/CalWORKs (a state program) and SSI (a federal program). Both programs provide cash and both automatically confer Medi-Cal eligibility. Thus, it would not be surprising if some people on SSI report to the CPS that they are on welfare. We would code those people as "false positives." Because SSI confers Medi-Cal eligibility, we can identify SSI in MEDS and, therefore, in the matched data. Tabulations on the validated years suggest that a sizable fraction of the welfare false positives in fact have SSI. Cell counts are too small too support more precise estimates or their release. Even with these explanations, some of the welfare false positives are still unexplained. Based on these results for SSI, however, it seems likely that some of the remaining false positives are related to receipt of General Relief (which we can not identify in our data).

Unfortunately, we cannot use similar methods to better understand the false positive reports for Medi-Cal, because we do not have similar administrative data that identify other sources of public insurance. By analogy to the results for welfare, however, it seems likely that some other public health insurance programs are being reported as Medicaid, again inducing false positive responses. For example, in California, Healthy Families is administered outside the Medicaid system and not recorded in MEDS. Thus, some of the Medi-Cal Only false positives are probably Healthy Families enrollees. However, given that we have dropped everyone under age 15 and Healthy Families in California does not cover adults, this is unlikely to be a major factor. More recently, several counties have put in place public, county-level Medicaid-like programs. Although these programs were implemented after the period our data cover and, thus, cannot explain the false positives in our data, this may be an issue in future CPS interviews, where such programs may incorrectly be reported as Medicaid.

Another possible explanation for the false positives in both of the programs is migration between states. As we discuss in detail at the end of Appendix A, we control for migration, but those controls are incomplete. For example, some people who we have classified as California residents in fact spent part of the reference year in another state. If they were enrolled in Medicaid/welfare there, we would have incorrectly labeled them as false positives.

We note, in contrast, that the false negative rates (i.e., the percent of people enrolled in the program based on the MEDS that do not report participation in the CPS) seem much too high to be explained away by any of these factors.

**CPS REFERENCE PERIOD**

Figure 3.1 and Figure 3.2 suggest that enrollment in the following March increases the probability of a positive response. Of particular note are the people who were not enrolled at any point during the last year, but are enrolled in March of this year. They should give a negative response to a question about enrollment last year. For Medi-Cal, about 20 percent of this group gives a positive response; for welfare the comparable figure is about 5 percent. These percentages are about half of the percentages for people enrolled in March of this year, but only enrolled one month of the previous year. These results are consistent with Swartz (1986) and Sudman, Bradburn, and Schwarz (1996) who argue that the CPS responses should be interpreted as referring not to the previous calendar year, but instead to the interview month. CBO seems to accept this argument (Bilheimer, 1997; CBO, 2003). Lewis, Ellwood, and Cazjka's (1998, p. 27) in their review conclude:

> The CPS is designed to measure the number of individuals uninsured throughout a given year. Yet most researchers believe the CPS estimates of the uninsured represent a mix of those uninsured *throughout* [emphasis in the original] the previous year and those uninsured at a point in time [i.e., as of the interview].

Interpreting the CPS responses as referring, not to the previous year, but instead to the interview month would in general yield both lower Medi-Cal enrollment rates (fewer people are enrolled in a given month than in an entire year) and higher uninsurance rates (fewer people are uninsured in an entire year than in a given month).[16]

A more in depth analysis of this issue provides additional evidence to support this argument. For Medi-Cal and welfare, we explore this question directly. We divide our sample into four groups, by their true (i.e., MEDS) enrollment status in the survey month (yes/no) crossed with their true enrollment status in the previous year (yes/no). Our interest focuses on the cases where the status last year diverges from the status at the interview. In that case, the question is whether people are more likely to respond with their status last year (i.e., what the CPS instructions call for) or their status as of the interview (i.e., the cognitively simpler task). For the matched sample, Table 3.4 reports the prevalence of each cell in the MEDS and the probability of an affirmative response. The table has two implications. First, the overwhelming share of people have the same outcome last year and in the survey month. Thus, using the wrong reference period is unlikely to explain much of the net under-reporting.

---

[16] See Bennefield (1996b) who argues that the problem is generic under-reporting because of the length of the recall period. See also Fronstin, 1996, on dual coverage; Kronick, 1991, on private health insurance; Beauregard et al., 1997, comparing to MEPS results; Bennefield, 1996c, using CPS experimental questions; and Long and Marquis, 1996, comparing to RWJF survey.

Second, among those who do not have the same outcome last year and in the survey month, the more common response is "no." Specifically, consider people enrolled last year, but not at the interview (i.e., rows C and D). The "correct" response is "yes"/last year (i.e., row D), but "no"/survey month (i.e., row C) is more than twice as common (for Medi-Cal 1.56% vs. 0.58%; for Welfare 1.16% vs. 0.36%). Alternatively, consider people enrolled at the interview, but not last year (i.e., rows E and F). The correct response is "no"/last year (i.e., row E). Compared to "yes"/survey month, it is five times more common for Medi-Cal (0.43% vs. 0.08%) and essentially no one gives the correct response for welfare.

The net effect is to induce too many false negatives. For Medi-Cal the net effect is slightly less than one percentage point (0.89=(1.56-0.58)+(0.43-0.08)). For welfare, the net effect is slightly more than one percentage point (1.33=(1.16-0.46)+(0.19-0.00)).

In terms of whether the reference period is more usefully viewed as the survey month or the previous year, the two effects approximately net out, yielding a small net effect. For Medi-Cal the net effect is about half a percentage point towards survey month over previous year (0.63=(1.56-0.58)+(0.08-0.43)). For welfare, the net effect is about half a percentage point towards survey month over previous year (0.51=(1.16-0.46)+(0.00-0.19)). Overall, the net effect is just too small to explain the large differences that are seen across the two data sets.

**Table 3.4**
**CPS Reference Period**

| | **MEDS Response** | | | | |
|---|---|---|---|---|---|
| | **Any Time Last Year** | **Survey Month** | **CPS Response** | **Medi-Cal** | **Welfare** |
| A | Y | Y | N | 6.85% | 1.91% |
| B | Y | Y | Y | 1.80% | 1.91% |
| C | Y | N | N | 1.56% | 1.16% |
| D | Y | N | Y | 0.58% | 0.46% |
| E | N | Y | N | 0.43% | 0.19% |
| F | N | Y | Y | 0.08% | 0.00% |
| G | N | N | N | 86.71% | 93.81% |
| H | N | N | Y | 1.99% | 0.56% |

Notes: Cells are percent responding enrolled in the program in the CPS given enrolled last year and/or enrolled in survey month (from the MEDS). Validated years only. Final row is net impact of last year versus survey month; entry greater than one implies last year is more important (as per CPS instructions). Computed from validated years only.

**TIME TRENDS**

The previous analyses have pooled results across the (validated) years for which we have matched data. However, two factors suggest the importance of studying change over time. First, based on the aggregate data there is some evidence of increases over time in under-reporting. Second, the Census Bureau has altered the CPS questions, most notably in 1996 and 2000, which appears to have increased reports of health insurance coverage.

Unfortunately, the matched samples in each year are small, about 5,000 adults per year. Medi-Cal enrollment rates average about 10 percent, about 500 people in the CPS; welfare and Medi-Cal only rates are about half that. Furthermore, false positives are rare, often a few percent or less. As a result, we cannot provide descriptive evidence on time trends in the probability of false positives and false negatives. The multivariate analyses described below, however, do provide some evidence on this issue.

**"BEHAVIORAL REGRESSIONS" AND "IMPUTATIONAL REGRESSIONS"**

Clearly from both a statistical perspective and a disclosure perspective, the sample sizes are only barely large enough to support contingency table analyses. Furthermore, we would like to understand how reporting has shifted over time and the potential effects of the changes in the CPS questions, most notably those in 1996 and 2000. Appendix A reports a total of 48 regressions:

- **Three outcomes**: Medi-Cal, welfare, Medi-Cal Only;
- **Four concepts**: Behavioral false negatives (the probability of answering "not enrolled" in the CPS, given that the MEDS indicates enrollment); behavioral false positives (the probability of answering "enrolled" in the CPS, given that the MEDS indicated not enrolled); imputational false negatives (the probability of being enrolled according to the MEDS, given answering "not enrolled" in the CPS); and imputational false positive (the probability of not being enrolled in the MEDS, given answering "enrolled" in the CPS). (See below for a discussion of the distinction between behavioral and imputational regressions.)
- **Four stepwise-regression specifications**: All the variables in the "levels," a pruned set of variables in the levels, interactions for

all the variables included in the levels, and a pruned set of variables in the interactions.

Appendix A provides details of the variable specifications, the stepwise strategy, and the regression coefficients.

These results are crucial for the imputation models that follow, but they are otherwise difficult to interpret. Crucially for our purposes, they confirm the aggregate data and show no trend in the imputation false negatives for Medi-Cal but confirm a strong increase in false negatives for welfare. This is consistent with the tabulations from the unmatched data in Table 2.2.

**CPS IMPUTATIONS**

The previous analyses in this chapter have used what we called the matched sample. To ensure that any lack of congruence resulted from true response errors, we deleted all imputed data from the analysis.

However, we can and did perform the match for all people for whom a SSN was available. We will use this sample to analyze the quality of CPS imputation of program participation. As in the main analysis, we drop those without SSNs, apparently bad matches, and movers. Table 3.5 reports the congruence of the MEDS data and CPS imputed responses. The table has three panels. The first panel considers what the Census refers to as “allocation” of welfare (i.e., imputation for item non-response). This allocation is done by the standard Census hot-deck procedure. Similarly, the second panel considers the “allocation” of Medicaid responses for item non-response. Finally, the third panel considers logical imputation of Medicaid. The CPS automatically imputes Medicaid to anyone who reports the receipt of welfare or SSI.

**Table 3.5**
**MEDS Data for CPS Imputed Records**

| | **MEDS Data** | | |
|---|---|---|---|
| **CPS Value** | **No** | **Yes** | **Total** |
| Welfare Allocated | | | |
| No | 93.96% | 2.85% | 96.82% |
| Yes | 0.66% | 2.52% | 3.18% |
| | 94.63% | 5.37% | 100.00% |
| Medi-Cal Allocated | | | |
| No | 56.00% | 30.27% | 86.28% |
| Yes | 6.24% | 7.48% | 13.72% |
| | 62.24% | 37.76% | 100.0% |
| Medi-Cal Logically Imputed | | | |
| Yes | 51.20% | 48.80% | 100.0% |
| Total | 51.20% | 48.80% | 100.0% |

Note: Cells are percent within the panel. Verified years only.

Because the CPS does not impute for program participation for many people, the sample sizes for the estimates presented in Table 3.5 are small (about 100 cases are imputed over all of our data). Consequently, these results need to be treated with caution. With that caveat, the results suggest that the hot-deck algorithm is under-estimating program enrollment. For welfare, the difference is small. The MEDS suggests that about 5.37 percent of the allocated cases are enrolled; the Census imputes welfare to only 3.18 percent of them. For Medi-Cal the differences are larger. In MEDS, 37.76 percent of the cases are enrolled in Medi-Cal whereas the CPS imputes only 13.72 percent.

The logical imputations add Medicaid to children whose parents report Medicaid. For older children, we can check this imputation against the MEDS. We find that only 38.80 percent (partially imputed to Medi-Cal) actually have Medi-Cal. The imputation is making things worse.

In the matched sample, the logical imputations are about two percent of all cases and the allocations about one percent of all cases. Relative to perfect imputations, the incorrect logical imputations therefore raise the Medi-Cal enrollment rate by about one percentage point. Relative to perfect imputations, the incorrect allocations lower the Medi-Cal enrollment rate by about a fifth of a percentage point.

For welfare, they represent less than a fifth of a percentage point. The effect on overall enrollments of any imputation errors is therefore trivial. In fact, the imputations of welfare participation appear to be quite good.

These results suggest that people who do not answer the Medicaid questions are substantially more likely to have Medicaid/Medi-Cal than the demographically similar households the CPS hot-deck procedure is using for its imputations. In short, item non-response is not random, but the effect on total estimated enrollment is trivial. In contrast, the Medicaid logical imputations are wrong about half the time, increasing estimated Medi-Cal enrollment by about one percentage point, which is about ten percent of true Medi-Cal enrollment. Some additional attention to the Medicaid logical imputations may be appropriate.

**CONCLUSION**

This chapter has considered congruence between the CPS and MEDS data in the best possible matching sample. Even in this sample, the level of congruence is distressingly low and appears to be getting worse over time for welfare.

# 4. EXTRAPOLATING TO THE FULL DATA

The previous chapter analyzed the congruence of responses among those who provided a SSN (what we referred to as the matched sample) and met a set of sample inclusion criteria. However, as Table 3.1 notes, many people do not provide a SSN. Furthermore, not providing a SSN is differential. People who are more disadvantaged are less likely to provide a SSN, but people who are on welfare are slightly more likely to provide a SSN.

The basic problem that is addressed in this chapter is that we do not have SSNs for about half of the sample. We do not want to assume that the responses in the unmatched sample are perfect. Instead, we want to use information from the reporting errors in the matched sample (where we have the MEDS information, treated as truth) to perform better imputations of program enrollment in the unmatched sample. The basic idea is that individuals with characteristics associated with under-reporting in the matched sample are more likely to under-report in the unmatched sample. We estimate a logistic regression model of such reporting errors (both under-reporting and over-reporting) on the matched sample. We then use that model to multiply impute a true response in the unmatched sample; where by multiple imputation we mean that we assign a probability of each response to each individual based on the regression model.

In practice, we have one more piece of information. We can estimate the total number of people in the unmatched sample who are enrolled in a program. To do so, we take the total estimates from the administrative data and subtract the estimates of enrollment in the matched sample (i.e., we use the CPS weights and the MEDS/administrative data information). Our logistic regression models in general under-predict the number of program enrollees in the matched sample. We therefore append a multiplicative adjustment factor. The effect of that adjustment factor is to force the imputed number of program enrollees to exactly match the administrative totals.

The balance of this chapter provides a precise mathematical discussion of the problem and our approach. The discussion in this chapter is extremely formal and technical. Many readers will want to skip to the next chapter where we provide the substantive results.

**THE IDENTIFICATION PROBLEM**

We can conceptualize the CPS matching problem as a table including eight "cells," in terms of total weighted counts. The columns distinguish whether the individual is on the program according to the MEDS (i.e., YES/NO). The rows distinguish both the CPS response and whether the record has a SSN (so it is potentially matchable). The letters name the cells to ease the discussion below.

| | | | **MEDS** | | |
|---|---|---|---|---|---|
| | | | **YES** | **NO** | **Total** |
| CPS | SSN | YES | $A=TP_S$ | $B=FP_S$ | $C=Y_S$ |
| | | NO | $D=FN_S$ | $E=TN_S$ | $F=N_S$ |
| | Absent | YES | $G=TP_A$ | $H=FP_A$ | $I=Y_A$ |
| | | NO | $J=FN_A$ | $K=TN_A$ | $L=N_A$ |
| | | Total | $M=Y_M$ | $N=N_M$ | $O=T$ |

Thus, the subscripts are:

- "S"—SSN present (i.e., a match was in principle possible; in practice, we drop the bad matches as well);
- "A"—SSN absent (i.e., a match is not possible);
- "M"—MEDS.

And the other codes are:

- TP—true positive;
- TN—true negative;
- FN—false negative;
- FP—false positive.

And finally:

- Y—"Yes" (on Medi-Cal/welfare);
- N—"No" (not on Medi-Cal/welfare);
- T—Total.

We treat the MEDS data as "truth." Thus, our goal is to use the MEDS data to "fix" the CPS data. From the records that provided SSNs, we know $TP_S$, $FP_S$, $FN_S$, and $TN_S$. So, we simply adjust the CPS answers to align with the MEDS answers.

The challenge therefore is the unmatchable data—those records for which no SSN was available in the CPS. There, we only know the row totals—$Y_A$, $N_A$,—

and the column totals by subtraction from the "S" sample—$TP_A+FN_A=Y_M-TP_S-FN_S$ and $FP_A-TN_A=Y_M-FP_S-TN_S$. However, there is some additional—we will see, not quite enough—information from the matched sample.

To understand our approach, begin by formally defining the *imputational false positive rate* and the *imputational false negative rate* as:

(4.1) $$\rho^i_{FP} = \frac{FP_S}{TP_S + FP_S} \qquad \rho^i_{FN} = \frac{FN_S}{TN_S + FN_S}$$

where the "i" is for imputation and these are the rates with respect to the CPS answers (as opposed to the behavioral rates in terms of the true behavior as measured in the MEDS that we also considered in the previous chapter).

These imputational rates are in contrast to the behavioral rates of the previous chapter. The tabulations there addressed the behavioral question. Given the true status, what is the probability of a false response? This is not a useful concept for imputation. In the CPS, we observe the potential false response and want to infer the true status. To do so, we want the imputational rates, i.e., the probability that the true status is different than the observed response, given the observed response. The two sets of rates are exactly related. From a complete 2x2 contingency table (i.e., TP, FP, TN, FN), we can compute both sets of rates. From one set of marginals and one set of rates, we can recover the other set of rates. Which rate is more insightful depends on whether we are addressing behavioral questions (as in the last chapter) or imputational questions (as in this chapter).

Then, if we knew these imputational error rates, we could probabilistically impute the data. We would create two pseudo-observations for each observation (dividing the sample weight between the pseudo-observations). So, for example, if an observation reported "Y" in the CPS, that observation would be assigned a "Y" with probability $1-\rho^i_{FP}$ and "N" with probability $\rho^i_{FP}$. Similarly, if an observation reported "N" in the CPS, that observation would be assigned a "N" with probability $1-\rho^i_{FN}$ and a "Y" with probability $\rho^i_{FN}$.

We do not know these rates in the unmatched sample. Furthermore, the rates from the matched data are not directly applicable in the unmatched data. If the rates from the matched sample applied in the unmatched sample, then applying those rates to the unmatched data would recover the actual number of people on Medi-Cal/welfare in the MEDS, i.e.:

(4.2) $$Y_M = TP_S + FN_S + Y_A\left(1-\rho^i_{FP}\right) + N_A$$

However, we have already noted that the under-reporting in the matched sample is not large enough to explain the under-reporting in the full sample.

We have a fundamental non-identification problem: One equation for $Y_M$ and two unknowns—the rates in the unmatched data. Setting one of the rates fixes the other rate.

Given that in net we have under-reporting of Medi-Cal/welfare and false positives are rare (and relatively stable through time), we adopt the simplest rule. We use the false positive rate from the matched data in the unmatched data. We then adjust the false negative rate (by a multiplicative factor, "α") until the implied total count of people on Medi-Cal/welfare in the CPS equals the count in the MEDS (assumed to be truth).

(4.3) $$Y_M = TP_S + FN_S + Y_A\left(1-\rho_{FP}^i\right) + N_A\alpha$$

The left-hand side is the "true" number of individuals on Medi-Cal/welfare from the MEDS. The right-hand side is the "fixed" number of individuals on Medi-Cal/welfare in the CPS. Considering each of those terms in turn:

- $TP_S + FN_S$: The number of people who have Medi-Cal/welfare in the matched sample (true positives plus false negatives).
- $Y_A\left(1-\rho_{FP}^i\right)$: The number of people who report having Medi-Cal/welfare who actually do. We know the number of people who report having Medi-Cal/welfare in the unmatched sample. We estimate the number of these people who actually do have Medi-Cal/welfare using the imputational false positive rate from the matched sample. This is the identifying assumption.
- $N_A\alpha$: The number of people who report not having Medi-Cal/welfare who actually do. Again, we know the number of people who report not having Medi-Cal/welfare in the unmatched sample. Finally, α gives the probability of a false negative in the unmatched data.

Solving for α, the false negative rate in the unmatched sample yields:

(4.4) $$\alpha = \frac{Y_M - TP_S - FN_S - Y_A\left(1-\rho_{FP}^i\right)}{N_A}$$

Except for the false positive rate, each of the terms on the right side is observable. In the numerator, the first term in parentheses is the number of people on Medi-Cal/welfare from the MEDS. The second and third terms are the number of people on Medi-Cal/welfare from the MEDS in the matched sample. The fourth term is the product of the number of number of people in the unmatched sample who claim to have Medi-Cal/welfare. The denominator is the (weighted) number of people in the CPS sample who do not provide a SSN who claim not to have Medi-Cal/welfare. The false positive rates for the unmatched data are not observed, but by assumption we use the value estimated from the CPS. Since the CPS is a sample, each of these concepts should be weighted.[17]

**IMPUTING THE DATA**

This analysis of identification suggests that we are missing one piece of information. However, once we assume that the false positive rate is common in the matched sample and the unmatched sample, we can solve for $\alpha$. Then, knowing $\alpha$ is enough to solve for each of the cells:

$$\begin{aligned} TP_A &= Y_A(1-\rho^i_{FP}) \\ FP_A &= Y_A\rho^i_{FP} \\ FN_A &= N_A\alpha \\ TN_A &= N_A(1-\alpha) \end{aligned} \tag{4.5}$$

Cell counts for the terms in individual years are often too small to allow public release. However, the totals over the whole 11-year period are releasable. To understand our methods, equation 4.7 shows the actual numbers for Medi-cal and welfare respectively, summing over all 11 years (rounded to hundreds of thousands).

$$\begin{aligned} \alpha &= \frac{Y_M - TP_S - FN_S - Y_A\left(1-\rho^i_{FP}\right)}{N_A} \\ &= \frac{10.4-1.2-1.2-1.0}{2.1} = 3.3 \approx 2.960 \end{aligned} \tag{4.6}$$

---

[17] We note that this is the analysis considering the concepts (Medi-Cal, welfare, Medi-Cal Only) separately. It would also be of interest to impute jointly welfare and Medi-Cal. Table 3.3 and the probit regressions reported in Table 3.8 provide the inputs for such an analysis.

We do not perform the full imputation here. The actual imputation would be more complicated than the single imputation attempted here. The single imputation considered here is for a 2x2 table, with two error rates. Fixing one of the error rates is enough to allow computation of the other one from the data.

In contrast, the joint response problem is a 3x3 table, with six distinct error rates. We need to fix four of them to be able to compute adjustment factors for the last two. By analogy, with the approach in the body of the paper, it would be natural to assume that the three upcoding error rates are common. However, that is not sufficient. We still need to fix either the P[W|N] or P[W|MO]. We have seen that both of these errors are common and changing over time, so it is not clear how to proceed.

(4.7)
$$\alpha = \frac{Y_M - TP_S - FN_S - Y_A\left(1-\rho_{FP}^{i}\right)}{N_A} = \frac{13.9-3.6-4.4-1.5}{3.1} = 1.4 \approx 1.415$$

Thus, over the full 11 years, the MEDS has 10.4 million adults on Medi-Cal. The matched CPS data have 1.2 million true positives and 1.2 million false negatives. Using the imputational false positive rates and false negative rates, we would estimate 1.0 million false positives and 2.1 million false negatives in the unmatched sample. To align the CPS totals with the MEDS totals, we need to increase the false negative count by a factor of 3.3/2.960 (i.e., from 2.1 to 6.9/7.0 million, where the first figure is implied by the rounded data and the second figure is implied by the unrounded data).

For welfare, the MEDS has 13.9 million adults on welfare. The matched CPS data have 3.6 million true positives and 4.4 million false negatives. Using the imputational false positive rates and false negative rates, we would estimate 1.5 million false positives and 3.1 million false negatives in the unmatched sample. To align the CPS totals with the MEDS totals, we need to increase the false negative count by a factor of 1.4 (i.e., from 3.1 million to 4.3 million; to the hundreds of thousands, the rounded and unrounded answers are identical, so we only need to report one figure).

**STRATIFYING**

The above analysis is applicable when the population is homogeneous. In reality, the population is heterogeneous. We are able to address this to some extent. We have a small number of variables—calendar year (in principle, also gender and age)—that are measured (nearly) consistently in the MEDS and the CPS. For these variables, we can totally stratify (i.e., we will compute a different value of $\alpha$ for every strata).

In addition to the small number of variables that are common to both data sets, we have many other covariates in the CPS. This allows us to estimate the false negative and false positive rates in the matched sample, not only in terms of the small set of common variables, but also in terms of the larger number of variables in the CPS alone. These are exactly the imputational regressions discussed in Chapter 3.

Using these multivariate models seems particularly important for two reasons. First, many of these variables are likely to be strongly related to Medi-Cal/welfare eligibility and therefore to true Medi-Cal/welfare coverage, e.g., marital status, presence of children in the household, and household earnings. Second, dual coverage (Medicaid and also other, usually private, health insurance) is an issue of substantive interest. As much as possible, we want to correctly impute in the sub-samples with and without private health insurance.

Suppose that within the strata, *s*, we can assign each individual his/her own $\rho$, then we can write our equation for $\alpha$ as:

$$Y_{M,s} = TP_{S,s} + FN_{S,s} + \sum_{j \in Y_A} w_j \left(1 - \rho^i_{FP}[j]\right) + \alpha_s \sum_{k \in N_A} w_k \rho^i_{FN}[k]$$

(4.8)

$$\alpha_s = \frac{Y_{M,s} - TP_{S,s} - FN_{S,s} - \sum_{j \in Y_A} w_j \left(1 - \rho^i_{FP}[j]\right)}{\sum_{k \in N_A} w_k \rho^i_{FN}[k]}$$

In practice, we use the predictions of the imputational probit regression models from the previous chapter to estimate the $\rho$s.[18]

For this project, we have the matched data. However, given the imputational probit model, this approach can also be applied to the CPS public use data by those who do not have the matched data. To see this write

$$Y_{M,s} = \sum_{j \in Y_S \cup Y_A} w_j \left(1 - \rho^i_{FP}[j]\right) + \gamma_s \sum_{k \in N_S \cup N_A} w_k \rho^i_{FN}[k]$$

(4.9)

$$\gamma_s = \frac{Y_{M,s} - \sum_{j \in Y_S \cup Y_A} w_j \left(1 - \rho^i_{fp}[j]\right)}{\sum_{k \in N_S \cup N_A} w_k \rho^i_{fn}[k]}$$

Below, we compute $\alpha$ and $\gamma$ for each strata. Thus, an analyst without access to the matched data could also create an imputed data set.

In what follows, we apply this approach directly to our CPS data. We stratify by year. In practice, the estimates within demographic sub-groups are too small to yield reliable estimates. Table 4.1 presents the resulting estimates for $\alpha$. For adults, the adjustment factors are quite large in the early years. For Medi-Cal, we need to triple or even quadruple the false

[18] The form of the equation in the text is computationally straight-forward. One could argue that it would be more consistent with the probit modeling strategy to include $\alpha$ inside the probit index. Doing so would require a non-linear optimization to compute $\alpha$.

negative rates in the early years, suggesting that the unmatched sample is very different from the matched sample. For welfare, despite the fact that the under-reporting is absolutely more severe, the unmatched sample is closer to the matched sample. The highest adjustment factors are only slightly greater than two. The Medi-Cal Only adjustment factors are even larger than those for Medi-Cal.

**Table 4.1**
**Adjustment Factors α**

| | **Adults** | | | **Children** | | |
|---|---|---|---|---|---|---|
| **Year** | **M** | **W** | **MO** | **M** | **W** | **MO** |
| 1990 | 3.0 | 1.8 | 3.3 | 5.6 | 5.8 | 2.9 |
| 1991 | 4.2 | 2.0 | 4.8 | 8.1 | 7.9 | 3.8 |
| 1992 | 3.7 | 2.2 | 4.1 | 7.0 | 6.9 | 3.5 |
| 1993 | 3.2 | 1.8 | 3.8 | 7.3 | 8.5 | 3.7 |
| 1994 | 3.6 | 1.3 | 4.7 | 5.0 | 5.7 | 2.2 |
| 1995 | 2.9 | 1.6 | 3.3 | 5.3 | 5.4 | 2.4 |
| 1996 | 2.1 | 1.2 | 2.5 | 5.5 | 5.0 | 2.5 |
| 1997 | 2.9 | 1.5 | 3.2 | 5.4 | 5.9 | 2.4 |
| 1998 | 2.1 | 1.3 | 2.9 | 5.2 | 5.5 | 2.7 |
| 1999 | 1.4 | 1.2 | 1.4 | 4.6 | 3.9 | 2.5 |
| 2000 | 1.4 | 0.7 | 1.8 | 7.2 | 6.8 | 3.4 |
| Average | 2.5 | 1.4 | 2.9 | 6.2 | 6.4 | 3.0 |

Over the 11 years covered by our analysis, the adjustment factors shrink. By 2000, the adjustment factor for Medi-Cal is under 1.5; for welfare, under 1; and for Medi-Cal Only, under 2. It is not clear whether these changes over time result from changes in the CPS instrument or from changes in who is receiving welfare. The large drop in 1995 is consistent with the desired effects of the change in the CPS instrument in that year. The drop in 2000 would also be consistent with the changes in the CPS instrument in that year. Unfortunately, the drop seems to date back to 1999, one year too early. The preceding discussion applies to adults, for whom we potentially have a SSN. We do not have SSNs for any children. Following Census practice, we impute from parents to children.[19] For Medi-Cal this is consistent with Census's

[19] See for example the March CPS documentation for 1990 (p. 9-8; http://www.census.gov/apsd/techdoc/cps/cpsmar00.pdf): "After data collection and creation of an initial microdata file, further refinements were made to assign Medicaid coverage to children. In this procedure all children under 21 years old in families were assumed to be covered by Medicaid if either the householder or spouse reported being covered by Medicaid (this procedure was required mainly because the Medicaid coverage question was asked only for persons 15 old and over). All adult AFDC recipients and their children, and

logical imputations. If parents have Medicaid, then children are imputed to have Medicaid. For welfare, this is definitional. Children are never asked about welfare in the CPS. Instead, we impute welfare to both adults and children based on the receipt of public assistance from a welfare program. We then follow the equivalent approach; in other words, we adjust the false negative rate until it aligns the imputed data with the MEDS totals.

We note that the adjustment factors for children are much higher. Furthermore, unlike the adjustment factors for adults, the adjustment factors for children do not fall through time. These adjustment factors are large enough to cast some doubt on the quality of the imputations for children. The adjustments will align the total number of children with the control counts. Our methods impute to children in proportion to the false negative rates. This continues to be a reasonable approach. However, the adjustment factors are so large as to suggest that there is some factor beyond false negatives explaining the under-reporting for children. Whatever it is, the matched data do not identify it.

Given the adjustment factors shown in Table 4.1, we create a multiply-imputed data set. For the matched data, we overwrite the CPS data with the MEDS data. For the unmatched "Yes" responses, we multiply impute based on the false positive rates implied by the probit regression coefficients from the matched sample. For the unmatched "No" responses, we multiply impute based on the product of the false negative rates implied by the probit coefficients from the matched sample and the adjustment factor, $\alpha$, for this survey year.

In practice, we create two data sets, one for the analysis of Medi-Cal (and health insurance) and a second for the analysis of welfare. We do not attempt the full joint imputation of Welfare and Medi-Cal Only. In the subsequent chapters we use the multiply-imputed data sets to obtain a better understanding of how the mis-reporting in the CPS can affect different types of analyses. Specifically, in Chapter 5, we examine how mis-reporting of Medi-Cal receipt affects estimates of the number of uninsured in California and in Chapter 6 we look at how mis-reporting of Medi-Cal and welfare participation affect estimates of program take-up.

---

SSI recipients living in States which legally require Medicaid coverage of all SSI recipients, were also assigned coverage."

# 5. NEW ESTIMATES OF THE UNINSURED

Having characterized the under-reporting problem and described our approach to estimating true rates from the matched data, in this chapter we present the first substantive results of this paper—adjusted health insurance rates, overall and by subgroups.

## DUAL REPORTING

To provide improved estimates of the number of people who are uninsured, the crucial issue concerns dual coverage. We know the number of people with Medi-Cal exactly from the MEDS administrative data. However, our matched data provides no new information on who has private health insurance coverage. If no one had both Medi-Cal and private health insurance, we could compute the number of uninsured as the total population less the MEDS estimate of those on welfare and the CPS estimate of those with private health insurance.

However, dual coverage is possible and not uncommon. First, it is possible that a person has both Medi-Cal and other health insurance in a given month. Second, over the course of a year, some months of Medi-Cal and some months of other insurance are even more likely. In the matched sample, about 23.9 percent of adults who report to the CPS that they have Medi-Cal also report private coverage. (See Table 5.1.) Over the entire sample, the figure for children is also 23.9 percent. For survey year 2000 (the last year for which we have matched data; note that this refers to calendar year 1999), slightly more adults are dually covered (27.6 percent) and slightly fewer children (20.3 percent).

These are the dual-coverage rates for everyone in the matched sample who reports having Medi-Cal. Figure 5.1 demonstrates that they are not the relevant population for the computation of the increase in health insurance when we correct for under-reporting. On net, our imputation moves people from the first row (does not have Medi-Cal) to the second row (has Medi-Cal). For people with other health insurance (OHI; the left column of Table 5.1), there is no net increase in health insurance/decrease in uninsurance. For people without other health insurance (the right column of Table 5.1), there is a net increase in health insurance/decrease in uninsurance.

The previous tabulations implicitly assume that those to whom we impute Medi-Cal are like those with Medi-Cal. However, this seems unlikely. We have already seen that reporting no Medi-Cal in the CPS when one actually had Medi-Cal varies with the intensity of Medi-Cal in the previous year. As the number of months of Medi-Cal enrollment drops, the probability of a false negative increases; and false negatives are the population to whom we are trying to impute Medi-Cal.

**Figure 5.1**
**Dual Coverage Rates and Adjusting Total Health Insurance Coverage**
**(OHI: Other—non-Medi-Cal—Health Insurance)**

| | | CPS Reports of OHI | |
|---|---|---|---|
| | | **No** | **Yes** |
| Medi-Cal Enrollment | No | A: Uncovered | B: OHI Only |
| | Yes | C: Medi-Cal Only | D: Dual Coverage |

Furthermore, it seems plausible that such people are more likely to have OHI. One reason to drop Medi-Cal is gaining private insurance. In the matched sample, we can identify such false negatives. Indeed, they have higher rates of OHI (see Table 5.1) and the difference is non-trivial. For example, for adults over the entire period, 23.9 percent of those with Medi-Cal also have OHI; for the false negatives, the figure is a quarter higher at 32.4 percent.

These tabulations are informative for three imputation methods possible with only the public-use file and aggregate tabulations from the administrative data. To understand the argument, we introduce some new notation:

(5.1) $$U = T - OHI - MC + DC$$

The number of uninsured individuals (*U*) can be computed as the total population (*T*) less the count of those with other (non-Medi-Cal) health insurance (*OHI*), less the count of those with Medicaid insurance (*MC*), and adding back in those with dual coverage—other health insurance and Medicaid *(DC)*. The results of the previous chapter imply that the CPS estimate of *MC* is much too small.

Given this formulation, the three imputation methods are:

Method 1: *Raw CPS Data* $U = T - OHI_{CPS} - MC_{CPS} + DC_{CPS}$: Since Medicaid/Medi-Cal is seriously under-reported, simply using the raw data will yield an estimate of the number of uninsured that is too high.

Method 2: *Simple Administrative Data Adjustment*

$U = T - OHI_{CPS} - MC_{CPS} + DC_{CPS} - (MC_{Admin} - MC_{CPS})$: Since there is significant dual coverage, estimating the number of uninsured as the total population less the CPS estimate of OHI and the administrative data estimate of Medicaid/Medi-Cal will yield an estimate of the number of uninsured that is too low.

Method 3: *Public-Use File Adjustment for Dual Coverage*

$U = T - OHI_{CPS} - MC_{CPS} + DC_{CPS} + (1 - \delta_{CPS})(MC_{Admin} - MC_{CPS})$: The simple administrative data adjustment implicitly assumes no dual coverage. However, we can generate a rough estimate of dual coverage from the Public Use File, $\delta_{CPS}$, as the fraction of those reporting Medicaid in the CPS who also report OHI.

This third Method will be appropriate if the dual-coverage rates among those reporting Medi-Cal (i.e., the union of the true positives and the false positives) equaled the dual-coverage rates among the false negatives. However, it seems plausible that it is exactly people who had private health insurance (perhaps at the end of the year) who would not report the Medi-Cal they had (perhaps at the beginning of the year). Tabulations from the matched sample are consistent with this hypothesis. (See Table 5.2). This suggests using the dual-coverage rates from the false negatives in the matched sample in the adjustment above and that doing so will yield a larger estimate of the number of uninsured.

This analysis is only an approximation. The full analysis has 16 cells: (matched/unmatched) x (private health insurance yes/no) x (TP, FN, FP, TN) and the total number of uninsured is:

$$U = TN[S,N] + FP[S,N] + TN[A,N] + FP[A,N]$$

where the function arguments are S/A for SSN present/SSN absent and Y/N for private health insurance/no private health insurance. Then, in both the matched and unmatched samples, there are two ways to be uninsured: (1) true negative for Medicaid and no private health insurance; or (2) false positive for Medi-Cal and no private health insurance.

Given the assumption that the MEDS data is truth for Medi-Cal and the CPS data is truth for private health insurance (we can do no better), we know the first two terms exactly and the rates from the matched sample are plausibly informative about the last two terms (i.e., rates in the unmatched sample). This analysis suggests two more estimators of the number of uninsured:

Method 4: *FN Adjustment for Dual Coverage* $U = T - OHI_{CPS} - MC_{CPS} - DC_{CPS} + \left(1-\delta_{FN}\right)\left(MC_{Admin} - MC_{CPS}\right)$: Since the major concern is dual coverage among the false negatives, it seems preferable to use the rate from the matched sample's FNs, $\delta_{FN}$. Of course, this is only possible with the matched data.

Method 5: *Full Imputation Model*: The imputation models we estimated in the previous chapter include (control for) private health insurance coverage. They thus control for dual coverage in all four cells (TP, FN, FP, TN).

**OUR APPROACH AND OUR RESULTS**

Table 5.2 reports the results of these five adjustments for adults and children, pooled over the entire file and for the last year (the 2000 survey year, corresponding to calendar year 1999). Consider first adults over the entire panel. For this group, under-reporting is about 4.1 percent of the total. The raw CPS estimate of uninsurance is 23.4 percent. Simply adding back in the under-reporting cuts the estimated fraction of uninsured to 19.3 percent.

**Table 5.1**
**Estimates of Dual Coverage and Uninsurance**

| | **Pooled** | | | | **2000/1999** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Adults** | | **Children** | | **Adults** | | **Children** | |
| | **DC** | **UI** | **DC** | **UI** | **DC** | **UI** | **DC** | **UI** |
| Under-reported | | 4.1% | | 12.7% | | 4.1% | | 11.8% |
| Method 1 | | 23.4% | | 23.2% | | 23.2% | | 21.9% |
| Method 2 | 0.0% | 19.3% | 0.0% | 10.6% | 0.0% | 19.1% | 0.0% | 10.1% |
| Method 3 | 23.9% | 20.3% | 17.0% | 13.6% | 27.6% | 20.2% | 20.3% | 12.5% |
| Method 4 | 32.4% | 20.6% | 32.4% | 14.7% | 31.2% | 20.4% | 40.2% | 14.9% |
| Method 5 | 32.1% | 20.6% | 37.0% | 15.3% | 34.4% | 20.5% | 45.0% | 15.4% |

Note: DC-estimate of dual coverage used to adjust estimates of uninsurance; UI-fraction of the population uninsured.

Rows are:

Under-reported: Fraction of the population under-reported (MEDS-CPS/Total Population).

Method 1: Raw CPS Data

Method 2: Medi-Cal from MEDS, implicitly assuming no double counting

Method 3: Medi-Cal from MEDS, using dual coverage rate among those in the matched sample who report Medi-Cal
Method 4: Medi-Cal from MEDS, using dual coverage rate among false negatives in the matched sample
Method 5: Full multivariate imputation (see below)

This estimate is clearly too small. It assumes no dual coverage. Using the rate of dual coverage among those reporting Medi-Cal (Method 3) yields a slightly higher estimate of the fraction uninsured, 20.3 percent. Using either the rate of dual coverage among the false negatives (Method 4) or the full imputation model (Method 5) yields estimates of uninsurance of 20.6 percent, much lower than the simple CPS estimate (23.4 percent) and slightly higher than simply adding back in the under-reporting (20.3 percent). Thus, for adults, all that matters is that we make some adjustment for under-reporting. Differences between the various adjustments are quite small.

For children over the entire period, under-reporting is a much bigger problem, 12.7 percent of all children (versus 4.1 percent of all adults), and the divergence between the estimates of dual coverage is larger. Therefore, the effect of the different correction methodologies varies. The raw CPS estimate of uninsurance is 23.2 percent. Simply subtracting off the under-reporting (Method 2) cuts that estimate to 10.6 percent. The three corrections for dual coverage (Method 3, Method 4, and Method 5) successively raise the uninsurance rates. Our preferred estimate is from Method 5 (the last row). It suggests true uninsurance rates of 15.3 percent. This estimate is considerably lower than the unadjusted estimate of 23.2, but considerably higher than the no dual coverage estimate of 10.6.

The right side of the table gives the equivalent figures for the last year of our data, survey year 2000 referring to calendar year 1999. Dual-coverage rates are slightly higher, but the qualitative story and the estimates of uninsurance are similar.

**DISCUSSION**

This analysis suggests that under-reporting of Medi-Cal seriously inflates our estimates of the size of the uninsured population. In addition, dual coverage is sufficiently common that ignoring it results in a significant underestimate of the size of the uninsured population even after we correct for under-reporting.

Thus, some correction for dual coverage is needed. On a priori grounds, this dual-coverage estimate from the false negatives in the matched sample seems to be a preferable estimate of the unmatched false negatives than the simple CPS public-use file estimate. Evidence from the matched sample suggests that dual coverage is slightly more common among false negatives and also more common than the simple CPS public-use file estimate. Using this plausibly better estimate yields a slightly higher estimate of the uninsured. Using the full imputation model yields an estimate that is slightly higher. The size of the corrections will vary with the magnitude of the under-reporting and the amount of dual coverage.

Our best estimates of dual coverage suggest that the raw CPS figures over-estimate uninsurance by about three percentage points for adults and eight percentage points for children. These are sizable over-estimates. They imply that the problems of lack of health insurance are non-trivial, but that they are considerably smaller than what would be implied by the simple CPS tabulations.

# 6. NEW ESTIMATES OF MEDI-CAL AND WELFARE TAKE-UP RATES

If the only question of interest was: "How many people are enrolled in welfare/Medicaid?", we could answer that question directly from the administrative data. However, both researchers' and policy makers' interest typically goes well beyond the number of people enrolled to concerns about take-up rates. The question of interest is what share of the target population is actually enrolled in the program of interest. This is a rate that cannot be measured with administrative data. While the numerator (i.e., the number of people enrolled) is available in the administrative data, the denominator (i.e., the number of people in the target population) is not, because the administrative data only includes information on those actually enrolled in the program.

Therefore, to estimate take-up rates, analysts generally turn to survey data for both pieces of information (i.e., the number of people in the target population and the number of people enrolled). Unfortunately, we have seen that actual enrollment is seriously under-reported in survey data, so take-up rates based on these data will also be under-reported. Furthermore, we have seen that under-reporting is not random. Non-reporting is more common among those who are covered for less of the year and closer to the border of eligibility.

In this chapter, we use the adjusted California CPS data based on the analyses and methods described in the previous chapters to generate new estimates of take-up rates. We note that the estimates here are not pure take-up rates. Pure take-up rate estimates would attempt to impute eligibility for Medi-Cal based on all of the survey information and Medi-Cal program rules. Here, we perform only rudimentary take-up computations (i.e., the fraction of a demographic sub-group enrolled) without attempting a full eligibility simulation.

## POOLED RESULTS

We begin by pooling across all of the years in our analysis, 1990-2000. Table 6.1 presents our basic results for Medi-Cal, in the format used for all the tables that follow. The left panel refers to adults (15-65); the right panel refers to children (0-14). The rows consider subgroups related to Medi-

Cal and welfare eligibility: everyone, males, females, single women with children—overall and by poverty status. For each group, we report the unadjusted CPS take-up rate (the ratio of enrollees to the population), the adjusted rate (after our imputations from the matched data), and the “Delta”—the increase with imputation (the ratio of the adjusted to the unadjusted take-up rates minus one; note that the ratio is computed from the underlying figures with more significant digits; it thus will differ from what would be computed using the “Raw” and “Imputed” columns).

**Table 6.1**
**Take-Up Rates: Unadjusted, Adjusted, Discrepancy**
**Medi-Cal, Pooled Years**

| | Adults | | | Children | | |
|---|---|---|---|---|---|---|
| | **Raw** | **Imputed** | **Delta** | **Raw** | **Imputed** | **Delta** |
| All | 10% | 14% | 44% | 24% | 36% | 54% |
| Male | 7% | 10% | 39% | 23% | 36% | 52% |
| Female | 12% | 18% | 46% | 24% | 37% | 55% |
| SW w/kids | 28% | 39% | 41% | 24% | 37% | 55% |
| SW w/kids <50% FPL | 48% | 65% | 36% | 45% | 70% | 54% |
| SW w/kids 50%-100% FP | 60% | 74% | 23% | 53% | 71% | 34% |
| SW w/kids 100%-150% FP | 40% | 55% | 39% | 34% | 52% | 52% |
| SW w/kids 150%-200% FP | 21% | 41% | 91% | 17% | 44% | 154% |
| SW w/kids >200% FPL | 8% | 14% | 79% | 6% | 10% | 70% |

Note: “Raw” is the unadjusted CPS estimate; “Imputed” is the adjusted CPS estimate, based on the multiply-imputed data set; and “Delta” is the percentage (not percentage point) increase in estimated take-up with imputation.

Overall Medi-Cal take-up increases by about half when we adjust for under-reporting using our imputation model. The increases are slightly smaller for adults (44 percent) and slightly larger for children (54 percent). Consistent with an explanation of under-reporting in terms of stigma, the increases are smallest for those in poverty, largest for those between one and two times poverty, and large for those at more than twice poverty.

We note the anomalous result that, even after adjustment, adults at less than half the poverty line have lower take-up rates than those between half the poverty line and the poverty line. Welfare is considered income in the computation of the poverty line. In California, welfare will take a family to more than half the poverty line. Therefore, the families at less than half the poverty line are unlikely to have contact with the welfare system.

Table 6.2 presents comparable estimates for welfare. Consistent with earlier results, compared to Medi-Cal, the levels of take-up are lower and the adjustments have a larger effect. The average adjustment more than doubles take-up rates. Again, the adjustments are larger for children, smaller for those in poverty, and larger for those out of poverty.

**Table 6.2**
**Take-Up Rates: Unadjusted, Adjusted, Discrepancy**
**Welfare, Pooled Years**

| | Adults | | | Children | | |
|---|---|---|---|---|---|---|
| | **Raw** | **Imputed** | **Delta** | **Raw** | **Imputed** | **Delta** |
| All | 3% | 6% | 109% | 11% | 25% | 132% |
| Male | 1% | 4% | 355% | 10% | 24% | 136% |
| Female | 5% | 9% | 70% | 11% | 25% | 129% |
| SW w/kids | 15% | 24% | 58% | 11% | 25% | 129% |
| SW w/kids <50% FPL | 25% | 43% | 67% | 22% | 57% | 159% |
| SW w/kids 50%-100% FP | 41% | 55% | 34% | 32% | 60% | 87% |
| SW w/kids 100%-150% FP | 20% | 33% | 60% | 14% | 32% | 128% |
| SW w/kids 150%-200% FP | 9% | 19% | 119% | 5% | 21% | 288% |
| SW w/kids >200% FPL | 3% | 6% | 134% | 1% | 4% | 287% |

Table 6.3 presents comparable estimates for Medi-Cal only. Our earlier analysis suggested little net under-count of Medi-Cal Only, apparently because many people with welfare, report Medi-Cal Only. Consistent with that earlier analysis, the adjustment deltas are smaller than for either welfare or Medi-Cal, on average 15 percent for adults and negative 6 percent for children. Consistent with the mis-reporting of welfare explanation, the corrections are small for the poor in single female families, the population likely to be eligible for welfare. For better-off families, the corrections are larger, consistent with true under-reporting of Medi-Cal only in the population that is usually ineligible for welfare.

**Table 6.3**
**Take-Up Rates: Unadjusted, Adjusted, Discrepancy**
**Medi-Cal Only, Pooled Years**

| | Adults | | | Children | | |
|---|---|---|---|---|---|---|
| | **Raw** | **Imputed** | **Delta** | **Raw** | **Imputed** | **Delta** |
| All | 7% | 8% | 15% | 15% | 14% | -6% |
| Male | 6% | 6% | 1% | 15% | 14% | -7% |
| Female | 7% | 9% | 27% | 15% | 14% | -6% |
| SW w/kids | 13% | 14% | 15% | 15% | 14% | -6% |
| SW w/kids <50% FPL | 23% | 22% | -4% | 32% | 29% | -7% |
| SW w/kids 50%-100% FP | 19% | 20% | 5% | 28% | 25% | -10% |
| SW w/kids 100%-150% FP | 19% | 23% | 18% | 24% | 22% | -6% |
| SW w/kids 150%-200% FP | 13% | 20% | 60% | 15% | 20% | 33% |
| SW w/kids >200% FPL | 5% | 6% | 22% | 5% | 4% | -20% |

Table 6.4, Table 6.5, and Table 6.6 present the equivalent results for the last year of our data, the 2000 survey referring to the 1999 calendar year. Consistent with our basic analysis of under-reporting in Chapter 2, the results are similar for Medi-Cal and Medi-Cal Only. However, for welfare, there has been a sharp increase in non-reporting. Correspondingly, the deltas are much higher in 2000 than in the pooled sample, both overall for adults (169 percent versus 109 percent) and for children (263 percent versus 132 percent).

**Table 6.4**
**Take-Up Rates: Unadjusted, Adjusted, Discrepancy**
**Medi-Cal, 2000 Survey/1999 Calendar Year**

| | **Adults** | | | **Children** | | |
|---|---|---|---|---|---|---|
| | **Raw** | **Imputed** | **Delta** | **Raw** | **Imputed** | **Delta** |
| All | 9% | 13% | 42% | 24% | 35% | 50% |
| Male | 7% | 9% | 37% | 24% | 35% | 49% |
| Female | 11% | 17% | 45% | 23% | 35% | 51% |
| SW w/kids | 25% | 36% | 46% | 23% | 35% | 51% |
| SW w/kids <50% FPL | 49% | 59% | 21% | 47% | 65% | 38% |
| SW w/kids 50%-100% FP | 48% | 63% | 31% | 52% | 67% | 30% |
| SW w/kids 100%-150% FP | 41% | 55% | 34% | 37% | 51% | 38% |
| SW w/kids 150%-200% FP | 24% | 45% | 84% | 22% | 54% | 139% |
| SW w/kids >200% FPL | 8% | 16% | 104% | 7% | 13% | 72% |

**Table 6.5**
**Take-Up Rates: Unadjusted, Adjusted, Discrepancy**
**Welfare, 2000 Survey/1999 Calendar Year**

| | **Adults** | | | **Children** | | |
|---|---|---|---|---|---|---|
| | **Raw** | **Imputed** | **Delta** | **Raw** | **Imputed** | **Delta** |
| All | 2% | 4% | 169% | 5% | 19% | 263% |
| Male | 0% | 3% | 523% | 6% | 19% | 228% |
| Female | 3% | 6% | 117% | 5% | 20% | 304% |
| SW w/kids | 8% | 18% | 123% | 5% | 20% | 304% |
| SW w/kids <50% FPL | 19% | 35% | 82% | 14% | 54% | 286% |
| SW w/kids 50%-100% FP | 17% | 37% | 114% | 10% | 45% | 339% |
| SW w/kids 100%-150% FP | 17% | 31% | 82% | 10% | 30% | 190% |
| SW w/kids 150%-200% FP | 5% | 16% | 193% | 6% | 22% | 246% |
| SW w/kids >200% FPL | 1% | 6% | 482% | 0% | 3% | **3967%[?]** |

**Table 6.6**
**Take-Up Rates: Unadjusted, Adjusted, Discrepancy**
**Medi-Cal Only, 2000 Survey/1999 Calendar Year**

| | **Adults** | | | **Children** | | |
|---|---|---|---|---|---|---|
| | **Raw** | **Imputed** | **Delta** | **Raw** | **Imputed** | **Delta** |
| All | 8% | 9% | 14% | 20% | 18% | -9% |
| Male | 6% | 7% | 7% | 20% | 18% | -6% |
| Female | 9% | 10% | 20% | 20% | 17% | -11% |
| SW w/kids | 17% | 17% | 2% | 20% | 17% | -11% |
| SW w/kids <50% FPL | 29% | 24% | -20% | 40% | 29% | -28% |
| SW w/kids 50%-100% FP | 31% | 27% | -13% | 46% | 38% | -17% |
| SW w/kids 100%-150% FP | 24% | 23% | -6% | 29% | 25% | -16% |
| SW w/kids 150%-200% FP | 19% | 27% | 44% | 19% | 28% | 46% |
| SW w/kids >200% FPL | 7% | 8% | 25% | 7% | 7% | -10% |

**DISCUSSION**

This chapter has considered the effect of correcting for under-reporting of program enrollment in sub-populations using our imputation model for program take-up rates. Our earlier chapters found substantial under-reporting. That under-reporting yields large under-estimates of take-up rates. Furthermore, the variation in take-up rates is not simple. Overall, the corrections appear to be smallest for those in deepest poverty and larger for those with only borderline eligibility. This pattern is consistent with greater stigma for the borderline eligible. It is also consistent with the borderline-eligible only being enrolled for part of the year; we have seen that those enrolled for part of the year are less likely to report enrollment in the CPS.

# 7. CONCLUSION

This report describes analyses of matched CPS-MEDS data. The CPS data are known to have substantial under-reporting of Medi-Cal enrollment and even larger under-reporting of welfare enrollment. The matched data—along with some auxiliary assumptions—generate adjusted and improved estimates of who is covered by Medi-Cal in the CPS and of true health insurance coverage rates. In brief, we find that adjusting substantially cuts the estimates of the uninsured population and substantially increases estimates of take-up rates.

Given that these results confirm that the CPS data significantly undercounts enrollment, do these results suggest any solution? Happily the answer appears to be "yes." Non-reporting is differential, but the differentials are second-order compared to the non-reporting itself. Simple ratio adjustments with a simple correction for dual coverage (e.g., from those in the CPS who report Medi-Cal coverage) are likely to eliminate most of the bias.

Such simple ratio adjustments can be computed from unmatched tabulations from the CPS and the MEDS. These under-count rates do vary over time. Thus, current official tabulations are needed. CDHS and CMS already publish some such tabulations. To correct the CPS, ideally one would use tabulations slightly different from those currently published. The necessary tabulations would consider any receipt in the past year, with consistent breaks by program and age. CDHS could easily generate the requisite tabulations.

# APPENDIX A. DETAILED NOTES ON FILE CONSTRUCTION AND MATCHING

This appendix describes in detail the procedures we used to create our analysis files. We begin by describing the raw data files we received, how we merged them together, and the results of our efforts to eliminate false matches. This appendix concludes with a discussion of the congruence of race/ethnicity coding across the two data sources.

## A.1. THE RAW DATA

This project's analyses are made possible by the availability of scrambled SSNs (referred to by Census as PIKs) on both the CPS data and on the MEDS data. Specifically, we received three files:

- *MEDS File*: We received an extract from the Medi-Cal Eligibility Data System (MEDS). The MEDS is the official roster of those on Medi-Cal in California. Conceptually the file contained one record for each person who appeared in the MEDS data in any month between January 1987 and December 2002.[20] Given its purpose, anyone who was covered by Medi-Cal during this period (including welfare/AFDC/TANF/CalWORKs) should have a record in the file. The record contains basic demographics (gender, date of birth, race/ethnicity, language) and for each month January 1989 to December 2000[21] the aid code (i.e., type of Medi-Cal coverage), eligibility status, case id, county-of-residence, and zip code.

---

[20] In fact, to minimize file size, the MEDS data arrived in three files. To understand the three files it is useful to note that the underlying MEDS file was created from archived version of the MEDS created every six months (the "December cut" and the "June cut"), each with 15 months of history. The first file had the time invariant data (gender, date of birth, race/ethnicity, language), taken from the most recent file. The second file had the invariant information from an underlying MEDS "cut file" (county-of-residence, case id, zip code), one record for each person, for each "cut file." The third file had the information that varied every month. The MEDS "cut files" are overlapping. We used the information from the most recent file.

[21] Note that we have monthly data for a narrower window (January 1999 to December 2000) than the period over which one would have needed to have been covered by Medi-Cal to be in the file at all (January 1997 to December 2002). The net result is that people who were covered by Medi-Cal in the broader window, but not in the narrower window, will have MEDS records that never show Medi-Cal coverage (all "zeros"). See the next section for a discussion of the implications of this distinction.

Finally, the file contained a Census generated PIK (Protected Identification Key). This file contained 22,848,715 persons.

- *CPS File*: We created the CPS file directly from the March Public Use Files. This file contained 1,588,115 observations from 1990 to 2000. The CPS rotation group structure implies that most people will appear in two successive CPS files. This observation count thus counts such people twice. We have made no correction for that correlation.
- *Cross-walk File*: For every person in the March CPS for interview years 1990 to 2000, the cross-walk file contained the PIK (Protected Identification Key), the CPS Household Number, the CPS Person Number. The file contained 1,347,282 observations (including interviews inside and outside of California).

Unfortunately, the Cross-walk Files we received were not consistent across years. For every individual 15 and older, the CPS interviewer requests a Social Security Number (SSN). In practice, not everyone supplies a SSN; and not all of those SSNs that are reported are correct (some due to simple errors of memory or transcription; some due to deliberate obfuscation). As far as we have been able to ascertain, over the CPS years we are analyzing (1990-2000), there have been no significant changes in the procedures for the collection of SSNs, nor for the verification of SSNs.

We note that the CPS files we use are of two types. Some of the files are "unvalidated." This apparently means that the SSNs from which the PIKs were constructed are as recorded by the Census interviewer.

This is in contrast to "validated" files. "Validated" files differ in two ways. First, those SSNs that were provided were checked against Social Security Administration SSN records (apparently checking for a match on name, gender, and birth date). When the CPS provided information that did not match the information in the SSA records for that SSN, the SSN was dropped (even though a SSN had been provided).

Second, those individuals who did not or could not provide a SSN were asked for permission to use SSA files to impute a SSN. When permission was granted, the name, gender, and birth date provided were matched against SSA records. When a match was found, that SSN was appended to the record (even though no SSN had been provided at the interview).

As we note in the body of the report. In validated files, a higher percentage of records have SSNs and a lower fraction of the matches are "bad."

We also note that our analysis would be easier and of higher quality if we had a consistent time series (ideally, all validated; but alternatively all unvalidated). Presumably, unvalidated versions of the files for each year once existed. We were, however, informed that unvalidated versions of the validated files were not available; and that, Census could not provide validated versions of the currently unvalidated data within the time frame of the project (i.e., in six to nine months).

**A.2. MATCHING THE CPS AND THE MEDS**

In principle, the cross-walk file should allow us to link MEDS records to CPS individuals. The link, however, will never be complete.

For example, in order to complete the link, we must have a SSN for each CPS individual. However, as noted above, not everyone has a SSN in the files we received. Individuals under 15 years of age were not asked to provide SSNs.[22] Many of those asked to provide SSNs were unable to do so or refused. Only some of those asked gave permission for SSA to impute a SSN, and only for some of those people who gave permission, did SSA actually provide a SSN.

**A.3. VERIFYING MATCHES**

The previous discussion concerns mechanical matches of records in the two files. As we discuss below in detail, there must be false non-matches in the matching process: People who were actually enrolled in Medi-Cal, but for whom we cannot match MEDS and CPS records. Given the non-trivial rates of not providing a SSN, it seems likely that most of these false non-matches are people who refused to provide a SSN. In addition, some people are likely to have provided an incorrect SSN that did not match to any MEDS record (and which was not caught by the SSA validation process, perhaps because this was a year in which no validation was done). There does not appear to be anything that we can do directly about this. However, we discuss below our indirect adjustments for this issue.

[22] Given the CPS rolling panel structure, it might be possible to recover SSNs for half of the 14 year olds, by using conventional CPS matching methods (to match the SSN provided at their second March interview (when they were 15) to their first March interview (when they were 14). We have not done so.

There is also the possibility of false matches: People who provided a SSN that matched to the MEDS, but for whom the two records do not represent the same person. In unvalidated years, this might represent people who gave the wrong SSN (a memory error, a transcription error, or deliberate obfuscation). In validated years, such errors should almost always have been caught by comparisons of name, gender, and age between the CPS responses and the SSA files.

To assure that both the CPS and MEDS records with the same SSN truly referred to the same individual, we verified the correspondence of gender and age between the MEDS data and the CPS data (where we use the term "verified" to refer to our cross-check and the term "validated" to refer to SSA's cross-check). We considered matches to be verified if gender matched and age differed by no more than one year (i.e., we did not require an *exact* match on age). Matches not meeting both of these criteria were deemed false or bad matches and dropped at this verification stage.

In such cases where the SSNs match, some (perhaps most) of the failure to "verify" is probably caused by incorrect recording of SSN, gender, or age in the MEDS or in the CPS. Note that in validated years, the validation process should have caused the SSN to be dropped from CPS records for which the SSN, gender, and age information did not match the SSA administrative data. The net result should have been higher quality SSN, gender, and age information in the CPS records with validated SSNs.

Table A.1 provides some additional information on the results of the verification. The table stratifies by validated and unvalidated years. For each set of years, we report the distribution of correspondence by gender and age. Note that the match rates improve with validation. For the validated data, we accept 95 percent of the matches; while for the unvalidated data, we accept 92 percent of the matches. The matches deemed rejected are entered in italics.

**Table A.1**
**Age Differential (CPS Age – MEDS Age)**
**(Percent within Validation Status)**

| **Gender** | **Validated** | | **Unvalidated** | |
|---|---|---|---|---|
| **Match** | **Yes** | **No** | **Yes** | **NO** |
| -- | *1%* | *0%* | *2%* | 2% |
| -1 | 2% | | 4% | |
| 0 | 91% | *1%* | 84% | 1% |
| +1 | 2% | | 4% | |
| ++ | *2%* | *1%* | *3%* | 0% |

Note: Rows are age difference (CPS Age – MEDS Age). CPS age is as per response at survey; MEDS age computed based on birth date and CPS interview date.

“–“ MEDS age is greater than CPS age by two or more years; “-1” MEDS age is greater than CPS age by exactly one year; “0” MEDS age equals CPS age; “+1” CPS age is older than MEDS age by exactly one year; “++” CPS age is older than MEDS age by more than two or more years for gender match and discrepancy for no gender match.

## BIBLIOGRAPHY


Alpha Center. "The Impact of Federal Welfare Reform on Medicaid: Medicaid Enrollment Declines as De-linking Yields Eligibility Confusion." State Initiative in Health Care Reform. R30, pp. 4-10, Washington, DC, 1999.

Alpha Center. A Survey of Surveys: What Does it Take to Obtain Accurate Estimates of the Uninsured? *State Coverage Initiatives*, No. 1, March 2000.

Beauregard, Karen M., Susan K. Drilea, and Jessica P. Vistnes. "The Uninsured in America–1996." *MEPS Highlights*, No. 1, May 1997.

Bennefield, R. L. "Health Insurance Coverage: 1995." *Current Population Reports*, pp. 60-199. Washington, DC: U.S. Bureau of the Census, 1996.

Bennefield, Robert L. "A Comparative Analysis of Health Insurance Coverage Estimates: Data from CPS and SIPP." Proceedings from the 1996 Joint Statistical Meetings, American Statistical Association, Chicago, August 1996c.

Bennefield, Robert L. "Dynamics of Economic Well-Being, Health Insurance, 1993 to 1995." *Current Population Reports*, Washington, DC: U.S. Department of Commerce, Economics and Statistics Administration, July 1998, pp. 70-64.

Bennefield, Robert L. "Who Loses Coverage and for How Long?" *Current Population Reports*, Washington, DC: Census Bureau, May 1996b, pp. 70-54.

Bennefield, Robert, L. "Health Insurance Coverage 1995." *Current Population Reports*, Washington, DC: Census Bureau, pp. 60-195, 1996a.

Bilheimer, Linda T. "CBO Testimony on Proposals to Expand Health Coverage for Children." Testimony before the Subcommittee on Health, U.S. House of Representatives, Committee on Ways and Means, Washington, DC, April 8, 1997.

Blumberg, S. J. and M. L. Cynamon. "Misreporting Medicaid Enrollment: Results of Three Studies Linking Telephone Surveys to State Administrative Records." Paper delivered at the 7th Conference on Health, Survey Research Methods, Williamsburg, VA, 1999.

Bradburn, N. M., L. J. Rips, and S. K. Shevell. "Answering Autobiographical Questions: The Impact of Memory and Inference on Surveys." *Science*, Vol. 236, No. 4798, pp. 157-161, 1987.

Call, K. T., A. S. Somrners, R. Feldman, T Rockwood, Y. Jonk, and B. Dowd. Minnesota Health Access Survey, 1999 Final Report. University of Minnesota, School of Public Health, Division of Health Services Research and Policy, Minneapolis, MN, 1999.

Call, Kathleen Thiede, Gestur Davidson, Anna Stauber Sommers, Roger Feldman, Paul Farseth, and Todd Rockwood. "Uncovering the Missing Medicaid Cases and Assessing their Bias for Estimates of the Uninsured." *Inquiry*, Vol. 38, No. 4, Winter 2001/2002, pp. 396-408.

Card, David, Andrew K. G. Hildreth, and Lara D. Shore-Sheppard. "The Measurement of Medicaid Coverage in the SIPP: Evidence from California, 1990-1996." Cambridge, MA: National Bureau of Economic Research (NBER), Working Paper No. 8514, October 2001. [http://papers.nber.org/papers/w8514.pdf]

Congressional Budget Office. *How Many People Lack Health Insurance and For How Long*, 2003.

Cutler, David and Jonathan Gruber. "Does Public Insurance Crowd out Private Insurance?" *Quarterly Journal of Economics*, CXI, 1996, pp. 391-430.

Cutler, David and Jonathan Gruber. "Medicaid and Private Insurance: Evidence and Implications." *Health Affairs*, Vol. 17, No. 1, 1997, pp. 194-200.

Dubay L. and G. Kenney. "Lessons from the Medicaid Expansions for Children and Pregnant Women: Implications for Current Policy." Statement for Hearing on Children's Access to Health Coverage, Subcommittee on Health, U.S. House Committee on Ways and Means, April 8, 1997.

Dubay, L., and G. Kenney. "Effects of Medicaid Expansions on Insurance Coverage of Children." *Future of Children*, Vol. 6, No. 1, pp. 152-161, 1996.

Fay, R. E. "An Analysis of Within-Household Undercoverage in the Current Population Survey." Paper delivered at the U.S. Bureau of the Census Annual Research Conference, Washington, DC, 1989.

Fronstin, P. "*Expanding Health Insurance for Children: Examining the Alternatives.*" Washington, DC: Employee Benefit Research Institute, 1997a.

Fronstin, P. "*Trends in Health Insurance Coverage.*" Washington, DC: Employee Benefit Research Institute, 1997b.

Fronstin, Paul and Rachel Christensen. "The Relationship Between Income and the Uninsured." *EBRI Notes*, No. 3, Employee Benefit Research Institute, March 2000, pp. 1-4.

Fronstin, Paul. "Sources of Health Insurance and Characteristics of the Uninsured: Analysis of the March 1996 Current Population Survey." *EBRI Issue Brief*, No. 179, Washington, DC: EBRI, November 1996.

Fronstin, Paul. "Trends in Health Insurance Coverage." *EBRI Issue Brief*, No. 185, Washington, DC: EBRI, May 1997.

Giannarelli, L. *An Analyst's Guide to TRIM2: The Transfer Income Model, Version 2.* Washington, DC: The Urban Institute, 1992.

Groves, R. M. *Survey Errors and Survey Costs.* New York, NY: Wiley and Sons, 1989.

Gruber, Jonathan. "Medicaid." Cambridge, MA: National Bureau of Economic Research (NBER), Working Paper 7829, August 2000. [http://papers.nber.org/papers/w7829.pdf]

Hainer, P., C. Hines, E. Martin, and G. Shapiro. “Research on Improving Coverage in Household Surveys.” Paper delivered at the U.S. Bureau of the Census Annual Research Conference, Washington, DC, 1988.

Hall, J., G. Kenney, G. Shapiro, and I. Flores-Cervantes. “Bias from Excluding Households without Telephones in Random Digit Dialing Surveys: Results of Two Surveys.” *Proceedings of the Survey Research Methods Section of the American Statistical Association,* 1999, pp. 382-387.

Hogan, H. “The 1990 Post-Enumeration Survey: Operations and Results.” *Journal of the American Statistical Association*, Vol. 88, No. 423, pp. 1047-1060, 1993.

Holahan, J., C. Winterbottom, and S. Rajan. “A Shifting Picture of Health Insurance Coverage.” *Health Affairs*, Vol. 14, No. 4, pp. 253-264, 1995.

Kronick, R. “Health Insurance, 1979–1989: The Frayed Connection between Employment and Insurance.” *Inquiry*, Vol. 28, 1991, pp. 318–332.

Krosnick, J. A., S. Narayan, and W. R. Smith. “Satisficing in Surveys: Initial Evidence.” *New Directions for Program Evaluation*, Vol. 70, pp. 29-44, 1996.

Levit, Katharine R., Gary L. Olin, and Suzanne W. Letsch. “Americans’ Health Insurance Coverage, 1980-91.” *Health Care Financing Review*, Vol. 14, No. 1, Fall 1992, pp. 31-57.

Lewis, K., M. Ellwood, and J. L. Czajka. *Counting the Uninsured: A Review of the Literature.* Washington, DC: The Urban Institute, 1998.

Lewis, Kimball, Marilyn Ellwood, and John L. Czajka. “Counting the Uninsured: A Review of the Literature.” Washington, DC: Urban Institute, Assessing the New Federalism, Occasional Paper No. 8, July 1998.

Long, Stephen H. and M. Susan Marquis. “Some Pitfalls in Making Cost Estimates of State Health Insurance Coverage Expansions.” *Inquiry*, Vol. 33, Spring 1996, pp. 85–91.

Marquis, K. and J. Moore. “Measurement Errors in the Survey of Income and Program Participation (SIPP) Program Reports.” Paper read at 1990 Annual Research Conference, August 1990.

Martini, A. “Seam Effect, Recall Bias, and the Estimation of Labor Force Transition Rates from SIPP.” Paper read at American Statistical Association, Section on Survey Research Methods, August 1989.

Moyer, M. Eugene. “A Revised Look at the Number of Uninsured Americans.” *Health Affairs*, Summer 1989, pp. 102-110.

Nadeau, R. and R. G. Niemi. “Educated Guesses: The Process of Answering Factual Knowledge Questions in Surveys.” *The Public Opinion Quarterly*, Vol. 59, No. 3, pp. 323-346, 1995.

Nelson, Charles T. and Robert J. Mills. “The March CPS Health Insurance Verification Question and Its Effect on Estimates of the Uninsured.” U.S.

Bureau of the Census, Housing and Household Economic Statistics Division, August 2001. [http://www.census.gov/hhes/hlthins/verif.html]

Perry, M., S. Kannel, R. B. Valdez, and C. Chang. "Medicaid and Children: Overcoming Barriers to Enrollment, Findings from a National Survey." Washington, DC: Kaiser Commission on Medicaid and the Uninsured, 2000.

Presser, S. "Is Inaccuracy on Factual Survey Items Item-Specific or Respondent-Specific?" *The Public Opinion Quarterly*, Vol. 48, No. 1B, pp. 344-355, 1984.

Rajan, Shruti, Stephen Zuckerman and Niall Brennan. "Confirming Insurance Coverage in a Telephone Survey: Evidence from the National Survey of America's Families." *Inquiry*, Vol. 37, No. 3, Fall 2000, pp. 317-327.

Schuman, H. and S. Presser. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context.* New York, NY: Academic Press, 1981.

Selden, T. M., J. S. Banthin, and J. W. Cohen. "Medicaid's Problem Children: Eligible but not Enrolled." *Health Affairs*, Vol. 17, No. 3, pp. 192-200, 1998.

Shapiro, G., G. Diffendal, and D. Cantor. "Survey Undercoverage: Major Causes and New Estimates for Magnitude." Paper delivered at the U.S. Bureau of the Census Annual Research Conference, Washington, DC, 1993.

Short, Pamela Farley. "Counting and Characterizing the Uninsured." ERIU Working Paper 2, 2001. [http://www.umich.edu/~eiru/pdf/wp2.pdf]

StatCorp. *Stata Statistical Software: Release 5.0.* College Station, Texas: Stata Corp., 1998.

Sudman, S., N. Bradburn, and S. Schwarz. *Thinking about Answers*. San Francisco, CA: Jossey-Bass, 1996.

Swartz, K. and J. Purcell. "Letter: Counting Uninsured Americans." *Health Affairs*, Vol. 8, No. 4, pp. 193-197, 1989.

Swartz, Katherine and Patrick J. Purcell. "Letter: Counting Uninsured American." *Health Affairs*, Winter 1989, pp. 193-196.

Swartz, Katherine. "Changes in the 1995 Current Population Survey and Estimates of Health Insurance Coverage." *Inquiry*, Vol. 34, No. 1, Spring 1997, pp. 70-79.

Swartz, Katherine. "Interpreting the Estimates from Four National Surveys of the Number of People without Health Insurance." *Journal of Economic and Social Measurement*, Vol. 14, 1986, pp. 233-243.

Tourangeau, R., G. Shapiro, et al. "Who Lives Here? Survey Undercoverage and Household Roster Questions." *Journal of Official Statistics*, Vol. 13, No. 1, pp. 1-18, 1997.

U.S. Bureau of the Census. "*1990 U.S. Census.*" Washington, DC, 1990.

U.S. Department of Health and Human Services, Health Care Financing Administration. *HCFA 2082 Reports*, various years (1992, 1993, 1994, 1995).

U.S. General Accounting Office. "Uninsured Children and Immigration, 1995." Publication No. GAO/HEHS-97-126R, Washington, DC: GAO, 1997.

United States Department of Commerce Bureau of the Census. "Health Insurance Historical Tables." [Available at http://www.census.gov/hhes/hlthins/historic/ (updated December 2000).]

United States Department of Commerce Bureau of the Census. "Survey of Income and Program Participation (SIPP) Quality Profile." [Available at http://www.census.gov/sipp/ (undated).]