From Multiple Modes for Surveys to Multiple Data Sources for Estimates— *The Role of Administrative Records in Federal Statistics*

Committee on National Statistics

THE NATIONAL ACADEMIES

Constance Citro, *Director,* CNSTAT

WSS Conference • September 18, 2014

Adaptation of WSS President's Invited Address (May 2013)*

Delighted that my talk last year led Mike Fleming and his colleagues to organize this very timely conference, which includes a great set of presentations on administrative records (AR) as an integral part of federal statistics

My job is to set the stage, by indicating why I argued for a shift from the probability sample paradigm as the bedrock of federal statistics to a multiple data sources paradigm, and why I think AR deserves priority among all of the varieties of data sources that are mushrooming out there

I will be much briefer than a year ago, so that I can get out of the way of the speakers who are actually putting AR to work!

*Full set of slides and references available from the author

Wider, Deeper, Quicker, Better, Cheaper Official Statistics (Holt, 2007)

My reference point was—and is—Tim Holt's enumeration of the challenges perennially facing official statistics; to his "*wider, deeper, quicker, better, cheaper*" list, I add "*more relevant*" and "*less burdensome*"

To achieve even some of these goals, I argue that official statistics need to move from the sample survey paradigm of the past 70 years to a ***mixed data source paradigm*** for the future; the official statistics mindset should start with user needs for information for policy development, monitoring, and evaluation, and understanding societal trends, and ***work backwards*** from concepts to the best combination of data sources

For the U.S., and for household surveys in particular, combining ***administrative records*** with survey data is the obvious and imperative next step

Outline of My Presentation*

- Background on NAS/CNSTAT (brief)
- Rise/benefits of probability sampling in U.S. (brief)
- Current challenges to surveys (brief)
- New paradigm of multiple data sources (a little less brief)
  - Which data sources to bolster surveys? AR prime candidate
  - Two ripe opportunities for mixed data sources in U.S. statistics—housing and income
- Barriers to innovation, particularly to changing paradigms, and suggestions for knocking them down

*My remarks are informed by my work at CNSTAT, but are my own; I apologize for frequent references to Census Bureau programs—what I know best; I offer critiques from a deep admiration and respect for the work of statistical agencies

The National Academies & CNSTAT

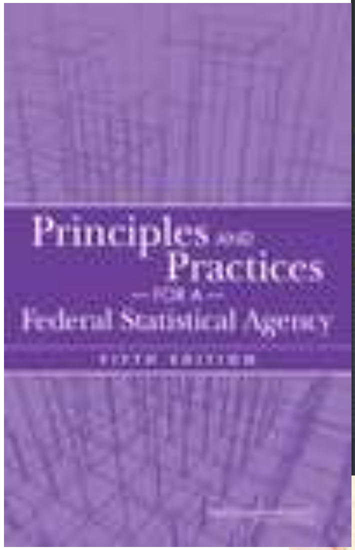
National Academy of Sciences—independent, nonprofit honorific society; congressional charter (1863) to advise government on science and “art” (NAE, 1964; IOM, 1970)

National Research Council – est. 1916 as NAS operating arm; 50 standing units like CNSTAT

CNSTAT’s mission is to improve the statistical methods and information on which public policy decisions are based; also serves as a coordinating force in decentralized U.S. statistical system; over its 40+-year history, CNSTAT has produced over 240 consensus, interim, and workshop reports

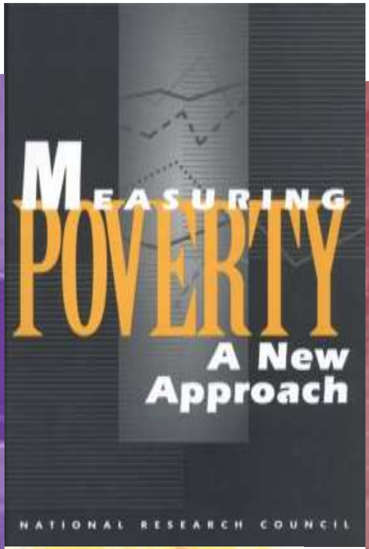
Committee on National Statistics
THE NATIONAL ACADEMIES

Sampling of CNSTAT Studies



Principles and Practices for a Federal Statistical Agency

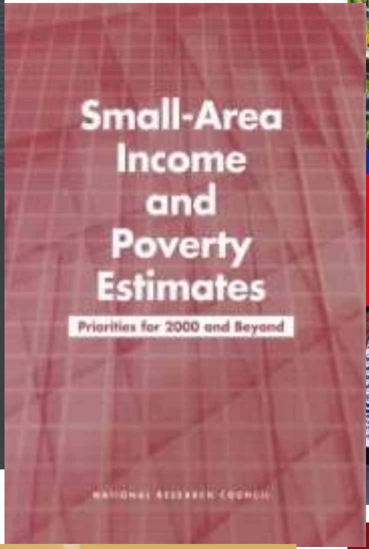
Fifth Edition



Measuring Poverty

A New Approach

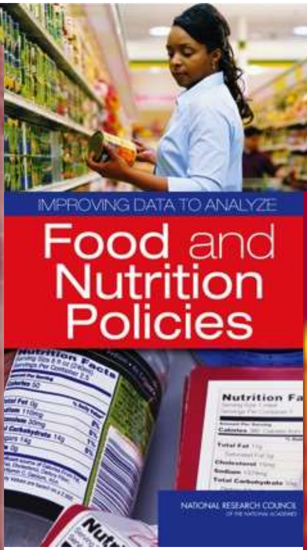
National Research Council



Small-Area Income and Poverty Estimates

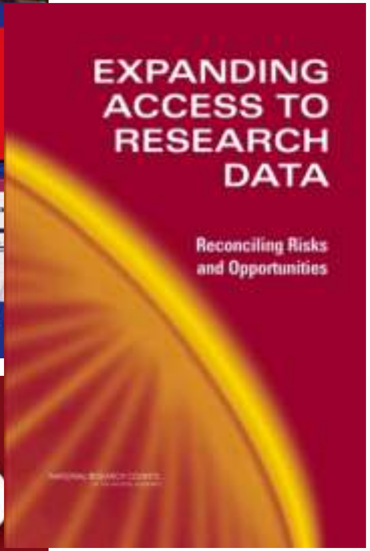
Priorities for 2000 and Beyond

National Research Council



Improving Data to Analyze Food and Nutrition Policies

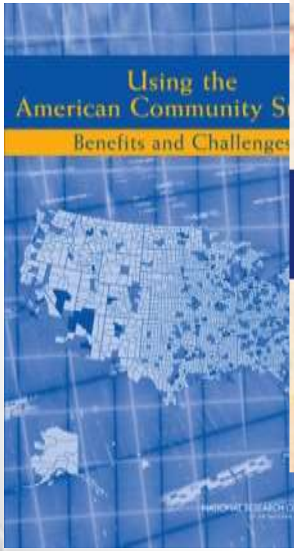
National Research Council of the National Academies



Expanding Access to Research Data

Reconciling Risks and Opportunities

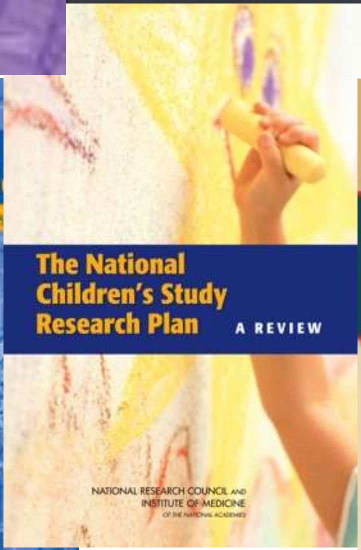
National Research Council of the National Academies



Using the American Community Su

Benefits and Challenge

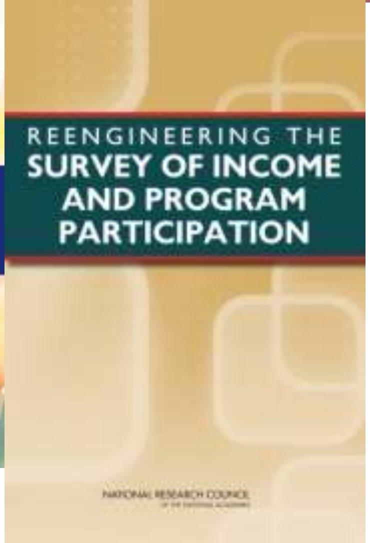
National Research Council of the National Academies



The National Children's Study Research Plan

A Review

National Research Council and Institute of Medicine of the National Academies



Reengineering the Survey of Income and Program Participation

National Research Council of the National Academies

National Research Council of the National Academies

Envisioning the 2020 Census

Committee on National Statistics

THE NATIONAL ACADEMIES

Probability Sample Surveys the 20th Century Answer to Quicker, et al.

Large-scale surveys in U.S. date to late 1930s (Harris-Kojetin, 2012)

- 1937 – 2% sample of households on non-business postal routes estimated much higher number of unemployed than a “complete” (voluntary) census of all residential addresses
- 1940 – Census asked six questions on 5% sample basis
- 1940 – Monthly Report on the Labor Force (became the CPS)
- 1950 – Census asked two-fifths of questions on sample basis—sampling fractions of 20% and 3.3% for population items; 20% for housing items, using matrix design

Gave much more reliable information (could estimate sampling error) than non-probability samples and at greatly reduced cost and increased timeliness compared with censuses

Sample results readily accepted by the public

Obvious Win-Win Like Sampling *Not Obvious* at the Time (Anderson, 1988)

The development of theory and practice of modern probability sampling for finite populations in the U.S. took time

- Jerzy Neyman's arrival in America in the mid-1930s gave tremendous boost to work of W. Edwards Deming, Cal Dedrick, Morris Hansen, and others to develop the needed theory
- Small-scale uses of sampling by academicians and agencies (e.g., 1930s surveys of consumer purchases, unemployment, urban housing, and health) provided proofs of concept and practical tips

The government's Young (statistical) Turks still had to surmount hurdles in the bureaucracy up to the White House before they could get sampling into the mainstream of federal statistics

- Political pressure on both sides regarding the need for hard numbers on unemployment
- Skepticism by "old-timers" at the Census Bureau

Federal Surveys Provide Rich Array of Policy-Relevant Information

- Consumer Expenditure Survey – *1888* (non-probability) (periodically conducted through *1972-73)* (continuous since *1980*) . . . *feeds the CPI*
- Current Employment Statistics – *1939* . . . *payroll employment, etc.*
- Current Population Survey – *1942* . . . *unemployment, poverty, etc., etc.*
- Monthly Wholesale Trade Survey – *1946* . . . *principal economic indicator*
- Survey of Industrial R&D – *1953* . . . *reinvented as BRDIS in 2008*
- National Survey of Fishing, Hunting, and Wildlife-Associated Recreation – *1955* . . . *vitally important to its constituency*
- NHANES – *1960* (continuous since *1999*) . . . *physical exams linked to survey data*
- National Crime Victimization Survey – *1972* . . . *crimes not reported to the police*

Federal Surveys Innovate Concepts, Methods, and Technology

- Punch cards invented for 1890 census
- Probability sampling used in CPS forerunner in 1940–1942; rotation design introduced in 1953
- UNIVAC I helped process 1950 census
- Census mailout-mailback tested in 1950s, 1960s
- Small business income tax records used in 1963 Economic Census
- Longitudinal surveys begun 1966–1973 (NLS, High School Class of 1972, SDR)
- Continuous measurement for small areas via American Community Survey tested in 1996, implemented in 2005; ACS designed to use sampling for nonresponse follow-up
- NASS offered Web response option in 2002

BUT… Federal Surveys Challenged

- Unit response in decline and costly to remedy

  NHIS: 91.8% household response in 1997; 75.7% in 2013

  CBECS: 91% establishment response in 1992; 82% in 2003

- Item nonresponse high and growing for key variables

	1993	1997	2002
Total income imputed: CPS	23.8%	27.8%	34.2%

- Socioeconomic coverage differences remain even after post-stratification with demographic estimates
- Measurement error often a problem

	1987	2005	
CPS estimate of SNAP (food stamps)	74.2%	54.6%	[ratio of
SIPP estimate	85.9	76.4	benchmark]

- Concepts too often out of date (e.g., "regular money income")
- Perceived/actual burden – ACS generating small but steady stream of complaints; CE respondents "satisficing"

Strategies to Combat and/or Compensate for Problems (NRC, 2013)

Survey researchers actively seeking ways to reduce and/or compensate for nonresponse, measurement error, and, more recently, burden—

- Throw money at the problem, but not viable with reduced budgets
- Use paradata, auxiliary information, state-of-the-art methods for more effective nonresponse bias identification and adjustment
- Use adaptive/responsive design to optimize cost/quality of response
- Major emphasis on multiple frames/multiple modes—ACS:

  Push for Mailout/Internet-back (as of 1/2013); Mailout/Mailback;

  CATI for mail nonrespondents; CAPI for sample of remaining NRs
- Research to address burden by optimizing follow-up calls/visits
- Efforts to document benefits of/needs for the data

ALL GOOD, BUT ENOUGH???

Committee on National Statistics

THE NATIONAL ACADEMIES

New Paradigm—Multiple Data Sources from the Get-Go

Survey problems cannot be fully addressed by holding fast to survey paradigm as single best source for official statistics

Indeed, **no single source** can stand alone—2011 German census revealed significant overestimation by registers (used since last census in 1987), due to failure to identify foreign-born emigrants

The operative paradigm should be to work **backwards** from policy and public information needs to best **combination of sources** for optimizing relevance, accuracy, costs, timeliness, and burden

For U.S. household surveys in particular, **AR** is the key go-to source—some programs, e.g., economic census, have long used AR to reduce costs/burden, but uses in household surveys (e.g., for pop. controls) have been at the margins

There are Increasingly Multiple Data Sources—Why AR?

For decades, available sources were essentially surveys and administrative records (AR)—federal, state, and local government agency records that may be useful for official statistics—e.g., tax returns, food stamp case files, Medicare claims files, vital records

In 1970s and 1980s, retail bar codes/scanners and satellite imagery generated data of potential use for official statistics

Since the 1990s, a flood of data became potentially available from Internet searches, social media, traffic camera feeds, etc., etc., etc.—"big data," or "organic data" (Groves, 2011a, 2011b, Keller et al., 2012)

How do we sort out and evaluate the various sources for their utility for official statistics, particularly for use with surveys?

Field Needs Something Akin to Total Error Concept for Surveys

Start from observation that many components of non-sampling error apply to non-survey data sources as well as surveys

Field can/should move toward an ALL-DATA-SOURCE set of metrics—as a starting point, consider the following dimensions:

(1) Accessibility to and control by statistical agency

(2) Ability to identify and measure components of error

(3) Data quality attributes (Biemer et al., 2014; see also Iwig et al., 2013; Daas et al., 2012a)—(a) relevance for policy/public; (b) relevance of covariates; (c) frequency of collection; (d) timeliness of release; (e) comparability/coherence; (f) accuracy *(frame error, nonresponse, processing error, measurement error, modeling/estimation error, specification error)*

(4) Burden (on respondent, or other data source, such as AR agency)

(5) Cost (to statistical agency)

Surveys Look Good on [Crude] Metrics–So Why Change Paradigm?

Comparing surveys, AR, commercial transactions, and autonomously-generated Internet interactions (e.g., social media) crudely on these metrics, surveys look good, social media look bad, with AR and commercial transactions in between

But many official surveys don’t look nearly as good as they would have 30, 40, or 50 years ago, particularly on components of accuracy, and they are increasingly high on burden and costs

Not all, but many AR, look good to fill in for survey weaknesses—at least some AR variables are definitely better than some survey responses (or nonresponses)

[Asking householders to provide their own alternative data by, e.g., consulting records, is a losing game—experiments with SIPP, CE interviewer debriefings document the difficulties]

Focus on Administrative Records

AR have problems (as do surveys), but are generated according to rules about eligible population (like a census), who must file what, who is entitled to what, etc. Should be easier to develop a conceptual framework for total error than would be for web-scrapings

Potential applications to household surveys:

Adjust for coverage errors (already common, could refine)

Improve sampling frames (already common, could refine)

Use in models for small-area estimates (need joint estimation)

Improve imputations (fertile ground here)

Replace responses for nonreporters and underreporters (common in microsimulation policy models, better if done in surveys)

Replace survey questions

Increasing utility but more difficult to protect confidentiality/obtain consent

Example 1—Housing in the American Community Survey

ACS better quality (e.g., more frequent/timely with less missing data) than 2000 long-form sample

But is threatened—prime threats are from two sources:

- User expectations for accuracy of small-area data not always met—modeling may help over the long run
- Perceived/actual respondent burden (letters to Congress—most objections to income, plumbing facilities, disability, time leave for work)
  - Questions could be dropped entirely—but (assuming needs are real)
  - Better yet to eliminate items from the questionnaire by integrating information from other sources

Housing a prime candidate for moving to an alternative data source such as AR:

- Respondents often don't know the answers (e.g., year structure built)
- Questions on financial characteristics and utilities are burdensome and difficult to answer accurately—homeowners (2/3 of U.S. population) are most burdened in ACS (along with large households)

Example 1—Not Kidding about ACS Housing/Homeowner Burden

(% imputed, 2012)

Tenure (own/rent) *(1.1)*
Units in structure *(1.4)*
Year built *(17.2)*
Lot size/agri. sales *(4.5)*
Business on property *(3.2)*
Number rooms *(5.3)*
Number bedrooms *(4.3)*
Running water *(1.9)*
Flush toilet *(2.0)*
Bathtub/shower *(2.0)*
Sink with a faucet *(2.0)*
Stove or range *(2.5)*
Refrigerator *(2.7)*
Heating fuel *(3.4)*

Last month electricity $ *(7.2)*
Last month gas $ *(10.2)*
Last 12 mos. water/sewer $ *(8.4)*
Last 12 mos. other fuel $ *(11.3)*
Annual property tax $ *(17.0)*
Annual insurance $ *(24.3)*
Mortgage status *(2.1)*
Monthly mortgage $ *(10.5)*
Whether second mortgage (3.2)
Whether home equity loan *(3.9)*
Other mortgage $ *(18.4)*
Annual mobile home $ *(21.2)*
Property value *(13.2)*
Monthly rent *(9.2)*

Example 1—Augmented MAF

A way to square the circle is for the Census Bureau to develop an *augmented Master Address File* with more and more housing variables included over time that would no longer, or less frequently, need to be asked in the ACS and other surveys

Many housing characteristics available, with some effort, from public local records, which do not require confidentiality protection or consent; utility companies another potential source

Some characteristics are even invariant over time (e.g., plumbing)

HUD, working with the Census Bureau, is already thinking along these lines in redesigning the American Housing Survey

Which is better—to work out the myriad kinks in moving toward an augmented MAF, or risk the loss of the ACS?

Example 2—Income in Surveys

CPS/ASEC is flagship for U.S. household income and poverty statistics, which influence policies and programs

ACS income estimates important for small areas and groups

SIPP income estimates key for more nuanced understanding of economic well-being and changes over periods of time

Other surveys need good income measures without adding burden

BUT high, increasing, and variable levels of imputation and underreporting for income

Leaving aside conceptual issues (such as whether/how to update regular money income concept), there is an imperative to improve income estimates for U.S., subnational geographic areas, and population groups

Example 2—Comparing CPS-NIPA $ Adjusted to CPS Concept

CPS/Adj. NIPA (%)	Household Mean	Median
1999	88.9%	89.9%
2001	88.2	90.2
2003	87.1	89.2
2005	84.6	88.8
2007	81.0	85.7
2009	84.0	86.4
2010	82.4	85.0

(Fixler and Johnson, 2012:Table 2—
CPS concept = regular money income
NIPA = National Income and Product Accounts)

Example 2—Better Income Estimates through Multiple Sources

Suggested steps forward:

- Improve population coverage adjustments in CPS et al., by, e.g., post-stratifying race/ethnicity categories by total income (from IRS? ACS?), so that coverage adjustments (partly) capture SES
- Plan for model-based small-area ACS estimates from the get-go
- Move strategically, source by source, to improve imputations of income amounts—and receipt—in major income surveys by use of administrative records; Census already has access to many records—although not all in one place—and is working to get more (e.g., SNAP) as part of 2020 census planning
- Move—carefully—toward Canadian model, whereby respondents can skip entire blocks of income questions by permitting access to their administrative records

Example 2—Pipedream?? Not with Vision and Long-term Planning

Making improvements in income estimates in surveys is daunting:

- Legal/bureaucratic/"big brother" difficulties of obtaining access
- Consent issues if actual records are substituted for questions
- Could risk timeliness due to lags in records receipt (modeling)
- Error structures of records not (yet) well known
- Strain on headquarters staff (not like hiring/laying off field staff)
- Need for multiple, linked data processing systems

BUT, levels and trends for key policy variables and public understanding are seriously in error from relying on surveys

A well-articulated, staged, strategic plan, starting from policy & public data needs, could empower statistical system to work toward quality gains for income estimates and reduction in burden and costs

The Challenge of Effecting Paradigm Change

Statistical systems have admirable record of innovation, but changing paradigms is always difficult [recall earlier slide about obstacles to sampling]; particularly hard to rethink long-lived, ongoing, "comfortable" programs

Rare for companies to hit one home run after another; est. Fortune 500 company average life is 40-50 years (*Business Week*); most innovation from newcomers

Even harder for government because Schumpeter's "creative destruction" not quite what one wants for government agencies serving public good– Trick is to make the harder work of government innovation a feature and not a bug

Effecting Paradigm Change—Six Barriers

Inertia—from coasting on earlier successes; enabled by data users, who want "their" series walled off from change

Monopoly (often associated with inadequate channels of communication and feedback)—private sector quasi-monopolies can disdain innovation until a newcomer outflanks them; govt. agencies, too, run risk of losing touch with users' changing needs

Fear of undercutting existing business lines

Overemphasis on "this is how we do things here"

Unclear/faulty concept of primary goals—the business of statistical agencies is serving user needs, ***not*** continuing long-standing data collection programs for their own sakes

Plus [for U.S.], decentralized statistical system

Effecting Paradigm Change—Ways to Move Forward

Leadership buy-in at all levels of stat agency essential to:

- **Instill /sustain laser focus on policy uses/concepts and from there to appropriate data sources**, and not on marginal changes to long-time survey *x* (or long-time AR system *y*)
- **Think strategically** about threats to current ways of data collection (e.g., burden, underreporting) and prioritize remedies
- **Bolster role and heft of subject-matter analysts** to interface with outside users and inside data producers
- **Staff operational programs with expertise in all relevant data sources**, such as surveys and AR, on equal footing
- **Rotate assignments within and without agency** to foster innovative thinking and diverse perspectives—give survey researchers some AR agency experience/vice versa

BOTTOM LINE

Official statistics innovated the probability survey paradigm in the second half of the 20th century; let's move with alacrity in this century to a paradigm of combining surveys with administrative records for policy and public needs (and then move on to yet additional data sources!) THANK YOU!

Committee on National Statistics
THE NATIONAL ACADEMIES

Constance Citro
Director • **Committee on National Statistics**
Full slides with references available from:
ccitro@nas.edu • (202) 334-3009