New Challenges and New Opportunities for the Federal Statistical System

Katharine G. Abraham, University of Maryland

Federal Committee on Statistical Methodology Conference

October 24, 2023

Envisioning the future of the Federal statistical system

- In December 2018, Congress passed the Foundations for Evidence-Based Policymaking Act (aka the Evidence Act)
  - Incorporated many of the recommendations of the Commission on Evidence-Based Policymaking related to data access and use
  - When making those recommendations, Commission members had the Federal statistical system very much in mind
- Take approaching 5-year Evidence Act anniversary as opportunity to think about the future of the Federal statistical system
- Will argue that there is both a need and, thanks in part to the Evidence Act, an opportunity to rethink and strengthen the federal statistical system infrastructure

Federal statistical system faces growing challenges

Foundations for U.S. social and economic statistics infrastructure laid in mid-20th century

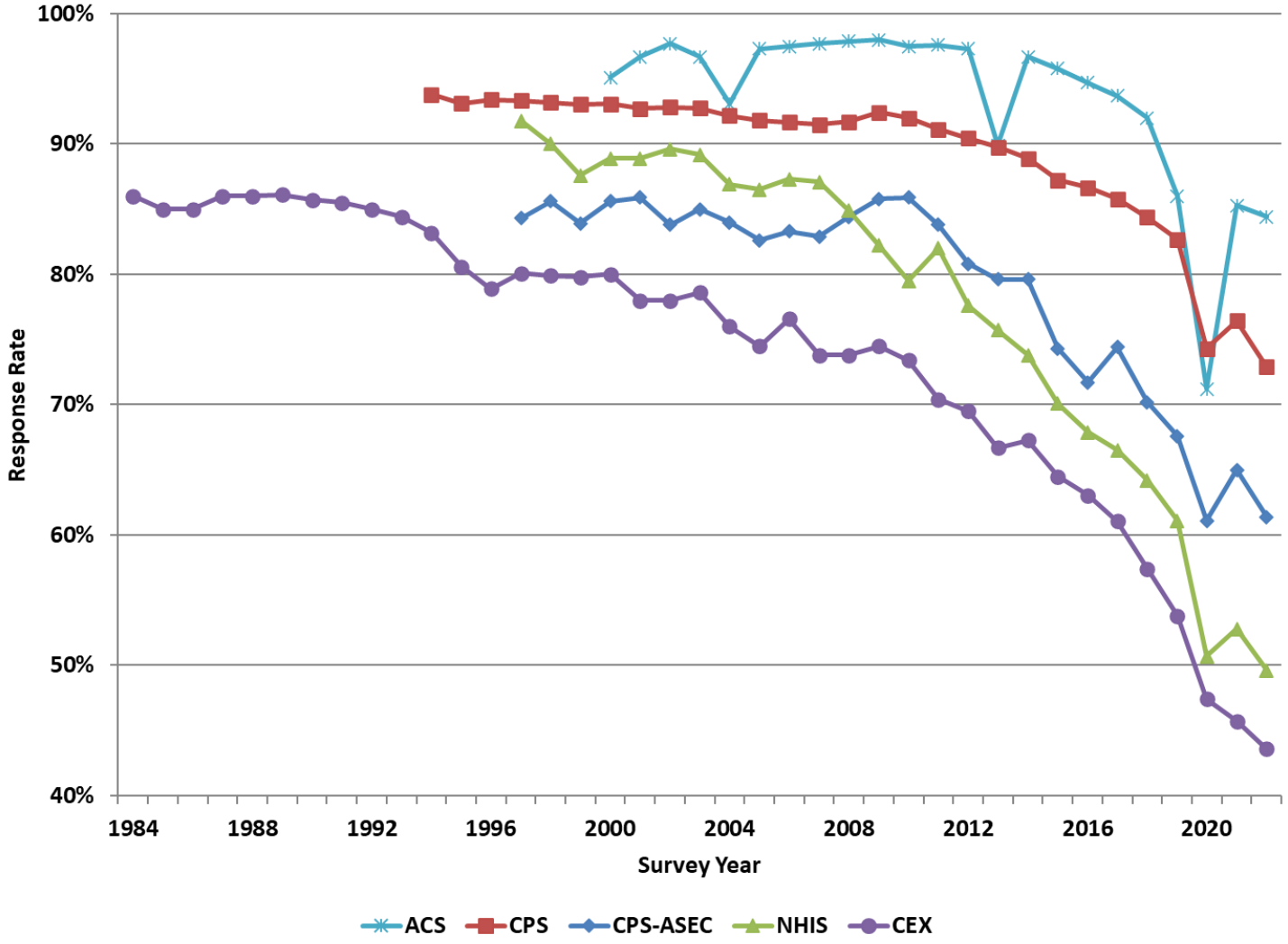
- Probability surveys of households and businesses at core
  - Samples designed to represent target population
  - Questionnaires designed to collect desired information
  - Standard measures produced on a regular schedule
- Census and administrative benchmarks for survey data
  - Decennial censuses for social statistics
  - Economic censuses and administrative data for economic statistics
- Tasks allocated across multiple statistical agencies, each with its own separate set of assigned responsibilities

Existing statistical infrastructure has served nation well, but subject to growing pressures

- Increasing difficulty of obtaining survey responses

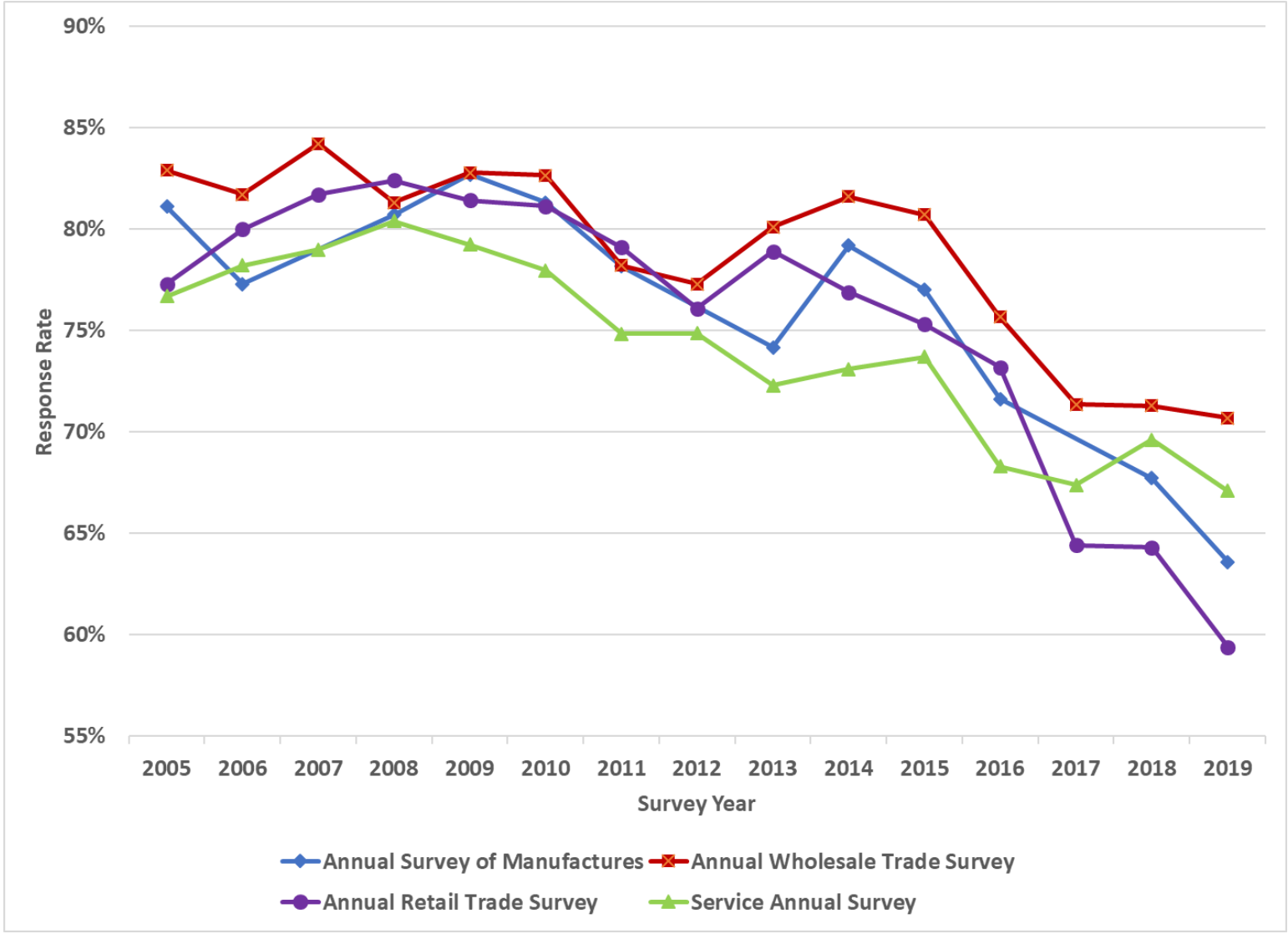
- Expanding demands from data users

Unit response rates, selected household surveys, 1984-2022



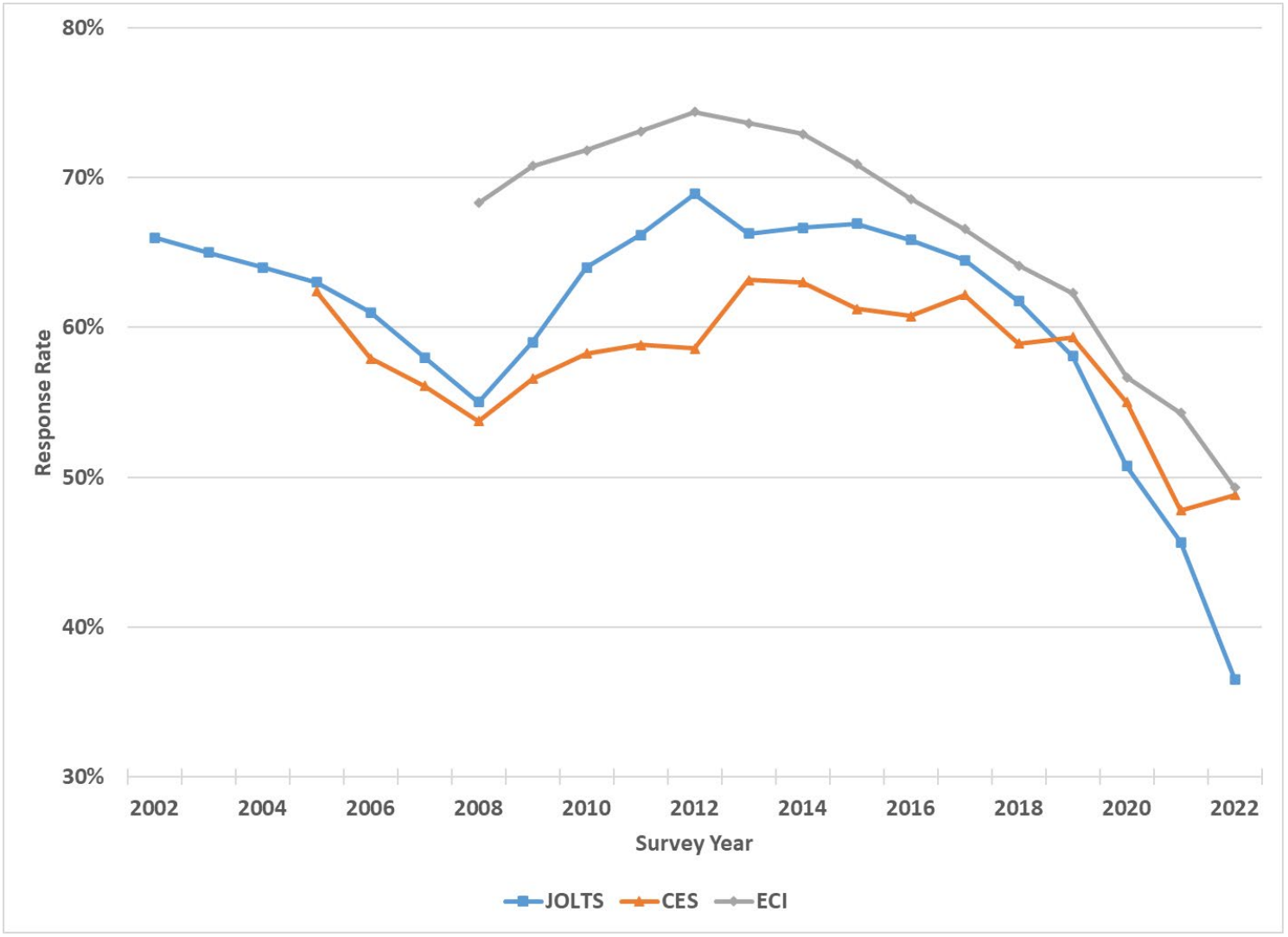
Source: Meyer, Mok and Sullivan (2015); agency websites

Unit response rates, selected Census Bureau business surveys, 2005-2019



Source: U.S. Census Bureau

Unit response rates, selected Bureau of Labor Statistics business surveys, 2002-2022



Source: Bureau of Labor Statistics.

Imputation rates for selected items, American Community Survey, 2000 and 2022

Item	2000 Allocation Rate	2022 Allocation Rate
Own versus rent, occupied housing units	1.4%	2.2%
Age	2.4%	1.4%
Marital status (age 15 plus)	2.3%	7.5%
Place of birth	6.4%	12.3%
Citizenship	0.5%	9.8%
Educational attainment (age 3 plus)	4.8%	11.9%
Weeks worked past 12 months (age 16 plus and worked)	9.6%	17.8%
Self-employment income (age 15 plus)	6.3%	14.4%
Some or all income imputed (age 15 plus)	23.9%	31.8%

Source: U.S. Census Bureau

Existing statistical infrastructure has served nation well, but subject to growing pressures

- Increasing difficulty of obtaining survey responses
- Expanding demands from data users

What do today's data users want?

- Timely data of appropriate quality
- Data that are compatible across agencies and programs
- Data more disaggregated along multiple dimensions
  - More detailed demographic groups
  - State, metropolitan area, county, customized geographies
- Data to answer both long-standing and emerging questions

Administrative data already helping statistical agencies meet users' needs

Administrative records

- Administrative records contain information collected for the purpose of administering government programs.
- Examples of administrative records:
  - Individual and business income tax returns
  - Unemployment insurance reports
  - Medicare and Medicaid records
  - SNAP applications and benefit receipt histories
  - Customs declarations

Long history of administrative data serving significant statistical purposes

- A number of statistical agencies use administrative data as sampling frames; others produce statistics based exclusively on administrative data as core part of mission
- Federal law directs that, where possible, Census Bureau use administrative and other existing data rather than survey data
  - Title 13(9)(c): "To the maximum extent possible and consistent with the kind, timeliness, quality and scope of the statistics required, the [Census Bureau] shall acquire and use information available from [existing federal, state or private data resources]… instead of conducting direct inquiries."
- OMB M-14-06: Guidance for Providing and Using Administrative Data for Statistical Purposes
  - "…the Federal Government has the opportunity to create … statistical information more efficiently through greater use of information that the Federal Government has already collected for programmatic and regulatory purposes…"

Census Bureau: Using administrative data to substitute for missing survey data

- For many business surveys, Census uses administrative data to fill in for missing survey data
- Preferred response rate cited for these surveys reflects data obtained both through surveys and from administrative records
  - Total quantity response rate (TQRR): Share of estimated (weighted) total for a data item represented by survey responses *or sources deemed to be of equivalent quality*
- Administrative data an important contributor to effective overall response rate

NCHS: Enhancing health surveys with linkages to administrative records

- National Center for Health Statistics (NCHS) data linkage program dates to the early 1990s
- Both National Health Interview Survey (NHIS) and National Health and Nutrition Examination Survey (NHANES) have been linked to:
  - National Death Index (date and cause of death)
  - Medicare and Medicaid claims records
  - Social Security benefit records
- Linked records have contributed to assessments of the accuracy of health statistics and a better understanding of the determinants of health outcomes

Census Bureau: Producing small area poverty estimates

- Census Bureau required to produce estimates of poverty for states, counties and school districts, but direct American Community Survey estimates for many jurisdictions are noisy
- In its Small Area Income and Poverty Estimates (SAIPE) program, Census Bureau has produced modeled estimates that rely in part on relationship between (noisy) poverty measure and other variables by geography. For example, county model includes logs of:
  - Number of exemptions claimed on tax returns where reported income places the unit below the poverty line
  - Total number of exemptions claimed on tax returns
  - Number of SNAP beneficiaries
  - Total number of people in poverty as of the 2000 Census
  - Estimated total population

Other places where administrative data could help to improve agency data products?

- Census: Use of administrative data to produce better estimates of income and program participation
  - Building on prior work by academic researchers, National Experimental Well-being Statistics (NEWS) project is actively exploring how to do this
- Bureau of Labor Statistics: Linkages of National Longitudinal Survey records to administrative records such as educational histories, earnings, program participation, and interactions with the criminal justice system

Use of private sector data becoming more important

Naturally occurring private data have proliferated

- Naturally occurring private data are generated by private actors for purposes other than statistical purposes.
- Many types of natively digital private data. A few examples:
  - Retail scanner data
  - Online information about prices and product characteristics
  - Credit card transactions data
  - Payroll processing and scheduling data
  - Electronic health records
  - Sensor data (e.g., satellite imaging, traffic cameras)

BLS: Using big data to improve the Consumer Price Index (CPI)

- CPI price data collected by surveying businesses and rental units
  - Commodities and Services Survey: ~94,000 prices per month
  - Housing Survey: ~8,000 rental housing unit quotes per month
  - Majority of data collected by personal visit
- Efforts underway to substitute data from alternative sources where feasible and cost-effective
  - Currently identified data sources could replace prices for up to 22 percent of the CPI market basket
  - Similar efforts underway in several countries

Census Bureau: Using big data to produce state-level retail sales estimates

- Monthly retail sales data collected from a survey sample of ~13,000 retail and food services businesses
  - Data collected at company level; no geographic component to design
- Census has explored use of point-of-sale data from 3rd party vendor NPD to reduce respondent burden and improve national estimates
- NPD data also used for new experimental monthly state-level estimates
  - Top-down estimates: Allocation of national sales
  - Bottom-up estimates: Sum of sales for survey reporters operating in a single state, pre-selected multi-unit businesses from NPD, and imputed values for other retailers
  - Composite estimates: Weighted sum of two estimates; weights based on relative variances

Bureau of Economic Analysis: Using private data to improve early estimates of GDP

- First estimates of GDP often rely on “judgmental trend”
- Where possible BEA uses an extrapolator until data until Census data on receipts become available. Some examples:
  - For purchased software, information on company receipts from SEC filings
  - For various components, data from trade sources (e.g., intercity bus transportation, intracity mass transportation, motion picture theaters, accommodations, mining exploration).
- BEA researchers have explored more sophisticated “nowcasting” models based on machine learning methods for projecting Quarterly Services Survey data not available for first estimates
  - Best models used BLS employment and credit card data
  - Nowcasting results have been incorporated into production; most often used for healthcare services and software investment

Other places where private data could help to improve agency data products?

- NCHS: Private sector health records could in principle augment or replace collection of some information through surveys
- Census Bureau: Exploring whether data from private sources plus online permit data could replace its Building Permit Survey
- Census Bureau: Exploring whether satellite images could be used to measure building starts, completions, and selected unit characteristics, replacing measurement through its Survey of Construction

Don’t mean to suggest that data integration is simple!

Different approaches to combining administrative or private data with survey data

- Record linkage
  - Deterministic or probabilistic matching; fill in missing or deficient survey values with information for same individual or unit
- Imputation
  - Statistical matching; fill in missing or deficient survey values with values for similar individuals or units
- Modeling
  - Extrapolate to produce current estimates
  - Exploit auxiliary information to produce estimates for population subgroups or small geographic areas
- Successful projects have had to address many technical challenges
  - Won't always be the case that use of alternative data is the best path forward

Beyond the technical challenges, existing infrastructure not conducive to data integration

- Administrative data files typically not structured for statistical analysis and often poorly documented
- Holders of administrative often reluctant to share
  - Concerns about legal and other risks, especially as related to data protection
  - Absence of benefit to sharing agency or state
- Owners of private data voice similar concerns
  - Data may be viewed as sensitive or proprietary
  - Data can be viewed as an asset to be monetized
- Negotiation of memoranda of understanding a time-consuming case-by-case process

Evidence Act an important step towards lowering barriers to data integration

Key Evidence Act provisions

- Enhance protections for confidential and sensitive data
  - Reinforce need for legal protection of confidential data
  - Create requirement for risk assessment prior to data release
- Enhance capacity for secure access to confidential data
  - Direct creation of data inventories
  - Direct creation of standard application process
  - Create presumption of access to data for evidence-building purposes unless prohibited by law
- Encourage use of data for evidence-building
  - Require learning agendas and evaluation plans

Key Evidence Act provisions (continued)

- Require the assignment of leaders to guide the development, sharing and use of data within the federal government
  - Chief Data Officers
  - Statistical Officials
  - Chief Evaluation Officers
- Mandate the formation of the Advisory Committee on Data for Evidence Building, charged with making recommendations for creation of a National Secure Data Service

How will all of this affect statistical agencies moving forward?

- Increased access to administrative data that can help them to
  - Reduce respondent burden
  - Assess and potentially improve the quality of survey estimates
  - Support the production of more disaggregated estimates
  - Augment data items collected on surveys to answer new questions
- New options for accessing private data
  - Research on access and linkage modalities carried out as part of the National Secure Data Service Demonstration

How will all of this affect statistical agencies moving forward? (continued)

- New roles *vis a vis* other government agencies
  - Chief Data Officers given responsibility for data stewardship, but Statistical Officials (and their staffs) have a critical role to play in ensuring that data are collected, captured and documented in a fashion that makes them useful for statistical analysis
  - New emphasis on evidence building will encourage partnerships between statistical agencies and program agencies, both within and outside of their departments

Example of collaboration with program agencies: NCHS-HUD partnership

- Department of Housing and Urban Development (HUD) approached the National Center for Health Statistics (NCHS) about linking their administrative data to NCHS' NHIS and NHANES survey data
  - Interested in learning about health of housing program participants
- Once NCHS expressed interest, several issues to work through
  - Was project consistent with both agencies' missions and with the consent obtained from NCHS survey respondents?
  - How would data be shared? Where would they be stored?
- Processing and cleaning HUD administrative data a cooperative effort
- Project has yielded information that is of interest to NCHS and invaluable to HUD
  - One important finding: Lead hazard control regulations in HUD properties associated with significantly lower blood lead levels in resident children; led HUD to seek to tighten those rules

Statistical agencies will need help to fully realize their expanded potential

- Leadership from the Office of Management and Budget (OMB) and the Interagency Council on Statistical Policy (ICSP)
  - In the near term, especially looking forward to the release of the draft "presumption of access" regulation
  - Development of a plan for a functioning National Secure Data Service that can augment existing capacities
- Necessary staffing and other resources
- Ultimately, changes in laws that constraint use of data for statistical purposes

How the statistical agencies view themselves also will be important

- Openness to innovation
  - Work at agencies including NCHS, Census, and the Bureau of Economic Analysis, among others, on data integration and experimental data products has produced exciting results
  - New data products that provide insights about important questions may be valuable even if they don't meet traditional quality standards
    - Effective communications about data quality will be important!
- System perspective versus silo perspective
  - Contributions to evidence-building that fall outside agencies' traditional role but can help to inform policy likely to grow in importance
- New opportunities are there to be seized. I hope you will do so!

Katharine G. Abraham
University of Maryland
kabraham@umd.edu