Federal Committee on Statistical Methodology
2020 Research and Policy Conference

DRAFT Program

Walter E. Washington Convention Center
April 14 - 16, 2020

Major Sponsors
Bureau of Economic Analysis
Bureau of Justice Statistics
Bureau of Transportation Statistics
Economic Research Service
National Agricultural Statistics Service
National Center for Education Statistics
National Center for Science and Engineering Statistics
Statistics of Income Division, Internal Revenue Service
U.S. Census Bureau

Also Sponsored By
DC AAPOR
American Statistical Association
Washington Statistical Society
NORC at the University of Chicago
RTI International
Westat

Hosted By: Council of Professional Associations on Federal Statistics
The Federal Committee on Statistical Methodology and the Council of Professional Associations on Federal Statistics recognize and are abundantly grateful to the following organizations for sponsoring this year’s FCSM Research and Policy Conference. Their commitment and support help to ensure the exchange of cutting edge statistical methods and findings among statisticians and other social scientists from the Federal government, academia, and the business community.
FCSM 2020 Research and Policy Conference
Program Committee

David Kashihara, Co-Chair, Agency for Healthcare Research and Quality
Tamara Rib, Internal Revenue Service

Darius Singpurwalla, Co-Chair, National Center for Science and Engineering Statistics
Gavin Corral, National Agricultural Statistics Service

Aaron Maitland, National Center of Health Statistics
Jennifer Beck, National Center for Science and Engineering Statistics

Jessica Graber, National Center of Health Statistics
Tiandong Li, Health Resources and Services Administration

Tala Fakhouri, National Center of Health Statistics
Michael Leonesio, Social Security Administration

Ellen Galantucci, Bureau of Labor Statistics
Scott Gibbons, Department of Labor

Parvati Krishnamurty, Bureau of Labor Statistics
Thomas Wei, Institute of Education Sciences

Erik Scherpf, National Opinion Research Center at the University of Chicago
Andrew Zuckerberg, National Center for Education Statistics

Tom Garin, Department of Veterans Affairs
Andrew White, National Center for Education Statistics

Scott Wentland, Bureau of Economic Analysis
Kevin Scott, Department of Justice

Denise Abreu, National Agricultural Research Service
E. Ann Carson, Department of Justice

Jacob Bournazian, Energy Information Administration
Janine McFadden, Bureau of Transportation Statistics

Dawn Nelson, U.S. Census Bureau
Grace Medley, Substance Abuse and Mental Health Services Administration
FCSM 2020 Research and Policy Conference

Rochelle (Shelly) Martinez, Office of Management and Budget (Chair)
Stephen Blumberg, National Center for Health Statistics
Chris Chapman, National Center for Education Statistics
Jennifer Edgar, Bureau of Labor Statistics
John Eltinge, U.S. Census Bureau
John Finamore, National Science Foundation, National Center for Science and Engineering Statistics
Dennis Fixler, Bureau of Economic Analysis
Michael Hawes, U.S. Census Bureau
Jennifer Madans, National Center for Health Statistics
Wendy Martinez, Bureau of Labor Statistics
Jaki McCarthy, National Agricultural Statistics Service
Jennifer Nielsen, National Center for Education Statistics (Project Manager)
Jennifer Ortman, U.S. Census Bureau
Ann Parker, National Center for Health Statistics
Jennifer Parker, National Center for Health Statistics
Polly Phipps, Bureau of Labor Statistics
Mark Prell, Economic Research Service
Joseph Schafer, U.S. Census Bureau
Rolf Schmitt, Bureau of Transportation Statistics
Marilyn Seastrom, National Center for Education Statistics
Joy Sharp, Energy Information Administration
Bob Siviniski, Office of Management and Budget
David Williamson, Centers for Disease Control and Prevention
Linda Young, National Agriculture Statistics Service
The Council of Professional Associations on Federal Statistics (COPAFS) is devoted to educational activities and to preserving the public good represented by federal statistical collections.

Since 1980, COPAFS has provided an open dialog between those who use federal statistics in professional contexts and the Federal statistical agencies that produce those statistics for the public good. Supporting organizations include professional associations, businesses, research institutes, and others that help to produce and/or use federal statistics.

Our Goal is: Advancing Excellence in Federal Statistics.

COPAFS’ objectives are to:

- Increase the level and scope of knowledge about developments affecting Federal statistics
- Encourage discussion within and among professional organizations to respond to important issues in Federal statistics and bring the views of professional associations to bear on decisions affecting Federal statistical programs.

In support of these objectives, COPAFS:

- Obtains information on developments in statistics through discussions with officials, attendance at congressional hearings and meetings of statistical advisory committees, and exchanges of documents
- Disseminates information and encourages discussion and action on developments in federal statistics through correspondence and presentations at Council and professional association meetings
- Plans and presents educational programs on uses of statistics in policy formulation, public and private decision-making, research, the distribution of products, and the allocation of resources.

COPAFS helps:

- Professional associations and other organizations obtain and share information about developments affecting federal statistical programs
- Federal agencies to disseminate information on developments of interest to the professional community and to obtain advice about professional societies’ concerns and priorities
- Congressional offices to clarify issues and questions about the federal statistical system, to plan hearings related to federal statistical programs, and to identify experts to testify
- The public to learn more about the federal statistical agencies, to communicate views of data users concerning Federal statistical activities, and to obtain a better understanding of how policy and budget are likely to affect the availability of federal statistics.

Member associations and affiliates appoint representatives to serve on the Council and to attend its quarterly meetings in Washington, DC. The representatives are responsible for establishing COPAFS’ priorities and guiding its activities. At each COPAFS meeting, members discuss significant cross-cutting issues, hear from statistical agencies and other producers and users of data, and make suggestions for further action.

The Board of Directors, comprised of elected officers and four at-large members, facilitates the work of COPAFS between meetings. The Executive Director is responsible for the day-to-day operations. Financial support for COPAFS’ on-going programs comes from annual member dues.
2020 FCSM Research and Policy Conference

The Federal Committee on Statistical Methodology (FCSM) an interagency committee dedicated to improving the quality of federal statistics. This conference helps the committee achieve their major goals, which are to:

- Communicate and disseminate information on statistical practice among all federal statistical agencies.
- Recommend the introduction of new methodologies in federal statistical programs to improve data quality.
- Provide a mechanism for statisticians in different federal agencies to meet and exchange ideas.

The 2020 FCSM Research and Policy Conference will focus on the Federal Statistical System’s role in helping agencies and the public meet the demands of evidence-based policymaking. Several sessions will provide attendees an overview of the Foundations for Evidenced-Based Policymaking Act of 2018, and share approaches to interagency data sharing as well as insights from program evaluators using statistical data in their evidence building.

The conference provides a forum for experts and practitioners from around the world to discuss and exchange current methodological knowledge and policy insights about topics of current and critical importance to the Federal Statistical System.

Each day of the conference will offer papers on a wide range of topics relevant to the production, quality and use of federal statistics. Attendees from a range of backgrounds will find sessions of interest, including statistical methods, administrative data, questionnaire design, program evaluation, policy making, and more.

Sessions feature presentations by government, private sector, and academic researchers from multiple countries. All sessions will include an open discussion and some sessions will include a formal discussion. Presentations will be made available on the conference website following the conference.

**KEYNOTE SPEAKER**

John Friedman from Brown University will be the conference’s keynote speaker. Dr. Friedman is a professor of economics, international affairs and public policy. With Raj Chetty, he is a founding co-director of Opportunity Insights at Harvard University. We look forward to hearing about the valuable research that he and his team have conducted leveraging secure access to a broad set of federal data. This body of work illustrates what is already possible and informs a vision for the role of statistical agencies and units as pivotal actors in the broader federal evidence-building ecosystem. This ecosystem offers the opportunity to create new insights to drive better policy while making the creative use of data a routine part of government activity. Many of the conference sessions will dive into different aspects of this vision.

**PLENARY PANEL**

A panel of pioneers in the field of Artificial Intelligence (AI) will discuss the past, present, and future of the use of AI and machine learning in public sector applications. Speakers will discuss the importance of establishing guiding principles in advance of the proliferation of AI in public sector applications and how machine learning can be used to help form, target, and optimize public programs as well as the importance of establishing guiding principles in advance of the proliferation of AI in public sector applications.
7:30 a.m.
Registration
(Concourse)

Continental Breakfast
(Concourse)

8:30 - 9:30 a.m.
Welcoming Remarks and
PLENARY SESSION
(Rooms 146A, 146B and 146C)

9:30 - 10:00 a.m.
Break
(Concourse)

10:00 - 11:45 a.m.
CONCURRENT SESSION
A-1: Room 146A
A-2: Room 146B
A-3: Room 146C
A-4: Room 145AB
A-5: Room 147AB

11:45 a.m. - 1:15 p.m.
Lunch on Your Own

1:15 – 3:00 p.m.
CONCURRENT SESSION
B-1: Room 146A
B-2: Room 146B
B-3: Room 146C
B-4: Room 145AB
B-5: Room 147AB

3:00 – 3:15 p.m.
Break
(Concourse)

3:15 – 5:00 p.m.
CONCURRENT SESSION
C-1: Room 146A
C-2: Room 146B
C-3: Room 146C
C-4: Room 145AB
C-5: Room 147AB

7:30 a.m.
Registration
(Concourse)

Continental Breakfast
(Concourse)

8:30 - 10:15 a.m.
CONCURRENT SESSION
D-1: Room 146A
D-2: Room 146B
D-3: Room 146C
D-4: Room 145AB
D-5: Room 147AB

10:30 a.m. – 12:15 p.m.
CONCURRENT SESSION
E-1: Room 146A
E-2: Room 146B
E-3: Room 146C
E-4: Room 145AB
E-5: Room 147AB

12:15 p.m. - 1:45 p.m.
Lunch on Your Own

1:45 – 3:30 p.m.
CONCURRENT SESSION
F-1: Room 146A
F-2: Room 146B
F-3: Room 146C
F-4: Room 145AB
F-5: Room 147AB

3:30 – 3:45 p.m.
Break
(Concourse)

3:45 - 5:30 p.m.
CONCURRENT SESSION
G-1: Room 146A
G-2: Room 146B
G-3: Room 146C
G-4: Room 145AB
G-5: Room 147AB

7:30 a.m.
Registration
(Concourse)

Continental Breakfast
(Concourse)
8:30 - 10:15 a.m.
CONCURRENT SESSION
H-1: Room 146A
H-2: Room 146B
H-3: Room 146C
H-4: Room 145AB
H-5: Room 147AB

8:30 - 10:15 a.m.
Break
(Concourse)

10:30 a.m. – 12:15 p.m.
CONCURRENT SESSION
I-1: Room 146A
I-2: Room 146B
I-3: Room 146C
I-4: Room 145AB
I-5: Room 147AB

12:15 p.m. - 1:45 p.m.
Lunch on Your Own

1:45 – 3:30 p.m.
CONCURRENT SESSION
J-1: Room 146A
J-2: Room 146B
J-3: Room 146C
J-4: Room 145AB
J-5: Room 147AB

1:30 – 3:15 p.m.
CONCURRENT SESSION
K-1: Room 146A
K-2: Room 146B
K-3: Room 146C
K-4: Room 145AB
K-5: Room 147AB

3:30 – 5:15 p.m.
CONCURRENT SESSION
L-1: Room 146A
L-2: Room 146B
L-3: Room 146C
L-4: Room 145AB
L-5: Room 147AB

Medical Assistance available in Room 149A.


**Session: A-1**

**Expanded Access and Issues with Administrative Data**

- **Room:** 146A
- **Organizer:** David Kashihara (Agency for Healthcare Research and Quality)
- **Session Chair:** Jeffrey Gonzalez (Economic Research Service)

**An Approach to Tiered Access in the Department of Veterans Affairs**

Michael Schwaber (U.S. Department of Veterans Affairs)

**Ethical Principles for the All Data Revolution – Repurposing Administrative and Opportunity Data**

Stephanie S. Shipp, Sallie Keller, and Aaron Schroeder (University of Virginia)

**Integrating Survey and Administrative Data Across Sources and Across Agencies to Create Statistical Products: A Case Study from Education**

Sarah Grady (National Center for Education Statistics) and Emily Isenberg (American Institutes for Research)

**Moving from Field Agents to Data Extracts: Anticipated Challenges and Benefits of Administrative Court Data**

Suzanne M. Strong (U.S. Department of Justice), KiDeuk Kim (Urban Institute), Cynthia Lee (National Center for State Courts)

---

**Session: A-2**

**Innovation in Health Insurance Data Collection**

- **Room:** 146B
- **Organizer and Session Chair:** Laryssa Mykyta (U.S. Census Bureau)


Edward Berchick, Heide Jackson and Laryssa Mykyta (U.S. Census Bureau)

**Two Times a Charm? Verifying Reports of Uninsurance in a National Survey**

Paul D. Jacobs and Patricia Keenan (Agency for Healthcare Research and Quality)

**Improving Measurement of VA Health Coverage among Military Veterans on the National Health Interview Survey**

Robin A. Cohen and Carla E. Zelaya (National Center for Health Statistics)

**Using Insurance Claims Data in the Medical Price Indexes**

John Bieler, Brian Parker and Daniel Wang (Bureau of Labor Statistics)

**Discussant:** Steven B. Cohen (Vice President, Statistical and Data Sciences, RTI International)
Session: A-3

**Education and the Workforce**
Room: 146C

**Organizer and Session Chair:** Andrew White (National Center for Education Statistics)

**Pathways to Jobs in the Skilled Technical Workforce**
J. Bayoan Santiago Calderon, Vicki Lancaster, and Sarah McDonald (University of Virginia)

**Demand and Supply of Skilled Technical Workers**
J. Bayoan Santiago Calderon, Vicki Lancaster, and Sarah McDonald (University of Virginia) and John Finamore (National Center for Science and Engineering Statistics)

**A Task-based Approach to Constructing Occupational Categories with Implications for Empirical Research in Labor Economics**
Julia Manzella, Gary Benedetto and Evan Totty (U.S. Census Bureau)

**High-School Dropout: Current Knowledge, Implications for Practice, and Future Directions**
Burhan Ogut, Ruhan Circi and Nevin Dizdari (American Institutes for Research)

**Bayesian Estimation of Program Effects and Their Heterogeneity between Programs in the Trade Adjustment Assistance Community College and Career Training Grants**
Stas Kolenikov and David Judkins (Abt Associates)

Session: A-4

**The Evidence Act 101**
Room: 145AB

**Organizers:** Jennifer Edgar (Bureau of Labor Statistics) and Keenan Dworak-Fisher (Office of Management and Budget)

**Moderator:** Katharine Abraham (University of Maryland)

**Panelists:**
- Diana Epstein (Office of Management and Budget)
- Sharon Boivin (Department of Education)
- Monique Eleby (U.S. Census Bureau)

**Discussant:** Emilda Rivers (Statistical Official, National Science Foundation and Director of the National Center for Science and Engineering Statistics)

Session: A-5

**Census 2020 Update**
Room: 147AB

**Organizer and Session Chair:** Gina Walejko (U.S. Census Bureau)

**2020 Census Self-Response Operations: Status Update One Month Into Data Collection**
Michael Bentley (U.S. Census Bureau)

**2020 Census Real-Time Analysis of Data**
Sarah Konya (U.S. Census Bureau)

**The 2020 Census Integrated Partnership and Communications Program in Action: Shape Your Future. START HERE.**
Monica Vines (U.S. Census Bureau)

**Mail Experiments During the 2020 Census**
Julia Coombs (U.S. Census Bureau)
Session: B-1
**Advances for Record Linkage for Science and Policy**
Room: 146A
**Organizer and Session Chair:** Jennifer Sinibaldi (National Center for Science and Engineering Statistics)

- **Advancing Survey Data for Research and Evaluation on Scientific Productivity** Wan-Ying Chang (National Center for Science and Engineering Statistics)
- **Constructing Patent Citations Networks of NSF Awards** Nicholas Daly (National Science Foundation)
- **Pre-Decisional Management of Overlapping Proposals at Funding Agencies** George Santangelo (National Institutes of Health)
- **Bayesian Record Linkage with Clustered Sub-Models (BRACS)** Jody Heck Wortman (Democratic National Committee)
- **Discussant:** Matthew Williams (National Center for Science and Engineering Statistics)

Session: B-2
**Privacy in the American Community Survey**
Room: 146B
**Organizer:** Rolando Rodriguez (U.S. Census Bureau)

- **A Roadmap for Privacy in the American Community Survey** Rolando Rodriguez (U.S. Census Bureau)
- **Evaluation and Tuning Insights from the ACS Fully Synthetic Data** Christine Task (Knexus Research Corporation)
- **Strengthening Privacy Protections in the American Community Survey - Lessons from the SIPP Synthetic Beta** Joanna Motro and Jordan Stanley (U.S. Census Bureau)
- **Challenges with Differential Privacy for Complex, Multi-stage Surveys** Chris Clifton (Purdue University), Eric Hanson (Brandeis University), Keith Merrill (Brandeis University), and Shawn Merrill (Purdue University)

Session: B-3
**How to Lie Inform with Statistics: Recent Advances in Communicating Research Findings**
Room: 146C
**Organizer:** Ignacio Martinez (Google)

- **The Challenges of Communicating Statistics to the Public: a Case Study of the Uganda Bureau of Statistics** Winny Nekesa Akullo (IASSIST, Public Procurement and Disposal of Public Assets Authority - Uganda)
- **Speaking on Data’s Behalf: What Researchers Say Affects How Audiences Choose** Dan Thal (Mathematica)
Bayesian Interpretation of Estimates (BASIE): A More Intuitive Way to Communicate Impact Findings than ‘Statistical Significance’
John Deke (Mathematica)

Applying the BASIE Framework to Interpret Democracy Prep Charter Schools’ Impact on Civic Participation
Mariel Finucane (Mathematica)

Discussant: Thomas Wei (Institute of Education Sciences)

Session: B-4

Implementing the Evidence Act: The Journey thus Far and the Road Ahead!
Room: 145AB
Organizers: Jennifer Edgar (Bureau of Labor Statistics) and Joe Parsons (National Agricultural Statistics Service)
Moderator: Hubert Hamer, Administrator, (National Agricultural Statistics Service)
Panelists:
- Ted Kaouk (Chief Data Officer, U.S. Department of Agriculture)
- Kelly Bidwell (Evaluation Officer and Statistical Official, General Services Administration)
- Samuel C. “Chris” Haffer (Chief Data Officer, U.S. Equal Employment Opportunity Commission)
- Matt Greene (Deputy Chief Data Officer for Governance, U.S. Department of Education)

Session: B-5

Working with Non-Probability Samples
Room: 147AB
Organizer and Session Chair: Tiandong Li (Health Resources and Services Administration)

Estimation of Population Characteristics from Web and Traditional Probability Samples in Case of Large Number of Potential Covariates
Vladislav Beresovsky (National Center for Health Statistics)

Adjusting for National Coverage and Selection Bias in Two Nonprobability Health Surveys: The Effect of Weighting Methods
Davia Moyse, Yangyang Deng, Matt Jans (ICF), Scott Worthge (Quest Mindshare), Sara Chung and Laura O’Campo (MFour)

Combining Probability and Non-Probability Samples using Propensity Modeling and Small Area Estimation
Nadarajasundaram Ganesh, Edward Mulrow, Michael Yang and Vicki Pineau (NORC at the University of Chicago)

Can Sample Surveys from Online Panels Support Evidence-Based Policy-Making Decisions?
Mansour Fahimi and Frances M. Barlas (Ipsos Public Affairs)

Is Panel Conditioning a Concern with Online Probability-based Panels?
Frances M. Barlas, Mansour Fahimi and Randall K. Thomas (Ipsos Public Affairs)

Concurrent Session C
Tuesday, April 14th
3:15 - 5:00 PM

Session: C-1
Leveraging Administrative Data
Room: 146A

Organizer and Session Chair: Erik Scherpf (NORC at the University of Chicago)

Using Administrative Records Data to Produce Business Statistics: The Nonemployer Statistics by Demographics Series (NES-D)
Adela Luque (U.S. Census Bureau)

Incorporating Administrative Data in Survey Weights for the Survey of Income and Program Participation
Ashley Westra and Jonathan Eggleston (U.S. Census Bureau)

The Impact of an Improved Sampling Frame - the 2018 National Sample Survey of Registered Nurses
Tiandong Li (Health Resources and Services Administration)

Blending Administrative Data with a Probability Sample of Nonparticipants to Produce National Estimates: The NCS-X NIBRS Estimation Project
Marcus Berzofsky, Dan Liao (RTI International) and Alexia Cooper (Bureau of Justice Statistics)

Characterizing Federal Funding of Research and Development Using Administrative Data
Samantha Cohen (University of Virginia), Sean Pietrowicz (University of Notre Dame) and Joel Thurston (University of Virginia)

Session: C-2

Applications of Machine Learning to Enhance Federal Statistics
Room: 146B
Organizer: Tala Fakhouri (National Center for Health Statistics)

Session Chair: Khair ElZarrad (Deputy Director of the Office of Medical Policy at the Food and Drug Administration’s Center for Drug Evaluation and Research)

Using Machine Learning to Improve Forecast Accuracy
Gianna Short (Economic Research Service)

Detecting Pharmaceutical Innovations in Text-Based Data Using Machine Learning
Gizem Korkmaz (University of Virginia), Gary Anderson (National Center for Science and Engineering Statistics), Devika Nair, Neil Kattampallil and Sallie Keller (University of Virginia)

Predicting Population or Subpopulation Estimates Using Machine Learning Algorithms: An Application using Survey and Administrative Records Data
Harold Gomes (Joint Program in Survey Methodology, University of Maryland)

A Data-Driven Method of Specifying Efficient Blocking Schemes for Record Linkage
Dean Resnick (NORC), Marc Roemer (National Center for Health Statistics), Scott Campbell (NORC) and Lisa Mirel (National Center for Health Statistics)

Exploring Non-Traditional Statistical Methods as Ways to Address Cultural, Social, and Linguistic Biases
Bradley Rentz (REL Pacific at McREL International) and Christina Tydeman (REL Pacific at McREL International, Assistant Secretary of Health and Human Services for Planning and Evaluation)

Session: C-3

Impact of Questionnaire Design on Data Quality
Room: 146C
Organizer and Session Chair: Jessica Graber (National Center for Health Statistics)

Empirically Assessing the Sensitivity of Survey Questions and Responses
Using Qualitative Data to Improve Surveys: Results from a Cognitive Interview Study of the National Study of Long-term Care Providers
Meredith Massey, Lauren Harris-Kojetin and Manisha Sengupta (National Center for Health Statistics)

Does the Wording Affect the Rates? An Experiment in Sexual Orientation and Gender Identity (SOGI) Measurement
Deirdre Middleton, Matt Jans, Yangyang Deng (ICF), Scott Worthge (Quest Mindshare), Sara Chung, Laura O’Campo (MFour) and Kerith Conron (UCLA Williams Institute)

Casting a Wider Net: Response Format Effects on Self-reported Individual and Household Disability
Randall K. Thomas and Frances M. Barlas (Ipsos Public Affairs)

Session: C-4
Using Data in New Ways: Leveraging the Evidence Act to Coordinate Evaluation, Statistics and Policy
Room: 145AB
Organizers: Jennifer Edgar (Bureau of Labor Statistics) and Erica Zielewski (Office of Management and Budget)
Session Chair: Jennifer Edgar (Bureau of Labor Statistics)

Framing the Evidence Act’s Vision for Coordination and Collaboration
Erica Zielewski (Office of Management and Budget)

U.S. Department of Housing and Urban Development’s Experience Supporting and Enhancing its Data Infrastructure and Use
Calvin Johnson (U.S. Department of Housing and Urban Development)

Linking State Medicaid Data and Child Welfare Data for Outcomes Research (ASPE and ACF/HHS)
Valeria Butler (Office of Planning, Research and Evaluation, Administration for Children and Families)

Department of Labor’s Data Exchange and Analysis Platform (DEAP)
Christina Yancey (Chief Evaluation Officer, Department of Labor), David Judkins (Abt Associates) and Scott Gibbons (Department of Labor)

Session: C-5
Reducing Barriers to Researcher Access
Room: 147AB
Time: April 14th 3:15-5:00
Organizer: Barbara Downs (U.S. Census Bureau)
Moderator: Michael Hawes (U.S. Census Bureau)
Panelists:
- Molly Dahl (Senior Advisor, Congressional Budget Office)
- Peter Meyer (Senior Fellow, HealthCare Programs, NORC at the University of Chicago and U.S. Office of Management and Budget)
- Kristen Monaco (Bureau of Labor Statistics)
- Barbara Downs (U.S. Census Bureau)
- Margaret Levenstein (Inter-University Consortium for Political and Social Research [ICPSR], University of Michigan)
Concurrent Session D
Wednesday, April 15th
8:30 - 10:15 AM

Session: D-4
Challenges and Advances in Measuring Sexual Orientation and Gender Identity in Federal Surveys: Best Practices, Recent Findings, and Public Use
Room: 145AB

Organizer and Session Chair: Nancy Bates (U.S. Census Bureau)

Improving Measurement of Sexual Orientation and Gender Identity in the Federal Statistical System
Sylvia Fisher (Health Resources and Services Administration) and Nancy Bates (U.S. Census)

How Would You Describe Yourself? Recent Developments in Sexual Orientation and Gender Identity Survey Measures
Maura Spiegelman (National Center for Education Statistics) and Christina Dragon (Department of Labor)

Pretesting SOGI Questions for Self and Proxy: How do In-Person Cognitive Interviews Compare to Online Testing?
Robin Kaplan (Bureau of Labor Statistics)

Accessing and Analyzing Sexual Orientation and Gender Identity in Public-Use Files
Andrew R. Flores (American University and the Williams Institute)

Session: D-5
Evaluating Data Quality
Room: 147AB

Organizer and Session Chair: Grace Medley (Substance Abuse and Mental Health Services Administration)

Nonresponse Bias Studies of the Consumer Expenditure Survey
Barry Steinberg, Sharon Krieger and Brett McBride (Bureau of Labor Statistics)

Using Process Data to Understand Non-Response in NAEP
Markus Broer, Ruhan Circi and Juanita Hicks (American Institutes for Research)

Rethinking Response Rate Calculations for Probability-based Samples from Online Panels
Mansour Fahimi and Frances M. Barlas (Ipsos)

Fitness for Use: Assessing Data Quality Through the Lens of Data Use
Daniel Dorfman and Adam Safir (Bureau of Labor Statistics)
Practical Tools for Evaluating Data Quality of Administration Records
Erin Tanenbaum and Zachary H. Seeskin
(NORC at the University of Chicago)

Session: D-6
Improvement of Opioid Use and Disorder Measurement: Putting Data into Action at the National and State Levels
Room: 156
Organizer: Lisa Wagner (National Center for Health Statistics)
Panelists:
- Jessica White (HHS Office of Assistant Secretary for Planning and Evaluation)
- Kristen Miller (National Center for Health Statistics)
- Stephanie Willson (National Center for Health Statistics)
- Fadia Shaya (University of Maryland School of Pharmacy)

Concurrent Session E
Wednesday, April 15th
10:45 - 12:30 PM

Session: E-1
Linking Extant Data to a National Survey for Official Statistics: The National Household Food Acquisition and Purchase Survey
Room: 146A

Organizer and Session Chair: Elina T. Page
(Economic Research Service)
Session: E-2

Measuring the Impact of Technology on the Economy
Room: 146B

Organizer and Session Chair: Javier Miranda
(U.S. Census Bureau)

An Analysis of Trade and Employment in Potentially ICT-Enabled Services in the United States
Badri Narayanan Gopalakrishnan (University of Washington Seattle) and Shounkie Nawani (Infinite Sum Modelling)

New Measures of Robot Use in U.S. Manufacturing
Catherine Buffington, Javier Miranda, and Rob Seamans (U.S. Census Bureau)

Measures of Robotic Expenditures from the Annual Capital Expenditures Survey
Valerie Mastalski, Catherine Buffington, Anne S. Russell and Javier Miranda (U.S. Census Bureau)

Business Dynamics Statistics of High Tech Industries
Nathan Goldschlag and Javier Miranda (U.S. Census Bureau)

Discussant: Emin Dinlersoz (U.S. Census Bureau)

Session: E-3

Using Data about the Data
Room: 146C

Organizer and Session Chair: Gavin Corral
(National Agricultural Statistics Service)

Lessons and Insights from the Use of Audit Trails to Analyze the Effect of Moving a Survey Question
Brandon Kopp and Lucilla Tan (Bureau of Labor Statistics)

Using Process Data to Study Interviewer Effects on Measurement Error and Nonresponse in the Consumer Expenditure Survey
John Dixon and Erica Yu (Bureau of Labor Statistics)

Improving Paradata Measures using Qualitative Analysis, Visualization and Machine Learning
Renee Ellis (U.S. Census Bureau)

Diving into Process Data: New Insights into NAEP Items
Ruhan Circi Kizil, Janita Hicks (American Institutes for Research) and Emmanuel Sikali (National Center for Education Statistics)

Metadata Projects Made Plain for U.S. Statistical Agencies
Peter B. Meyer, Daniel Gillman (Bureau of Labor Statistics) and Kathryn McNamara (U.S. Census Bureau)

Session: E-4

Data Ethics Frameworks
Room: 145AB

Organizer and Session Chair: Jessica Graber
(National Center for Health Statistics)

Ethical Considerations for Data Access and Use
Amy O’Hara (Georgetown University)

Policy and Technology: Ensuring Ethics in the Submission and Access of Biomedical Research Data
Dina N. Paltou (National Institutes of Health)

Ethical Issues in the Development of Complex Machine Learning Algorithms
Sara R. Jordan (Policy Counsel, Artificial Intelligence, Future of Privacy Forum)
Session: E-5

**Information Quality Frameworks for Alternative and Integrated Data**
Room: 147AB

**Organizers:** Joy Sharp (Energy Information Administration) and Dennis Fixler (Bureau of Economic Analysis)

**Session Chair:** Joy Sharp (Energy Information Administration)

**A Framework and Best Practices for Measuring and Reporting Data Quality**
Keenan Dworak-Fisher (Bureau of Labor Statistics)

**Total Error Frameworks for Integrated Survey and Found Data**
Paul Biemer (RTI International)

**Discussants:** Colm O’Muircheartaigh (University of Chicago) and Frauke Kreuter (University of Maryland)

---

Session: E-6

**Privacy and Accuracy for 2020 Census Data Products: The CNSTAT Workshop**
Room: 156

**Organizers:** Connie Citro and Daniel Cork (Committee on National Statistics)

**Session Chair:** Daniel Cork (Committee on National Statistics)

**Implications of Differential Privacy in 2020 for State and Local Revenues**
Nicholas Nagle (University of Tennessee)

**Effects of Differentially Private Noise Injection on Survey Operations**

Quentin Brummet (NORC at the University of Chicago)

**Differential Privacy and Mortality Rates in the United States**
Alexis Santos-Lozada (Pennsylvania State University)

**Demographic Findings of the 2010 Census Demonstration Product and Subsequent Refinement of the TopDown Algorithm for 2020**
Matthew Spence (U.S. Census Bureau)

**Summary of What Data Users Heard at the CNSTAT Workshop and What Kinds of Census-User Interactions Are Needed Going Forward**
V. Joseph Hotz (Duke University)

---

**Concurrent Session F**
**Wednesday, April 15th**
**1:45 - 3:30 PM**

Session: F-1

**Blended Data for Evidence-Based Research and Evaluation**
Room: 146A

**Organizer and Session Chair:** Matthew Williams (National Science Foundation)

**Blending Privately-Provided Payroll Data and Government Statistics**
Leland Crane (Federal Reserve Board)

**Public Libraries and Collective Efficacy: An Exploratory Study of Blending Data from the Public Libraries Survey and the American Housing Survey**
Lisa Frehill (Institute of Museum and Library Services)

Luca Sartore (National Agricultural Statistics Service)

**Measuring the cost and impact of open source software innovation on GitHub**
José Bayoán Santiago Calderón*(UVA), Brandon Kramer*(UVA) , Gizem Korkmaz (UVA), Carol Robbins (NCSES), Aaron Schroeder (UVA) , Sallie Keller (UVA)

**Discussant:** Nathan Cruze (National Agricultural Statistics Service)

**Session: F-2**
**Advances in Disclosure Avoidance**
Room: 146B
**Organizer:** Harrison Quick (Drexel University)

**Session Chair:** Monika Hu (Vassar College)

**Incorporating Economic Conditions in Synthetic Microdata for Business Programs: A Case Study**
Katherine Jenny Thompson (U.S. Census Bureau)

**Risk-Efficient Bayesian Data Synthesis for Privacy Protection**
Terrance Savitsky (Bureau of Labor Statistics)

**Comparative Study of Differentially Private Synthetic Data Algorithms and Evaluation Standards**
Joshua Snoke (RAND)

**Generating Poisson-Distributed Differentially Private Synthetic Data**
Harrison Quick (Drexel University)

**Discussant:** Tom Krenzke(Westat)

**Session: F-3**
**The National Corrections Reporting Program: Methodological Issues and Research**
Room: 146C
**Organizer:** Ryan Kling (Abt Associates)

**Moderator:** Ryan Kling (Abt Associates)

**Panelists:**
- Ryan Kling (Abt Associates)
- Christopher Cutler (Abt Associates)
- Walter Campbell (Abt Associates)
- Gerald Gaes (Independent Consultant)

**Session: F-4**
**Data Science: Capacity Building to Solve Real World Problems**
Room: 145AB

**Organizers:** Jennifer Edgar (Bureau of Labor Statistics) and Frauke Kreuter (University of Maryland)

**Moderator:** Frauke Kreuter (University of Maryland)

**Panelists:**
- Yulei He (National Center for Health Statistics)
- Alex Measure (Bureau of Labor Statistics)
- Mark Denbaly (Economic Research Service)
- Lisa Frid (U.S. Census Bureau)

**Session: F-5**
**Communicating Fitness for Use**
WEDNESDAY, APRIL 15

Room: 147AB
Organizer: Jennifer Parker (National Center for Health Statistics)
Moderator: Jennifer Parker (National Center for Health Statistics)
Panelists:
- Amy Branum (National Center for Health Statistics)
- Marilyn Seastrom (National Center for Education Statistics)
- Regina Nuzzo (American Statistical Association)
- Robert Sivinski (U.S. Office of Management and Budget)
- Samantha Tyner (Bureau of Labor Statistics)

Session: F-6
Applications of Non-Probability Samples
Room: 156
Organizer and Session Chair: Aaron Maitland (National Center for Health Statistics)
Lock Sampling, or: Yes, Panels are Different - Now What?
Jake Soffronoff (USPS Office of Inspector General)
Crowdsourcing for Recruiting Hard-to-Reach Populations: An Example of Recruiting Military Veterans
Y. Patrick Hsieh, Leyla Stambaugh and Herschel Sanders (RTI International)
Can We Do This Another Way? Potential Nonprobability Sample Sources for Social and Health Surveys
Matt Jans, Davia Moyse and Yang Yang Deng (ICF)
Systematic Evaluation of Respondent Driven Sampling Implementation

Sunghee Lee, Ai Rene Ong and R.J. Batas (University of Michigan)

Concurrent Session G
Wednesday, April 15th
3:45 - 5:30 PM

Session: G-1
Linkage Applications of ACS Data
Room: 146A
Organizer and Session Chair: Scott Wentland (Bureau of Economic Analysis)
Augmenting ACS Microdata with CPS using Machine Learning
Kendra Asher, Peter Meyer and Matthew Russell and Jay Stewart (Bureau of Labor Statistics)
Impacts of Broadband Development on Rural Property Values
Joshua Goldstein, Teja Pristavec and Devika T. Mahoney-Nair (University of Virginia)
Counting (on) Journalists: Using Federal Statistical Data to Estimate the Size, Characteristics and Geographic Distribution of Newsroom Employees in the United States
Elizabeth M. Grieco (Pew Research Center)
Veteran Population Projection Based on the Blended Data from the Administrative Records and American Community Survey
Jin Kim, Hyo Park, Charles Lin, and Tom Garin (U.S. Department of Veterans Affairs)

Session: G-2
Measuring Self-Employment Status and Income
Room: 146B
Organizer: Robert Munk (U.S. Census Bureau)
Session Chair: Mark Klee (U.S. Census Bureau)
Improving Self-Employment Imputations with Administrative Data and Model Based Imputations: An Analysis Using the Survey of Income and Program Participation
Jonathan Eggleston, Mark Klee and Robert Munk (U.S. Census Bureau)
Tax Burdens and Barriers for Self-Employed Taxpayers: Results from the IRS Estimated Tax Survey
Janet Li, Yan K. Liu, Brett Collins, Alicia Miller and Tomas Wind (Internal Revenue Service)
The Effect of Structural and Cyclical Changes on Trends Across Time in the Number of Independent Contractors
Anne Polivka (Bureau of Labor Statistics)
Reconciling Survey and Administrative Measures of Self-Employment
Katharine G. Abraham, John C. Haltiwanger (University of Maryland), Claire Hou (University of Maryland and U.S. Census Bureau), Kristin Sandusky and James R. Spletzer (U.S. Census Bureau)

Michael Jackson, Danielle Battle and Rebecca Medway (American Institutes for Research)
Recruiting a Probability Sample of 18 Year Olds for a Longitudinal Study on Interpersonal Violence
David Cantor, Reanne Townsend and Gail Thomas (Westat)
Mode Effects or Measurement Reliability: Differences in Estimates from Same Individuals between Web and Mail Survey Administration
John Boyle, Ronaldo Iachan and Matt Jans (ICF)
Using SMS as a Survey Recruitment Tool – Challenges Convincing Respondents and Carriers of Legitimacy
James Dayton, Robynne Locke and Rachel Kinder (ICF)
Survey Improvement Recommendations from a Field Staff
Graham Jones, Daniel Dorfman and Yezzi Angi Lee (Bureau of Labor Statistics)

Session: G-3
Innovations in Survey Design
Room: 146C
Organizer and Session Chair: Grace Medley (Substance Abuse and Mental Health Services Administration)
Optimizing Data Collection Procedures For Web-push Survey Designs: Evidence From A National Study
Javier Miranda (U.S. Census Bureau)
Automating Response Evaluation for Franchising Questions on the 2017 Economic Census
Joseph Staudt (U.S. Census Bureau)

Machine Learning and the Commodity Flow Survey
Christian Moscardi (U.S. Census Bureau)

Session: G-5
Transparent Reporting of Statistics
Room: 147AB
Organizer and Session Chair: Mark Prell (Economic Research Service)

Overview of the Transparent Reporting Project and Assessing the Data User’s Perspective
Mark Prell (Economic Research Service)

Quality Considerations when Integrating Alternative Data into the Consumer Price Index
Crystal Konny (Bureau of Labor Statistics)

Reporting on Integrated Data Quality for Linkage between the National Hospital Care Survey and the National Death Index
Lisa Mirel (National Center for Health Statistics)

Tom Garin (National Center for Veterans Analysis and Statistics)

Reporting on Integrated Data Quality for the National Postsecondary Student Aid Study
Chris Chapman (National Center for Education Statistics)
Session: G-6

The Evidence Act and Evaluation in Action

Room: 156

Organizer: Matthew Williams (National Center for Science and Engineering Statistics, NSF)

Session Chair: Jennifer Sinibaldi (National Center for Science and Engineering Statistics)

Panelists:

- Melissa Abelev (USDA Food and Nutrition Service)
- Russ Burnett (General Services Administration)
- Cynthia Phillips (National Center for Science and Engineering Statistics)
- Susan Queen (National Center for Health Statistics)
Concurrent Session H  
Thursday, April 16th  
8:30 - 10:15 AM

Session: H-1
Leveraging Multiple Health Care Data Sources to Generate Statistical Information in the U.S. Health Care Delivery System
Room: 146A  
Organizer and Session Chair: Denys Lau (National Center for Health Statistics)
Benefits and Challenges in Using Survey and Administrative Data in the National Post-Acute and Long-Term Care Study (NPALS)  
Lauren Harris-Kojetin (National Center for Health Statistics)
Lessons Learned from using Integrated Modes of Abstracted Medical Records and Electronic Health Records for Collection of Physician Data: National Ambulatory Medical Care Survey, 2016-2017  
Joseph Staudt (National Center for Health Statistics)
National Hospital Care Survey: Trials and Tribulations of Integrating Claims and Electronic Health Records Data  
Geoffrey Jackson (National Center for Health Statistics)
Discussant: Susan Queen (National Center for Health Statistics)

Session: H-2
Innovative Measures for Traditional Concepts
Room: 146B
Organizer and Session Chair: E. Ann Carson (Department of Justice)
Building a New Urban, Suburban, and Rural Indicator  
Emily Molfino (U.S. Census Bureau), Shawn Bucholtz (Housing and Urban Development) and Jed Kolko (Indeed)
Measuring the Small Business Economy  
Tina Highfill and Richard Cao (U.S. Bureau of Economic Analysis), Richard Schwinn (Small Business Administration), Richard Prisinzano (University of Pennsylvania) and Danny Leung (Statistics Canada)
Poverty in the U.S. Using the Comprehensive Income Dataset  
Bruce Meyer (University of Chicago, NBER, AEI, and U.S. Census Bureau), Derek Wu (University of Chicago) and Carla Medalia (U.S. Census Bureau)
A Nontraditional Data Approach to the CPI Gasoline Index  
Sarah Niedergall, John Bieler and David Popko (Bureau of Labor Statistics)
The Unemployment Rate Using Alternative Data: Estimating Labor Market Status with Consumer Bank Flows  
Benjamin Mandel and Boqiu Lu (J.P. Morgan Asset Management)
Session: H-3
Innovations in Small Area Estimation Models for Official Statistics Programs
Room: 146C
Session Organizer: Andreea Erciulescu (Westat)
Session Chair: Jean D. Opsomer (Westat)

Preserving Acreage Relationships in Small Area Agricultural Models
Lu Chen (National Institute of Statistical Sciences and National Agricultural Statistics Service)

Small Domain Estimation in the National Compensation Survey
Daniel Ayasse (Bureau of Labor Statistics)

Adopting Prior Distributions for the Variance-Covariance Matrices to Fully Specify an Area-Level Bivariate Hierarchical Bayes Three-Fold Model for Proportions of Adult Competency
Andreea Erciulescu (Westat)

A Bayesian and Spatial Modeling Approach for Multiple Correlated Health Outcomes: Application to Teen Births Data by Race and Hispanic Origin Groups
Diba Khan (National Center for Health Statistics)

Discussant: Robert Fay (Westat)

Session: H-4
Assessing Privacy Risk: Reconstruction & Re-Identification
Room: 145AB
Organizer: Michael Hawes (U.S. Census Bureau)
Moderator: Michael Hawes (U.S. Census Bureau)

Panelists:
- Simson Garfinkel (U.S. Census Bureau)
- Aref Dajani (U.S. Census Bureau)
- Ellen Galantucci (Bureau of Labor Statistics)
- Shawn Bucholtz (U.S. Department of Housing and Urban Development)

Session: H-5
Data Quality: Communication of Uncertainty in Official Statistics
Room: 145AB
Organizer: John Eltinge (U.S. Census Bureau)
Session Chair: Jennifer Parker (National Center for Health Statistics)

More Fully Capturing Uncertainty Associated with Official Estimates
Linda Young (National Agricultural Statistics Service)

Evaluating Uncertainty in Multiple Dimensions of Data Quality
John Eltinge (U.S. Census Bureau)

Tailored Transparency: Public Trust vs. Reproducibility
Peter Miller (Professor Emeritus at Northwestern University and U.S. Census Bureau, Retired)

Discussant: Jeffrey Gonzalez (Economic Research Service)
Concurrent Session J  
Thursday, April 16th  
10:30 - 12:15 PM

Session: J-1  
Applications in Survey Methodology  
Room: 146A  
Organizer and Session Chair: Jennifer Beck (National Center for Science and Engineering Statistics)

Improving the Quality of Data and Reducing Burden of the Public Libraries Survey  
Lisa M. Frehill (Institute of Museum and Library Services)

A Secondary Analysis of Interviewer Effects in the BRFSS  
Antonia Warren, Ting Yan (Westat) and Carol Pierannunzi (National Center for Health Statistics)

Interviewer Training for Classroom versus Distance Learning: Initial Skill Gains and Measures of Drift  
Hanyu Sun, Angie Kistler and Ryan Hubbard (Westat)

The Impact of the Pregnancy Checkbox and Misclassification on Maternal Mortality Trends in the US, 1999-2017  
Lauren Rossen, Lindsay Womack and Sayeedha Uddin (National Center for Health Statistics)

Lessons Learned from Statistical Analysis of Elections Returns to Help Improve Ballot Design: The Case of Name Order  
Jon A. Krosnick (Stanford University)

Session: J-2  
Improving Methodologies in Establishment Surveys  
Room: 146B  
Organizer and Session Chair: Ellen Galantucci (Bureau of Labor Statistics)

Towards Developing a Quantitative Measure for Response Burden for Coordinated Sampling of Annual Business Surveys  
Laura Bechtel and Diane Willimack (U.S. Census Bureau)

The Record-Keeping Practices of Medium-sized Multi-unit Businesses and Organizations  
Diane K. Willimack, Erica Marquette (U.S. Census Bureau), Struther Van Horn (Bureau of Labor Statistics) and Demetria Hanna (U.S. Census Bureau)

Understanding the Characteristics of Unresolved Matched Records in Capture-Recapture Methodology  
Denise A. Abreu (National Agricultural Statistics Service)

Sampling Strategies for the QCEW Business Supplement  
Sharon S. Stang and Emily deWolf (Bureau of Labor Statistics)

Session: J-3  
Better Than a “Guess-Estimate”: Methods for Estimation and Analysis  
Room: 146C  
Organizer and Session Chair: Janine McFadden (Bureau of Transportation Statistics)
Constructing Better Coverage Intervals for Some Estimators Computed from a Complex Probability Sample
Phillip S. Kott (RTI International)

A Decomposition Analysis of the Food Expenditure Series
Eliana Zeballos and Timothy Park (Economic Research Service)

Using Relative Distribution Methods to Analyze Federal Justice Data
Mark Motivans (Bureau of Justice Statistics)

An Easy Way to Calibrate on Partly Known Multiple Totals in Frequency Tables with Application to Real Data
Michael Sverchkov (Bureau of Labor Statistics)

The Relationship Between the Seasonal Regression Model-based F Test and a Diagnosis of Residual Seasonality
Demetra Lytras and Kathleen McDonald-Johnson (U.S. Census Bureau)

Session: J-4
Disclosure Review Boards: Design, Governance, Modernization
Room: 145AB
Organizer: Michael Hawes (U.S. Census Bureau)

Moderator: Michael Hawes (U.S. Census Bureau)

Panelists:
- Rob Sienkiewicz (U.S. Census Bureau)
- Daniell Toth (Bureau of Labor Statistics)
- Darius Singpurwalla (National Center for Science and Engineering Statistics)
- Greg Fortelny (U.S. Department of Education)

Session: J-5
Nonresponse Bias in Federal Surveys – Gaps in Knowledge and Future Opportunities
Room: 147AB
Organizer: Tala Fakhouri (National Center for Health Statistics)

Moderator: Tala Fakhouri (National Center for Health Statistics)

Constructing an Inventory of Non-response Bias Studies in Federal Surveys
Peter Miller (Professor Emeritus at Northwestern University and U.S. Census Bureau, Retired)

Developing and Assessing Weighting Methods for the Redesigned National health Interview Survey
James Dahlhamer (National Center for Health Statistics)

Finding the Right Auxiliary Information for Non-response Adjustment Models: In Search of Zs with Desirable Properties
Andy Peytchev (RTI International)

Estimating Survey Non-response Bias Using Tax Records Bruce Meyer (University of Chicago, NBER, AEI, and U.S. Census Bureau)
Concurrent Session K  
Thursday, April 16th  
1:30 - 3:15 PM

Session: K-1  
Linked Data from the Census Bureau for Evidence Building: Accessing the Data and Recent Results  
Room: 146A  
Organizer and Session Chair: Katie Genadek (U.S. Bureau of the Census)

Criminal Justice in the US and Economic Inequality: Results from the Criminal Justice Administrative Records System  
Keith Finlay (U.S. Census Bureau)

UMETRICS: Data For Examining How Research is Produced and How it Affects the Broader Economy  
Joseph Staudt (U.S. Census Bureau)

Results from the Evidence Building Project Series: Health at Birth, Later Life Achievement, and the Intergenerational Transmission of Advantage  
Sarah Miller (University of Michigan)

The Census Longitudinal Infrastructure Project – Linked Census Data and Results from the Impact of Preschool on Later-Life Outcomes  
Katie Genadek, U.S. Census Bureau

Session: K-2  
Integrating Information from Multiple Data Sources to Support Policies to Reduce Rural Health Disparities  
Room: 146B  
Organizer and Session Chair: Ernest Moy (U.S. Department of Veterans Affairs)

Healthy People: Exploring Urban/Rural Health Disparities in the Nation’s Health  
David T. Huang and Johanna Alfier (National Center for Health Statistics)

Integrating Information from Multiple Data Sources to Support Policies to Reduce Rural Health Disparities in AHRQ Quality and Disparities Report  
Barbara Barton and Karen Chaves (Agency for Healthcare Research and Quality)

Employing Geographic Concepts and Methods to Maximize the Effectiveness of Rural Health Programs  
Mark F. Guagliardo (Department of Veterans Affairs Office of Enterprise Integration)

Integrating Information from Multiple Data Sources by the VHA Office of Health Equity to Support Policies to Reduce Rural Health Disparities Among Veterans  
Ernest Moy and Kenneth Jones (Veterans Health Administration Office of Health Equity)

Discussant: Sarah Heppner (Federal Office of Rural Health Policy)
THURSDAY, APRIL 16

Session: K-3

**Innovative Uses of Imputation**

Room: 146C  
**Organizer:** Tamara Rib (Internal Revenue Service)  
**Session Chair:** Kimberly Henry (Internal Revenue Service)

**An Imputation Solution for Differentiating between Unreported Attitudes and Genuine Nonattitudes in Survey Data**

Jeff Gill (American University) and Natalie Jackson (PRRI)

**Applying Cluster Analysis to Improve the American Housing Survey Hot Deck**

Brian Shaffer, Stephen Ash and Kathy Zha (U.S. Census Bureau)

**Model Selection within Sequential Imputation: A New Treatment for Missing Values in High-Dimensional (Survey) Data**

Micha Fischer (University of Michigan)

**Survey Data - Editing and Imputation (E&I) Methods**

Menuka Ban (EY)

**Using Machine-Learning Algorithms to Improve Imputation in the Medical Expenditure Panel Survey**

Chandler McClellan, Emily Mitchell and Jerrod Anderson (Agency for Healthcare Research and Quality)

---

Moderator: Michael Hawes (U.S. Census Bureau)

Panelists:
- Barry Johnson (IRS Statistics of Income Division)
- Ashley Landreth (U.S. Census Bureau)
- Shelly Martinez (Office of Management and Budget)
- Jennifer Madans (National Center for Health Statistics)

Session: K-5

**Geospatial Act and You**

Room: 147AB  
**Organizer:** Wendy Martinez (Bureau of Labor Statistics)  
**Moderator:** Ed Strocko (Department of Transportation)

Panelists:
- Ivan Deloatch (U.S. Geological Survey, Federal Geographic Data Committee)
- Mike Ratcliffe (U.S. Census Bureau)
- Derald Dudley (Department of Transportation)
- Douglas Geverdt (National Center for Education Statistics)

---

Session: K-4

**Disclosure in an Era of Administrative Records and Data Sharing**

Room: 145AB  
**Organizer:** Michael Hawes (U.S. Census Bureau)
Concurrent Session L
Thursday, April 16th
3:30 - 5:15 PM

Session: L-1

Using Record Linkage to Assess a Social Outcome
Room: 146A
Organizer and Session Chair: Kevin Scott (Department of Justice)

The Project Talent-National Death Index Repository
Kelly Peters, Susan Lapham (American Institutes for Research) and Benjamin Chapman (University of Rochester Medical Center)

The Promises and Challenges of Linked Rent Data from the Consumer Expenditure Survey and Housing and Urban Development
John Voorheis, Garret Christensen and Nikolas Pharris-Ciurej (U.S. Census Bureau)

Linking ACS and IRS Data to Assess College Attendance and Completion by Family Income
Leah Clark, Jennifer Ortman, Nikolas Pharris-Ciurej and John Voorheis (U.S. Census Bureau)

Health and Retirement Study: Weight Adjustment for Linkage with Earnings Records from the Social Security Administration
Sirin Yaemsiri, Jeff M. Tessin, Michael J. Collins, Jennifer A. Gregory, Kathleen C. McQueeney and Christopher Zbrozek (U.S. Government Accountability Office)

Linking State Medicaid and Child Welfare Data for Parental Behavioral Health Services and Child Welfare Outcomes
Emily Madden (Assistant Secretary of Health and Human Services for Planning and Evaluation), Valeria Butler (Office of Planning, Research and Evaluation, HHS) and Robin Ghertner (Assistant Secretary of Health and Human Services for Planning and Evaluation)

Session: L-2

Data Modernization at the Equal Employment Opportunity Commission
Room: 146B
Organizer and Session Chair: Jiashen You (U.S. Equal Employment Opportunity Commission)

Improving Employer Data Collection- The Journey to Modernization of the U.S Equal Opportunity Commission’s Employer Information EEO-1 Survey
Margaret Noonan (U.S. Equal Employment Opportunity Commission)

Estimating, Describing and Locating U.S. Workers Vulnerable to Workplace Discrimination
Benjamin Overholt (U.S. Equal Employment Opportunity Commission)

Recent Advances in Data Access for EEO-1 Survey
Justin West (U.S. Equal Employment Opportunity Commission)

Discussant: Nick Hart (Data Coalition)

Discussant: Chris Haffer (Chief Data Officer, U.S. Equal Employment Opportunity Commission)
Session: L-3
Translating High-Profile Statistics for Policymakers and the Public: Lessons from the Drug Overdose Epidemic
Room: 146C
Organizer: Renee Gindi (National Center for Health Statistics)
Moderator: Renee Gindi (National Center for Health Statistics)
Panelists:
- Renee Gindi (National Center for Health Statistics)
- Holly Hedegaard (National Center for Health Statistics Division of Analysis and Epidemiology)
- Jeff Lancashire (National Center for Health Statistics Public Affairs Office)
- Susan Queen (National Center for Health Statistics Office of Planning, Budget and Legislation)

Assessing the FJSP Linking Algorithms using Synthetic Data
Ryan Kling and David Izrael (Abt Associates Inc.)

Steps Toward Procedures for Disclosure Review of Qualitative Research Reports
Joanne Pascale, Nancy Bates and Diane Willimack (U.S. Census Bureau)

Session: L-4
Disclosure Review Risk Assessments
Room: 145AB
Organizer and Session Chair: Jacob Bournazian (Energy Information Administration)

Quantifying the Disclosure Risk for Public Data Products: Theories and Practices
Jianzhu Li and Tom Krenzke (Westat)

Reliance on Goodness-of-Fit Criteria to Make an Accurate Assessment of Re-Identification Risk for Survey Microdata
Lin Li, Jianzhu Li and Tom Krenzke (Westat)

Session: L-5
Big Survey Meets Big Data: Integrating Administrative Data into the American Community Survey
Room: 147AB
Organizer: Jennifer M. Ortman (U.S. Census Bureau)
Moderator: Jennifer M. Ortman (U.S. Census Bureau)
Discussant: Quentin Brummet (NORC at the University of Chicago)
Panelists:
- Sandra L. Clark (U.S. Census Bureau)
- R. Chase Sawyer (U.S. Census Bureau)
- Jonathan Rothbaum (U.S. Census Bureau)
- Nikolas D. Pharris-Ciurej (U.S. Census Bureau)
Abstract Booklet

This section represents abstracts received as of March 5, 2020.

The following abstracts have not been edited for content.
An Approach to Tiered Access in the U.S. Department of Veterans Affairs Office of Enterprise Integration’s Office of Data Governance and Analytics (DGA)
Michael Schwaber (U.S. Department of Veterans Affairs)

This paper describes DGA’s development of tiered data access in an evolving environment of privacy, as well as future plans to expand access while protecting confidentiality. DGA is an organization that stores, links, processes, and distributes large amounts of veteran data, some of which contain personally identifiable information (PII). Recent updates to data policy and new laws encourage more data access across agencies and to the public. This has led to DGA’s examination of its disclosure risk mitigation strategies to better protect its data. DGA is developing access tiers defined by combinations of protection levels on each of the elements of the “Five Safes” framework.

Ethical Principles and Data Science – Repurposing Administrative and Opportunity Data
Stephanie S. Shipp (University of Virginia), Sallie Keller (University of Virginia), Aaron Schroeder (University of Virginia)

The data revolution has transformed the conduct of social science research through the incorporation of data science, but ethical dimensions should not be compromised. Researchers can now observe behavior based on repurposing existing administrative and opportunity data without consent or awareness by those providing the data. The principles set forth in the Belmont Report on Ethical Principles and Guidelines for the Protection of Human Subjects of Research are still as applicable as when these principles were first established in 1978. Discussions about ethics need to be a natural part of every research project, especially when repurposing data for analytical purposes. A publicly-shared ethical checklist at each research stage can help researchers identify and frame any potential concerns and evaluate their relative impacts. A key part of this checklist is the assessment of implicit biases. Ethical principles require the implementation of everyday practices around documentation, transparency, ongoing discussion, questioning, and constructive criticism. We will discuss the history of these ethical principles and our experiences implementing them into our research.

Integrating Survey and Administrative Data Across Sources and Across Agencies to Create Statistical Products: A Case Study from Education
Sarah Grady (National Center for Education Statistics), Emily Isenberg (American Institutes for Research)

The National Center for Education Statistics (NCES), within the U.S. Department of Education (ED), developed supplementary geocode data files for the Early Childhood Program Participation, Parent and Family Involvement in Education, and Adult Training and Education surveys of the 2016 National Household Education Surveys Program. The geocode files use sample members’ addresses to integrate data from other federal agencies and ED administrative data collections. The data files include new radius-based measures of household proximity to educational opportunities and job search assistance. The presentation will provide an overview of how the geocode files demonstrate some of the goals of evidence-based policymaking. The presentation will also discuss some of the challenges inherent in creating the files. It will discuss the challenges encountered in identifying auxiliary data sources, evaluating them for appropriateness, and in assessing disclosure risk of the resulting files. Data timeliness and cost will also be discussed.
Moving from Field Agents to Data Extracts: Anticipated Challenges and Benefits of Administrative Court Data
Suzanne M. Strong (U.S. Department of Justice), KiDeuk Kim (Urban Institute), Cynthia Lee (National Center for State Courts)

This presentation will focus on the Bureau of Justice Statistics (BJS) moving from a field agent model of courts data collection for the National Judicial Reporting Program (NJRP) to the request for administrative data extracts in the Analysis of Publicly Available Court Data (APACD) model. BJS and their partners, Urban Institute and the National Center for State Courts (NCSC), will discuss the pilot test for the APACD and outline the expected changes from the NJRP model to the APACD model. The discussion will include the increased availability of data elements not collected in NJRP, the use of a census-sample strategy, and strategies to maximize court participation.

CONCURRENT SESSION A-2
Innovations in Health Insurance Data Collection

Edward Berchick (U.S. Census Bureau), Heide Jackson (U.S. Census Bureau), Laryssa Mykyta (U.S. Census Bureau)

After a decade of research suggesting that the CPS ASEC captured less health insurance coverage than other federal surveys, the Census Bureau implemented a two-stage redesign. A new questionnaire was introduced in 2014, and a new “processing system” for extracting and imputing data was introduced in 2019. Rising nonresponse makes it critical to evaluate how post-collection survey procedures—not just the questionnaire—contribute to accurate estimates and improved data quality.

In this paper, we use data from the 2017 CPS ASEC Production and Research Files and the 2018 CPS ASEC Production and Bridge Files, to assess the effect of processing system changes on CPS ASEC health insurance estimates. To do so, we ask: (1) How do health insurance estimates with the updated processing system compare with existing estimates? and (2) Do new imputation methods improve estimates of the uninsured rate in the United States?

Results from linear probability models reveal that the updated data processing system improves estimates of health insurance coverage and addresses previously noted limitations of the CPS ASEC. Updates to data extraction and imputation captured additional coverage and reduced the overestimate of multiple forms of coverage. These results highlight the contribution of post-collection data processing for improving data quality in surveys that captures health insurance information.

Two Times a Charm? Verifying Reports of Uninsurance in a National Survey
Paul Jacobs (Agency for Healthcare Research and Quality), Patricia Keenan (Agency for Healthcare Research and Quality)

To improve health insurance coverage estimates, the Medical Expenditure Panel Survey Household Component (MEPS-HC) added a verification series. The series confirms whether individuals who did not initially report coverage actually had any health insurance coverage during the survey round. The MEPS verification series was based on the Current Population Survey (CPS) verification questions and adapted to the MEPS context. In this analysis, we use 2017 MEPS-HC data to examine the impact of the verification questions on coverage. Using the same respondents, we compare insurance coverage rates before and after taking into account responses to the verification series. We examine the percent reporting coverage through verification, the difference in coverage rates, overall and by coverage types (public versus private, employer, nongroup,
Medicaid/CHIP, and Medicare). In particular, we explore whether the verification series helped address the Medicaid undercount, or whether it resulted in increases in other categories, such as private coverage rates (as was found for the CPS verification question). We also examine whether responses varied by sub groups, such as educational attainment, income, age, and family size or structure.

**Improving Measurement of VA Health Coverage Among Military Veterans on the National Health Interview Survey**
Robin A. Cohen (National Center for Health Statistics), Carla E. Zelaya (National Center for Health Statistics)

In 2017, the Department of Veterans Affairs (VA) estimated that approximately 8.8 million veterans enrolled in, and 6.0 million utilized the VA health care system. However, National Health Interview Survey (NHIS) estimates of VA health care coverage through self-report fell short (3.0 million) of these administrative statistics. Since 1993, VA health care coverage has been included as a response option in the NHIS health insurance section. However, veterans who do not consider the VA a primary source of care, use VA health care for specific health services (e.g., mental health care), or an enrolled veteran who has never utilized VA health care, may not report VA coverage.

In 2018, mirroring a similar approach to address undercounts of Medicare and Medicaid coverage, we added a new question to the veteran section of the NHIS: “[Have you/has {person}] ever used or enrolled in VA health care?” Results show an increase in the reporting of VA health care coverage to 8.8 million, the same as the 2018 VA estimate. Respondents responding “Yes” to the probe were more likely to have private or public health coverage, be under age 65, employed, or in fair or poor health, and were less likely to be divorced or separated or to be the family respondent. Our results suggest that having more than one question on the same concept may minimize measurement error of complex concepts (e.g., linked to social and other identities).

**Using Insurance Claims Data in the Medical Price Indexes**

The use of medical claims data in the construction of the medical price indexes presents many opportunities and challenges for the Bureau of Labor Statistics. This project seeks to develop a feasible methodology for supplementing manual price collection in the medical CPI using insurance claims data. As part of an initial feasibility study, we constructed price indexes using data purchased from an insurance company for a large city and compare them to the CPI medical indexes. Some of the practical issues we consider are the effect of the time lag involved in using claims data, weighting issues related to the use of both claims data and traditional manually collected data, and the high variability of prices in the claims data at the level of specific provider and service. The results of our preliminary analysis show promise for the use of claims data. We have purchased data representing claims nationwide and are developing methods for incorporating monthly updates from the data provider into the production framework of the CPI.

**CONCURRENT SESSION A-3**
Tasks in Education and Workforce

**Pathways to Jobs in the Skilled Technical Workforce**
Samantha E Cohen, J. Bayoán Santiago-Calderón and Devika Nair (University of Virginia)

Skilled Technical Workforce (STW) jobs require a high level of knowledge in a technical domain but do not require a bachelor’s degree; they are essential to the U.S economy and can provide a path to the middle class. The pathways for producing STW candidates is poorly understood, preventing the development of evidence-based policies needed to grow this crucial class of workers. Currently there are no federal surveys that collect comprehensive education, training, and employment data on the STW. Our research explores the feasibility of using online resume data to fill this data gap.
We used Burning Glass Technology (BGT) resume data to develop a methodology for describing pathways to a job in STW. BGT resume data are constructed from millions of online resumes, the data include non-degree credentials, college degrees, employment history, and skills. Using these data, we have developed methodologies for characterizing how job seekers enter, remain relevant, and advance within the STW. The goal is to provide policymakers with the information needed to create and implement policies that encourage job seekers to train for jobs in the STW.

**Demand and Supply of Skilled Technical Workers**  
Vicki Lancaster (University of Virginia), John Finamore (National Center for Science and Engineering Statistics)

This talk explores the supply and demand of the Skilled Technical Workforce (STW) using real-time labor market information collected by Burning Glass Technologies (BGT). A job in the STW does not require a bachelor's degree for entry but requires knowledge of a technical domain. STW jobs can provide job seekers with the opportunity to enter the middle class without incurring the debt of a four-year college degree while filling the jobs needed to keep the U.S. competitive in the global market. Due to a shortage of skilled technical workers, policymakers want knowledge of those skills and credentials demanded by employers that are in short supply. This talk describes our approach for determining the fitness-for-use of the job-ad and resume data scraped from the web by BGT. Based on these results we selected and derived variables to construct occupation match matrices which aligned the skills demanded by employers and supplied by job seekers. The matrices allowed us to identify the skills and skill combinations in short supply for STW occupations. This proof-of-concept was developed using BGT data from the state of Virginia for the years 2016 and 2017.

**A Task-based Approach to Constructing Occupational Categories with Implications for Empirical Research in Labor Economics**  
Julia Manzella (US Census Bureau), Gary Benedetto (US Census Bureau), Evan Totty (US Census Bureau)

Researchers analyzing labor market outcomes typically account for the role of occupations by controlling for occupational categories. Such categories are often derived from the Standard Occupation Classification (SOC) coding scheme, which is based largely on narratives or qualitative measures of workers' tasks. Alternatively, we propose two quantitative task-based approaches to constructing occupational categories by using factor analysis with O*NET job task descriptors. We find that our task-based approach outperforms the SOC-based approach in terms of lower occupation distance measures. We also replicate a recent analysis and find that our task-based occupational categories explain more of the gender wage gap than the SOC-based approaches. Our study enhances the Federal Statistical System’s understanding of the SOC codes, investigates ways to use third-party data to construct useful research variables that can potentially be added to Census Bureau data products to improve their quality and versatility, and sheds light on how the use of alternative occupational categories in economics research may lead to different empirical results and deeper understanding of economic outcomes.

**High-School Dropout: Current Knowledge, Implications for Practice, and Future Directions**  
Burhan Ogut (American Institutes for Research), Ruhan Circi, (American Institutes for Research), Nevin Dizdari (American Institutes for Research)

High school dropout is a significant issue that impacts both the individual and the greater society. It negatively affects a persons’ well-being and their overall productivity in society. More research is needed to understand potential causes and possible remedies for high school completion. We used data from the High School Longitudinal Study of 2009 (HSLS:09) which is a nationally representative, longitudinal study of 24,000 9th graders in 944 schools. Transcript data – which were collected as part of the HSLS:09 – provided course-level information such as the
time/semester that students attended a course and student performance in the respective class. This creates robust data to explore and model students’ drop out tendency. In this study, we engineer a comprehensive set of aggregated variables at varying levels of granularity and compare multiple models to predict student drop out; logistic regression, random forests, gradient boosting, and survival analysis were used. Our results show that we can reliably predict high school dropout.

Bayesian Estimation of Program Effects and Their Heterogeneity between Programs in the Trade Adjustment Assistance Community College and Career Training Grants
Stas Kolenikov (Abt Associates), David Judkins (Abt Associates)

We analyze the DOL Evaluation of Round 4 Trade Adjustment Assistance Community College and Career Training (TAACCCT) data to provide estimates of the characteristics of participants, their service utilization, and their outcomes related to education, employment, earnings and receipt of public benefits. One of research questions of the outcomes study is to address variability of the success rates (program completion, employment in targeted field and earnings) across programs. Using Bayesian hierarchical framework, with programs as random effects, we analyze both continuous outcomes (income) and binary outcomes (program completion, college credit, etc.). We adjust for overshrinkage typical for Bayesian estimates, where the estimates (in this case, average outcomes at the level of individual programs) are shrunk towards the overall mean using Ghosh’s (1992) constrained Bayes estimation approach, where the constraints are placed on the first two moments of the Bayes estimates. We incorporate constrained Bayes approach into our workflow, and demonstrate how it affects the estimates, especially those for small programs, and those of the range of outcomes in super population of programs.

CONCURRENT SESSION A-4
The Evidence Act 101

The Evidence Act 101 session will provide a high-level overview of the main components of the Foundations for Evidence-Based Policymaking Act of 2018 (the Evidence Act), as well as the motivation and vision behind it. With an extended introduction from Dr. Katharine Abraham, who was involved with the original Commission for Evidence-Based Policymaking, the audience will hear about learning agendas and how to cultivate plans for evidence building, data governance and data inventories, and the presumption of accessibility to data that the Act provides. The session will conclude with a discussion of what the Act means for statistical agencies.

CONCURRENT SESSION A-5
Census 2020 Update

2020 Census Self-Response Operations: Status Update One Month Into Data Collection
Michael Bentley, U.S. Census Bureau

The purpose of the 2020 Census is to conduct a census of population and housing and disseminate the results to the President, the states, and the American people. The goal of the 2020 Census is to count everyone once, only once, and in the right place. This decennial census includes numerous innovations to modernize the census, including an online response option and a tailored contact strategy with up to five mailings to each housing unit. The Census Bureau expects the plurality of households to respond online, though other self-response options include calling Census Questionnaire Assistance, or returning a paper questionnaire through the mail. Self-response reduces the need to conduct expensive in-person follow-up for the enumeration.
This presentation will provide an overview and status update on the 2020 Census, about one-month into self-response data collection. The self-response mailings are scheduled to begin on March 12, 2020. This talk will include metrics on the preliminary self-response rate, discussion on how this compares to projections, and other operational considerations and findings to date.

2020 Real Time Analysis of Data
Sarah Konya, U.S. Census Bureau

The Real Time Analysis of Data (RTAD) project will support the operations and management of the 2020 Census by monitoring potential data quality issues, identifying areas of concern, and providing recommendations to intervene as appropriate. RTAD will produce several metrics including response rates, average time to complete survey, completeness of Address Canvassing workload, and number of Transitory Units identified. The analysis will be done by different characteristics as well as at several levels of geography, ranging from the nation down to the tract level. RTAD will compare actual events to benchmarks such as projected response rates, population demographics, and past decennial test results for metrics like average time to complete the survey. Additionally, the data will be monitored for outliers. This presentation will discuss some of the analyses from the first weeks of 2020 Census data collection.

The 2020 Census Integrated Partnership and Communications Program in Action: Shape your future. START HERE.
Monica Vines, U.S. Census Bureau

The 2020 Census is the third benefitting from Congressional funding to conduct a communications campaign to raise awareness and drive participation. At the time of this conference, the 2020 Census Integrated Partnership and Communications program (IPCP) will be fully operational with a completed education phase and amidst the motivation phase. In these phases, the IPC program first defined what the census is and why it is important to participate and is now urging everyone to respond immediately. In order to appeal to an increasingly diverse population and to break through today's crowded media environment, the IPCP, founded on large scale research efforts and rigorous testing, contains a plethora of components, all connected through a tagline (Shape your future. START HERE) and tailored to reach each audience with the right message, in the right place, and at the right time. These include paid advertising, both traditional (e.g., TV, radio, and print) and digital (e.g., online advertising and within social media platforms); a rich website with tailored content; a strong social media presence across many platforms (e.g., Facebook, Twitter, Instagram, etc.); local, regional, and national partners to help the Census Bureau connect with the public through “trusted voices”; and a Statistics in Schools (SIS) program to reach people through their school aged children. This presentation will give a real time glimpse into the IPCP through example advertisements highlighting customization across audiences and media types; a look at the 2020 Census website and social media channels; and materials from the SIS program. Furthermore, there will be a discussion of how Census is managing the campaign through multiple inputs like real time response rates, findings from tracking social media, data from an attitudinal survey, partner feedback and insights, and more, while optimizing the deployment of resources and tactics as possible based upon aggregated analysis.
Mail Experiments during the 2020 Census
Julia Coombs, U.S. Census Bureau

The U.S. Census Bureau has historically taken advantage of the once-in-a-decade awareness and environment that occurs during the decennial census to test new methods of conducting the census. This testing helps the Census Bureau evaluate promising techniques that may not otherwise be feasible to include in mid-decade tests. This presentation will describe two experiments planned for the 2020 Census that will occur concurrently with the production 2020 Census self-response campaign. These two experiments include panels that alter the content, format, and timing of the 2020 Census self-response invitations mailed to most housing units in the United States with the primary goal of improving self-response rates. In addition to a review of the treatments and alternative mail materials, sampling procedures and operational mailout plans for these experiments will also be discussed.

CONCURRENT SESSION B-1
Advances for Record Linkage for Science and Policy

Advancing Survey Data for Research and Evaluation on Scientific Productivity
Wan-Ying Chang (National Center for Science and Engineering Statistics)

The National Center for Science and Engineering Statistics (NCSES) is collaborating with bibliometric data providers and survey data stakeholders to build a network of linked data for research and evaluation. The key questions pertain to the impact of scientific productivity on career transitions and advancements of the U.S.-trained science, engineering and health (SEH) doctorate population. A broad range of NCSES surveys on education, workforce, research facilities, and R&D funding and expenditure are linked and then further enhanced by additional linkage to multiple scientific publication and citation databases. The network of linked data is used to investigate research output, co-authorship and collaboration patterns among the early career doctorate holders. In addition, we assess the impact of scientific publication on employment outcomes while controlling for rich sociodemographic and training data unique in the NCSES surveys. Linkage methodology and data quality evaluation processes will also be discussed.

Constructing Patent Citations Networks of NSF Awards
Nicholas Daly (National Center for Science and Engineering Statistics)

Using publicly available data from the USPTO we extracted references to National Science Foundation funded awards over a ten-year period. By linking the patent citations with NSF data, we were able examine the number of awards by directorate, and time between an award was granted, and a patent application was filed. By combining NSF data and patent citation data a social network was developed which showed the citation of NSF awards in USPTO data, and continued this to identify patents that these NSF related patents were later cited in. A single NSF award may be cited in dozens of patents, and are later cited in several future patents. NSF funded research concentrates on fundamental science, while some of the later cited patents may be found in consumer-based products. Over 4,000 awards were found to be cited across over 7,000 patents. By looking at the citation of patents that directly cite an NSF award it is possible to look at the wider impact of NSF related patents.
Pre-decisional Management of Overlapping Proposals at Funding Agencies
George Santangelo (National Institutes of Health)

The GSA 10X program has funded a collaborative effort between the National Institutes of Health and the National Science Foundation to enable sharing of pre-decisional grant application data while protecting personally identifiable information and intellectual property of the respective agencies, in accordance with the Privacy Act of 1974. We are developing a method that takes advantage of deep-learning content analysis (word2vec) and blockchain technology to support data sharing, identify topically-similar applications, and reduce potential duplication in funding across the government. The method and policy implications will be discussed.

Bayesian Record Linkage With Clustered Sub-models (BRACS)
Jody Heck Wortman (Democratic National Committee)

In this presentation, we apply an algorithm for Bayesian record linkage with clustered sub-models (BRACS) to 2016 voting data from North Carolina. We describe the process of provisional voting and the list of provisional voters provided by the North Carolina Board of Elections. We provide background on the North Carolina voter file, of which we use a snapshot from November 2016. We outline the limitations of exact-matching the two files using only the state-provided identifiers. Finally, we use BRACS to link the two files in order to estimate the number of removed voters who cast provisional ballots in the November 2016 election in North Carolina. We estimate a higher number of removed voters attempting to vote with BRACS than using exact-matching, which suggests that aggressive purging of voter rolls may have a higher impact on reducing voter turnout than previously thought.

CONCURRENT SESSION B-2
Privacy in the American Community Survey

A Roadmap for Privacy in the American Community Survey
Rolando Rodríguez (U.S. Census Bureau)

The 2020 Census will be the first census to use formal privacy methods to protect respondent identities and attributes. While the use of formal privacy methods for the American Community Survey (ACS) will not occur before 2025, the U.S. Census Bureau will continue to investigate the use of other methods to bolster privacy in ACS public data releases. Such efforts are crucial against the backdrop of increasing availability of personal data and the computational power to harness that data for re-identification. We give a roadmap for ACS privacy efforts, focusing on those efforts aimed at creating a synthetic ACS microdata file and an accompanying validation server. These products will provide ACS microdata users with several improvements and opportunities, including a larger sample, greater geographic detail, transparency about privacy practices, and the ability to validate results.

Evaluation and Tuning Insights from the ACS Fully Synthetic Data
Christine Task (Knexus Research Corporation)

As part of its effort to strengthen privacy protections and improve access and transparency for the American Community Survey (ACS), the U.S. Census Bureau is considering the creation of a synthetic ACS file and accompanying validation server at full sample size and with finer geographic detail. In this work, we provide details regarding the modeling process used to create a synthetic ACS file. We begin by discussing several techniques for evaluating the accuracy of the synthetic data, reviewing distributional similarity, propensity metrics, and analytic results. Then, starting from a baseline algorithm using a sequence of classification trees, we produce significant improvements in terms of runtime, quality, and privacy, through enhancements to data encoding, models, and synthesis flow.
Strengthening Privacy Protections in the American Community Survey - Lessons from the SIPP Synthetic Beta
Joanna Motro (U.S. Census Bureau), Jordan Stanley (U.S. Census Bureau)

In pursuing disclosure avoidance in the American Community Survey, synthetic data presents one avenue for protecting privacy. Using internal data to validate statistical output based on synthetic data is one approach to getting output the data user wants while mitigating disclosure risk. This paper summarizes our experiences with an internal validation system for the SIPP Synthetic Beta (SSB), a publicly available product of the U.S. Census Bureau. The underlying data for the SSB links select information from the Survey of Income and Program Participation (SIPP) to select administrative records. Data users apply for access to the SSB, and, after gaining access, develop programs to perform statistical analysis on this synthetic data set. Users can send well-functioning programs to internal Census employees who run the programs on a corresponding internal file to validate the requested statistical output. Any output requested for release is subject to standard Census Bureau disclosure requirements before being released to the data user. This paper summarizes the design of the synthetic data and validation system and provides a quantitative comparison of the synthetic to internally validated output from past research. We discuss lessons learned and key principles that could guide future implementations of internal validation systems for externally available synthetic data.

Challenges with Differential Privacy for Complex, Multi-stage Surveys
Chris Clifton (Purdue University), Eric Hanson (Brandeis University), Keith Merrill (Brandeis University), Shawn Merrill (Purdue University)

Survey-based statistical products such as the American Community Survey pose additional challenges for differential privacy beyond those seen in enumeration-based products such as the decennial census. Some of these are obvious, such as increased dimensionality, but others are more subtle. We will detail two such issues, missing data imputation and post-stratification, and show why they pose difficulties for differential privacy. We will discuss how improvements in methodology can both improve accuracy, and enable formal privacy with reduced impact on the outcomes. We end with recommendations for those developing new statistical methodologies who want these to be amenable to formal privacy.

CONCURRENT SESSION B-3
How to Inform with Statistics: Recent Advances in Communicating Research Findings

Bayesian Interpretation of Estimates (BASIE): A More Intuitive Way to Communicate Impact Findings than "Statistical Significance"
John Deke (Mathematica)

Determining which impact estimates are worthy of attention is challenging. Historically, statistically significant estimates have been deemed worthy of attention. But statistical significance is often misinterpreted. For example, a statistically significant impact estimate is often misinterpreted to mean that there is a very high probability (say, 95 percent) that an intervention works. In this presentation we demonstrate the potential size of these misinterpretations and describe the BASIE framework for interpreting impact estimates. In federal impact evaluations, we expect findings to be based on quantifiable evidence, not personal belief. While statistical significance is not based on personal beliefs, it does not answer the question policy-makers most likely want to know: what is the probability that an intervention was effective? Bayesian methods can answer this question, but they often do so by drawing on prior beliefs about the intervention being studied. The advantage of BASIE is that it answers the question of interest to policy-makers while only relying on evidence, not belief. BASIE also has the advantage of being applicable to extant findings generated using conventional research methods – it is not necessary to use computationally intensive, specialized software.
Speaking on Data’s Behalf: What Researchers Say and How Audiences Choose
Ignacio Martinez (Mathematica)

In this paper we assess how presenting results in different ways can influence the decisions that people make. We present people with the exact same data described using either a traditional frequentist test of the null hypothesis (default frequentist), or alternative Bayesian tests that describes potential outcomes in probabilistic terms. We also describe results either in relation to only the null hypothesis, or also in relation to the minimum effect of policy interest. We provide the first evidence that we are aware of that different statistical methods lead to different policy recommendations. In our experiments (conducted on a convenience sample of adult Americans), we show that people are more likely to switch to a new educational technology and were more confident in their choice when they see the results of Bayesian null hypothesis testing than when they see a more conventional frequentist test. Importantly, people are also sensitive to information about the minimum effect of interest, and are reluctant to advocate change for effects that fall below this threshold. These findings highlight the importance of using statistical tests that are aligned to the decision that practitioners must make.

Applying the BASIE Framework to Interpret Democracy Prep Charter Schools’ Impact on Civic Participation
Mariel Finucane (Mathematica)

Democracy Prep is a charter school network with an explicit mission to prepare students for effective citizenship. We estimated the causal impact of Democracy Prep on voter registration and participation in elections by leveraging Democracy Prep’s randomized admissions lotteries. Although our randomized design guards against selection bias, sampling variability is still a concern: Our traditional estimated impacts were dramatically larger than the impacts found in previous literature on the effects of education on registration and voting, suggesting that random variation could have led to an overestimate of the size of the impacts. We therefore conducted a complementary impact analysis using the BASIE framework that grounded our traditional impact estimates in the findings of previous research.

Given the chances of having a positive effect that we see from past studies, combined with the traditional estimate of Democracy Prep’s impact that we calculated in this study, we find that even a conservative Bayesian analysis (which accounts for the fact that positive impacts are more likely to be published) suggests that enrolling in Democracy Prep has large positive effects on students’ democratic participation in adulthood. This finding is robust to the selection of prior evidence.

Challenges of communicating statistics to the public in Uganda; a case of Uganda Bureau of Statistics
Winny Nekesa Akullo (IASSIST, Public Procurement and Disposal of Public Assets Authority – Uganda)

Statistics have become continuously useful in different domains and are considered as the milestone for an informed decision making in both private and public institutions. Many countries, especially developing ones suffer from bad politics, poor governance, corruption, lack of accountability and poor planning. This leads to unnecessary spending and investment in ‘white elephant’ projects. UBOS is the principal data collecting, processing, analysing and disseminating agency responsible for coordinating and supervising the National Statistical System. The Bureau coordinates the development and maintenance of a National Statistical System which will ensure collection, analysis and dissemination of integrated, reliable and timely statistical information. This paper focuses on the user statistical information needs of the public, the challenges the face in accessing the statistical information, challenges Uganda Bureau of Statistics faces in communicating this statistics to the public and strategies to improve on communication of statistics.
CONCURRENT SESSION B-4
Implementing the Evidence Act: The Journey thus Far and the Road Ahead

The theme for the 2020 FCSM Research and Policy Conference is the Federal Statistical System’s role in equipping agencies and the public to leverage data resources for evidence-based policymaking. There is a three-session track on the Implications of the Evidence Act. The goal of the track is to provide background information about the Act and highlight the opportunities the Act will create. This second session in the track brings together Chief Data and Statistical Officers to discuss their experiences implementing the Evidence Act thus far, and their thoughts about the road ahead. The Evidence-Based Policymaking Act of 2018 became law in January 2019. The four titles of the act include Federal Evidence Building Plans, the Open Government Data Act, Confidential Information Protection and Statistical Efficiency Act (CIPSEA) and general provisions. Chief Data and Statistical Officers are formally designated in the Act and are have several new responsibilities.

CONCURRENT SESSION B-5
Working with Non-Probability Samples

Can Sample Surveys from Online Panels Support Evidence-Based Policy-making Decisions?
Mansour Fahimi (Ipsos Public Affairs), Ronaldo Iachan (Ipsos Public Affairs), (Ipsos Public Affairs)

For decades, traditional methods of survey sampling have served as a foundation for data-driven decision-making processes. However, the main tenets supporting such surveys are becoming exceedingly difficult to secure. Adding to this, the growing constraints for budgetary and time resources as well as the evolving informational needs of decision-makers that are beyond the reach of structured surveys, are among the reasons why traditional methods are struggling to remain as viable options in the digital age. It is in this context that sample surveys from online panels are dominating the survey research landscape as the method of choice. However, the vast majority of online panels are constructed using convenience methods of recruitment, rendering their survey estimates void of inferential integrity. As part of an ongoing research for improving survey quality from our online panel (Knowledge Panel) we have conducted investigations to identify any differential nonresponse patterns that our recruitment process may be subject to. This research provides a summary of comparisons between recruited and nonrecruited households based on a long and diverse list of ancillary data.

Estimation of Population Characteristics From Web and Traditional Probability Samples in Case of Large Number of Potential Covariates
Vladislav Beresovsky (National Center for Health Statistics)

Traditional probability-based sample surveys can be expensive to conduct and experience growing nonresponse rates. Web surveys provide a viable alternative because of lower costs and expedient data collection, although web samples may be less accurate than traditional survey samples. This explains interest in statistical techniques for valid inferences from web samples. We consider augmented inverse propensity weighted estimators for inferences of finite population means, using variables of interest from the web sample, randomized design of an auxiliary traditional survey sample and covariates shared between the two samples. Parameters of the outcome variable and web response models are simultaneously estimated with the group Lasso, which is also an efficient model selection technique. Robustness of the proposed method to model misspecification is tested in a simulated problem with a large number of covariates. It is applied for the estimation of the prevalence of chronic health indicators for the U.S. civilian, noninstitutionalized population of adults, using real data from the web and auxiliary samples from experimental web survey and the National Health Interview Survey.
Adjusting for National Coverage and Selection Bias in Two Nonprobability Health Surveys: The Effect of Weighting Methods
Davia Moyse (ICF), Yangyang Deng (ICF), Matt Jans (ICF)

Whether nonprobability samples will ever replace probability samples depends on how well they replicate findings from established surveys. This study uses two nonprobability sample sources to evaluate the role of weighting methods in producing statistical estimates similar to findings from the Behavioral Risk Factor Surveillance System (BRFSS). The first is the Surveys-on-the-Go mobile panel, which only includes people who have a smart phone. The second source is Amazon Mechanical Turk (MTurk), which also includes members who have access to a traditional computer. Respondents from both sources were asked health questions selected from BRFSS. Four raking approaches are compared: basic demographics alone; expanded demographics alone; demographic and geographic information as independent margins; and cross-classified margins of geographic by demographic characteristics. The four methods are evaluated based on the distribution of the raked weights and by benchmarking weighted estimates of survey responses to the comparable BRFSS survey. Initial results show that biases resulting from the nonprobability sampling methods may be mitigated through weighting for some key health behaviors.

Combining Probability and Non-probability Samples Using Propensity Modeling and Small Area Estimation
Nadarajasundaram Ganesh (University of Chicago), Edward Mulrow (University of Chicago), Vicki Pineau (University of Chicago)

Probability sampling has been the standard basis for design-based inference from a sample to a target population. In the era of big data and increasing data collection costs, however, there has been growing demand for methods to combine data from probability and nonprobability samples in order to improve the cost efficiency of survey estimation without loss of statistical accuracy. In prior presentations (2018 BigSurv, 2019 ESRA, 2019 JSM), we discussed using models commonly used in small area estimation to generate unbiased model-based estimates when combining probability and non-probability samples assuming the smaller probability sample yields unbiased estimates. In this presentation, we discuss combining a propensity approach that estimates the probability of being a non-probability sample respondent along with small area modeling to generate more efficient estimates compared to our previous approach of calibrating the non-probability sample and then using small area modeling. We investigate the properties of our estimators using several general population studies and a simulation study.

Is Panel Conditioning a Concern with Online Probability-based Panels?
Frances Barlas (Ipsos Public Affairs), Mansour Fahimi (Ipsos Public Affairs), Randall Thomas (Ipsos Public Affairs)

Prior investigations of panel conditioning have had difficulty teasing apart the effects of repeated exposure to surveys from potential non-random panel attrition. To investigate panel conditioning, we used Ipsos’ probability-based online KnowledgePanel® and segmented the panel into five tenure groups from those having completed fewer than 5 studies to those who have been with KnowledgePanel for more than 6 years. For each segment, we selected a sample using probability-proportional to size (PPS) sample methodology to ensure that the samples selected would be representative and self-weighting and administered the same questionnaire which included measures that could be compared to benchmarks from high-quality, largely Federally-funded studies. Weighting each segment to consistent targets, we then compared the segments across demographics, attitudinal and behavioral measures to assess the extent of panel conditioning. Controlling for tenure/attrition and weighting to the same targets, we found very little evidence for panel conditioning across the segments. We discuss the implications for research that uses online panels.
Incorporating Administrative Data in Survey Weights for the Survey of Income and Program Participation
Ashley Westra (U.S. Census Bureau), Jonathan Eggleston (U.S. Census Bureau)
Declining survey response rates in federal surveys has led to concerns of increasing nonresponse bias in key government statistics. A potential solution is to leverage administrative data from federal agencies when constructing survey weights. This project performs initial research on incorporating administrative data into the weighting algorithm for the Survey of Income and Program Participation (SIPP). Specifically, we match income data from IRS tax forms and demographic data from the Social Security Administration and the Decennial Census to both respondents and nonrespondents. We then use this matched data in the household nonresponse adjustment of the SIPP weighting algorithm, which adjusts the weights of respondents to account for differential nonresponse rates among subpopulations and reduce nonresponse bias in survey estimates. We show how these new weights affect estimates of wealth, income, poverty, health insurance coverage, and participation in government assistance programs and their impact on nonresponse bias compared to the traditional weights. In conclusion, we discuss how this method can be extended to other household surveys in the federal government.

The Impact of an Improved Sampling Frame - the 2018 National Sample Survey of Registered Nurses
Tiandong Li (Health Resources and Services Administration)
The National Sample Survey of Registered Nurses (NSSRN) has been criticized for estimation biases by demographic variables. Research suggests that the bias is largely due to differential response tendencies by demographic variables. However, with high missing rates in demographic variables in the sampling frame, no demographic information could be used in sample selection and weight adjustments. During the redesign of the 2018 survey, the sampling frame was linked to the Census Bureau’s Person Identification Validation System (PVS) and administrative records data to fill in the missing information in the frame. This study provides an overview of the frame construction and how the improved sampling frame helps the sample selection and weighting of the RN sample. The estimation bias are evaluated by comparing to other key national data source on the nursing workforce, the Current Population Survey (CPS) and the American Communities Survey (ACS). Differential nonresponse rates are identified across demographic characteristics. Initial findings show that the biases on demographic estimates identified in the prior rounds are corrected through weight adjustments.

Using Administrative Records Data to produce Business Statistics: the Nonemployer Statistics by Demographics Series (NES-D)
Adela Luque (U.S. Census Bureau)
The Survey of Business Owners (SBO) was the only comprehensive source of information on business demographics. To address increasing nonresponse rates and costs, and a rising demand for more frequent and timely data, the Census Bureau has consolidated three business surveys. One of the consolidated surveys is the SBO. The nonemployer component of the SBO will be accomplished through a new blended-data approach that leverages existing administrative (AR) and census records to assign demographic characteristics to the universe of nonemployers, and produce an annual series that will become the only source of nonemployer demographics estimates. This new series is the Nonemployer Statistics by Demographics series or NES-D. Meeting the public's needs, NES-D will provide reliable estimates with no respondent burden on a more frequent and timely basis than the SBO. Using the 2014-2016 vintages of nonemployer businesses and demographic information from the decennial census, the American Community Survey, the Census Numident and AR from the Department of Veteran Affairs, we discuss preliminary results, the challenges encountered along the way, and next steps.
Blending Administrative Data with a Probability Sample of Nonparticipants to Produce National Estimates: The NCS-X NIBRS Estimation Project
Marcus Berzofsky (RTI International), Dan Liao (RTI International), Alexia Cooper (Bureau of Justice Statistics)

Administrative data collected through a set of agencies (e.g., law enforcement, schools) can be a rich source of information, but misleading, if the data suffer from quality issues such as item missingness or incomplete coverage. When the data source suffers from incomplete coverage, the data are not representative of the population. If a census is not possible, one alternative is to select a probability sample of nonparticipating agencies, collect their data, and blend them with the reporting agencies. The FBI’s National Incident-Based Reporting System (NIBRS) collects incident-based information on all crimes reported to the police. Currently, 33% of law enforcement agencies in the US submit to NIBRS, but these agencies mainly represent less populated parts of the country. The National Crime Statistics Exchange program is recruiting a probability sample of 400 agencies designed to produce nationally representative estimates when blended with the existing reporting agencies. However, the methodology for addressing quality issues and producing estimates is complex. We describe how we intend to address these issues and the plan for developing the appropriate estimation methodology.

Characterizing Federal Funding of Research and Development Using Administrative Data
Samantha Cohen (University of Virginia), Joel Thurston (University of Virginia), Sean Pietrowicz (University of Notre Dame)

The Federal Government accounts for about one-fourth of total Research and Development (R&D) funding in the United States—but what exactly does this public funding support? While the National Center of Science and Engineering Statistics (NCSES) provides high-level data from surveys on the disposition of federal obligations for R&D, more granular research characteristics (e.g., project topic) remain untapped. Federal agencies also release publicly available administrative data that describe projects in far greater detail (e.g., USA Spending). This presentation documents the usefulness of these administrative data to enhance and supplement NCSES surveys of federal funding. Using several such databases, we estimate total U.S. Science and Engineering (S&E) funding for six federal agencies. Initial results suggest that using administrative records for understanding Federal S&E support has promise in that our findings do resemble aggregate values reported on NCSES’s Federal Support Survey. We will further examine the available administrative data with respect to S&E data and the viability of differentiating R&D funding in administrative data from other spending categories.

CONCURRENT SESSION C-2
Applications of Machine Learning to Enhance Federal Statistics

Predicting Population or Subpopulation Estimates Using Machine Learning Algorithms: An Application using Survey and Administrative Records Data
Harold Gomes (University of Maryland)

There is a high demand to produce population estimates of various groups by Statistical Agencies, the United Nations, OECD & other organizations because of high-stakes impacts in demography. This research investigates how machine learning algorithms perform in predicting population estimates by a group based on some predictors. It attempts to predict the SSA disability-population count by State, based on two formulations: 1) as a function of trend of each State; 2) as a function of 10 causes of death, poverty & geolocation. It also demonstrates how administrative data from SSA is merged with the Survey data from Census, NCHS, & metadata of State information. Four datasets are joined to construct the final data. LOESS, LASSO, CART, CTREE & Random Forest models are implemented to predict the disability-population by State. The best model(s) is selected with
highest predictive power on a future test dataset. Results suggest machine learning prediction could be useful when response data are not readily available, or a different dataset is available with some predictors, or predicted estimates are necessary in order to make a resource allocation in advance without the observed data.

**Using Machine Learning to Improve Forecast Accuracy**  
Gianna Short (Economic Research Service)

The USDA’s Economic Research Service publishes the Food Price Outlook which consists of monthly forecasts of the annual percent change in the food series of the Producer Price Index and the Consumer Price Index. The existing pass-through modeling approach builds forecasts from crop level – to wholesale – and then to retail. However, new research indicates that food price shock pass-through has reduced in volatility and duration over recent decades, due in part to increased market concentration in the food supply chain (Brown and Tousey 2019). This research uses machine learning and cross-validation to test forecast accuracy of current and new models for the Food Price Outlook. Testing new machine learning models outside of the pass-through framework allows the potential incorporation of additional data sources including demand-side price indicators.

**A Data-Driven Method of Specifying Efficient Blocking Schemes for Record Linkage**  
Marc Roemer (National Center for Health Statistics), Dean Resnick (NORC), Scott Campbell (NORC)

Record linkage is the process of identifying records from distinct data files that refer to the same entity. Because a detailed comparison between every record in one data file with every record in another data file is computationally impractical, it is important to first determine which record pairs are worth considering as possible links, a process known as “blocking.” Under demands for evidence-based policymaking and the availability of multiple big data sources, the need for a systematic method of determining an efficient blocking scheme is increasing. We propose a data-driven method to define blocks and compare our proposed method to a previously utilized machine learning, sequential coverage algorithm. We link National Center for Health Statistics survey data with the National Death Index and with Centers for Medicare and Medicaid Services data and compare the results of the two blocking methods on three dimensions: the size of resultant candidate set, production processing time, and the blocking coverage rate of true matches. We also analyze and present rates of linkage error across blocking schemes.

**Exploring Non-Traditional Statistical Methods as Ways to Address Cultural, Social, and Linguistic Biases**  
Bradley Rentz (REL Pacific at McREL Internationala), Christina Tydeman (REL Pacific at McREL Internationala)

Students who have graduated high school but are considered academically underprepared are often placed in developmental courses. When making course placements using a single indicator such as placement scores, the threat of bias is substantial. Institutions assume placement exams are predictive of college success, but they may rarely examine predictive validity. Multiple measures may not result in unassailable placement decisions but may paint a more complete picture of past successes and motivation indicators. However, how this placement system is constructed matters. Without critical evaluation, it is easy to replicate the same inequitable placements that reify implicit biases. Evaluation of developmental education decision-making typically uses multiple linear regression, which may not capture the complexity of student patterns and imposes a top-down structure on the data that can obfuscate underlying patterns. This session will describe two machine-learning analytic models, classification and regression trees (CART) and cluster analysis, that can be applied to student college placements to examine emergent data structures invisible with traditional analyses.
Detecting Pharmaceutical Innovations in Text-Based Data Using Machine Learning
Gizem Korkmaz (University of Virginia), Gary Anderson (National Center for Science & Engineering Statistics), Devika Nair (University of Virginia)

Innovation is traditionally measured through surveys of selected companies such as the Business R&D and Innovation Survey (BRDIS) that focuses on innovation incidence, i.e., the number of innovating firms. This paper aims to develop machine learning methods to measure business innovation using non-traditional data sources to enrich and complement innovation measures obtained through these surveys. We focus on product innovation in the pharmaceutical sector (drugs and medical devices) that is heavily regulated by the Food and Drug Administration (FDA). The non-traditional data sources include publicly available opportunity and administrative data such as financial filings, and news articles obtained from Dow Jones, a business news and data provider. We collect and scrape around 3K filings, and parse 2M news articles, and develop text-mining methods to link datasets by companies. We develop machine learning methods to estimate the number of new products companies introduce to the market. Our findings are compared to the publicly available approval data provided by the FDA to study the fraction of innovation activity that could be captured using these novel data sources.

CONCURRENT SESSION C-3
Impact of Questionnaire Design on Data Quality

Does the Wording Affect the Rates? An Experiment in Sexual Orientation and Gender Identity (SOGI) Measurement
Deirdre Middleton (ICF), Matt Jans (ICF), Yangyang Deng (ICF)

Increasing visibility of sexual and gender minorities has led to heightened demand for accurate population statistics. Large-scale surveys increasingly measure sexual orientation and gender identity (SOGI). A unique SOGI measurement challenge is that questions must be understood by the non-LGBTQ population and meaningful to and respectful of the LGBTQ community. Poorly-worded questions incur measurement and nonresponse error risks. This study assesses the impact of SOGI question wording on self-identification rates and item nonresponse. Respondents from two nonprobability surveys (MFour's Surveys on the Go® and Amazon Mechanical Turk), were randomly assigned to one of two sexual orientation and one of two gender identity measures used by three highly-influential health surveys: the California Health Interview Survey, Behavioral Risk Factor Surveillance System, and National Health Interview Survey. Results of the experiment are benchmarked to estimates from these surveys and other high-profile surveys. In addition to the experimental question wording results, this study demonstrates the potential for collecting SOGI data from nonprobability samples.

Empirically Assessing the sensitivity of Survey Questions and Responses
Shelley Feuer (U.S. Census Bureau), Stefanie Fail (Nuance Communications, Inc.), Michael F. Schober (New School for Social Research)

It is well known that survey questions vary in their sensitivity. But the same questions may be differently sensitive for different respondents. This paper reports on a method for empirically assessing the extent to which specific questions and response options are considered sensitive (embarrassing or socially undesirable) at a particular moment or among a particular population. To test the method, online ratings of the sensitivity of survey questions and potential responses were collected from two samples of 100 US respondents. Participants rated how embarrassed they thought most people would be to answer each question and give each response option during an interview. Results demonstrate clearly that sensitive questions can have nonsensitive responses and that nonsensitive questions can have sensitive responses. Depending on the threshold one picks for judging a question or response as sensitive (e.g., 40% or 50% of respondents think most
people would find it embarrassing to be asked or answer), different questions and responses emerge as sensitive. Given that the sensitivity profile changes, this type of assessment (rather than relying on researcher judgment) may be useful.

Casting a Wider Net: Response Format Effects on Self-reported Individual and Household Disability
Randall Thomas (Ipsos Public Affairs, Frances M.Barlas (Ipsos Public Affairs)

Disabilities have been measured differently in national surveys, both at individual and household (HH) levels. This paper reports on 2 different experiments using 3 different response formats to measure disability for both self and other HH members (for visual/auditory/cognitive/physical impairments). In a web-based survey, respondents were randomly assigned to 1 of 3 groups: 1) presented disability items separately for self and HH members using a yes-no grid (YNG); 2) presented disability items separately for self and HH member using a multiple response format (MRF); or 3) presented disability items and asked to select for self, another HH member, self and another HH member, or no one in HH had each impairment (Combination Grid – CG). The CG format was completed fastest overall, though the YNG and CG formats generally resulted in similar and higher prevalence levels of most disabilities for both self and HH levels than MRF. We discuss a number of results that indicated that disability salience may be a more important determinant of response selection with MRF than other response formats but MRF may incompletely describe prevalence.

Using Qualitative Data to Improve Surveys: Results from a Cognitive Interview Study of the National Study of Long-term Care Providers
Lauren Harris-Kojetin (National Center for Health Statistics), Manisha Sengupta (National Center for Health Statistics)

The National Study of Long-Term Care Providers (NSLTCP) monitors trends in long-term care services. NSLTCP uses survey data on the residential care community (RCC) and adult day services sectors (ADSC), and administrative data on the home health, nursing home, and hospice sectors. When responding this type of establishment survey, respondents are often asked to provide information that is maintained in the form of administrative records. However, there is often no standard for what information is maintained, how it is maintained or who has access to it. This presentation will explore the impact of record keeping practices on both data quality and response burden using the results from a cognitive interview study evaluating NSLTCP. The overall purpose of this project was to 1) determine whether there are data quality (measurement and response generation) challenges with the aggregate-level services users surveys questions and 2) investigate provider perceptions of the burden in looking up the records, and the impact of the perceived burden on data quality. This presentation focus on how the results of this study were used to improve the survey.
CONCURRENT SESSION C-4
Using Data in New Ways: Leveraging the Evidence Act to Coordinate Evaluation, Statistics, and Policy

Framing the Evidence Act’s Vision for Coordination and Collaboration
Erica Zielewski, OMB

In brief introductory remarks, this presentation will summarize key elements of the Act with an emphasis on those places that highlight coordination and collaboration between evaluation and statistics functions and roles, such as the creation of multi-year learning agendas and Annual Evaluation Plans. These remarks will also introduce the examples that follow each of which brings something different to the discussion: one example of an agency that generally makes data available; an example of how an evaluation shop and statistical unit work together within one organization; an example focused on state data; and an example of leveraging administrative data from an agency for research purposes.

HUD’s Approach to Making Data Available for Research and Evaluation
Calvin Johnson, HUD

HUD traditionally prioritizes its data capacity (see for example the data section in their Research Roadmap) and has spent effort to get their data into the hands of people that can use it HUD has also done some cool unique data linkages (for example, with Census and with NCHS), and more broadly tried to make data more accessible for use. We would like you to discuss how this has worked in practice.

Linking State Medicaid Data and Child Welfare Data for Outcomes Research
Valeria Butler, ACF/HHS & Emily Maddens, ASPE/HHS

HHS recently launched this effort to link and build datasets that can be used for research and policy. It is unique and innovative, but nascent, so the presentation would focus more on intentions/plans vs. concrete activities/outcomes

The Department of Labor’s Data Exchange and Analysis Platform (DEAP)
Christina Yancey, Chief Evaluation Officer, DOL; Scott Gibbons, Chief Data Officer, DOL; and David Judkins, ABT Associates

This presentation will include speakers from DOL discussing the DEAP tool, and will begin with a discussion of the critical role of capacity-building for ongoing evidence building, followed by an introduction and overview of the DEAP tool. A contractor with experience collaborating with DEAP will provide a specific use case.

CONCURRENT SESSION C-5
Tiered Data Access

Work by the Commission on Evidence-Based Policymaking, and subsequent related legislation, have led numerous stakeholders to explore practical ways in which to provide researchers with secure, efficient access to confidential statistical data. In response to this need, some proposals have centered on “tiered access” to data in ways that are more flexible than previous approaches that often placed data into two distinct classes: “restricted” or “open public access.” In-depth work in this area requires development of new regulatory and management structures; related process innovations; and extensive outreach to a wide range of stakeholders. This session reports on work
by several groups to address these issues. The chair will begin the session with an overview of the ways in which tiered access fits into a broader initiative to improve privacy protections for federal data while improving responsible and secure data access for researchers. The first speaker will provide a user’s perspective on tiered data access. The second speaker will outline some recently developed OMB guidelines related to tiered access. The third speaker will outline efforts by the Bureau of Labor Statistics to explore multiple access modes for its restricted-use microdata. The fourth speaker will discuss two related responses to the stakeholder needs identified in the previous presentations: development of a pilot project to explore management of remote access within the Federal Statistical Research Data Center system; and an OMB initiative to develop a single application process for restricted microdata access requests. The fifth speaker will discuss opportunities and challenges encountered in the development of efficient channels for establishing standardized credentialing systems for researcher access to microdata.

CONCURRENT SESSION D-4
Challenges and Advances in Measuring Sexual Orientation and Gender Identity in Federal Surveys: Best Practices, Recent Findings, and Public Use

How Would You Describe Yourself? Recent Developments in Sexual Orientation and Gender Identity Survey Measures
Maura Spiegelman, National Center for Education Statistics & Christina Dragon, Department of Labor

In 2016, a federal research group focused on measuring sexual orientation and gender identity (SOGI) released three working papers to address the dearth of data on sexual and gender minority (SGM) populations and the methodological considerations for collecting such data. One of these working papers, Toward a Research Agenda for Measuring Sexual Orientation and Gender Identity in Federal Surveys, addresses the knowledge gaps by proposing research priorities and strategies that will make the greatest impact on improving SOGI measurement in Federal surveys. A primary priority was the need for further research on the wording and ordering of response options for SOGI questions. Since 2016, research and data collection efforts have continued to expand, laying the groundwork to address this priority. This presentation will review the most recent key terms (e.g., sexual identity, gender identity, “something else”) used in collecting data on sexual and gender minorities, as well as the current data collection methodologies. In addition, this presentation summarizes the results of open-ended responses provided by respondents who identify as “something else” or another sexual orientation or gender identity, along with item nonresponse, the use of don’t know/refused response options, and recent research examining question format and comprehension of related questions and terminology. This presentation concludes with a discussion of findings surrounding SGM terminology and recommendations for future research.

Pretesting SOGI questions for self and proxy: How do in-person cognitive interviews compare to online testing?
Robin Kaplan, Bureau of Labor Statistics

Survey organizations have shown increased interest in measuring respondents’ sexual orientation and gender identity (SOGI) via self and proxy response. SOGI questions have been pretested using traditional, in-person cognitive interviews (Willis, 2005) to understand and identify sources of measurement error. However, participants who agree to participate in SOGI research may differ from those who do not, both demographically and in their experience and attitudes toward the topic.
Additionally, in-person pretesting has been found to elicit more socially desirable responses than other modes (Behr et al., 2017), and in terms of proxy response, data quality can vary depending on the relationship between respondent and proxy (e.g., spouses tend to provide the most accurate responses; Cobb, 2018). This research was designed to assess potential mode differences in pretesting SOGI questions using traditional, in-person cognitive interviews versus online pretesting in evaluating respondents' level of sensitivity, difficulty, willingness, and ability to self and proxy-report for other household members on SOGI questions. We compare and contrast data from a traditional, in-person cognitive pretesting study (n=132) to a self-administered, online study (n=154). Responses from both studies were content coded and systematically analyzed for sensitivity, difficulty, ability, and willingness to self and proxy-report SOGI. We found in both studies, most respondents indicated that they were willing and able to report on SOGI information for themselves and others in their household, with little sensitivity or difficulty. One exception to this was respondents who identified as or lived with sexual and gender minorities. Contrary to prior research on proxy response, we found evidence that some spouses and partners disagreed with one another on their SOGI status. In comparison, online respondents indicated more sensitivity and overall concern in their open-ended responses to the idea of collecting SOGI data, in terms of both self and proxy-reports. Additionally, we found a higher prevalence rate of self-reports of identifying as bisexual in the online versus in-person sample. We discuss the results and implications for pretesting SOGI questions, differences between the two samples that may have contributed to the results, and the importance of pretesting in different modes.

**Improving Measurement of Sexual Orientation (SO) and Gender Identity (GI) in the Federal Statistical System**

Sylvia Fisher, HRSA and Nancy Bates, Census Bureau

The Interagency Research Group on Measuring Sexual Orientation and Gender Identity (SOGI) explores and promotes accurate and reliable measurement of SOGI in the context of Federal information collections. The collection of accurate SOGI survey data presents a number of measurement and methodological challenges. The comparatively small population of LGBT persons in the U.S. suggests that relatively small sampling or reporting errors can lead to significant errors in estimation, requiring an emphasis on precision and accuracy. Furthermore, the constructs of gender identity and sexual orientation are becoming more fluid in the U.S. as terms used by individuals to describe themselves are rapidly changing. Finally, gender and sexuality can be sensitive topics, and the privacy and confidentiality of respondents and respondent data must be handled with care. This presentation will discuss common challenges, best practices, current research, and outstanding research gaps regarding the measurement and collection of SOGI data across the federal statistical system.

**Accessing and Analyzing Sexual Orientation and Gender Identity in Public-Use Files**

Andrew R. Flores, American University and the Williams Institute

The inclusion of sexual orientation and gender identity measures in federally sponsored surveys means that scholars can make new insights into the lived experiences of sexual and gender diverse people in the United States. Government and other data that are accessible to scholars and contain sexual orientation and gender identity measures have been underutilized. Accessing and analyzing such data can be burdensome. The National Inmate Survey and the National Survey of Youth in Custody require significant effort to apply for, gain access to, and obtain results from those data. Other surveys such as the National Health Interview Survey and the National Crime Victimization Survey, while far more accessible, introduce methodological complications that require careful attention to survey design and implementation. Differences in prevalence estimates and demographic
characteristics of sexual and gender diverse people across these surveys requires careful attention to methodology and communicating research findings to academics, media, and the general public. Finally, even if accessible, public-use files seldom are accompanied with sufficient supporting documentation to replicate and extend analyses produced by federal agencies. This potentially limits the number of people accessing and analyzing these data.

CONCURRENT SESSION D-5
Evaluating Data Quality

Using Process Data to Understand Non-Response in NAEP
Markus Broer (American Institutes for Research), Ruhan Cirici (American Institutes for Research), Juanita Hicks (American Institutes for Research)

In paper-pencil tests, all items are presented to the student and responses are observed for each item. Yet, students may not always respond to all the items in the assessment. Generally, items not answered beyond the last responded item are considered not reached (NR) and items left not-answered before the last attempted item are recorded as omitted (Omit). Both decisions for NR and Omit are based on the item response sequence for paper-pencil tests.

The introduction of digitally based assessments (DBA) has enriched the data collected. The availability of process data makes it possible to understand non-responses and evaluate the appropriateness of the conventional definitions for NR and Omit. Given that the impact of item non-responses on scoring and parameter estimates is non-ignorable, it is important to examine NR and Omit occurrence with the help of process data.

This study used NAEP 2017 mathematics process data and scoring files to explore current analysis methods of NR and Omit, while also exploring new, alternative methods to define NR and Omit. Results of this study can enhance scoring and parameter estimation, and update test paradigms based on paper-pencil tests.

Nonresponse Bias Studies of the Consumer Expenditure Survey

The Consumer Expenditure Survey (CE) is the source of data for the relative importance of items in the Consumer Price Index. Like many federal surveys, the CE survey is a nationwide household survey whose response rate has been decreasing in recent years. Over the past ten years it decreased more than ten percentage points, which is a concern because if respondents and nonrespondents have different kinds of expenditures then there may be an increased bias in favor of the respondents’ expenditure patterns. Two nonresponse bias studies were performed: one to determine whether the nonrespondents are "missing completely at random," (MCAR) and another to estimate the amount of nonresponse bias in the survey’s published expenditure estimates. Both studies consisted of four sub-studies, whose methodologies included a new cross-survey response rate “relativity” measure. The studies found the CE data were not MCAR and there were only slight levels of nonresponse bias. It does not affect the accuracy of the CE data and it provides a counterexample to the commonly held belief that if a survey’s data are not MCAR, then its estimates are subject to nonresponse bias.

Rethinking Response Rate Calculations for Probability-based Samples from Online Panels
Mansour Fahimi (Ipsos Public Affairs), Frances M. Barlas (Ipsos Public Affairs)

Traditionally, response rates have been relied upon as a singular metric for gaging the quality of survey estimates vis-à-vis bias. However, the growing recognition for the Total Survey Error framework highlights that nonresponse is only one of the many potential threats to survey quality.
Yet concerned about the legacy stigma attributed to low response rates, researchers often look for creative loopholes for reporting higher rates. Whether or not response rates are telling metrics for survey quality, it is important for their operational definitions and methods of calculations to be based on sound and transparent principles. While a growing number of researchers are now relying on online panels for survey administration, we continue to use an outdated methodology for computing response rates for such surveys. This research aims to define a robust and coherent definition, and an associated calculation methodology, for response rates when the underlying sample is probability-based and selected from online panels that are not subject to any systematic exclusions. Results are cross-examined across a number of samples for validations and empirical justifications.

**Using Metadata to Evaluate Survey Response**
Cleo Redline (National Center for Education Statistics), Dylan Clark-Boucher (George Washington University)

The housing of metadata was originally conceived to provide the information that is required for a user to effectively analyze a dataset. The primary focus on metadata in the federal statistical system has been to make substantive data accessible, discoverable, and usable by the public. We examine whether metadata might hold promise as a diagnostic tool for evaluating the survey practices that are employed by agencies to improve data quality. It is well documented that response rates have been falling over time. When response rates fall, nonresponse, which is the failure to collect data from all persons in a sample survey, threatens data quality. Using the US Department of Education's ED Data Inventory, we examine whether, in line with other studies, survey response has decreased over time at the (NCES). We examine whether this is the case for all NCES data collections, including administrative data collections and student assessments, or if it is only true for the sample surveys. We look at whether other variables in the inventory may explain response behavior as well and conclude that using metadata to evaluate an agency’s data collection methods does appear promising.

**Fitness for use: Assessing Data Quality Through the Lens of Data Use**
Daniel Dorfman (Bureau of Labor Statistics), Adam Safir (Bureau of Labor Statistics)

The Consumer Expenditure Survey is a nationwide household survey sponsored by the U.S. Bureau of Labor Statistics to measure the spending patterns of U.S. consumers. It is the only federal government survey that provides information on the complete range of consumers’ expenditures as well as their incomes and demographic characteristics. While these data are critical to the construction of relative importances for item groups in the Consumer Price Index, there are many other federal agencies which make use of CE data, each with their own unique considerations regarding data quality. For example, factors such as response rates, measurement error, aggregation, data level, and timeliness differ by data user constituent. This presentation will provide an overview of the different uses of expenditure data at the federal level, and discuss the program’s approach to assessing fitness for use within the context of Juran and Gryna’s paradigm; i.e., survey quality in terms of the interests of both data producers and data users. The discussion will be of interest to data producers, survey methodologists, and data users.
Enriching Analysis by Linking Data: The National Household Food Acquisition and Purchase Survey (FoodAPS)
Elina Page (Economic Research Service)

The National Household Food Acquisition and Purchase Survey (FoodAPS), co-sponsored by the Economic Research Service and Food and Nutrition Service of the US Department of Agriculture, was the first nationally representative survey of American households to collect unique and comprehensive data on household food purchases and acquisitions. Data collected through FoodAPS have already enabled critical research related to food purchase and acquisition patterns of U.S. households, particularly of low-income households and households participating in food assistance programs; the nutritional quality of foods purchased by Americans; and the relationship between food access and food acquisitions. This presentation will detail the impetus for the survey, highlight the unique features of the data (including the use of multiple sources of extant data and State administrative records to improve and enhance usability of the data collected through survey instruments), and demonstrate data visualizations of several key findings and statistics.

Do Supplemental List Frames for Subpopulations Increase Subpopulation Sampling Efficiency? Evidence from the National Household Food Acquisition and Purchase Survey
Shiyu Zhang (University of Michigan)

Multiple-frame sampling has been regarded as a device for increasing efficiency in identifying small subpopulations. However, there has been a lack of empirical evidence in supporting the efficiency of the multiple-frame approach and in guiding best practices. Using paradata from the National Household Food Acquisition and Purchase Survey (FoodAPS), the current analyses were built on a case of recruiting households that received Supplementary Nutrition Assistance Program (SNAP) benefits as a sub-goal of the survey sampling. The SNAP households account for around one-fifth of the general population, compared to a survey goal of 30 percent of responding households. Our findings confirm the theoretical expectations that having and using the additional SNAP list frames improved the efficiency of identifying SNAP households as opposed to screening a general address-based sample frame.

Estimation of Underreporting in the National Household Food Acquisition and Purchase Survey
John A. Kirlin (Kirlin Analytic Services)

The National Household Food Acquisition and Purchase Survey (FoodAPS) is a nationally representative diary survey, providing an important data source for decision-makers to design policies and programs for promoting healthy lifestyles. Unfortunately, a multiday diary survey like the FoodAPS can be subject to various survey errors, especially item nonresponse error occurring at the day level. The FoodAPS public-use data set provides survey weights that adjust only for unit nonresponse. Due to the lack of day-level weights (which could possibly adjust for the item nonresponse that arises from refusals on particular days), the adjustments for unit nonresponse are unlikely to correct any bias in estimates arising from households that initially agree to participate in FoodAPS but then fail to report on particular days. This paper develops a general methodology for estimating the extent of underreporting due to this type of item nonresponse error in diary surveys, using FoodAPS as a case study.
Assessing the Nutritional Quality of Household Food Acquisitions Using the National Household Food Acquisition and Purchase Survey
Linda Kantor (Economic Research Service)

A key strength of the National Household Food Acquisition and Purchase Survey is that it collected detailed information about all foods acquired, from grocery stores and restaurants, as well as free or subsidized foods (from schools, work, own production, food pantries, community centers, and family and friends). However, large variations in the quality of available nutrient information for these items resulted in an extensive post-survey data processing and cleaning effort. This paper provides an overview of the methodology and challenges associated with linking extant nutrition databases to food acquisitions reported in FoodAPS. Appending nutrition data to FoodAPS items supports the assessment of diet quality using the Health Eating Index (HEI), which quantifies the healthfulness of food acquisitions in terms of conformance with Federal dietary guidance.

Multilevel Regression and Poststratification (MRP) for Small Area Estimation with Geocoded FoodAPS Data
Xingyou Zhang (Substance Abuse and Mental Health Services Administration)

National social, economic, and health survey direct estimates are often not reliable for small geographic areas. This study explores a multilevel regression and poststratification (MRP) approach for small area estimation with the first National Household Food Acquisition and Purchase Survey (FoodAPS-1) that involves four major steps: 1) construct and fit multilevel models that link an outcome with both individual characteristics and area-level factors; 2) select a final multilevel model that could explain major geographic variations; 3) apply the final model to make a prediction with the census 2010 population; and 4) summarize small area estimates (SAEs) at census tract, county, state, and national levels. Preliminary results suggest that a multilevel logistic model that can explain more than 80% county-level variations in adult obesity status. Model-based estimates of obesity prevalence ranged from 14.1% to 54.7% with a median of 34.5% at county level and from 3.9% to 63.1% with a median of 31.3% at census-tract level. Additionally, FoodAPS model-based estimates were correlated and consistent with BRFSS model-based estimates at both county and census-tract levels.

CONCURRENT SESSION E-2
Measuring the Impact of Technology on the Economy

New Measures of Robot Use in U.S. Manufacturing
Catherine Buffington (U.S. Census Bureau), Javier Miranda (U.S. Census Bureau), Rob Seamans (U.S. Census Bureau)

The Annual Survey of Manufacturers (ASM) is a survey collected by the U.S. Census in order to capture detailed information about key inputs to and outputs of production by manufacturing establishments. The 2018 ASM included three questions designed to measure the purchase and use of industrial robotic equipment by manufacturers. In this presentation, we will discuss the development of the questions, potential issues with the collected data based on interviews with survey participants, and plans for public use data products based on these questions.
Measures of Robotic Expenditures from the Annual Capital Expenditures Survey
Valerie Mastalski (U.S. Census Bureau), Catherine Buffington (U.S. Census Bureau), Javier Miranda (U.S. Census Bureau)

The Annual Capital Expenditures Survey (ACES) is a survey of the U.S. Census Bureau designed to provide data on capital spending for new and used structures and equipment by U.S. nonfarm businesses with and without employees thus covering all domestic, private, non-farm businesses. In this paper we describe our efforts to expand the survey to include the collection of capital spending in robotic equipment. This firm level survey will provide capital expenditures information on both industrial and service robots across industries. We discuss results from the cognitive testing interviews, the challenges faced in the data collection, and what we expect for future years. We also discuss plans to produce statistics for new and used robotic spending at the national level by 3-digit and selected 4-digit NAICS industries. Data on the amount of business expenditures on robotics are critical to evaluate productivity growth, changes in industrial capacity, and measures of overall economic performance.

Business Dynamics Statistics of High Tech Industries
Nathan Goldschlag (U.S. Census Bureau), Javier Miranda (U.S. Census Bureau)

Modern market economies are characterized by the reallocation of resources from less productive, less valuable activities to more productive, more valuable ones. Businesses in the High Tech sector play a particularly important role in this reallocation by introducing new products and services that impact the entire economy. In this paper we describe an extension to the Census Bureau's Business Dynamics Statistics (BDS) that tracks job creation, job destruction, startups, and exits by firm and establishment characteristics including sector, firm age, and firm size in the High Tech sector. We preview the resulting statistics, showing the structural shifts in the High Tech sector over the past 30 years including the surge of entry and young firm activity in the 1990s that reversed abruptly in the early 2000s.

An Analysis of Trade and Employment in Potentially ICT-Enabled Services in the United States
Badri Narayanan Gopalakrishnan (University of Washington Seattle)

Global trade in services rose as it increasingly capitalised from ICT revolution and of late, there is evidence of a growing trend in the “servicification” of manufacturing. To this extent, the trade in potentially ICT-Enabled Services is gaining prominence in USA, as is evident from the available trade statistics. We analyse the R&D expenditures in these sectors and examine the relationship between exports and employment in the United States with respect to the potentially ICT-Enabled Services provided by the U.S. companies, based on the past data, using ex-post econometric analysis and further extended for an ex-ante analysis using the widely used global economic modelling framework, GTAP (Global Trade Analysis Project). This is followed by an attempt to understand the industry concentration, structure and the future potential of these services for the United States. One data issue that we face in this analysis is that of disclosure avoidance, which limits our ability to uncover all the statistics related to certain services sectors, such as design services. This issue needs attention from policymakers.
Using Process Data to Study Interviewer Effects on Measurement Error and Nonresponse in the Consumer Expenditure Survey

This study investigates the relationship between survey, household, and interviewer characteristics to estimate interviewer effects in the Consumer Expenditure Interview Survey (CE). The focus of this study is on understanding the impact of interviewer effects on two data quality indicators: measurement error, approximated using item missing data, edit indicators during post-interview processing, and changes to recorded answers during the interview; and unit nonresponse bias, explored through details of contact attempt outcomes and expressions of respondent concerns, as recorded by interviewers in the Contact History Instrument (CHI). Multilevel models use indicators such as interviewer characteristics (e.g., tenure, workload) and household-level interview process characteristics (e.g., interview length, collection mode) to model measurement error and a separate model for noncontact and refusal. Findings from this study will strengthen our understanding of the role that interviewers play in nonresponse bias, with possible implications for how survey programs and field offices manage interviewer training, workload, and contact procedures.

Improving Paradata Measures Using Qualitative Analysis, Visualization and Machine learning
Renee Ellis (U.S. Census Bureau)

Web paradata are data about the process of collecting information via a web survey. These data have many uses, including survey monitoring, question evaluation, instrument usability evaluation, and detecting suspicious reporting activity. At the Census we have created standard measures that can be used to help improve survey design of web surveys. Some of these measures have proven somewhat difficult to interpret. For example, although we can collect information about respondents backing up, it is hard to say how that impacts data quality or what it means for survey improvement. This current project examines how we can use qualitative review of data to find patterns in respondent behaviors across survey pathways. These behavior patterns can be used to create new measures that are more straightforward to interpret. The purpose of this presentation is to show some of the qualitative patterns found in backing, errors triggers, and overall pathway navigation and discuss how these patterns can be used to create programs, data visualizations and machine learning models to use for survey improvement.

Diving into Process Data: New Insights into NAEP Items
Ruhan Circi Kizil (American Institutes for Research), Juanita Hicks (American Institutes for Research), Emmanuel Sikali (National Center for Education Statistics)

The use of digitally based assessments (DBA) has enriched the data about students' interactions with assessment items (e.g., answer changes). Digitizing students' interactions provides useful information, which various stakeholders can feed into interpretation of assessment results, test/item development, test assembly, and communication with the public. Process data, as a new data type, can provide insights on how items with similar characteristics elicit responses from students and how policy makers can use this information to communicate better with teachers, parents, and students about student knowledge. This study analyzed NAEP 2017 mathematics process data using exploratory and advanced statistical methods (i.e., Mixture Modeling, Neural Networks) to understand how items function; particularly, for new item types (e.g., multiple selection). This study attempts to profile items by examining student-item interactions and communicate results via data visualization. Preliminary results show promise in advancing the use of this new data for insights into test development. The communication of these results to the public can also enhance the public's understanding of assessments.
Metadata projects made plain for U.S. statistical agencies
Peter B. Meyer (Bureau of Labor Statistics), Daniel Gillman (Bureau of Labor Statistics), Kathryn McNamara (U.S. Census Bureau)

Linking data between data sets calls for some coordinated standards among the data-issuing agencies. This paper shows the purpose of metadata systems for statistics organizations with examples from various other organizations. Metadata efforts can aid the discovery, understanding, and use of statistical data and improve interoperability, comparability, and ability to harmonize across sources, agencies, and countries. Enhanced metadata practices can contribute to the data quality, transparency, dissemination, and integration called for in the Evidence-Based Policymaking Act of 2018 and OMB guidance. Relatedly, machine learning applications create new opportunities and new requirements. This paper gives an overview of metadata concepts, standards, tools, and selected domestic and international projects. We discuss: (1) basic metadata definitions with examples from libraries, museums, and geographic datasets to illustrate the goals to be achieved in the realm of aggregate statistics; (2) relevant US federal laws, orders, and regulations; (3) how metadata systems and standards are put into practice; and (4) recommendations and reflections for Federal statistical agencies.

Lessons and insights from the use of audit trails to analyze the effect of moving a survey question
Brandon Kopp (Bureau of Labor Statistics), Lucilla Tan (Bureau of Labor Statistics)

Changes to survey design features are to be expected. The Consumer Expenditure Survey (CE) Program is developing a set of metrics, based on a wide range of data sources, that can be routinely re-used to evaluate the impact of survey changes. One such data source is Blaise audit trails. Audit trails provide a wealth of information on the data collection process for each sample unit – such as the time of entries made, resources that were referenced, edit checks triggered, and the navigation pattern through the questionnaire. Audit trails in their native format are unstructured text files. The CE Program has developed a pipeline to transform audit trails into a structured tabular format, yet there still remains a learning curve to understand how information from the questionnaire are represented in the audit trails. This understanding and its documentation are needed to accomplish the goal of developing reusable and interpretable metrics. We share the lessons learned from using audit trails to examine the effect of moving a block/grid-format question from one section of the survey to another section, as well as insights about the interview experience from this type of survey change.

Ethical Considerations for Data Access and Use
Amy O’Hara, Georgetown University

Project leads and data owners typically focus on legal and policy requirements when sharing data for research and evaluation, relying on written laws, regulations, standards, and policies. Ethical issues are seldom addressed in the same manner. Limited guidance exists to span the sectors, domains, and disciplines involved. We review materials available to guide decisions that data owners, controllers, analysts, and regulators face about whether and how data can be used responsibly. These decisions address concerns about possible or likely harms affecting individuals and groups, at present and into the future. We discuss cross-sector and interdisciplinary projects that are developing ethical guidelines and identifying best practices, and we identify the role that data intermediaries can play in establishing transparent practices that facilitate ethical data sharing.
Data Ethics & the Fourth Industrial Revolution
Christopher S. Lee, JD, CIPP, Chief Privacy Officer; United States Senate, Sergeant at Arms

The Fourth Industrial Revolution (4IR) has started. 5G networks, Cloud Computing and Quantum Computing are being integrated using artificial intelligence, machine learning and software. 4IR will be bigger than the dotcom revolution of the 1990s, and creates opportunities to process and use exponentially more data to make better, informed decisions faster than ever before. The 4IR will usher in a new wave of technical products and services. It will also create tools that can be used to benefit society or infringe upon privacy and civil liberties. This session will introduce the concept of 4IR and lay the ground work for identifying and addressing associated data ethics issues.

Policy and Technology: Ensuring Ethics in the Submission and Access of Biomedical Research Data
Dina N. Paltoo, National Institutes of Health

NIH has a large and growing number of valuable data repositories for human research data. Facilitating access to these data safely and in a manner that honors the privacy of the research participant requires creating innovative approaches to facilitate data submission and access. Accelerating data-driven discovery and providing the best return-on-investment on existing data and resources, in order to accelerate and improve science and build trust in the research enterprise, necessitates that the National Institutes of Health (NIH) facilitate the reuse of data collected in one study for use in future research. The first step is to ensure responsible stewardship of data, through policy and technology, such that data are submitted and made available in a manner that is consistent with the original conditions (e.g., consent) under which the data were collected, as well as in accordance with Federal regulations for de-identifying data and protecting participant privacy. Effective models exist for sharing data while providing necessary protections, such as controlled-access (e.g., to large-scale human genomic data) or results dissemination (e.g., of registered clinical trials), in addition to new models of data stewardship that use cloud-based approaches.

Ethical Issues in the Development of Complex Machine Learning Algorithms
Sara R. Jordan, Virginia Tech

Many statements of ethics for machine learning and artificial intelligence (AI/ML) are written at a high level that does not acknowledge fully the complexity of developing machine learning algorithms. Specifically, while statements that AI/ML ought to be “transparent” or “explicable” are easily laudable, they are not technically feasible except when coupled with some extraordinary steps taken by programmer teams and their team leaders. In this presentation, I take ethics for AI/ML down from the high level statements to explanations of in medias res techniques for how to build explicable and accountable AI. I will focus on the development of neural network models in the realm of natural language processing algorithms alone in order to demonstrate where the intersections of ethical norms and technical practices will change conventional technical practices, such as data collection, transformation, model building, and model testing. Model deployment in consumer products will be discussed briefly.
CONCURRENT SESSION E-5
Information Quality Frameworks for Alternative and Integrated Data

A Framework and Best Practices for Measuring and Reporting Data Quality
Keenan Dworak-Fisher, Bureau of Labor Statistics

In the US, statistical agencies have increasingly adopted approaches that combine administrative and other "alternative" data sources with the survey and census data upon which they have traditionally relied. At the same time, the increasing availability of data from a wide variety of sources and the ongoing advance of analytic innovations has created opportunities for many other actors to create powerful data products. As the range of potential applications of data expands, it is increasingly important for data users to have information about the data quality that enables them to determine the data's fitness for particular purposes. In this paper, we propose a data quality framework for US Federal agencies to use for data quality measurement and reporting. Like the frameworks developed in several other countries, the proposed framework incorporates many data quality dimensions over which users and producers may trade-offs. Rooted in emerging international precedents, existing Federal guidance, and the longstanding values of the Federal statistical system, the framework emphasizes the importance of meeting user needs and protecting the privacy of subjects while achieving high levels of accuracy. It is applicable to data that is designed for particular uses (e.g., sample surveys) as well as data that has been adapted from one use to another (e.g., statistical applications of administrative data). Using the framework, we describe factors that may threaten data quality along one or more of its dimensions. Within the accuracy dimension of quality, this includes threats covered in the survey data literature, similar threats that are confronted in alternative sources, and threats that pertain specifically to the combination of multiple sources. Data users need to be aware of the potential impact of these threats and how data producers may have addressed them. We build on our discussion of threats to identify best practices for Federal agencies to use in data quality measurement and reporting.

Total Error Frameworks for Integrated Survey and Found Data
Paul Beimer, RTI

The survey world is relying more heavily on "found" data for inference and decision making rather than survey or "design" data. Found data are data that are not primarily collected for statistical purposes but contain information that might be useful for inference. Data become "found" when they are used to achieve some statistical objective through data mining or analysis. This paper focuses on the quality of data produced by integrating two or more datasets, particularly when one of those datasets is from a survey and the other is found. First it considers the processes by which two or more data sources are integrated as well as the processes by which (hybrid) estimates are derived from integrated datasets. It then describes several total error frameworks that have been proposed for evaluating the quality of the integrated dataset itself and several related frameworks for evaluating the quality of the hybrid estimates that may be produced from such datasets. One of these data quality frameworks is illustrated for an application where an administrative data set is being integrated with a data set from a national survey.

CONCURRENT SESSION F-1
Blended Data for Evidence-Based Research and Evaluation

Blending Privately-Provided Payroll Data and Government Statistics
Leland Crane (Federal Reserve Board)

We combine employment information from three sources in order to accurately measure the U.S. labor market in real time. The data sources are (1) microdata from the payroll processing firm ADP, (2) the CES survey from BLS, and (3) administrative QCEW data, also from the BLS. While the
ADP data cover a very large fraction of U.S. employment (the sample is similar in size to the CES), they are a convenience sample, not a probability sample. We use QCEW data to reweight the ADP data based on a set of observables to attenuate selection bias. We argue that pooling the CES and ADP data can reduce the sampling error inherent in both data sources. In particular, we infer “true” unobserved payroll employment growth using a state-space model that treats both series as noisy signals, and evaluate the forecasting power of various measures derived from this setup. We also incorporate the QCEW by benchmarking our ADP data to the QCEW at low frequencies, taking advantage of its comprehensive coverage of employment.

**Public Libraries and Collective Efficacy: An Exploratory Study of Blending Data from the Public Libraries Survey and the American Housing Survey**
Lisa Frehill (Institute of Museum and Library Services)

There has been much research about the roles played by public libraries in enhancing social capital and civic engagement (Audunson et al 2007, Gong et al 2008, Johnson 2010 and 2012, Horrigan 2016, Vårheim 2014 and 2016). Most of these studies use analyses of relatively small numbers of individuals or libraries that provide often rich details about these cases. Our study explores using data from two separate Federal data collections, the Public Libraries Survey and the 2013 data in the Collective Efficacy module of the American Housing Survey to explore the extent to which these data sources can be blended to take a more macro view of the connection between libraries, social capital and civic engagement. Such an approach trades rich detail for broader generalizability and, therefore, potentially greater utility for more data users in the library data community. The paper reviews the methodological challenges associated with blending data from these disparate sources and provides lessons learned on how to overcome these challenges.

**Linking Official Statistics and Remote Sensing Data for Training Crop-yield Regression Models**
Luca Sartore (National Agricultural Statistics Service)

USDA's National Agricultural Statistics Service (NASS) publishes more than 500 reports every year. Such publications include monthly and annual yield forecasts and estimates for major crops. To produce the forecasts, several surveys are conducted during the growing season. To enhance the accuracy of these forecasts, NASS uses external sources of information such as weather and remote sensing data obtained at different temporal and spatial resolutions by other national agencies. In fact, weather stations and satellites have provided additional information for the estimation of overall crop productivity, i.e. yield and planted area. The measurements from the Moderate Resolution Imaging Spectroradiometer (MODIS) project provide multispectral composite data at 250 m spatial resolution on crop regions throughout the growing season. An aggregation method based on empirical densities is proposed to reduce the loss of information. These densities are then linked to historical NASS official statistics with the aim to train regression models.

**Measuring the Cost and impact of Open Source Software Innovation on GitHub**
Carol Robbins (National Center for Science and Engineering Statistics)

Open Source Software (OSS), defined by Open Source Initiative, is computer software with its source code shared with a license in which the copyright holder provides the rights to study, change, and distribute the software to anyone and for any purpose. OSS is developed, maintained, and extended both within and outside of the private sector, through the contribution of independent developers as well as people from universities, government research institutions, businesses, and nonprofits. Examples include Apache server software, and R statistical programming software. Despite its ubiquity and extensive use, reliable measures of the scope and impact of OSS developed outside of the business sector are scarce. Activities around OSS development, a vital component of science activity, are not well-measured in existing federal statistics on innovation. Many of the OSS projects are developed and maintained in free
repositories, such as GitHub, and information embedded in these repositories, including the code, contributors, and development activity, is publicly available. In this paper, we use data from GitHub, the largest platform with 31 million users and developers worldwide, obtaining information about OSS projects. We collect 5.2 million project repositories, containing metadata such as author, license, commits (approved code edits), and lines of code. We adopt methods used code as the measure of effort to estimate the time spent on software development and calculate the monetary value using the average compensation for computer programmers from Bureau of Labor Statistics wage data and other costs based on national accounts methodologies. Finally, use network analysis methods developed for bibliometrics and patent analysis to study the impact of these projects.

CONCURRENT SESSION F-2
Advances in Disclosure Avoidance

Incorporating Economic Conditions in Synthetic Microdata for Business Programs: A Case Study
Katherine Jenny Thompson (U.S. Census Bureau)

Many agencies are currently investigating whether releasing synthetic microdata could be a viable dissemination strategy for highly sensitive data, such as business data, for which disclosure avoidance regulations would otherwise prohibit the release of public use microdata. The U.S. Census Bureau has identified the Economic Census as a candidate program and has been developing synthetic data generators accounting for skewed and irregular distributions that preserve selected privacy features and satisfy predetermined edit constraints. In this context, high-utility synthetic data should (1) retain multivariate relationships between items, (2) produce marginal totals that closely correspond to their published official statistics, and (3) reflect the state of economic expansion or contraction within the input data. Our prior research has been restricted to businesses that were in operation for the full year, ignoring the special features of births and deaths in the models. This paper introduces models that generate partially synthetic data that fully address all three utility requirements and provides preliminary results using selected industry data from the 2012 Economic Census.

Risk-Efficient Bayesian Data Synthesis for Privacy Protection
Terrance Savitsky (Bureau of Labor Statistics)

High-utility and low-risks synthetic data facilitates microdata dissemination by statistical agencies. In a previous work, we induced privacy protection into any Bayesian data synthesis model by employing a pseudo posterior likelihood that exponentiates each contribution by an observation record-indexed weight in [0, 1], defined to be inversely proportional to the marginal identification risk for that record. Relatively risky records with high marginal probabilities of identification risk tend to be isolated from other records. The downweighting of their likelihood contribution will tend to shrink the synthetic data value for those high-risk records, which in turn often tends to increase the isolation of other moderate-risk records. The result is that the identification risk actually increases for some moderate-risk records after risk-weighted pseudo posterior estimation synthesis, compared to an unweighted synthesis; a phenomenon we label "whack-a-mole". This paper constructs a weight for each record from a collection of pairwise identification risk probabilities with other records, where each pairwise probability measures the joint probability of re-identification of the pair of records. The by-record weights constructed from the pairwise identification risk probabilities tie together the identification risk probabilities across the data records and compresses the distribution of by-record risks, which mitigates the whack-a-mole and produces a more efficient set of synthetic data with lower risk and higher utility. We illustrate our method with an application to the Consumer Expenditure Surveys of the U.S. Bureau of Labor Statistics. We provide general guidelines to statistical agencies to achieve their desired utility-risk trade-off balance when disseminating public use microdata files through synthetic data.
Comparative Study of Differentially Private Synthetic Data Algorithms and Evaluation Standards
Joshua Snoke (RAND)

Differentially private synthetic data generation is becoming a popular solution to releasing analytically useful data while preserving the privacy of individuals in the data. In order to utilize these algorithms for informed public policy decisions, policymakers need an accurate understanding of these algorithms’ comparative performance. Correspondingly, practitioners also require standard metrics for evaluating the analytic qualities of the synthetic data. In this talk, we present an in-depth evaluation of several differentially private synthetic data algorithms from the recent National Institute of Standards and Technology’s (NIST) “Differentially Private Synthetic Data Challenge”, offering both theoretical and practical analyses of these methods. In addition, we implement two evaluation metric algorithms on the differentially private synthetic data and compare the results to the NIST data challenge outcome. We also investigate two of our own original, previously published differentially private synthetic data algorithms to evaluate their comparative performance, framing both the NIST data challenge methods and our own algorithms within the broader differentially private synthetic data literature. Our assessment of the differentially private data synthesis methods and the quality metric algorithms evaluates the relative usefulness (ranging from specific measures of model accuracy to general measures of distributional similarity), general strengths and weaknesses, preferred choices of the algorithms and metrics, as well as implications of our evaluation for policymakers seeking to implement differentially private synthetic data algorithms on future data products.

Generating Poisson-Distributed Differentially Private Synthetic Data
Harrison Quick (Drexel University)

The dissemination of synthetic data can be an effective means of making information from sensitive data publicly available while reducing the risk of disclosure associated with releasing the sensitive data directly. While mechanisms exist for synthesizing data that satisfy formal privacy guarantees, the utility of the synthetic data is often an afterthought. More recently, the use of methods from the disease mapping literature has been proposed to generate spatially-referenced synthetic data with high utility, albeit without formal privacy guarantees. The objective for this paper is to help bridge the gap between the disease mapping and the formal privacy literatures. In particular, we extend an existing approach for generating formally private synthetic data to the case of Poisson-distributed count data in a way that allows for the infusion of prior information. To evaluate the utility of the synthetic data, we conducted a simulation study inspired by publicly available, county-level heart disease-related death counts. The results of this study demonstrate that the proposed approach for generating differentially private synthetic data outperforms a popular technique when the counts correspond to events arising from subgroups with unequal population sizes or unequal event rates.

CONCURRENT SESSION F-4
Data Science: Capacity Building to Solve Real World Problems

With the passage of the Foundations for Evidence-Based Policy Making Act, the establishment of a Federal Data Strategy, and an environment of increasing data demands with stable or shrinking budgets, the need for a data and programming savvy workforce has never been more important. This diverse set of skills related to data access, transformation, modeling, and communication is often summarized as “data science.” This session will cover the challenges in building data science capacity in the federal system, particularly in terms of training, and solutions now underway at the Census Bureau and OMB. The session will also provide three motivating examples where data science was used in federal statistical agencies to solve real world problems. The talks are non-technical and geared toward managers and staff looking to expand the use of data science at their agency through building skills amongst their existing workforce.
Federal agencies in the United States produce a wide range of estimates from increasing sources of data to inform evidence-based policy decisions. Communicating the uncertainty of these estimates and the uncertainty of associated inferences (e.g. trends, comparisons) is essential to transparent quality reporting and making informed decisions. In 2016, the American Statistical Association (ASA) released a statement on the use of significance testing, one tool used for interpreting and communicating the uncertainty of statistical data, recommending a decreased reliance on p-values for decision making. This session brings together a panel to discuss communicating statistical uncertainty for federal agencies, including implications of the 2016 ASA statement, information needs of data users and stakeholders, and some alternatives for communicating statistical uncertainty for evidence-based policy decisions.

**CONCURRENT SESSION F-6**
Applications of Non-Probability Samples

*Lock Sampling, or: Yes, Panels are Different - Now What?*
Jeff Gill (American University)

Most survey analyses treat “don’t know” or nonattitude responses as missing values and drop them from analysis with case wise (list wise) deletion. There are problems with this approach. In this work we demonstrate first that nonattitudes and “don’t know” responses are not random, but rather come from a distinct group of survey respondents. This is shown by modeling relevant missingness as a dichotomous outcome variable explained by various characteristics, including demographic attributes, other attitudinal questions, and group level contexts. This model allows us to produce an imputational model to predict missingness due to ignorance versus intransigence.

*Crowdsourcing for recruiting hard-to-reach populations: An example of recruiting military veterans*
Y. Patrick Hsieh (RTI International), Leyla Stambaugh (RTI International), Herschel Sanders (RTI International),

Conducting surveys with military veterans is challenging, from accessing representative samples to gaining adequate response rates that yield valid, data-based conclusions. Consequently, US military veterans could be an operationally hard-to-reach population for survey research. This study addresses the issue of sample access while also piloting innovative recruitment strategy for increasing response rates and sample representativeness of veteran surveys by using Amazon’s Mechanical Turk (MTurk), the established crowdsourcing platform with broad population coverage in its user base. We started the sample construction by recruiting military veterans from MTurk since 2018. Our effort produced a diverse online sample of more than 600 unique veterans who passed our validation questions. We assessed the extent of its representation of the demographics of military veterans currently living in US households and describe the survey results relating to the corresponding estimates from national surveys of veterans. we will also discuss best practices and lessons learned for building future representative samples of hard-to-reach populations using online recruiting methodologies.
Systematic Evaluation of Respondent Driven Sampling Implementation
Sunghee Lee (University of Michigan), Ai Rene Ong (University of Michigan), RJ Batas (University of Michigan)

Absence of practical solution for sampling rare, elusive and hard-to-reach population subgroups made respondent driven sampling (RDS) gain popularity and be adopted at the federal level. In traditional sampling, methodological terminologies (e.g., area-probability sampling) convey some standardized meaning; and certain reporting elements (e.g., response rates) are required for transparency. In RDS, however, little is standardized; and its implementation may take vastly different forms.

This study attempts to assess the current state of RDS study implementation through a systematic review. We will locate peer-reviewed RDS articles, RDS technical documents (e.g., method reports) as well as grant-funded RDS projects through various search engines. With the located information, we will create a database by capturing study characteristics (e.g., target group) and design elements (e.g., mode) from each study. We will analyze the database to describe types of the population, modes, seed recruitment, incentive structures and coupon usage currently practiced with RDS. This will allow us to identify critical areas for assessing data quality and examine their treatment in the RDS literature.

CURRENT SESSION G-1
Linkage Applications of ACS Data

Augmenting ACS microdata with CPS using machine learning

The BLS’ Productivity Program uses CPS microdata to supplement estimates of hours worked and employment, and estimates of workforce composition using age, gender, education, and industry. The CPS’s timeliness and demographic detail make it an excellent source for measuring these statistics in larger industries. But the sample size is often too small in more-detailed NAICS industries. In this paper, we use ACS data to improve these estimates. We use random forest algorithms with benchmarking to known totals to create alternate industry labor hour and composition estimates with predicted values for employment, hours worked, and hourly wage. Similar data elements in the CPS Basic, CPS ASEC, and ACS surveys are matched and adjusted to account for their differences. The ACS values are benchmarked on demographic, geographic, major occupation, and sector, to CPS totals. For predicting hours worked and hourly wages we train our algorithms on CPS data to predict values for each ACS observation. The resulting matrix is used to compute hours worked and labor composition, and compared to official estimates. Our methods can be extended and used in many other applications.

Impacts of Broadband Development on Rural Property Values
Joshua Goldstein (University of Virginia), Teja Pristavec (University of Virginia), Devika Mahoney-Nair (University of Virginia)

Broadband expansion to underserved rural areas provides communities with many economic and social advantages. We combine multiple data sources to assess the effects of developing broadband infrastructure on quality of life across the rural-urban spectrum. Using data from the United States Department of Agriculture’s Rural Utilities Service (RUS), we investigate the impacts of loan and grant programs for broadband investments on rural property values. We capture broadband availability and speed using Federal Communications Commission (FCC) Form 477 data, and use American Community Survey (ACS) responses on broadband accessibility to address known FCC data biases and limitations. The analysis includes data linkage, quality assessment, and spatial
integration for RUS, FCC, ACS, and housing information from tax assessments nationally at the U.S. Census tract level. We use a spatial hedonic model to estimate the impact of broadband expansion on rural property values.

**Counting (on) Journalists: Using Federal Statistical Data to Estimate the Size, Characteristics and Geographic Distribution of Newsroom Employees in the United States**

Elizabeth Grieco (Pew Research Center)

Journalists play an important role in the democratic decision-making process, as they oversee the work of government officials on behalf of citizens. But we are losing our newsroom employees, as mergers, closures and layoffs continue to pummel media organizations. These recent trends in media employment raise an important question: Do we know how many newsroom employees work in the United States – and how many we’ve lost? Media researchers have traditionally relied on surveys by professional organizations for information on newsroom employees, but they do not produce estimates of the total number of newsroom employees and, due to methodological constraints, produce data that are not statistically representative. Two government surveys, while less-used by media researchers, include industry-wide as well as sector-specific estimates that are generalizable. This paper uses the OES and ACS to estimate the size, characteristics and distribution of newsroom employees, demonstrating that government surveys are a good alternative source for information on this important occupation.

**Veteran Population Projection Model: Projecting the Demographics and Trends to Better Serve the Veterans**

Jin Kim (U.S. Department of Veterans Affairs), Hyo Park (U.S. Department of Veterans Affairs), Charles Lin (U.S. Department of Veterans Affairs)

This paper provides an overview of the methodology and data sources the Department of Veterans Affairs (VA) uses to produce the official estimates and projections of the veteran population from the Veteran Population Projection model, as well as a summary of future veteran population trends. For veterans and their families, VA provides a broad range of benefits and services. To the extent that VA managers, planners, and program administrators have access to data on the size and geographic distribution of the veteran population now and in the future, they will be able to make evidence-based decisions regarding the most effective allocation of scarce resources. Through periodical model updates for improved methodology and more recent national survey data and more comprehensive administrative records on demographics and military service history, the model provides a 30-year projection of the size of veteran population by demographic characteristics such as age, sex, period of service, and geography. Such information for the present and future serves an important role in strategic and long-range planning within VA and by external organizations.

**CONCURRENT SESSION G-2**

**Measuring Self-Employment Status and Income**

**Tax Burdens and Barriers for Self-Employed Taxpayers: Results from the IRS Estimated Tax Survey**

Janet Li (Internal Revenue Service)

The IRS Estimated Tax Payment (ETP) Survey is a new survey instrument developed by the Research, Applied Analytics & Statistics division of the IRS. The survey samples individual taxpayers with estimated tax obligations: taxpayers who earn income outside of wages, retirement income, and government payments, including self-employed taxpayers. Self-employed taxpayers generally have to pay quarterly estimated taxes to cover their yearly income tax and self-employment tax obligations. The ETP survey was designed in alignment with other federal tax burden surveys conducted by the IRS under the scope of the Paperwork Reduction Act, which produce time-and-money estimates of the burden of tax compliance for individuals, businesses, and
other tax-paying entities in meeting various types of tax obligations. The data used for burden measurements include the time spent on recordkeeping, tax planning, gathering materials, and using tax services; and the money spent on tax preparation software or professional accounting services. The ETP Survey, launching in the fall of 2019, expands the scope beyond tax burden for estimated tax payers to include taxpayers with estimated tax obligations who incur estimated tax penalties or who otherwise do not pay estimated taxes on their non-wage income. Questions—mostly quantitative, with some open-response options—are organized around awareness of estimated tax requirements, income variability and predictability, preferences around payment scheduling, salience and utility of estimated tax penalties, identification and motivation around self-employment work, and attitudes towards taxes. A sample of 20,000 taxpayers are mailed letters with customized links to a web survey, and are comprised both estimated tax payers (70 percent of the sample) and non-payers (30 percent). The development of sampling criteria to identify taxpayers who are not meeting their estimated tax obligations is discussed. Survey responses are linked back to administrative tax data to measure, among other things, self-employed taxpayers’ awareness of their own behavior around estimated taxes. Survey results will ultimately be used to understand the operational and policy barriers most salient in estimated tax noncompliance and to investigate which of these can be addressed to minimize future burden on both self-employed taxpayers and the IRS.

Improving Self-Employment Imputations with Administrative Data and Model Based Imputations: An Analysis Using the Survey of Income and Program Participation
Robert Munk (U.S. Census Bureau)

Since 1984 the Survey of Income and Program Participation has used hot-deck imputation to impute variables related to self-employment—self-employment status and incorporation status. Given the computing resources available in 1984, hot-deck imputation was the best imputation technique for that period. However, the last 35 years have brought drastically improved computing resources, a deeper understanding of the limitations of hot-deck imputation, and increased availability of administrative data that provides information about the employment situation of survey respondents and non-respondents alike. Therefore, in this paper, we impute the variables related to self-employment using a model-based imputation technique that takes advantage of administrative tax data. This approach results in an increase in imputed data quality by producing a correlation between administrative and imputed self-employment variables that is more reflective of the correlations in reported data. This improvement then enables us to improve self-employment income imputations. We demonstrate this by running a pseudo-experiment on individuals that report self-employment: we impute data for the reporters and compare it to their reported values. We show how model-based imputation produces regression estimates, which are of interest to self-employment researchers that are more consistent to reported data than the estimates produced via hot-deck. We conclude with recommendations for data users and fellow survey administrators.

The Effect of Structural and Cyclical Changes on Trends Across Time in the Number of Independent Contractors
Anne Polivka (Bureau of Labor Statistics)

Despite numerous media reports that the prevalence of independent contractors have increased overtime, existing evidence does not always support this contention. This paper explores structural and cyclical influences that could be affecting the aggregate estimates of the proportion of the employed who are independent contractors. Because industries use independent contractors at different rates and because occupations differ in the degree to which this arrangement can be utilized, changes in both the industrial and occupational composition of the U.S. economy could influence trends in the proportion of workers who are independent contractors. The effect of these structural shifts are analyzed in this paper using shift-share analysis. Similarly, workers’ desire and need to be independent contractors could vary over the business cycle. The effect of the business cycle is examined in a regression context using the unemployment rate at both the national and state level as an indicator of cyclical variation. The effect of other demand shocks are also explored.
Finally, in addition to an examination of influences on the aggregate estimate, effects within specific industries—such as the transportation industry—are examined to determine if some industries are undergoing transformation, even if there are no observed effects in the aggregate estimate for the entire economy.

**Reconciling Survey and Administrative Measures of Self-Employment**  
James Spletzer (U.S. Census Bureau)

This paper addresses key questions about self-employment: whether and how much such work has grown over the past several decades, the characteristics of self-employment work and workers, and how self-employment interacts with wage and salary employment. Existing household survey data do a poor job of capturing self-employment, whereas administrative data typically contain limited information about the people who are doing it. We combine records from the Annual Social and Economic Supplement (ASEC) to the Current Population Survey (CPS) with earnings information contained in the Social Security Administration’s Detailed Earnings Record (DER) files. The information in the DER on self-employment earnings comes from Schedule SE filings; the DER also includes Form W-2 information on wage and salary earnings. This linked data file covers the 1996-2015 period. Analysis of the linked CPS-ASEC and DER records has produced several key findings:

- Over the 1996-2015 period, two-thirds of those with self-employment earnings in the DER had no self-employment earnings in the CPS-ASEC.
- Between 1996 and 2015, the number of people with self-employment earnings in the DER but not the CPS-ASEC grew by 6.2 million; the number with CPS-ASEC self-employment earnings who did not have DER self-employment earnings changed relatively little and the gap between the two self-employment series grew by 5.5 million. This paper investigates alternative explanations for the growing number of DER self-employed with no CPS-ASEC self-employment earnings. Possible contributing factors include changes in workforce characteristics (e.g. growth in the number of older self-employed persons), changes in the persistence of self-employment, and changes in reporting behavior.

**CONCURRENT SESSION G-3**  
Innovations in Survey Design

**Survey Improvement Recommendations from a Field Staff**  

The Bureau of Labor Statistics’ Consumer Expenditure Survey program periodically asks Census field staff to provide feedback about data collection procedures (e.g. use of a diary, and telephone interview in lieu of in-person interview), questionnaire design, and general topics of survey methodological interest. Field staff provide a unique perspective on the effectiveness of survey protocols, and potential improvements to the questionnaire and protocols. Their feedback is used to improve training and the data collection process. The 2019 Field Staff Survey was conducted from August to September 30, 2019 through Survey Monkey, and sought feedback on (i) implementation of recently introduced questions that ask where an item was purchased, (ii) changes to the diary, (iii) experience with telephone interviews, (iv) use of a show card booklet, records, and other respondents materials, (v) use of a Contact History Instrument for tracking contact attempt-based observations, and (vi) burden related research questions (e.g. perceived burden). These recommendations will be of interest to survey methodologists, data producers, and field staff.
Mode Effects or Measurement Reliability: Differences in Estimates from Same Individuals between Web and Mail Survey Administration
John Boyle (ICF), Ronaldo Iachan (ICF), Matt Jans (ICF)

As the number of multi-mode surveys increase, greater attention is being paid to potential mode effects introduced by these methods. Unfortunately, most of the literature on mode effects compare interviewer assisted modes to self-administered modes, while current multi-mode surveys tend to rely primarily on web and mail modes. We have very little test/retest data from a large national sample that would verify differences in survey estimates between self-administered modes. By chance, a sample of respondents in a national multi-mode ABS survey who initially responded by web were subsequently sent a request to complete a paper questionnaire. A total of 764 unique respondents at the same address completed the same interview by mail. Differences in the estimates from the same question by the same individual between web and mail samples provide a test of potential mode effects. However, these differences could also reflect response variability over time unrelated to mode. We assess measurement reliability (ignoring mode effects) by comparing estimates between the mail and web samples; we test mode effects by the consistency of the same respondent’s responses between the two modes.

Optimizing Data Collection Procedures for Web-Push Survey Designs: Evidence from a National Study
Michael Jackson (ICF), Danielle Battle (ICF), Rebecca Medway (American Institutes for Research)

Web-push mixed-mode survey designs are becoming increasingly common in the federal statistical system. This paper aims to advance the federal statistical community's understanding of best practices for web-push designs by reporting the results of several experiments incorporated into the 2019 cycle of the National Household Education Survey (NHES). A previous proof-of-concept experiment (in 2016) established the feasibility of using a web-push design for the NHES. The 2019 cycle therefore incorporated several randomized experiments that aimed to hone the web-push data collection procedures. These included tests of: (1) a prenotification letter, relative to no prenotification letter; (2) alternative timings of a FedEx mailing; and (3) a specially designed pressure-sealed reminder mailing that included web login information, vs. a more traditional postcard reminder that did not include login information. This paper will discuss the results of these experiments and their implications for web-push survey designs. Analysis will focus on the effects on response rates, the distribution of responses across the available modes, and data collection costs.

Using SMS as a Survey Recruitment Tool – Challenges Convincing Respondents and Carriers of Legitimacy
James Dayton (ICF), Robynne Locke (ICF), Rachel Kinder (ICF)

There is a high demand to produce population estimates of various groups by Statistical Agencies, the United Nations, OECD & other organizations because of high-stakes impacts in demography. This research investigates how machine learning algorithms perform in predicting population estimates by a group based on some predictors. It attempts to predict the SSA disability-population count by State, based on two formulations: 1) as a function of trend of each State; 2) as a function of 10 causes of death, poverty & geolocation. It also demonstrates how administrative data from SSA is merged with the Survey data from Census, NCHS, & metadata of State information. Four datasets are joined to construct the final data. LOESS, LASSO, CART, CTREE & Random Forest models are implemented to predict the disability-population by State. The best model(s) is selected with highest predictive power on a future test dataset. Results suggest machine learning prediction could be useful when response data are not readily available, or a different dataset is available with some predictors, or predicted estimates are necessary in order to make a resource allocation in advance without the observed data.
Recruiting a probability sample of 18 year olds for a longitudinal study on interpersonal violence
David Cantor (Westat), Reanne Townsend (Westat), Gail Thomas (Westat)

Recruiting a general population sample of young adults in the current survey environment is difficult. Short of doing in-person screening, using contact modes such as mail or the internet all have inherent difficulties. Young people do not readily respond to requests by mail or the internet. This presentation describes a recruitment of 18 year olds into a longitudinal study of inter-personal violence. The goal was to recruit a probability sample that can provide generalizable data on long-term trajectories of risk for and experiences with violence as young adults are transitioning out of the home to independent living. The design uses an address based sample (ABS) that is supplemented with a list of high school seniors to stratify and oversample households that were likely to have an 18 year old. The study has successfully recruited 1800 young adults by using a combination of postal requests, pushing respondents to the internet, incentives and gamification methods. This presentation will provide an overview of key design features, the response rate and the efficiency of using these methods to recruit this difficult-to-survey group.

CONCURRENT SESSION G-4
How I Learned to Stop Worrying and Love Data Science

Automating Public-Sector Frame Maintenance and Item Replacement
Keith Finlay (U.S. Census Bureau)

The Census Bureau collects public-sector data to produce statistics on governments and their activities such as those of the criminal justice system. A key step in this data collection is the manual, resource-intensive maintenance of survey frames, which involves updating frame information (e.g., contact information, unit characteristics, etc.) and hand matching new data tables to existing frames. Much of the information used to update public-sector frames is publicly available on the web, including content that could be used to replace response items in surveys. The Census Bureau is developing a frame maintenance and item replacement automation system that relies on web crawling, web scraping, algorithmic document parsing, and probabilistic matching to bring in unstructured data, identify useful structured information, and link those new data with existing frame records. This process will save resources and increase quality, and could be applied to many data collections at the Census Bureau and across government. This presentation will discuss the work accomplished by a Civic Digital Fellow during Summer 2019.

Blended data approach for estimating business owners’ characteristics
Aneta Erdie (U.S. Census Bureau)

In a climate of low and declining response rates, and high and increasing survey costs, it is imperative to innovate in order to continue to produce high quality data. Fortunately, the Census Bureau maintains a large data linkage infrastructure, which includes survey, census, and administrative data. This data infrastructure is now being used to replace the non-employer portion of the Survey of Business Owners. Using data from the American Community Survey, decennial census, and administrative records from the Veterans’ Administration and other sources, the Census Bureau is able to produce non-employer characteristics by demographics without collecting new information from respondents. In addition to reducing respondent burden and survey costs, the use of blended data will enable annual estimates on non-employers to be produced, much more frequent than the five-year estimates produced under the Survey of Business Owners. Ultimately, this project provides a model for federal agencies to use data science methods to overcome issues that threaten the business-as-usual survey approach.
Industry Classification using Machine Learning: an Application to the Economic Census
Javier Miranda (U.S. Census Bureau)

The US Census Bureau spends considerable time and resources identifying and classifying the industry of establishments in the U.S using the North American Industrial Classification System (NAICS). This information is critical to Census Bureau statistical products. The burden of this collection to businesses is considerable as well. We have developed a natural language processing and machine learning pipeline that uses a novel combination of proprietary and public data for the purpose of predicting complete NAICS codes. We focus on how this new approach can be implemented into the typical workflow for analysts, how it can be used to automate the easy-to-solve cases leaving analysts with more resources to tackle difficult to classify establishments, and how it could be used to eliminate entire survey operations while maintaining, and potentially, improving data quality while lowering costs.

Automating Response Evaluation for Franchising Questions on the 2017 Economic Census
Joseph Staudt (U.S. Census Bureau)

Between the 2007 and 2012 Economic Censuses (EC), the count of franchise-affiliated establishments declined by 9.8%. One reason for this decline was a reduction in resources that the Census Bureau was able to dedicate to the manual evaluation of survey responses in the franchise section of the EC. Extensive manual evaluation in 2007 resulted in many establishments, whose survey forms indicated they were not franchise-affiliated, being recoded as franchise-affiliated. No such evaluation could be undertaken in 2012. In this paper, we examine the potential of using external data harvested from the web in combination with machine learning methods to mostly automate the process of evaluating responses to the franchise section of the 2017 EC. Our method allows us to quickly and accurately identify and recode establishments that have been mistakenly classified as not being franchise-affiliated, increasing the unweighted number of franchise-affiliated establishments in the 2017 EC by 22%-42%.

Machine Learning and the Commodity Flow Survey
Christian Moscardi (U.S. Census Bureau)

The Commodity Flow Survey (CFS), a joint effort between the Bureau of Transportation Statistics and the Census Bureau, has implemented a machine learning process to improve data quality as well as reduce operational costs and respondent burden. In particular, we have implemented a machine-learning based “autocoder” to classify free-text product descriptions provided by respondents. Using this autocoder, we have imputed or relabeled codes for approximately 200,000 records from the 2017 CFS. This has improved production data quality and the resulting estimates, while simultaneously reducing respondent burden and costs. We will discuss the applied research process that led to our implementing this methodology into our production workflow. While this project has been implemented into production, we are working on ways to make a bigger impact with data science approaches. First, we are exploring how we can use Amazon’s MTurk to improve develop a better training dataset to improve the model’s performance. Second, we are investigating how we can deploy this model in future CFS collections, to replace a burdensome survey question and improve data quality. This subsequently unlocks our ability to collect significantly more data from respondents, enabling more granular and higher-quality estimates without significantly increasing respondent burden.
Overview of the Transparent Reporting Project and Assessing the Data User's Perspective
Mark Prell, Economic Research Service

The Transparent Reporting Project examined current practices on transparent reporting for selected integrated data products. The project developed a web-based customer survey that solicited user appraisals of the quality of the data, user satisfaction with the transparency of the agency reporting on data quality, and user suggestions for improving data quality and documentation. The framework adopted by the project and its customer survey consisted of eight dimensions of data quality: relevance, accuracy, reliability, timeliness, punctuality, consistency, comparability, and access. The customer survey represents a pioneering effort—the first known survey designed to collect assessments of transparent reporting on integrated data using common items across users of several statistical products produced by different agencies.

Quality Considerations when Integrating Alternative Data into the Consumer Price Index
Crystal Konny, Bureau of Labor Statistics

The Consumer Price Index (CPI) is a widely-used measure of inflation that measures average price change for the consumption sector of the American economy. BLS uses data from several surveys to select a sample, and field staff collect approximately 95,000 commodities and services prices and 8,000 rents monthly. BLS is exploring the use of alternative data, such as third-party private-sector datasets, to replace and supplement this traditional price and rent collection, with the goal of realizing several benefits, including cost savings, larger sample sizes, increased accuracy, quicker introduction of new goods into the index, and reduced respondent burden. This data would be blended into the CPI along with traditional survey data. Substantial work is underway in the Bureau of Labor Statistics to identify and research alternative data sources. This talk will cover the BLS process of assessing quality considerations when BLS evaluates third party datasets for use in the CPI.

Reporting on Integrated Data Quality for Linkage between the National Hospital Care Survey and the National Death Index
Lisa B. Mirel, National Center for Health Statistics

Survey data provide information on a wide range of topics; however, they often lack information on longitudinal outcomes. Through its Data Linkage Program, the National Center for Health Statistics (NCHS) has been able to enhance the analytic utility of the data collected in its surveys, by augmenting them with information from health-related administrative records. The survey data are linked to administrative data on a periodic basis that balances the timeliness and relevance of individual data sources with the availability of resources to undertake large data integration projects. Despite the benefits of combining these sources, the potential to introduce additional sources of error increases when integrating multiple data sources. This talk will describe the process of linking data from one of the NCHS surveys, the National Hospital Care Survey, to the National Death Index. It will outline the information provided to prospective users so they can assess the quality of the integrated data, including assessing potential bias due to linkage eligibility, linkage error, and other sources of error including systematic and random missing data. In addition, results from a customer survey about the documentation of the integrated data will be presented.
Reporting on the Veteran Population Model Documentation
Thomas A. Garin, Department of Veterans Affairs

The National Center for Veterans Analysis and Statistics (NCVAS) used mixed data to create the Veteran Population (VetPop) Projection Model. The Department of Veterans Affairs (VA) uses VetPop’s results as the official estimates and projections of the Veteran population and to project medical and financial resources needed down to county-level service areas. While the administrative data contains extensive amounts of demographic and military service data, information on older Veterans who separated prior to the mid-1970s is no longer available to be used in the VetPop model. The model, instead, relies on information drawn from the existing data sources, including the Veteran record-level administrative data, and blends it with the nationally representative survey data to account for the missing Veteran records. The VetPop model’s documentation defines the data quality and describes it in a setting that relies heavily on the administrative data. By placing the documentation on a public-facing website, it communicates data quality to the VetPop users. This session highlights some of the results from a customer survey about the documentation.

Reporting on Integrated Data Quality for the National Postsecondary Student Aid Study
Chris Chapman, National Center for Education Statistics

One type of integrated data product regularly produced by federal statistical agencies are micro-data sets, or data sets that a few to a few thousand unique variables for individuals, institutions, organizations, etc. Integration for such products can include a range of linkage types including linking survey data to administrative data, linking administrative to other types of administrative data, linking survey data to other survey data, etc. The National Postsecondary Student Aid Study (NPSAS) uses several different types of linking, but relies most heavily on linking survey data provided by sampled student respondents to administrative data about those same students. The purpose of NPSAS is to study how students and their families pay for college. It fulfills a legal mandate in the Higher Education Act for the National Center for Education Statistics (within the U.S. Department of Education’s Institute of Education Sciences) to collect and disseminate such information. NPSAS provides detailed measures on student financial aid and borrowing as well as other key indicators of postsecondary education (e.g., enrollment) and demographics. While the student interview has traditionally played a significant role in creating relevant measures in the study, the increasing quality and availability of administrative data has facilitated a shift over time, to where many constructs rely less on self-reported information and more exclusively on administrative records. The presentation will provide information about how data from six administrative sources are used to improve the data collection and resulting data, and information made available to consumers of NPSAS data to facilitate understanding of the data product.

CONCURRENT SESSION H-1
Leveraging Multiple Health Care Data Sources to Generate Statistical Information in the U.S. Health Care Delivery System

Benefits and Challenges in Using Survey and Administrative Data in the National Post-Acute and Long-Term Care Study (NPALS)
Lauren Harris-Kojetin (National Center for Health Statistics)

Post-acute and long-term care (PALTC) provided by paid regulated providers is an important part of personal health care spending. NCHS launched the biennial NPALS (formerly NSLTCP) in 2012—a unique resource that produces nationally representative statistical information on providers and services users to support and inform PALTC policy, research, and practice. NPALS covers seven PALTC sectors, using only survey data for two sectors (adult day services and assisted living/similar residential care) and only existing administrative data for the other sectors (home
health, hospice, inpatient rehabilitation, long-term care hospitals, and nursing homes). This efficient strategy allows coverage of more sectors, at more granular geography, and more frequently than possible if using only survey data, while introducing challenges (e.g., measurement harmonization across data sources, assessing data quality beyond the total survey error framework, disclosure risk, and communicating results). This presentation highlights: NCHS’s approach to these challenges; study features and benefits; illustrative findings; and, statistical policy implications of using multiple data sources.

**Lessons learned from using integrated modes of abstracted medical records and electronic health records for collection of physician data: National Ambulatory Medical Care Survey, 2016-2017**
Joseph Staudt (National Center for Health Statistics)

As the health care industry transitions to electronic health records (EHRs), NCHS has been moving towards an approach that combines manual abstraction and EHRs to collect data from visits to physicians and other health care providers. The use of these integrated modes of data collection for a nationally-representative physician survey is unique. It has the potential for both richer clinical data on physician office visits and boosting response rates in the National Ambulatory Medical Care Survey (NAMCS) by using existing EHR data, which can help ease participant burden. NAMCS, which has historically been fielded via manual abstraction, used this integrated approach for data collection in 2016-2017. This presentation provides a discussion of the successes and challenges experienced while collecting and processing these integrated data, particularly from a data quality perspective. Statistical policy implications of the use of integrated data, including guidance for standards and implications for confidentiality are also discussed.

**National Hospital Care Survey: Trials and Tribulations of Integrating Claims and Electronic Health Records Data**
Geoffrey Jackson (National Center for Health Statistics)

The National Hospital Care Survey collects electronic inpatient and emergency department data from a nationally representative sample of hospitals to produce estimates on healthcare utilization. A unique aspect of participation is that hospitals have the option of submitting either claims data or electronic health records (EHR). There are benefits and data quality issues inherent for each source. Claims data are standardized across hospitals but are optimized for billing and do not contain clinical information such as labs and medications. EHR data provide clinical depth and breadth, allowing for a detailed analysis of hospital encounters. For example, claims data can identify opioid-involved hospital visits, only from a diagnosis code, while EHR clinical notes can provide information on the specific type of opioid involved with a hospital visit, independent of the diagnosis code. However, EHRs are not standard across hospitals, resulting in the collection of data from disparate coding systems, text instead of coded data, and missing data elements. This presentation discusses the trials and tribulations of collecting claims and EHR hospital data to produce an integrated data set.
Measuring the Small Business Economy
Tina Highfill (U.S. Bureau of Economic Analysis)

Small businesses employ millions of Americans and represent the majority of businesses in the US. Despite their importance to the US economy, there is no comprehensive measure of economic growth for small businesses. To better track the overall growth and health of small business in the US, the Bureau of Economic Analysis is developing a Small Business Satellite Account, including a new Small Business Gross Domestic Product (GDP) measure. The new account will offer insight into the industries that comprise and contribute to small business economic production, and how those relationships have changed over time. Medium and large businesses will also be part of the statistics to provide a comprehensive representation of economic activity by business size. This paper presents initial estimates of components of GDP by business size for 1998-2016. The methodology and data used to construct these estimates are also described, which rely heavily on data from Census and other federal statistical agencies. Measurement challenges remain related to defining business size classes and accessibility of data needed to develop a full suite of economic statistics for small businesses.

Building an New Urban, Suburban, and Rural Indicator
Emily Molfino (U.S. Census Bureau), Shawn Bucholtz (Housing and Urban Development), Jed Kolko (Indeed)

Classifying areas as urban, suburban, and rural is of interest to researchers and policymakers alike. Yet, there is no objective division between these groups. Local residents may disagree given their own context and lived experience. Existing indicators rely on socio-demographic data. Yet, subjective features, we argue, are key in classifying areas as urban, suburban, or rural. The 2017 American Housing Survey (AHS) asked a national representative sample what type of area they reside. These data allow us to build an urban, suburban, and rural indicator at a granular level that uses socioeconomic and respondent-level features, combining objective metrics with subjective perceptions. We create small area estimates of urban, suburban, and rural by training a random forest classifier on weighted AHS respondent data and tract level estimates. We use this trained model to predict 2013-2017 ACS microdata. Predictions are aggregated to provide an indicator of how the majority of households in an area (tract, county, and congressional district) view where they live. This approach provides an illustrative example of the use of existing federal data to create innovative new data products.

Poverty in the U.S. Using the Comprehensive Income Dataset
Bruce Meyer (University of Chicago), Derek Wu (University of Chicago), Carla Medalia (U.S. Census Bureau)

This paper provides new estimates of poverty in the U.S. using a groundbreaking set of linked survey and administrative data, part of a larger project at the Census Bureau to develop a Comprehensive Income Dataset. The administrative data cover earnings and asset income from IRS tax records and transfer income safety net programs including Social Security, Supplemental Security Income, Supplemental Nutrition Assistance Program, Veterans’ Benefits, Public Assistance, housing assistance, and other programs. We link these data to two household surveys - the Current Population Survey and the Survey of Income and Program Participation – which is vital given that a large and rising share of benefits and other income sources is not recorded in the surveys. Using these linked data, we examine the extent to which misreporting of various survey income sources biases the reported poverty status of households, document how the demographics of those in poverty is changed by the improved data, and improve measurement of the resources of the low-income population. These results will provide a prototype for the combination of survey, program and tax data to improve poverty and income measurement.
**The Unemployment Rate Using Alternative Data: Estimating Labor Market Status with Consumer Bank Flows**
Benjamin Mande (J.P. Morgan Asset Management), Boqiu Lu (J.P. Morgan Asset Management)

Reduction in sampling variability in employment statistics is an inherently desirable goal insofar as it provides a higher signal-to-noise ratio about the state of the economy to firms, consumers and policymakers. In this paper, we explore the extent to which big data can help to achieve this goal. Using a battery of classification tree-based models we replicate the implicit logic used by the Bureau of Labor Statistics (BLS) in categorizing individuals in the Current Population Survey (CPS) as employed, unemployed or out of the labor force. We then deploy the classification models on a vast dataset of consumer bank inflows and use those estimates of labor force status to compute unemployment rates at the state and national levels. We find that for states with significantly more bank data relative to CPS, the level of the unemployment rate is quite similar to official measures. The alternative data measures are almost ubiquitously less volatile than official ones.

**A Nontraditional Data Approach to the CPI Gasoline Index**

The Bureau of Labor Statistics (BLS) has traditionally relied on manual collection of prices for commodities and services. However, given the trend of declining survey response rates in recent years and the high cost of manual collection, it has become essential for the BLS to employ alternative data collection processes to ensure indexes are accurate. One current study is the Crowd-Sourced Motor Fuels Data Analysis project, which may lead to the replacement of the current Consumer Price Index (CPI) gasoline sample. Motor fuel price data are obtained via web collection, with the site’s permission. Our interest stems from improving efficiency, but another objective is to increase the number of observations beyond what CPI survey and sampling methods can provide. Indexes were constructed using the crowd-sourced data to measure monthly price trends, and our preliminary results indicate the crowd-sourced data may be a suitable replacement for the CPI collected data. Additionally, the larger sample size of the crowd-sourced data captures the price change variance exhibited within a geographic zone. We propose a different weighting scheme than what is used in current CPI methodology.

**CONCURRENT SESSION H-3**
Innovations in Small Area Estimation Models for Official Statistics Programs

**Preserving Acreage Relationships in Small Area Agricultural Models**
Lu Chen (National Institute of Statistical Sciences and National Agricultural Statistics)

When National Agricultural Statistics Services (NASS) publishes county-level crop totals, certain factual relationships must be satisfied. For example, the acreages of crop area harvested within any given geographic boundary (e.g. county or state) should not exceed the acreages that were planted. While NASS’s County Agricultural Production Survey (CAPS) is administered and edited to ensure this consistency, the direct survey estimates may lack sufficient precision. For many crops, NASS can also access administrative sources on planted acreage and harvest loss. Current hierarchical Bayesian small area models incorporate those auxiliary data and improve county-level survey estimation. However, preserving the relationship (e.g. harvested acreages should not be larger than planted acreages) at all required levels may require additional steps. Therefore as an alternative to modeling harvested acreages directly, we proposed modeling the proportions of crop area harvested, again using a hierarchical Bayesian model. We comment on the challenges of specifying a fully joint model and comment on the effects on precision of estimates when conditioning harvested area on planted area, and incorporating other publication requirements.
Small Domain Estimation in the National Compensation Survey
Daniel Ayasse (Bureau of Labor Statistics)

The National Compensation Survey (NCS) provides measures of employer costs, compensation trends, and benefits among workers via statistical products like the Employment Cost Index (ECI), the Employer Costs for Employee Compensation (ECEC) series, and the incidence and provisions of employee benefits. Additionally, the NCS program leverages the geographical detail for which employment estimates are produced by the Occupational Employment Statistics (OES) program to produce the Modeled Wage Estimates (MWE) product. The MWE provides estimates for average hourly wages for domains of interest defined by geography, occupation, work level, bargaining unit status (union/non-union), full-time/part-time work status, and incentive-based/time-based pay. The current methodology for MWE creates estimates based off of observed data, focuses only on the wage component of total compensation, and does not include employer costs for employee benefits. Since the domains of interest can be rare, there are often domains with a sample size of zero, thus affecting the ability to produce reliable design-based estimates. Small domain estimation methods may be a solution to this problem as they aim to improve on estimates that rely only on observed data within a small domain. By borrowing information across all domains and by using auxiliary data, such as the OES, total compensation estimates, with potentially greater utility than the current MWE, can be produced by incorporating both wages and employer costs for all domains. The purpose of this paper is to present an overview of the new methodology, preliminary results, and a discussion of lessons learned and future research.

Adopting prior distributions for the variance-covariance matrices to fully specify an area-level bivariate hierarchical Bayes three-fold model for proportions of adult competency
Andreea Erciulescu (Westat)

The Program for the International Assessment of Adult Competencies (PIAAC) is a multicycle survey of adult skills and competencies sponsored by the Organization for Economic Cooperation and Development (OECD). The first cycle of PIAAC included three rounds: 24 countries participated in 2011–12 (round 1); 9 additional countries participated in 2014–15 (round 2); and 5 additional countries participated in 2017–18 (round 3). The United States (U.S.) participated in all three rounds of the first cycle of PIAAC. While the combined PIAAC 2012/2014 sample is nationally representative of the U.S. adult population 16-74 years old, it can only be used to produce county-level estimates when combined with the PIAAC 2017. Even after combining the samples from the three rounds (2012/2014/2017), survey data alone is not sufficient to produce reliable county-level estimates for all the counties in the U.S., the county-level sample sizes being as small as zero. On the other hand, auxiliary information with good predictive power for proficiency levels is available from multiple sources. Therefore, Westat has investigated small area estimation model-based techniques, as useful approaches to improve the precision of the estimates for counties with survey data, and to produce estimates for counties with no survey data. This paper focuses on internal model validation techniques that lead to the development of a hierarchical Bayes model with priors adopted for the components of the variance-covariance matrices decomposed into a vector of standard deviations and a correlation matrix: half-Cauchy distribution adopted as prior for the standard deviations and LKJ (named for developers Lewandowski, Kurowicka, and Joe) distribution adopted as prior for the correlation matrix (LKJ- type priors), for the 2012/2014/2017 U.S. PIAAC data. A model is developed using sparse county- level survey data, by accounting for the nested structure of counties within census divisions within states. Direct survey estimates for two competency proportions (Level 1 and below and Level 3 and above) are jointly modeled, using survey information across the nation and auxiliary information from multiple data sources. The performance of LKJ-type priors is compared to the performance of a traditional prior, the inverse-Wishart distribution.
A Bayesian and Spatial modeling approach for multiple correlated health outcomes: Application to teen births data by race and Hispanic origin groups
Diba Khan (National Center for Health Statistics)

Hierarchical Bayesian modeling techniques can be applied to a large number of rare causes of health outcomes to enable examination of spatial variations on smaller geographic scales such as counties. When multiple correlated health outcomes are analyzed, the relation between the outcomes needs to be considered. Hierarchical Bayesian Multivariate Conditional Autoregressive (MVCAR) models have been used in disease mapping to analyze multiple correlated health outcomes. In this study, the MVCAR model is used to account for spatial structure by smoothing a county's predicted teen birth rate (TBR) towards its neighboring counties TBR (also known as the shrinkage effect) while simultaneously modeling the correlations between the three race and Hispanic origin groups via the spatially structured multivariate random effects. A major methodological challenge of this research stems from the small number of teen births and small number of teens at the county level. Due to concerns about statistical reliability, county-level natality rates based on fewer than 20 counts are suppressed, precluding an examination of spatial variation in less common natality events at the county level using direct estimates based on actual natality data from the birth certificates. Such suppression would prevent any meaningful examination of county-level disparities in TBRs by race and Hispanic origin groups. We investigated county level teen birth rates for the year 2018 for the three largest race groups (Non-Hispanic White, Non-Hispanic Black and Hispanics) as one particular application of the MVCAR modeling technique to predict more reliable county-level estimates by borrowing strength across counties and race and Hispanic origin groups to explore geographic variation. Evidence of elevated risk of teen birth rates across counties are demonstrated via the exceedance probabilities.

CONCURRENT SESSION H-4
Assessing Privacy Risk: Reconstruction and Re-Identification

The Evidence Act requires statistical agencies to conduct a comprehensive disclosure risk assessment prior to the public release of any data product. While quantifying the disclosure risk of data products can be challenging, reconstruction and re-identification experiments can be useful tools for agencies to use. This session will discuss how and when to conduct reconstruction and re-identification studies, their benefits and limitations, and how agencies can effectively incorporate them into their broader disclosure avoidance programs.

CONCURRENT SESSION H-5
Communication of Uncertainty in Official Statistics

Evaluating Uncertainty in Multiple Dimensions of Data Quality
John L. Eltinge, United States Census Bureau

Evaluation of the quality of official statistics generally requires balanced consideration of multiple dimensions, e.g., accuracy, relevance, timeliness, granularity, comparability, interpretability and accessibility. In-depth study of the “accuracy” dimension for sample surveys has led to a substantial literature on “Total Survey Error” (TSE) models. More recently, several authors have explored extensions of TSE models to cases involving the integration of data from multiple sources, e.g., administrative records and commercial transactions, as well as sample surveys. This paper moves beyond the “accuracy” dimension to consider further extensions of the TSE approach that account for the other abovementioned dimensions of data quality in the integration of multiple sources. Strengths and limitations of both qualitative (descriptive) and quantitative (model oriented) approaches are considered. These ideas are illustrated with examples that involve appending
administrative record data to sample survey units (“relevance” and “comparability” dimensions); extensions of multiple-frame, multiple-mode methods to the integration of administrative or commercial records (“timeliness,” “granularity” and “interpretability” dimensions); and data subject to disclosure-protection methods (“accessibility” dimension).

More Fully Capturing Uncertainty Associated with Official Estimates
Linda J. Young, NASS

The Total Survey Error approach has made practitioners aware of sources survey error, such as sampling variability, interviewer effects, frame errors, response bias, and non-response bias. However, most measures of uncertainty published with estimates, including official estimates, fail to capture all of these sources of error, which generally results in overly optimistic reports of uncertainty. For the 2012 Census of Agriculture, a capture-recapture methodology was adopted to produce official estimates at all levels of geography. The estimates were based on four logistic models. In addition, the modeled results were calibrated to known population values, and the final results were rounded to integers. The uncertainty associated with fitting each model and the modeling error were captured, but the additional uncertainty from calibration and rounding was not. In 2017, the impact of calibration and rounding was incorporated in all measures of uncertainty. In this paper, the effect of accounting for these additional sources of variability is presented. Additional sources of uncertainty that have yet to be incorporated in the published results are reviewed, and the potential for including these in the future is discussed.

Tailored Transparency: Public Trust vs. Reproducibility
Peter V. Miller, Northwestern University; United States Census Bureau, Retired

The AAPOR Code states that, “Good professional practice imposes the obligation on all public opinion and survey researchers to disclose sufficient information about how the research was conducted to allow for independent review and verification of research claims.” Enabling expert review is a key purpose of the Code’s provision, but disclosure also is intended to facilitate better general understanding of surveys among less sophisticated audiences, and to foster trust in the research enterprise. The different purposes and audiences for methodological information suggest a tailored approach to transparency. Such an approach becomes increasingly useful as survey methods evolve and incorporate other data sources, increasing the scope and complexity of disclosure practice. The very different transparency demands of the general public and a scientific community focused on information needed to reproduce survey estimates also impel a tailored approach to disclosure. This paper outlines a procedure for tailored disclosure of survey information that takes into account different purposes and audiences for the documentation. The discussion considers surveys conducted to produce official statistics. The outline includes examples of customized disclosure and a possible process for achieving it. The suggested process includes fostering communities of data producers and consumers to define tailored methodological content. The utility of the total survey error model and other templates for guiding tailored disclosure practices will be considered.
Improving the Quality of Data and Reducing Burden of the Public Libraries Survey
Lisa Frehil, (Institute of Museum and Library Services)

The Public Libraries Survey (PLS) is an annual voluntary data collection initiated in 1988 in collaboration with state libraries. State Data Coordinators (SDCs) in each U.S. jurisdiction administer an annual data collection and serve as a point of contact with the Institute of Museum and Library Services (IMLS). Since its inception, the survey has had a 97 percent or better response rate among the nation’s more than 9,000 public library systems that operate more than 17,000 branch locations. This paper reviews the qualitative and quantitative methods IMLS used in a three-year PLS data quality research project. There were several purposes: (1) examine and determine ways to improve data quality; (2) explore methods to reduce respondent burden; and (3) better understand stakeholder data needs given changes in library services in the past decade. Our paper reviews how IMLS worked with key stakeholders, including the SDCs and the Chief Officers of state libraries. Finally, we conclude with potential challenges of balancing the needs of different stakeholders while maintaining the historically high response rate and long-term trend data in the PLS.

Interviewer Training for Classroom versus Distance Learning: Initial Skill Gains and Measures of Drift
Hanyu Sun (Westat), Angie Kistler (Westat), Ryan Hubbard (Westat)

There is abundant literature about interviewer effects on the survey process but studies of interviewer training are quite limited (e.g. Groves & McGonagle 2001). We conducted two training experiments with a group of 262 experienced field interviewers working on the Medical Expenditure Panel Survey’s Household Component (MEPS-HC) to assess training effectiveness. (1) The interviewers were stratified by performance on gaining cooperation and data quality. We selected 40% of the interviewers at random to attend a 2.5 day “refresher” training in a classroom setting. Peer learning was an important feature of the curriculum. We compare the performance of each interviewer before and after training, and compare the treatment group’s performance with the remaining interviewers who serve as a control group. (2) We selected two classroom modules and developed two distance learning: one for delivery in a virtual classroom setting, and the other for delivery in PowerPoint slides. Interviewers who were not invited to the classroom training were assigned to one of the distance learning modes. We compare performance post-training on specific skills to explore the effectiveness of distance learning.
The Impact of the Pregnancy Checkbox and Misclassification on Maternal Mortality Trends in the US, 1999-2017
Lauren Rossen (National Center for Health Statistics), Lindsay Womack (Centers for Disease Control and Prevention), Sayeedha Uddin (National Center for Health Statistics)

States incrementally adopted a pregnancy checkbox on death certificates between 2003-2017, to improve ascertainment of maternal mortality. Inconsistent measurement by states over time has made it difficult to estimate national maternal mortality rates (MMRs) and trends.

We used log binomial regression models to examine the impact of the pregnancy checkbox and predict MMRs from 1999-2017 under two alternative scenarios: 1) assuming all states had the checkbox from 1999-2017; and 2) assuming no states had the checkbox. We estimated the impact of the checkbox and related trends over time by age, race/ethnicity, state of occurrence, and cause of maternal death. Sensitivity analyses examined the impact of outcome misclassification. The implementation of the checkbox increased identification of maternal deaths by an average of 9.6 deaths per 100,000 live births (95% CI: 8.6–10.6). The average impact of the checkbox was greater for women over 40 years of age, non-Hispanic black women, and for certain causes of death. Estimated trends suggested that MMRs did not change appreciably from 1999-2017, though trends varied by subgroup and cause of death.

A Secondary Analysis of Interviewer Effects in the BRFSS
Antonia Warren (Westat), Ting Yan (Westat), Carol Pierannunzi (Centers for Disease Control and Prevention)

Interviewers can be a potential source of measurement error and non-response error in surveys which makes it important for researchers to examine these effects and look for ways to mitigate them. As such, this study is an extension of previous work that used CATI data from three states that participated in the 2017 BRFSS to examine interviewer effects on the variance components of survey estimates by interviewer productivity level. Results from this first set of analyses found interviewer effects across the selected items to be relatively small but there were relevant differences in the size of effects by interviewer productivity level. Given this, in the current study we examine if interviewer effects persist when controlling for respondent characteristics such as gender, age, or race. In addition, we compare the results from multi-level models using the 2017 data to 2018. Furthermore, using the 2018 data we plan to examine interviewer effects with call history level data (e.g. interviewer level response rates, number of contacts, refusals). We are interested in assessing if there is a correlation between interviewer effects on recruitment and survey estimates.

Towards Developing a Quantitative Measure for Response Burden for Coordinated Sampling of Annual Business Surveys
Laura Bechtel (U.S. Census Bureau), Diane Willimack (U.S. Census Bureau)

Traditionally, sample selection for business surveys at the U.S. Census Bureau is conducted independently. This has advantages in terms of optimal sample design, but can have response burden implications. Business populations are highly skewed, with a small number of very large businesses. Therefore, many companies are included in more than one survey. Response burden is a function of several factors, such as questionnaire length and complexity, accessibility of data, and frequency of collection. However, the potential of being included in multiple surveys compounds the response burden.
Sampling Strategies for the QCEW Business Supplement
Sharon S. Stang (Bureau of Labor Statistics), Emily deWolf (Bureau of Labor Statistics)

Every year, the Quarterly Census of Employment and Wages (QCEW) program conducts the Annual Refiling Survey (ARS) asking approximately 1.2 million business establishments to verify their main business activity, and their mailing and physical location addresses. This survey is collected fully on-line through the BLS ARSWeb system. For the 2020 survey, QCEW is piloting two sample designs for collecting a special-topic supplement to the ARS. ARSWeb is a low-cost platform for conducting Quick Business Surveys (QBS). These QBSs accompanying the ARS have almost no data collection overhead i.e., address refinement, printing, and mailing survey forms or letters. Respondents that are already logged into the ARSWeb secure website are asked to answer additional survey questions after completing the ARS. Utilizing the ARS in this manner provides BLS with the platform and the opportunity to field quick response surveys covering topics of interest to our data users. This presentation will outline the two sample designs tested in the 2020 QBS pilot, review the survey's results, and discuss how both sampling methods can be used to collect additional information from respondents.

The Record-Keeping Practices of Medium-sized Multi-unit Businesses and Organizations
Diane Willimack (U.S. Census Bureau), Struther Van Horn (U.S. Census Bureau), Erica Marquette (U.S. Census Bureau)

The U.S. Census Bureau is undertaking research to determine feasibility and implementation strategies for recommendations from the National Academy of Sciences (NAS) panel examining the Bureau’s annual appropriated economic survey programs (NAS, 2018). To address recommendations to manage respondent burden by harmonizing questionnaire content and streamlining data collection processes, we conducted a study of record-keeping practices and response behaviors of medium-sized companies with ten or more locations that operate in at least two industries, making up roughly half of multi-unit companies with employees. Using in-depth interviews with survey respondents, we investigated data definitions relative to companies’ accounting systems, organizational levels and timing of available data, and level of effort needed to report to Census Bureau surveys. In this paper, we associate accounting structures with companies’ organizational structures, management decision-making, and external financial reporting requirements. We describe how companies map their records to our current survey requests, and propose alternative data reporting strategies more aligned with record-keeping practices.

Understanding the Characteristics of Unresolved Matched Records in Capture-Recapture Methodology
Denise A. Abreu (National Agricultural Statistics Service)

The National Agricultural Statistics Service (NASS) conducts a Census of Agriculture (COA) every 5 years, in years ending in 2 and 7. The census uses a list frame. The 2017 COA used capture-recapture to adjust the COA for under coverage, nonresponse, and misclassification of farms/non-farms. NASS’s June Area Survey (JAS) was used as the independent survey in the capture-recapture approach. The JAS is conducted annually in June. It is based on an area frame and the data are collected via in-person interviews. Capture-recapture requires a matched dataset consisting of all matches of a COA record to a JAS record. This dataset is the foundation for modeling the probability that a JAS farm is captured by the COA. A farm is a place with $1000 or more of sales or potential sales. In the dataset, the farm status based on the JAS and the COA agree in most cases. However, in other cases, a record is identified as a farm (non-farm) on the JAS and a non-farm (farm) on the Census. These records have unresolved farm status. Resolving the farm status is important to the accuracy of the COA published estimates. The characteristics of the records with unresolved farm status are described.
CONCURRENT SESSION J-3
Better Than a “Guess-Estimate”: Methods for Estimation and Analysis

Constructing Better Coverage Intervals for Some Estimators Computed from a Complex Probability Sample
Phillip S. Kott (RTI International)

Coverage intervals for a parameter estimate computed using a complex probability sample are often constructed by assuming that the parameter estimate has an asymptotically normal distribution and the measure of the estimator’s variance is roughly chi-squared. The size of the sample and the nature of the parameter being estimated render this conventional “Wald” methodology dubious in many applications. A revised method of coverage-interval construction has been developed in the literature that “speeds up the asymptotics” by incorporating an estimated measure of skewness. This presentation will discuss how skewness-adjusted intervals can be computed for ratios, differences between domain means, and regression coefficients.

Using Relative Distribution Methods to Analyze Federal Justice Data
Mark Motivans (Bureau of Justice Statistics)

Using federal justice administrative data collected by the Bureau of Justice Statistics, this paper applies relative distribution methods as a tool for understanding gender differences in the amount of time served in prison. Traditional methods used to describe data (e.g., mean and median) leave much information about the distribution untapped, particularly in the upper and lower tails. Relative distribution methods follow a nonparametric statistical framework for analyzing and visualizing data in a fully distributional context (Handcock and Morris, 1999). Questions include: To what extent are relative distribution methods useful (above and beyond traditional methods) for understanding the effect that policy changes have had on overall and gender differences in the distribution of time served? Have the upper and lower tails of time served changed at similar rates for men and women? Responses to these questions will offer some insight into the utility of relative distribution methods as an additional tool for exploring and describing administrative data.

An easy way to calibrate on partly known multiple totals in frequency tables with application to real data
Michael Sverchkov (Bureau of Labor Statistics)

Deville and Sarndall (1992, Section 4) considered calibration on the known counts (cell counts or marginal counts) of a frequency table in any number of dimensions (generalized raking procedure). In this paper, we show that a similar procedure can be applied to the case of partly known overlapping counts. As an example we consider calibration of area-month-year unemployment estimates to month-year totals from a time series model of State estimates from the Current Population Survey and area-year totals from the American Community Survey.

Key words: calibration, raking, generalized raking procedure

The Relationship Between the Seasonal Regression Model-based F Test and a Diagnosis of Residual Seasonality
Demetra Lytras (U.S. Census Bureau), Kathleen McDonald-Johnson (U.S. Census Bureau)

The problem of an identifiable seasonal pattern that remains after seasonal adjustment—that is, residual seasonality—has long been a concern at the U.S. Census Bureau. In recent years, with scrutiny of a perceived residual pattern in the gross domestic product, attention has focused on the best ways to accurately identify this phenomenon so that analysts can minimize or eliminate it via the seasonal adjustment procedure. One of the most promising diagnostics according to recent research is the model-based F test from fitting seasonal regressors to the seasonally adjusted time series. In practice, however, results from the same series could contradict each other, depending on
the inclusion of other regressors and the choice of autoregressive integrated moving average (ARIMA) model. To investigate the appropriate approach for choosing the model to accompany the seasonal regressors and to understand the behavior of the diagnostic, we fit the seasonal regressors to simulated monthly time series and their seasonally adjusted counterparts with X-13ARIMA-SEATS using a variety of ARIMA models and the automatic modeling procedure. This paper presents the test results.

**A Decomposition Analysis of US Food Expenditure**
Eliana Zeballos (Economic Research Service), Timothy Park (Economic Research Service)

U.S. consumers, businesses, and government entities spent $1.71 trillion on food and beverages in 2018. Spending at food-away-from-home (FAFH) establishments—restaurants, school cafeterias, and other eating places—accounted for 54.4 percent of these expenditures, and the remaining 45.6 percent took place at grocery stores, convenience stores, and other retailers (FAH). The purpose of this paper is to attribute changes in food expenditure to various causes from 1997 to 2018 and assess the relative effects in periods before and after the Great Recession. In particular, we look at how the GDP, the share of retail spending, the share of retail and food spending, and the share of FAH has on the changes of FAFH. We use the food expenditure series (FES) created by the Economic Research Service of USDA, which measures the value of the U.S. food system over time, by outlet and product type, and by final purchasers and users. Preliminary results suggest that nominal FAFH grew at an annual rate of 5 percent from 1997-2018. Out of the four components that contributed to the growth of FAFH, the share of retail spending is the one that contributed the most.

**CONCURRENT SESSION J-4**
Disclosure Review Boards: Design, Governance, Modernization

The proliferation of third-party data sources and advances in computing power make protecting privacy in public data releases increasingly difficult. Now, more than ever, it is critical that agencies have effective, agency-wide mechanisms for assessing disclosure risk and approving the statistical disclosure limitation methods used to protect privacy and confidentiality in public data releases. This panel will discuss the importance of centralized Disclosure Review Boards (DRBs), examine various structures for ensuring effective and efficient governance of the disclosure review process, and explore options for keeping DRBs nimble in the face of a rapidly changing privacy landscape.

**CONCURRENT SESSION J-5**
Nonresponse Bias on Federal Surveys – Gaps in Knowledge and Future Opportunities

Peter Miller, Northwestern University and U.S. Census Bureau, Retired.

This presentation summarizes the first systematic review of nonresponse bias (NRB) studies involving Federal Surveys since the release of the 2006 OMB Standards and Guidelines for Statistical Surveys. NRB reports were identified through searches on PubMed, Google Scholar, Current Index to Statistics, Joint Statistical Meeting proceedings and through an open call to Federal statistical agencies and associated professional organizations. Some 165 studies were identified - 89 concerning establishment surveys and 76 involving household surveys. About 40 percent of the NRB studies were done during the period shortly after the 2006 OMB guidance. The methods employed for assessing NRB differed for establishment and household surveys. Studies involving establishment surveys mostly compared survey estimates to external (frame) data, while those involving household surveys mostly examined variations of estimates within the response set (e.g. early and late responders). A majority of studies reported some NRB in estimates prior to weighting and some reduction in bias after adjustment. The efficacy of weighting was often not explicitly documented in the reports. This systematic review is a first step in continuing research on NRB in Federal surveys.
Developing and Assessing Weighting Methods for the Redesigned National Health Interview Survey
Ronaldo Iachan ICF, National Center for Health Statistics.

In 2019, the National Health Interview Survey (NHIS) released its first redesigned instrument since 1997. In this paper, we present the results of a collaboration between the National Center for Health Statistics (NCHS) and ICF to evaluate weighting methods for use with the redesigned NHIS. The evaluation focused on the use of machine learning and multilevel logistic regression to assess nonresponse (NR) bias in key NHIS health estimates and develop NR bias adjustments for household, adult, and child sample weights.

We start by reviewing the data sources which provided potential predictors at different levels. The nonresponse models incorporated predictors from the NHIS Contact History Instrument and Neighborhood Observation Instrument, and auxiliary data from the Area Health Resource File and the Census Planning Database. The analysis employed machine learning methods such as lasso, random forest, and decision trees, as well as more traditional single-level and multilevel logistic regression, to find best-fitting models to use in NHIS nonresponse bias adjustments for household, adult, and child weights. We define key health indicators used in the nonresponse bias analysis at these different levels, and discuss data sources, decision processes, methodology and results focused on bias reduction. We also present results from capping the NR adjustment factors to limit variance inflation, as well as overlaying raking on top of NR adjustments to extend the simple demographic post-stratification currently used in the NHIS.

Finding the Right Auxiliary Information for Nonresponse Adjustment Models: In Search of Zs with Desirable Properties
Andy Peytchev, RTI International

Declining response rates increase the dependency of survey estimates on postsurvey adjustments. The identification of auxiliary information is becoming increasingly important. This presentation starts with a discussion of flaws in common current practice with regard to nonresponse adjustment models. It is followed by an overview of desirable properties of auxiliary information, and related challenges. In the third part, promising avenues for improvement are introduced, along with several illustrative examples from the research literature.

Estimating Survey Nonresponse Bias Using Tax Records
Bruce Meyer, National Bureau of Economic Research, American Enterprise Institute, and U.S. Census Bureau

Declining survey response rates is a widespread and troubling problem that raises the possibility of bias in key statistics. We propose and implement a new method to determine nonresponse bias by linking income tax records to respondents and nonrespondents by address. In light of the importance of income in assessing poverty, inequality, and material well-being, we focus on income but also examine bias along other dimensions measured on tax returns such as marital status and family size. To provide a framework, we first describe a theory of testing for differences between populations when linkage to validation data is incomplete. We then apply the methods to the Current Population Survey (CPS), the most used economic survey and the source of official employment, income, poverty, and inequality statistics. We link the CPS to IRS Form 1040 records, comparing several characteristics of respondents and nonrespondents, including income, its components, self-employment status, marital status, number of children, and the receipt of social security. We find little evidence of differences between the percentiles of the income distributions of the linked respondents and nonrespondents. We also find little difference between the income distributions of ASEC respondents.
and CPS Basic respondents who decline to participate in the ASEC (whole imputes). However, we find significant differences between respondent and nonrespondents in marital status, the number of children, and other characteristics.

**CONCURRENT SESSION K-1**

Linked Data from the Census Bureau for Evidence Building: Accessing the Data and Recent Results

**Criminal Justice in the US and Economic Inequality: Results from the Criminal Justice Administrative Records System**
Keith Finlay (U.S. Census Bureau)

CJARS is a joint Census Bureau-University of Michigan project started in 2016 to create a national, integrated, harmonized collection of criminal justice microdata at the Census Bureau. The project has three fundamental goals: (1) improve Census Bureau operations, (2) provide valuable aggregate statistical information to criminal justice agencies, and (3) increase the quality and quantity of criminal justice research by making the data available through the Federal Statistical Research Data Centers. The project highlights the opportunities provided by the Census Bureau’s Data Linkage Infrastructure. This paper provides new evidence on how felony conviction and imprisonment rates have changed for 30+ birth year cohorts over 185 distinct commuting zones in the U.S. using a novel piece of data infrastructure we have created called the Criminal Justice Administrative Records System (CJARS). We document striking variation in cumulative exposure to the justice system over geography, between birth cohorts, and across demographic groups, and leverage this newly documented variation to assess how changing risk of contact with the justice system correlates with economic outcomes in the U.S.

**UMETRICS: Data For Examining How Research is Produced and How it Affects the Broader Economy**
Joseph Staudt (U.S. Census Bureau)

The IMI UMETRICS data include information on awards, wage payments from awards to university research employees, vendor purchases, subcontracts, and the unit performing the funded research for 26 universities. These data can be linked to internal Census Bureau data products, such as the Decennial Census, American Communities Survey, Longitudinal Employee-Employer Household Dynamics database (LEHD), and the integrated Longitudinal Business Database, providing researchers with a comprehensive view on the businesses associated with the production of scientific research. This paper provides information on the data available, how researchers can access the data, and results from work in progress by researchers.

**Results from the Evidence Building Project Series: Health at Birth, Later Life Achievement, and the Intergenerational Transmission of Advantage**
Sarah Miller (University of Michigan)

This paper provides evidence on the long-run and intergenerational impacts of initial health endowments. We link detailed birth certificate records to federally-held survey and administrative data on earnings, educational attainment, and public assistance for all individuals born in California between 1960 and 2014, allowing us to observe measures of health at birth and long-run economic outcomes for over 25 million individuals. For a large subset of these individuals, we are also able to observe outcomes for their children, allowing us to trace the transmissions of health and advantage across generations. Our analysis is the first to document these effects in the United States using data of this size and scope. We use this data to analyze how health at birth within twin pairs, and within siblings, affects long-run and intergenerational health and achievement. We find that individuals with higher birth weights are better off in adulthood along a number of dimensions, and some evidence that this advantage transfers to the next generation in the form of higher birthweights and better economic and health outcomes.
The Census Longitudinal Infrastructure Project – Linked Census Data and Results from the Impact of Preschool on Later-Life Outcomes
Katie Genadek (U.S. Census Bureau)

The Census Longitudinal Infrastructure Project (CLIP) was created to support research using the linked data at the Census Bureau, including linked mandatory-response census and survey data, and to further develop the linked data infrastructure with expansion to historical data. There are currently more than 12 projects using the linked data at the Census Bureau through the FSRDC network. This paper will describe the linked data available and explain how researchers can access this data. Recent research using this data to analyze the effect of the Lanham Act preschools in the 1940s on later life outcomes will also be discussed.

CONCURRENT SESSION K-2

Integrating Information from Multiple Data Sources to Support Policies to Reduce Rural Health Disparities

Healthy People: Exploring Urban/Rural Disparities in the Nation’s Health
David Huang National Center for Health Statistics), Johanna Alfier (National Center for Health Statistics)

Healthy People is a federally-led national public health initiative which has provided overarching goals and objectives for improving the health of all Americans each decade since 1979. The past three iterations of Healthy People (2000, 2010, and 2020) have included health disparities and equity goals, the nature of which has changed over time. Specifically, Healthy People 2020 (HP2020), which is drawing to a close, includes an overarching goal to achieve health equity, eliminate disparities, and improve the health of all groups. HP2020 provides population-level national data for over 600 of the 1,100 measurable objectives tracked by the initiative, including breakouts by urban/rural status, where data are available. This presentation will provide a summary of Healthy People 2020 findings related to urban/rural health disparities including areas of success and areas for improvement. Lastly, we will describe implications for urban/rural health disparities for the latest iteration of national health objectives, Healthy People 2030.

Integrating Information from Multiple Data Sources to Support Policies to Reduce Rural Health Disparities in AHRQ Quality and Disparities Report
Barbara Barton (Agency for Healthcare Research and Policy), Karen Chaves (Agency for Healthcare Research and Policy). The annual National Healthcare Quality and Disparities Report (QDR) provides a comprehensive overview of the quality of healthcare received by the general U.S. population and disparities in care experienced by different groups including rural populations. More than 250 measures used in these reports span a wide range of structure, process, and outcome measures. Quality measures are grouped by priority areas, including person-centered care, patient safety, healthy living, effective treatment, care coordination, and affordable care.

Employing Geographic Concepts and Methods to Maximize the Effectiveness of Rural Health Programs
Mark Guagliardo (Department of Veterans Affairs Office of Enterprise Integration)

In FY 2016 eight federal departments spent $31 billion to improve the lives of rural Americans through 55 programs dedicated solely to rural communities. While some of these resources are allocated at the state or community level, many such as healthcare programs entail determining individual eligibility based on geographic location, and usually require delivery at precise service locations. This presentation considers important but often overlooked assumptions, concepts and hazards in making eligibility determinations and planning rural service locations. Topics covered will include how using various definitions of rurality can impact accessibility, why it is important to consider the decay of resource utility with increasing distance, and when basing a program on the notion of rurality may not be the best approach to planning service delivery. We
will demonstrate how geographic information system (GIS) methods can both expose weaknesses in distribution programs and guide decision makers to solutions. Examples will employ the USVETS data set, a unique product assembled by the National Center for Veterans Analysis and Statistics from 31 federal and commercial data sources.

**Integrating Information from Multiple Data Sources by the VHA Office of Health Equity to Support Policies to Reduce Rural Health Disparities Among Veterans**
Ernest Moy (Veterans Health Administration Office of Health Equity), Kenneth Jones (Veterans Health Administration Office of Health Equity)

The Veteran Health Administration (VHA) Office of Health Equity (OHE) is charged with reducing disparities in health and health care affecting Veterans and enabling all Veterans to achieve equitable health outcomes. It does this by integrating information across multiple datasets to develop tools that enable policy makers to understand and intervene to reduce health disparities among Veterans. Rural Veterans experience a variety of barriers to high quality health care related to culture, social determinants, and geography. This presentation will describe OHE work to examine and link multiple datasets to improve understanding of the health and care of rural Veterans. In the National Veteran Health Equity Report, Department of Defense, Medicare, and VHA Inpatient and Outpatient data are linked to examine health and care of rural Veterans treated in VA facilities. OHE also analyzes Survey of Veteran Enrollees’ Health and Use of Health Care data to understand use of non-VA facilities by rural Veterans enrolled in VHA and American Community Survey and Area Health Resources File data to understand barriers to care experienced by rural Veterans, including those not enrolled in VHA.

**CONCURRENT SESSION K-3**
Innovative Uses of Imputation

**Applying Cluster Analysis to Improve the American Housing Survey Hot Deck**
Brian Shaffer (U.S. Census Bureau), Stephen Ash (U.S. Census Bureau), Kathy Zha (U.S. Census Bureau)

The American Housing Survey currently uses hot deck methods to impute missing values for approximately 120 variables. Hot deck methodology involves imputing values for nonrespondents with values of respondents. The imputation process is completed within disjoint subsets of the universe, which we refer to as donor pools. We define the donor pools with auxiliary variables that are available for both the respondents and nonrespondents. In our paper, we introduce new auxiliary variables and apply cluster analysis to produce improved donor pools that minimize within-pool variation across all variables that use each set of donor pools. We also generate donor pools for imputing a single variable. We describe the clustering methods used to define the donor pools; the methods include classification and regression trees (CART), hierarchical agglomerative clustering, and k-means clustering. We compare the donor pools by measuring the within-pool variation of the imputed variables using the current donor pools and the alternative donor pools. We also will compare our results with imputed values generated with multivariate multiple imputation methods.

**Model Selection within Sequential Imputation: A New Treatment for Missing Values in High-Dimensional (Survey) Data**
Micha Fischer (University of Michigan)

Multiple sequential imputation is often used to impute missing values in data sets. The procedure leads to unbiased results if the missing data is missing at random and models are properly specified. However, in data sets where many variables are affected by missing values, appropriate specifications of sequential regression models can be burdensome and time consuming. Even available software packages for automated imputation procedures require model specifications for
each variable containing missing values. Additionally, their default models can lead to bias in imputed values. This research aims to automate sequential imputation of incomplete variables of any distribution in high-dimensional data sets with potentially complex and non-linear interactions. The proposed algorithm performs model selection from a pool of supervised learning techniques in each step of the sequential imputation procedure. The criterion is based on prediction accuracy and the similarity between imputed and observed values. A simulation study based on a real data set investigates in which situations this automated procedure can outperform current software (MICE, IVWare).

Survey Data - Editing and Imputation (E&I) Methods
Menuka Ban (EY)

Editing and Imputation (E&I) is the process of detecting erroneous and missing data and treating these data accordingly. E&I provides a set of activities for gathering knowledge and information on errors and error sources, for improving the overall survey process, and for providing feedback to the other survey phases, to reduce the amount of errors arising at the earlier stages of the data collection and processing. The presentation/paper will provide highlights on the recent advancement in statistical methodology for the detection of errors and treatment of errors. The research presentation will focus on the four main principles of the E&I process. 1) The quality of the data must be assessed at the beginning of the E&I process to understand if the E&I is required. 2) The E&I method has to be designed and implemented such that it does not introduce additional error in the data; 3) The quality of the data must be assessed at the end of the E&I process to make sure meets the users' expectation; and 4) The E&I process should be simple, cheap, easy, and swift to implement.

Using machine-learning algorithms to improve imputation in the Medical Expenditure Panel Survey
Chandler McClellan (Agency for Healthcare Research and Quality), Emily Mitchell (Agency for Healthcare Research and Quality), Jerrod Anderson (Agency for Healthcare Research and Quality)

Survey data collection often results in incomplete or missing responses. Imputation methods are widely used to complete survey data, and improvements to those methods are an active area of research. We study the feasibility and performance of various imputation methods applied to the Medical Expenditure Panel Survey (MEPS). The MEPS is an annual survey that collects nationally representative data on healthcare use and expenditures for the civilian, non-institutionalized U.S. population and is one of the United States' primary sources for data on medical expenditures. The MEPS contacts both households and their medical providers to gather as much accurate expenditure information as possible. However, a significant portion of this expenditure data must be imputed. Currently, MEPS expenditure data are imputed with a predictive mean matching (PMM) algorithm in which a linear regression model is used to predict total expenditures for recipients and donors. Recipients and donors are then matched based on the smallest distance between predicted total expenditures. We assess whether more sophisticated machine-learning (ML) algorithms can replace the predictive mean matching framework to predict total expenditures. After training both the PMM and ML algorithms, we apply both methods to a test set of MEPS data constructed to resemble real-world missingness patterns. We then compare the machine-learning algorithms and baseline PMM method to determine which approach generates the most accurate imputations. Better imputations have implications beyond just improving the MEPS expenditure data. A demonstrable improvement in predictive accuracy may be of value in other national surveys that currently rely on methods similar to PMM for imputation.
An Imputation Solution for Differentiating between Unreported Attitudes and Genuine Nonattitudes in Survey Data
Jeff Gill (American University), Natalie Jackson (PRRI)

Most survey analyses treat “don’t know” or nonattitude responses as missing values and drop them from analysis with case wise (list wise) deletion. There are problems with this approach. In this work we demonstrate first that nonattitudes and “don’t know” responses are not random, but rather come from a distinct group of survey respondents. This is shown by modeling relevant missingness as a dichotomous outcome variable explained by various characteristics, including demographic attributes, other attitudinal questions, and group level contexts. This model allows us to produce an imputational model to predict missingness due to ignorance versus intransigence.

CONCURRENT SESSION K-4
Disclosure in an Era of Administrative Records and Data Sharing

Effective assessment and management of disclosure risk requires taking a holistic view of all releases of related data. A major objective of the Evidence Act is to increase the availability of administrative and program data for use by statistical agencies across the government. Increased data sharing between agencies in support of this objective poses new challenges for coordinating the assessment and mitigation of disclosure risk by different agencies releasing data products derived from the same administrative data sources. This session will discuss these challenges, and will explore how some agencies have addressed these issues.

CONCURRENT SESSION K-5
Geospatial Act and You

The Federal statistical community is aware of the Foundations for Evidence-Based Policy Act of 2018, but many might not know about another data act that was recently passed – the Geospatial Data Act (GDA). The GDA is a component of the FAA Reauthorization Act (https://www.fgdc.gov/gda), and it was signed into law in the fall of 2018. This session will start with an overview of the main sections, items, and responsibilities outlined in the GDA given by a member from the Federal Geographic Data Committee (FGDC), which is the primary entity overseeing GDA activity. Remaining panel members come from various agencies impacted by the GDA. Each will give comments on the GDA and how it affects their agencies and the statistical data they produce. The Census Bureau is the largest statistical agency and producer of legal and statistical geography in the federal statistical system. It is also the agency leading the discussion about the potential for geography to compromise privacy and confidentiality. The Bureau of Transportation Statistics has a large and mature geodata footprint, and it is relatively well positioned to respond to the GDA. The National Center for Education Statistics is a small statistical agency with a small/new geodata footprint that might not be well-resourced to respond to the requirements of the GDA. Each of the panel members will give brief remarks on the GDA, which will be followed by an audience discussion.
Linking State Medicaid and Child Welfare Data for Parental Behavioral Health Services and Child Welfare Outcomes
Emily Madden (Health and Human Services), Valeria Butler (Health and Human Services), Robin Ghertner (Health and Human Services)

The nation’s opioid crisis has precipitated a child welfare crisis, as foster care caseloads saw a large increase over the past 6 years. Recent legislation, including the Family First Prevention Services Act of 2018, has recognized the impact on children of substance use disorders. These changes have transformed how federal dollars can support families and child safety by providing greater access to behavioral health treatment services. Opportunities for new investments are accompanied by increased demands for evidence-based practices, yet developing the evidence requires data infrastructure that do not exist. Currently, no state is able to identify behavioral health assessments and services for parents with children in child welfare systems. This presentation will describe a new HHS project aimed at developing the necessary data capacity for outcomes research and, eventually, rigorous evaluations. The project will produce new longitudinal, client-level datasets linking Medicaid records with child welfare records for parents and children involved in the child welfare system in several states. These datasets will allow tracking Medicaid enrollment, patient diagnoses, services, and claims, along with child welfare characteristics and outcomes. We will discuss current state of child welfare administrative data linkage efforts, lessons learned in overcoming bureaucratic hurdles to permit case-level data linking, insight into developing common data models across systems and states, and approaches to developing support for data linkages among a wide range of stakeholders.

The Project Talent-National Death Index Repository
Ashley Kaiser (American Institutes for Research), Kelly Peters (American Institutes for Research), (Alec Becker (American Institutes for Research)

Disparities in mortality by socioeconomic status and by race/ethnicity appear to be widening in the U.S. More research is needed to understand how early in life factors such as SES, IQ, and education impart risk on later-life mortality, as well as how adult indicators mediate this risk. Project Talent (PT) is a longitudinal study of 5% of U.S. high school students in 1960 (N=377,016). The first wave of data collection (1960) included a comprehensive assessment of adolescent characteristics. Three follow-up surveys (1961-1975) assessed education, occupation, military history, and family formation. We have recently linked PT decedents (N=50,150) to the National Death Index. This paper describes the PT sample and dataset, the method for linking PT individuals to NDI records, and the utility of the linked dataset. Project Talent is unprecedented in combining a large, nationally representative sample with detailed information on adolescent SES, cognitive ability, personality and environment, and mid-life SES, education and occupation. The PT-NDI dataset will allow researchers to investigate early-life predictors, and their relationship with mid-life indicators of mortality disparities.

The Promises and Challenges of Linked Rent Data from the Consumer Expenditure Survey and Housing and Urban Development
John Voorheis (U.S. Census Bureau), Garret Christensen (U.S. Census Bureau), Nikolas Pharris-Ciurej (U.S. Census Bureau)

This paper matches administrative data from Housing and Urban Development (HUD) to CE data to investigate alignment between survey responses to questions about rental assistance receipt and monthly rent paid with administrative records data. We first link all CE sample units to the HUD data, and find that sampled HUD participants are more likely to respond to the CE than are non-HUD participants. We compare the amount of monthly rent reported in the CE compared to HUD administrative records for CE respondents. The distribution of the difference between HUD
reported total tenant payment (the amount less any subsidy) and CE reported rent is skewed toward negative values, suggesting that CE rent may be over-reported for the linked sample. We investigate how this misreporting may affect calculations of the proportion of individuals experiencing rent burden, finding these to be substantially smaller when using linked CE-HUD data than when using CE survey data alone. Additionally, in work in progress, we investigate how estimation of Supplemental Poverty Measure poverty thresholds are affected by replacing survey-reported rent with HUD administrative data on gross rent.

**Linking ACS and IRS Data to Assess College Attendance and Completion by Family Income**
Leah Clark (U.S. Census Bureau), John Voorheis (U.S. Census Bureau), Nikolas Pharris-Ciurej (U.S. Census Bureau)

Family income during childhood is strongly correlated with high school completion, college attendance, and college completion. We link young adults in the American Community Survey to their families’ tax records while in high school, improving on traditional cross-sectional surveys and enabling study of the impact of family income on outcomes at a later life course stage. For cohorts born between 1985 and 1998, young adults from low-income families increasingly completed high school, narrowing the gap between the highest- and lowest-income quintiles by more than two-fifths. For cohorts born in the late 1980s (most of whom entered college prior to the Great Recession), high school graduates from low-income families increasingly attended college, narrowing the income gap in college attendance by nearly one-fifth. Notably, college attendance gains do not coincide with increases in college completion among low-income students. Post-recession, gaps in college enrollment held steady. We find a more modest decline in the income gap in college attendance than that published by NCES (using Current Population Survey data).

**Health and Retirement Study: Weight adjustment for linkage with earnings records from the Social Security Administration**

GAO analyzed the Health and Retirement Study (HRS), a major longitudinal study of aging in the U.S., to assess how mid-career earnings are linked to life expectancy and financial well-being in retirement. About 75% of the HRS Original and War Babies Cohorts were linked with earnings records from the Social Security Administration. To account for bias from non-linkage, we adjusted the public HRS household and respondent weights using models of the inverse consent and linkage probabilities. Linkage varied slightly across demographic, administrative, health, and financial characteristics, but varied substantially by whether a respondent's spouse had linked records. Over 90% of respondents with linked spouses also had their own records linked, compared to just over 70% of respondents without linked spouses. Our adjusted weights changed income and wealth estimates and their standard errors by small amounts, but did not affect observed trends across waves. The analysis demonstrates how non-response adjustments methods may be used to mitigate possible linkage bias in surveys joined to administrative data, and the extent of linkage bias in selected waves and cohorts of the HRS.
Improving Employer Data Collection- The Journey to Modernization of the U.S Equal Opportunity Commission’s Employer Information EEO-1 Survey
Margaret Noonan (U.S. Equal Employment Opportunity Commission)

EEO-1 data are used by the EEOC to investigate charges of discrimination against private employers, and the data serves as standards in the assessment of the utilization of minorities and women in the private industry. The Employer Information EEO-1 report (Standard Form 100) is collected annually under the authority of Section 709(c) of Title VII of the Civil Rights Act of 1964, as amended, and accompanying Equal Employment Opportunity Commission (EEOC) regulations, Title 29, Chapter XIV of the Code of Federal Regulations, §1602.7 through §1602.14. The EEO-1 data collection and reporting approaches employed by the agency have largely remained unchanged for over 2 decades. As a result, there are many opportunities to introduce innovation, efficiency, and improve quality. As such, the EEOC through the Office of Enterprise Data and Analytics, is evaluating current methods of data collection, reporting and access of the EEO-1 Survey. Improvements are being made to improve statistical rigor, reduce burden on respondents, enhance use of survey data in support of the EEOC mission and business needs, and improve public access to the data.

Estimating, Describing and Locating U.S. Workers Vulnerable to Workplace Discrimination
Benjamin Overholt (U.S. Equal Employment Opportunity Commission)

The U.S. Equal Employment Opportunity Commission is responsible for enforcing federal laws that make it illegal to discriminate against job applicants and employees. In carrying out its mission, the Commission conducts outreach to workers that are particularly vulnerable to workplace discrimination and are often unaware of their rights under the equal employment laws. In its Strategic Enforcement Plan for 2017-2021, the Commission identified immigrants and migrants as a key segment of the vulnerable worker population. To protect such workers effectively, district offices must understand the national origins and other demographic characteristics of the immigrant groups that reside in their district, what languages they speak, and where they tend to residentially concentrate within the district. Using data from the American Community Survey, this presentation presents infographics that the Office of Enterprise Data and Analytics is creating to assist EEOC district offices in their outreach efforts to immigrant and migrant workers. We will present and compare national origin, demographic, linguistic, and geographic descriptions of district immigrant populations.

Recent Advances in Data Access for EEO-1 Survey
Justin West (U.S. Equal Employment Opportunity Commission)

The U.S. EEOC’s Office Of Enterprise Data and Analytics produces annual aggregates of private industries’ employment statistics at National, State, and CBSA levels based on EEO-1 surveys for companies that meet the reporting standard. Employment data are reported for each establishment worksites by race, gender, occupation, and industry. Approximately 69,800 employers, representing 54.4 million employees, filed EEO-1 reports in 2017. In this paper, we present our work on creating a public use file for EEO-1 establishment data as well as a mapping and data query tool.
Quantifying the Disclosure Risk for Public Data Products: Theories and Practices
Jianzhu Li (Westat), Tom Krenzke (Westat)

In the past decades, federal agencies and other data collectors have devoted enormous effort to the protection of data confidentiality to ensure that the data products they prepare to release do not enable the identification of individuals or entities, given the information that has been previously released to public. Evaluating and controlling the disclosure risk becomes an indispensable and crucial step in the process of disseminating statistical products or results. In this paper, we depict the potential data breaches in different types of statistical products to be released, lay out a general process for risk assessment and estimation, and review several ways to quantify risk at record level or at the file level. The theories under which the approaches have been developed are based on specific models and assumptions. However, these assumptions may not be well respected when applying the approaches to survey data, and they also have limitations in their coverage of risk. This paper discusses the practical issues one may encounter in risk estimation and provides guidance and insights on how to set up risk assessments and make decisions in this process.

Assessing the FJSP linking algorithms using synthetic data
Ryan Kling (Abt Associates Inc.), David Izrael (Abt Associates Inc.)

The Bureau of Justice Statistics Federal Justice Statistics Program (FJSP) collects, standardizes, and links administrative data across the federal justice system, from investigation, arrest, prosecution, pretrial activity, adjudication, and corrections. However, the administrative data for each stage of the system is maintained by a separate agency, and the data are not designed to interact. Although the FJSP currently links individual and case data across each stage of the system, a common recurring challenge is to report the validity of our linking algorithms for data with lower-quality personal identifiers. This is particularly difficult since two stages in the system, prosecution and adjudication, report many fewer identifiers than other parts of the system. In this paper, we use known good linkages from other parts of the system to test the Type I and Type II error produced by our linking algorithms. By degrading the identifying data from the known links to match the quality of identifiers in the data with poorer identifiers, we can benchmark the linking algorithms and estimate the threat to statistical validity that poorer reporting of identifiers pose to FJSP linking.

Reliance on goodness-of-fit criteria to make an accurate assessment of re-identification risk for survey microdata
Lin Li (Westat), Jianzhu Li (Westat), Tom Krenzke (Westat)

In this presentation we discuss some practical issues encountered when estimating file-level disclosure risk measures of re-identification in survey microdata. We typically use the log-linear modeling approach (Skinner and Shlomo (2008)) to estimate disclosure risk in survey microdata files. Challenges emerge that relate to satisfying goodness-of-fit criteria of the log-linear models in the presence of model assumption violations. We evaluated the performance of the goodness-of-fit criteria of log-linear models particularly for the case of complex survey designs and differential survey weights through a simulation study. Results show that although the goodness-of-fit criteria are met, the risk estimates may be biased. The simulation results lead to guidance for a sensitivity analysis that helps to provide a better estimate of file-level risk of re-identification in survey microdata.
Formal procedures for disclosure avoidance have long been in place for quantitative research, but for qualitative research there is little published in the way of prescriptive guidelines or standards. Furthermore, relatively recent and evolving real-world changes (e.g., the availability of large, public datasets) raise new questions about qualitative research products that were previously considered less vulnerable to re-identification and other attacks on privacy and confidentiality. To address this, the Census Bureau set out to develop guidelines specifically for disclosure review of qualitative research reports. We canvassed our counterparts in other federal statistical agencies, examined policies of relevant professional associations, and consulted published literature. We then consulted seasoned disclosure avoidance experts within the Census Bureau, and we leveraged techniques commonly employed in quantitative disclosure avoidance, adapted for the unique nature of qualitative research methods, samples, and data. These guidelines are in the midst of a vetting process within the agency, hence our interest in sharing with the wider research community for feedback and input.