2023 Federal Computer Assisted Survey Information Collection Workshops

April 11 – 12, 2023

PROGRAM

Day 1, April 11

9:00 am – 10:00 am, April 11

Opening Session

Welcome Remarks

Keynote Address: **Modernizing Agricultural Statistics through Technology and Innovation**

Joe Parsons, USDA National Agricultural Statistics Service

Every federal statistical agency is on a journey to advance and modernize the business of collecting, processing, and disseminating statistics. That journey is shaped by the context in which each organization operates. How does an agency's operating environment impact the choices for technology and methodology modernization given scarce resources? How do characteristics of the populations of interest shape modernization priorities? How can technology adoption best alleviate agency and stakeholder pain points? How does feedback from stakeholders/audits/review panels develop into opportunities for innovative improvement? How can multiple modernization efforts be integrated into a cohesive strategic effort? How can the federal statistical community be more effective in renovating the "data factory" through sharing and collaboration? This talk seeks to address these questions from the perspective of an executive at the National Agricultural Statistics Service.

10:15 am – 11:45 am, April 11

Session 1

Session 1A: **Advances in Data Linkage**

Enhancing Record Linkage Production Data Quality

Douglass Huang, Todd Johnsson

ADI LLC

We have devised a method, Record Linkage Production Data Quality (RL-PDQ), to enhance the data quality of any record linkage system without disrupting the core design or operation of that system. As the Census Bureau increases its use of administrative records, one potential scenario involves using record linkage to incorporate administrative records as primary survey data for the 2030 Census, reducing respondent burden and operational costs. In this session we will review the theory and supporting research that indicates a further 70-80% reduction in false positives (impacting survey data quality) and a further 96-99% reduction in false negatives (impacting traditional enumeration efforts) when applying RL-PDQ in such a scenario. This outcome would drive production data quality improvement and lead to substantial enumeration cost savings, with measurable results.

Georeferencing the NASS List Frame Using New Crop Sequence Boundaries and Georeferenced Administrative Datasets

Kevin A. Hunt, Rachael V. Jennings, Denise A. Abreu

National Agricultural Statistics Service

The USDA National Agricultural Statistics Service (NASS) collects data from farm operators using a list frame. A farm is any place with $1000 or more in sales or potential sales. The list frame is not complete. In 2017, a report by the National Academies of Sciences Engineering Medicine (NASEM) recommended that NASS update the list frame to a georeferenced farm-level database and develop linkages to other administrative sources to improve the coverage of the list frame. The Agency has developed an approach using an alternative geospatial field-level boundaries dataset called Crop Sequence Boundaries (CSBs). The CSBs are developed from stacking 8 years of Crop Land Data Layers (CDLs) using an algorithm-based approach to identify field boundaries of homogenously cropped areas overtime. The approach overlays the CSBs with georeferenced data called Common Land Units (CLUs) from the USDA Farm Service Agency (FSA) to identify areas without CLU coverage. CSBs without CLU coverage are used to retrieve additional tax assessor landowner parcel information from CoreLogic Inc. The parcel records are linked to the NASS list frame and any new records are selected and surveyed to determine their status as a farm. This resulted in thousands of new farm operator records being added to the NASS list frame. The results from 12 states are presented here.

Preferred Capabilities of Record Linkage Systems for Facilitating Research Record Linkage

Krista Park[1], Mishal Ahmed[1], W. Glenn Ambill[1], John Cuffe[1], Khoa Dong[1], Suzanne Dorinski[1], Juan Humud[1], Shawn Klimek[1], Daniel Moshinsky[1], Kevin Shaw[1], Damon Smith[1], Yves Thibaudeau[1], Christine Tomaszewski[1], Victoria Udalova[1], Daniel Weinberg[1], Daniel Whitehead[1], Casey Blalock[1], Steven Nesbit[2]

[1]US Census Bureau, [2]PORTAL Technologies

Over 17 months, a team of more than 25 Census Bureau record linkage practitioners from across the Decennial, Economic, Demographic, and Research directorates met to define and describe requirements for record linkage systems. These requirements will facilitate moving record linkage activities from specialist production or research into the every-day production environment as the demand increases for new data products to enable evidence-based policymaking and alternative data sources are increasingly harnessed to develop more products previously created through survey data collection. The team defined two assessment tools over the course of their work to conduct a landscape analysis of record linkage solutions. This presentation and accompanying paper disseminate the teams two assessment tools which that take as input, solution requirements and associated weighting definitions. The first tool, the Technical Solutions Assessment (TSA), is a comprehensive framework designed to analyze robust, commercial solutions. It consists of 378 capability requirements across six categories: Technical, Operations, Performance, Cost Factors, User Experience, and Security. After review of 23 internal, commercial, open-source, academic, and federal record linkage systems, the team produced a narrower tool, the Condensed TSA, designed to better assess the non-commercial solutions which often do not provide stand-alone, end-to-end data handling throughout the record linkage pipeline. The team then used this Condensed TSA, consisting of 194 requirements, to assess internal Census Bureau record linkage and open-source record linkage solutions in preparation for a validating benchmarking test.

Ancillary Data Record Linkage to characterize the completeness of data for the All of Us Research Program

Yuyang Yang[1], Kelsey Rodriguez[2], Melissa Basford[2], Abel Kho[1], Lew Berman[3]

[1]Northwestern University, [2]Vanderbilt University Medical Center, [3]National Institutes of Health

The All of Us Research Program (AoURP) is a federal initiative to support biomedical research by creating a consented patient cohort of over 1 million Americans. AoURP health data is collected via survey questions, electronic health records (EHR), and biosamples to create a comprehensive picture of participant health. Despite this goal, investigations within the AoURP have found that participant health information is oftentimes incomplete. To improve data quality within the AoURP, we describe work to supplement AoU patient information using insurance claims data via a privacy preserving record linkage (PPRL) pipeline. We linked EHR data for 400,000 consented AoURP participants with claims data provided by IPM.AI (Swoop Analytics), an analytics company with estimated private and public claims records for over 90% of the U.S. population. We found that claims data contained 1.5x more service dates and diagnoses and 3.2x more procedures compared to AoU data for the 95% of matched participants. We found variation between AoURP and claims data for different patient profiles, such as patients with certain chronic conditions. Overall, the union of AoU and IPM.AI data greatly increases data completeness compared to either source alone. Our study has implications for data collection in the AoURP and shows that supplementary linkages can improve data completeness in national research initiatives.

Creating a New Data Infrastructure for Foreign-Born Scientist and Engineers: Data, Analysis and Use

Yunie Le, Nathan Barrett, Allison Nunez, Ekaterina Levitskaya

Coleridge Initiative

The U.S. government seeks evidence to support a more comprehensive understanding of the availability and demand for global science and engineering training and talent. The objective of this project is to establish the foundations of a national data infrastructure that will augment the federal survey data with administrative data linkages and help address unanswered questions about foreign-born scientists and engineers in the United States, beginning with estimating the economic return on investment on U.S. training of Foreign-Born Scientists and Engineers (FBSEs). The goal of the project is to establish the feasibility of constructing a linked data infrastructure, which can help answer important research questions about this population. As part of the project, the team explores and documents the coverage using federal government surveys, including Census, NCES, and NCSES data, and benchmarks it against administrative data from three U.S. states on educational enrollments and completions linked to individual and firm level employment data.

Session 1B: **Transitioning Data Collection Modes**

Video Interviewing: Applying a New Paradigm for a National Behavioral Health Survey

Ramasubramanian Suresh, Heidi Guyer, Christine Carr, Lilia Filippenko, Preethi Jayaram, Curry Spain, Matthew Check

RTI International

The Mental and Substance Use Disorders Prevalence Study (MDPS), funded by SAMHSA, is a pioneering study to estimate the prevalence of serious mental and substance use disorders in U.S. households, prisons, homeless shelters, and state psychiatric hospitals. MDPS hired clinicians with clinical training in mental health, including experience conducting the Structured Clinical Interview for the DSM-5 (SCID-5), to conduct clinical interviews. The clinical interview was programmed in Blaise and included a link to the NetSCID-5, a web-based version of the SCID-5. Video interviewing was planned for a large subset of the household

sample but the COVID-19 pandemic forced us to switch to this new paradigm for all of the household clinical interviews and for clinical interviews conducted in half of the institutional facilities. It quickly became obvious that this mode of interviewing is the ideal mode for certain populations, such as those with mental health issues, regardless of whether they are in households or institutional settings. Furthermore, the additional benefit of being able to record the interviews for expert review improved the quality of the diagnoses and the data. We conducted over 3,700 video interviews and 1,600 phone interviews for this study, in addition to in-person interviews within the facility settings. In this presentation, we will describe the steps we took and the challenges we faced as we developed the systems to conduct these interviews by video, including scheduling the appointments, mechanisms for reviewing and updating the diagnoses in the NetSCID-5, and the benefits and drawbacks of video interviewing in general.

NSHAP Remote Data Collection: Transitioning to Mixed Mode Data Collection among Older Adults

Lauren Sedlak

NORC at the University of Chicago

The National Social Life, Health, and Aging Project (NSHAP) is a longitudinal, population-based study of adults born between 1920 and 1965. Data are traditionally collected by interviewers every five years in the respondent's home. For NSHAP's fourth round (2021-2023), we developed ways to collect complex data remotely to accommodate pandemic restrictions and concerns, reduce costs, allow for participation from respondents who prefer remote modes, and to consider a hybrid remote/face-to-face approach for future work. A combination of web, phone, and paper was used to administer the traditionally in-person interview to half the existing sample, with in-person follow up conducted for non-interview respondents. The other half of the sample was approached starting in 2022 for face-to-face interviewing. This presentation focuses on the methodological and operational challenges in adapting a lengthy, complicated longitudinal survey for remote administration in a manner that balances methodological rigor while minimizing respondent burden. We compare data from the remote data collection against NSHAP's face-to-face benchmarks to examine overall data quality and item-level non-response. The presentation concludes with a discussion of lessons learned about asking a longitudinal sample of older adults to try new data collection protocols as well as methods for ensuring data quality.

2030 Census Research: Self-Response Options for In-Field Enumeration

Christine Borman

US Census Bureau

The Census Bureau is currently conducting research to improve data quality and reduce respondent burden associated with in-field enumeration of housing units for the 2030 Census. As part of this research, we are exploring ways to introduce the concept of self-response to in-field enumeration. In-field self-response options that are being considered include having respondents complete their questionnaire on the enumerator's Census Bureau-issued device and providing respondents with a QR code to complete their questionnaire on the respondent???s personal device. This presentation will explore the assumptions we are researching to determine the efficacy of these potential in-field self-response options and how they could be operationalized during field data collection for the 2030 Census.

Introducing Web to "No Answer" Cell Phone Numbers: A Pilot for the Behavioral Risk Factor Surveillance System (BRFSS)

Ting Yan[1], Machell Town[2], William Garvin[2], Gina Shkodriani[1], Reanne Townsend[1]

[1]Westat, [2]CDC

The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. and territorial residents regarding their health-related

risk behaviors, chronic health conditions, and use of preventive services. The BRFSS sample includes telephone numbers using landline and cell phones. Data are collected via computer-assisted telephone interviewing (CATI). A pilot study was conducted on a sample of cell phone numbers finalized as "No Answer" from five states. These cell phone numbers were matched with an address by a vendor and sent three mailings to invite them to complete the BRFSS via a web survey. The purpose of this pilot study is to examine whether it is feasible to use matched addresses to recruit respondents to complete the BRFSS by web. The pilot study includes two experiments. The first experiment assesses the effect on response rate of the framing of the survey request, which emphasizes either the respondent's positive contribution or lack of contribution. The second experiment addresses the effect of the framing of incentive, which emphasizes either the respondent's opportunity to receive an incentive or the loss of opportunity to receive an incentive. We will report the findings of this pilot study and discuss implications of the findings for BRFSS and other studies using a cell phone frame.

The Use of Video Chat and Receptivity of Live Video Interviewing of Web Panelists Over Time

Hanyu Sun, Ipek Bilgen, Leah Christian, Rene Bautista

NORC at the University of Chicago

The use of video chat in everyday life has increased substantially during the COVID-19 pandemic. Survey researchers and organizations have also started exploring the use of two-way live video interviewing as a mode of data collection (e.g., DeBell et al. 2022). Previous research has used non-probability samples to understand respondents' use of video chat and their receptivity of live video interviewing (e.g., Conrad et al. 2022). It is unclear to what extent findings on the use and receptivity of respondents from non-probably sample sources could apply to the general public. To address this research gap, we collected and monitored the respondents' use and comfort with video chat and their willingness to participate in a live video interview through a probability-based panel, specifically, AmeriSpeak Omnibus surveys conducted in June 2019, June 2020, and April 2021. Funded and operated by NORC at the University of Chicago, AmeriSpeak is a probability-based panel designed to be representative of the U.S. household population. Randomly selected U.S. households are sampled with a known, non-zero probability of selection from the NORC National Frame, and then recruited into AmeriSpeak by U.S. mail, telephone interviewers, overnight express mailers, and field interviewers (face-to-face). In the presentation, we will present how the key measures change over time among the AmeriSpeak panelists on their use of video chat, including frequency of self-reported use of video chat, demographic differences in use of video chat, reasons for video chat, types of platforms used for video chats, and receptivity of live video interviewing. We will also discuss the use of live video interviewing with web panelists as a mode of data collection.

Session 1C: **Data Visualization and Dissemination**

Stakeholder and User Engagement When Redesigning a Survey Data Dissemination Website: Lessons Learned

Kathryn Downey Piscopo, Herman Alvarado

SAMHSA

The process of redesigning a data dissemination website for household and establishment survey data from a recognized federal statistical agency or unit poses unique challenges, regardless of agency size. Best practices recommend identifying stakeholders and users of the website, and engaging them at all stages of the redesign, especially when gathering technical requirements. In a recent project to redesign SAMHSA's Data webpage, two major groups of survey data users stakeholders were identified: internal (statisticians, managers, data analysts) and external (other federal/state agencies, researchers, academics). Although the methodology of involving data users is relatively simple, through user testing and listening sessions, it is not

exempt of particular challenges. Issues like decision-making, overlaps in reviews, stakeholder trust, limited technical expertise, contract involvement, and scheduling meetings represented major challenges to this project. This presentation will cover the process of stakeholder and data user engagement, their respective challenges, and the proposed solutions, with particular emphasis on the lessons learned from this project that could be applicable to other federal agencies.

Design Strategies for Website-Embedded Data Visualization Tools

Ning Wang, Tiffany Julia

National Center for Science and Engineering Statistics

Data visualization is a fundamental tool used among data scientists and within organizations to allow users to visually explore data to better analyze relationships, gain relevant information, and develop their understanding. Data visualization needs have transformed from basic displays of information to tools that enable analysis and conclusions to match the increasing importance of timely and accurate insights. Understanding how to design more user-centered, accessible, and effective data visualization tools is the key to better serving the needs of data reviewers, agency analysts, policy experts and more. Here, we present a comprehensive framework for designing website-embedded data visualization tools. The proposed framework considers the design from three aspects: 1) the effectiveness of the information communicated; 2) the emotional engagement of users; and 3) the inclusiveness of the tool. For each aspect, three recommended design strategies are provided. The framework's strategies are based on evidence-based research in cognitive psychology, learning through multimedia formats, data visualization studies, user experience research, as well as industry best practices.

Bringing Automated Reports into 508 Compliance - Bulk Document Updating

Francis Smart[1], Issa Abboud[2]

[1]Censeo Consulting Group Inc, [2]General Services Administration

In this paper we discuss how we programmatically took nearly a thousand documents each containing numerous tables spread over dozens of pages and converted them to be 508 compliant. This process involved the scraping of source documents from Sam.gov, the ingestion of their text and tables, the appropriate recognizing and tagging of table headers, as well as the subsequent production of 508 compliant documents with embedded accessible tables as well as table of contents. The source reports were generated through a system which publishes database information in a slightly more readable format in the form of automated report generation. While we did not have access to the system which generated the reports, we did have the means to reprocess the reports bringing them into 508 compliance.

Visualizing Data Through the Mapping Features In data.census.gov

Jessica Barnett

US Census Bureau

Data.census.gov is the U.S. Census Bureau's premier data dissemination platform to access economic and demographic data in one single place. This platform is based on overwhelming feedback to streamline the way you get data and digital content from the Census Bureau. Through this approach, we continuously update and improve the customer experience. With a concept of using one centralized platform for data users, they spend more time using Census Bureau data and less time searching for data and content. In this session, we will provide a short demonstration of the latest mapping features of the site.

New Beginnings: User Interface Design in a new National Longitudinal Survey of Youth, NLSY26

Safia Abdirizak

Bureau of Labor Statistics

The National Longitudinal Surveys (NLS) program at the Bureau of Labor Statistics (BLS) has been tasked with developing a new cohort survey, the National Longitudinal Survey of Youth 2026 (NLSY26), including developing a dissemination system for the data to be collected. NLS will base the initial requirements for the new dissemination system on two sources of information. First, NLS will draw upon its experience with two ongoing cohorts, the National Longitudinal Survey of Youth 1979 (NLSY79) and the National Longitudinal Survey of Youth 1997 (NLSY97). During its stakeholder outreach in Spring 2022, NLS asked users their experience with the current NLS user interface. NLS will analyze this feedback to understand the features liked by users, as well as the obstacles for users with the current user interface. In this outreach, NLS has sought to understand what elements and features are important to users, such as layout, visualization, accessibility, and navigation. Second, NLS will examine the data dissemination tools, design features, and site navigation of other longitudinal studies, including the Panel Study of Income Dynamics (PSID), the Health and Retirement Study (HRS), Understanding Society: the UK Household Longitudinal Study (UKHLS), and the National Health and Aging Trends Study (NHATS). For each of these inputs, this paper will explain the implications for a new dissemination system. Based on these lessons, NLS will develop a list of requirements for the dissemination system for the new NLSY cohort.

11:45 am – 12:45 pm, April 11

Lunch Break

12:45 pm – 2:15 pm, April 11

Session 2

Session 2A: **Big Data, Big Decisions: Roundtable on Using Data Science in the Public Sector**

Erica Yu[1], Shalise S. Ayromloo[2], Josue DeLaRosa[3], Travis Hoppe[4], Brandon Kopp[1], Hiroaki Minato[5]

[1]Bureau of Labor Statistics, [2]US Census Bureau, [3]National Center for Education Statistics, [4]National Center for Health Statistics, [5]Energy Information Administration

Across federal statistical agencies, data science capabilities and challenges continue to grow - developing alternative data sources, innovating new ways to collect traditional data, changing the way we protect data, and more. To do this, data scientists are facing new challenges and new responsibilities. In this roundtable session, panelists will discuss the role of data science at their agencies and their perspective on how data science will continue to adapt at their agency. Discussion will include how AI can create new features for statistics and surveys, data stewardship best practices, how to prove the organizational value of incorporating data science, and how to support current and attract new and data scientists across the agency.

Session 2B: **Respondent-Centered Innovations in Data Collection Instruments for Establishment Surveys at the Census Bureau**

BERD Reporting Tools - Facilitating the Reporting Process

Michael Flaherty

US Census Bureau

One of the key characteristics to organizational non-response is the capability of the respondent to obtain the information requested (Tomaskovic-Devey 1994). Typical respondents to business surveys reside in accounting departments. If the information requested is not readily available in the company's records, it may require the respondent to reach out to someone else, in another department of the business. This presentation will explore strategies used to help facilitate this and the ongoing challenges the Business Enterprise Research and Development Survey faces in trying to get respondents to pass sections of the form off to others within their organization.

Rows and Columns and Respondents: Designing a Respond-by-Spreadsheet Option with Respondents for the Annual Integrated Economic Survey

Melissa Cidade, Heidi St Onge

US Census Bureau

The US Census Bureau's approach to annual economic data collection has historically been along established trade categories, with surveys being fielded by type of business. In 2024, the Census Bureau will launch the Annual Integrated Economic Survey (AIES), a cross-sector, harmonized annual survey designed using respondent-centered survey design principles. In this presentation, I will outline how respondent-centered research has informed instrument design decisions, particularly with the implementation of a respond-by-spreadsheet response mode. I will start by laying out how we incorporated the findings of formative research into a small-scale pilot. I will then discuss how we took findings from that pilot and turned them into recommendations for the final instrument. I will conclude with early renderings of the new AIES instrument.

Using Multiple Methodologies to Test Innovative Changes to the Commodity Flow Survey

Rebecca Keega

US Census Bureau

The Commodity Flow Survey (CFS) collects detailed data on the movement of goods in the United States. Several innovative changes have been made to the survey instrument. The CFS introduced a new method of data collection allowing respondents to upload all their shipment records, as opposed to creating a sample, which was a burdensome process. Additionally, a machine learning tool was implemented which allowed respondents to forgo entering in Census Bureau specific product codes by applying the code on the backend. Finally, a new feature allowed respondents to correct any errors within the instrument, and with one click, fix all identical errors. These new features were tested with several rounds of qualitative research through cognitive testing, usability testing and post-production debriefing interviews. This presentation will explore how traditional qualitative methodologies can be applied to innovative efforts, by helping to refine and confirm that changes to an instrument will benefit respondents and ease their reporting burden. As more surveys incorporate automation, the fundamental value of these qualitative methodologies remains a valuable part of the survey life cycle.

A Respondent-driven Web Portal for Establishment Surveys

Aryn Hernandez, Jason Baue

US Census Bureau

In 2016, the U.S. Census Bureau introduced the Respondent Portal to act as a hub for the bureau's economic surveys. The portal was developed in order to modernize and streamline the reporting process for business respondents, and the Census Bureau has since migrated its establishment surveys to this new portal. The development of the portal was based on many years of previous research into the response process and issues faced by business respondents. Development included multiple rounds of usability testing, respondent debriefings, paradata analysis, and other feedback provided by respondents throughout the portal's years of use. The Respondent Portal's features were designed to ease respondent burden by facilitating common business survey response processes, such as coordinating response timing with data availability and gathering data from multiple people and sources throughout the business. Previously, each survey was accessed through its own URL and required separate login credentials provided to respondents via letters. In the portal, respondents only need to manage one set of login credentials of their own making to access multiple establishment surveys. In addition, the portal provides respondents with a secure messaging center, survey help, and self-service options such as checking filing status, requesting due date extensions, delegating response tasks and data collection activities to colleagues, viewing reporting history, resetting passwords, and more. More recent improvements to the portal include automated emails with useful information and a content canister designed to point respondents to helpful information or our many data products. The Respondent Portal is a respondent-driven instrument that continues to evolve and improve with use.

Session 2C: **Improving Field Operations with Technology**

Developing a Manual Spanish Transcription Baseline to Evaluate Machine Transcription of Call Center Calls

Marcus Berger[1], Betsar Otero Class[1], Crystal Hernandez[2]

[1]US Census Bureau, [2]Connecticut College

With machine language transcription as a fast-evolving field, one important task is to evaluate the accuracy of different transcription models. As part of 2020 Census operations, we collected audio data from calls made to our Census Questionnaire Assistance (CQA) call centers with live agents. Calls were supported in English and 11 other languages, including Spanish. The volume of calls to these centers makes machine transcription of calls an attractive option. To determine which machine transcription model best transcribed the calls, we needed a high quality, human generated transcript to use as a baseline to compare against. As part of an earlier phase of the project, we found the best model for calls in English. For Spanish, we used a team of three Spanish-speaking researchers to transcribe the call segments. In this talk, we describe the method used to transcribe Spanish calls and the review process used to ensure transcription quality. We also discuss challenges encountered along the way, including how the method of transcription had to be adapted from more traditional transcription methods used in linguistics to fit the objective of comparing to machine transcriptions. Using this human generated transcription as a baseline allows us to be able to compare machine transcription models and determine the best fit.

Evaluating the Impact of Speech Analytics and Interviewer Self-Monitoring on Telephone Survey Metrics

Lauren Hartsough, Jason Rajan, Zhao Guo, Kate Hobson, Erin Criste

NORC

The use of speech analytics has become increasingly more common in call centers as a means of enhancing quality monitoring processes. NORC at the University of Chicago invested in an existing speech analytics system for the Telephone Surveys and Support Operations department to evaluate its feasibility as a tool for improving quality assurance and survey interviewer performance. The system included dashboards to give interviewers the ability to self-monitor by reviewing automated scores in addition to receiving feedback from a quality coach. We sought to determine the impact of incorporating speech analytics and self-monitoring into NORC's quality program. While the costs associated with implementing such a program can be substantial, there is the potential for a significant return-on-investment with savings in reduced monitoring time, greater efficiency, and higher response rates. NORC conducted an experiment to assess the value in giving telephone interviewers access to different components of the speech analytics software (i.e. automated scores, coach feedback). We will discuss key findings and implications for driving improvements in different areas of interviewer performance.

Modernizing Survey Software, or How We Get From Here to There

Jennifer Maiwurm

NASS

Many agencies are still using old tools in their survey processes. Even if an agency has managed to get rid of legacy tools, technology is advancing fast enough to require agencies to constantly evaluate and enhance their tools. There are two ways to advance from the status quo. Progress can be made in a series of "baby steps," marginal enhancements or changes to existing tools or moving to new tools that are well established and allow for similar processing. You can also advance by dramatically overhauling your software and processes in one "giant leap forward." Both ways can be helpful for improving processes, but they have different resource needs, risks, and benefits. I will examine those differences using the test case of editing software at USDA-NASS.

The In-Office Enumeration Vision for the 2030 Decennial Census

Syed Ali, Mary Kraetsch

US Census Bureau

This presentation will outline the U.S. Census Bureau's current plans for the In-Office Enumeration Operation for the 2030 Decennial Census. The In-Office Enumeration Operation aims to use high-quality administrative records data to enumerate housing units that do not self-respond, reducing the need for in-field enumeration and helping to inform mail strategy. The purpose of the operation is to use administrative data as well as additional data inputs gathered both pre-production and during data collection, to reduce costs resulting from mailings and in-field enumeration while maintaining or improving data quality. The concept of an In-Office Enumeration Operation was born from the combination of two components of the 2020 Decennial Census: (a) the use of administrative records to limit contact attempts during In-Field Enumeration and (b) the successful implementation of the In-Office Address Canvassing operation. The main idea is that the Census Bureau can use administrative records to enumerate non-responding housing units. As such, those housing units would be subject to fewer mail and field contact attempts during the 2030 Decennial Census. This presentation will describe current plans for the operation, key functions and activities, inputs and outputs, timeframe, current assumptions, and areas for further research.

2:15 pm – 2:30 pm, April 11

Break

2:30 pm – 4:00 pm, April 11

Session 3

Session 3A: **Top Three Challenges Organizations are Encountering in Technology and Survey Computing** (Roundtable)

Jane Shepherd[1], Amrit Kohli[2], Suresh[1], Gregg Bailey[3],

[1]Westat, [2]Bureau of Labor Statistics, [3]US Census Bureau

Panelists will identify the top challenges facing their organizations today given the changing survey technology, data systems, and hybrid work environments for systems managers and programmers. Projects today often include innovative survey technologies, enterprise software solutions, cloud deployments, the use of specialized programming customizations, incorporate administrative and extant data sources, and the integration of different devices and technologies to support research data collection. The panelists will discuss the ways that their organizations are dealing with the environmental changes that they have identified, and offer examples and lessons learned in addressing these challenges.

Session 3B: **Data Science Applications: Automation**

Programmatic PDF extraction: An applied case study in federal government application portfolio analysis

Kate Burdekin

RTI International

With the federal government's move into data modernization, our data sources are expanding and changing rapidly; and in today's modern data landscape, PDFs are increasingly becoming a source of information or downstream analyses and machine learning applications. In this presentation, we'll cover open-source methods for automating PDF content extraction alongside an applied case study: portfolio analysis for select federal government applications (Ex. Notice of Special Interest (NOSI), Research Opportunity Announcement (ROA)). Manual portfolio analysis is a very time intensive process, requiring hours to distill application information into meaningful insights. Automated PDF extraction enables the rapid identification of key information such as scientific area of focus, partner institutions, cost, and content themes. This modern approach allows for a programmatic, data-driven review process of applications, resulting in critical time savings for the federal government.

Transitory location frame enhancement and validation with web scraping and predictive modeling

Haley Hunter-Zinck, Louis Avenilla

US Census Bureau

Like housing units or other more permanent residence structures, transitory locations (TLs) require the development of a frame to ensure enumeration coverage. Mining of alternative data sources could facilitate and enhance in-person verification of TLs for frame augmentation and validation. We developed a web scraping tool to extract information on recreational vehicle (RV) resorts, campgrounds, and related facilities from the Good Sam website (www.goodsam.com). Good Sam provides information on campgrounds in 49 states across the United States including capacity, occupancy, address, and other attributes of these campgrounds. Although information on each campground is structured in a standardized

format, important data elements, for example capacity, are sometimes but not always present for each campground. To increase the utility of the dataset generated from the web scraping tool, we also developed a predictive model to impute the campground capacity. For campground webpages for which the capacity was not listed, we then predicted capacity using other pieces of data scraped from the same page for the campground. We report on data payloads from scraping and parsing all campgrounds on the Good Sam website and on the predictive model performance. Results show that even if required data is not explicitly available on a webpage for scraping, other data elements on the page may be used as proxies to impute the desired information. This project demonstrates the value of web scraping for harnessing alternative data sources for frame development.

Applying AI Techniques for Records Abstraction in MEPS-MPC

Ramasubramanian Suresh, Anna Godwin, Robert McCracken, Brandon Peele

RTI International

For the Medical Expenditure Panel Survey / Medical Provider Component (MEPS-MPC), data are abstracted from medical and billing records received via fax, mail, or provider portals. These records range from a few to hundreds of pages for each patient-provider pair. The data to be abstracted are not always in consistent formats, requiring the abstractor to browse through pages upon pages of electronic files in PDF format. For the Abstractor's first step in this record abstraction process, they manually highlight key items that need to be captured for entry into electronic data collection forms. These Items are highlighted in different hues based on the type of item such as dates, names, charges, etc. In addition, the whole process needs to be conducted in a FIPS-Moderate compliant environment. To reduce abstractor burden in this process, we developed an OCR-based AI technique to automatically highlight these key items. The Abstractor's manual highlight step now becomes a review step before extracting the data into collection forms. In this presentation, we will discuss the steps we took to develop the AI techniques, the challenges we faced in implementing this process, lessons learned, and future enhancements envisioned.

Harnessing Open-Source Python Tools to Match and Deduplicate Individuals Across Disparate Data Sources

M Daniel Brannock[1], Ed Preble[1], Lynn Langton[1], Marguerite DeLiema[2]

[1]RTI, [2]University of Minnesota

Millions of Americans are victimized by mass marketing scams every year, with a significant quantity of those scams being delivered through the United States (US) Postal Service. The US Postal Inspection Service identified four criminal mail fraud enterprises and seized their "customer" relationship management databases which contain invaluable insights on victimization patterns in mail fraud. To unlock those insights, individuals represented in the four databases had to be deduplicated and merged in a FIPS-certified secure computing environment. We used the open-source text mining tool Dedupe.io to identify individuals that appeared multiple times within and across the four scams. Active learning was applied to train models with as few as 300 labeled examples. Despite significant differences in the availability and reliability of identifiers, we were able to label transactions as belonging to the same individual with an estimated false linkage rate of 0.4% and missing linkage rate of 3.8%.

Composite Weighting for Hybrid Samples

Mansour Fahimi

Marketing Systems Group

Increasingly, survey researchers rely on hybrid samples to improve coverage or secure the needed number of respondents in a cost-effective manner by combining two or more independent samples. For instance, it is possible to combine two probability samples with one relying on RDD and another on ABS. More

commonly, however, researchers are compelled to supplement expensive probability-based samples with those from online panels that are substantially less costly. If carried out effectively, such samples may address both cost and coverage challenges of single-frame surveys. Traditionally, the conventional method of Composite Estimation has been used to blend results from different surveys to improve the robustness of the resulting estimates. This means individual point estimates from different surveys are produced separately and then pooled together, one estimate at a time. Given that for a typical study one has to produce dozens of estimates for key outcome measures, this computationally intensive methodology can require serious time and resources. Moreover, component point estimates used for composition are subject to the inferential limitations of the individual surveys that are used in this process. During this presentation the author will start with a quick review of the traditional method of composite estimation and then introduces the method of Composite Weighting that is significantly more efficient, both computationally and inferentially when pooling data from multiple surveys. For empirical illustrations, results from three surveys will be presented with each survey relying on hybrid samples comprised of probability-based components from the USPS address database and supplemental samples from online panels.

Session 3C: **Innovations in Survey Design**

A Web-based Tool for Establishment Survey Assessment and Evaluation

Bryan B. Rhodes

RTI

Establishment surveys differ in several important ways for household surveys, however, the research into establishment survey methods is small compared to research on household surveys. Because of this, study designers may have trouble identifying methods to improve their establishment survey designs and methodologies. To address this, RTI have developed a web-based system called the Tool for Establishment Survey Assessment (TESA). TESA walks the user through an assessment of several fundamental aspects of an establishment survey, such as survey saliency, instrument burden, and data quality control procedures. The tool identifies potential challenges or shortcomings in the design and points the user toward literature that could provide input on design improvements. For example, if the survey design is determined by the tool to have a particularly high potential burden, the tool would then help identify literature that has addressed the issue of burden in establishment surveys. This presentation will describe the methodology used to create the generalizable framework of establishment survey design, the process used to develop the web-based tool and demonstrate its use.

Grid Displays in an Establishment Web Survey

Megan Waggy[1], Rebecca Powell[1], David Heller[1], Steve Gomori[1], Hope Smiley-McDonald[1], DeMia Pressley[2]

[1]RTI, [2]Drug Enforcement Administration

Researchers often present questions with the same stem in grids to decrease reading time and shorten the document (Couper et al. 2013). However, grids can increase survey burden because respondents must work both horizontally and vertically (Dillman, Smyth and Christian 2009). While we assume establishment respondents are more adept than the general population at completing complex response grids, there is limited research on formatting grids for establishments specifically. The National Forensic Laboratory Information System (NFLIS) Medical Examiner and Coroner (MEC) Survey, sponsored by the U.S. Drug Enforcement Administration, asks offices to indicate toxicology request (Column 1) and quantitative analysis frequencies (Column 2) across 26 rows of drug classes. We tested versions of web visual display to determine if one resulted in better data quality. These were: (A) Full Grid View; (B) Chunked Grid View (split into three screens with no more than nine rows visible per screen); or (C) Single Item View (split into 26

screens with only one row visible per screen). We present findings on survey timing and data quality across the three randomly assigned groups.

Incorporating Evaluation into Digital Forms

Stephanie Permut, Blair Read

Office of Evaluation Sciences

Americans spend over 11 billion hours per year filling out federal forms. Form complexity can result in lack of completion, and errors on forms have far-reaching consequences. Complex federal forms also burden the agencies responsible for processing and verifying form responses. This evaluation took an initial step to build evidence and capacity for testing digital federal forms. A first of its kind in the federal government, this effort brought together multiple GSA offices and the American public to learn about the feasibility of incorporating A/B testing in this domain. The intervention intended to improve instruction clarity in federal forms by embedding instructions alongside form questions. These changes made forms??? instructions more accessible, more relevant, and easier to process to readers. OES evaluated two versions of a sample digital form: one with form instructions displayed on the first page, and the other with form instructions embedded on every page. We found that where instructions are placed impacts form submission. Moving forward, other agencies might adapt our A/B testing workflow to study other survey design insights.

An Electronic Life History Calendar as a Web Survey Memory Aid

Joseph Nofziger, Lilia Filippenko, Emilia Peytcheva

RTI International

In surveys that collect data of experiences over multiple years of their lives, respondents benefit from visual guides. A life history calendar is one such tool which many organizations have used over time, including the National Survey of Family Growth (NSFG). Using the existing paper Life History Calendar as a starting point, RTI developed an electronic life history calendar and integrated it with a Blaise 5 instrument so that web respondents have the option to view significant events in a timeline format. When subsequently asked to place other events in time, they may refer to the calendar for context. The calendar can be displayed on demand at any time - or in specific sections if so configured - either in full screen mode or along with the Blaise question. Its responsive design displays on any screen in portrait or landscape orientation. In iterative collaboration with our clients at the National Center for Health Statistics (NCHS), we heavily customized an off the shelf chart and automatically populated the calendar with responses in real time. Events are put into temporal context with year, month, respondent age, and previously entered responses to help orient the user. Descriptive hover text is available where label space is limited. The calendar and training materials were designed with input from cognitive testing.

Programming to Pass Survey to Another Respondent and to Create Dynamic Question Headers

Matthew Bensen, Ansu Koshy

RTI

RTI used Voxco Survey Software to program a web survey for a federal client towards gathering administrative data for its programs. We employed creative methods to achieve the requested survey objectives. The client anticipated that the person to whom the survey was sent might not be able to answer all the questions. To address this, Voxco has available a mechanism so that the initial survey taker can email the survey link to another person for it to be completed. We will show how we harnessed some Voxco features to do this. RTI also selectively created the question headers so that they are based on responses to previous questions. During the presentation, we will show how this was accomplished.

End of Day 1 Program

Day 2 Program Starts on the Next Page

Day 2, April 12

9:00 am – 10:30 am, April 12

Session 4

Session 4A: **Management challenges associated with survey project management** (Roundtable)

Karen Davis[1], Adam Safir[2], Kyle Fennell[3], Carolyn Pickering[4]

[1]RTI, [2]Bureau of Labor Statistics, [3]NORC, [4]US Census Bureau

Panelists will discuss the challenges facing managers with regard to resource planning, technology and talent acquisition, technical staff training and retention, budgeting, and risk management in today's inflationary environment and its impact on projects. Some of the specifics may include discussion of innovative approaches to technical staff retention, agile and hybrid methods of workflow management, use of data analytics to inform project decision-making, managing execution risk during multimode data collections, and monitoring the performance and productivity of a hybrid technical team designing and implementing surveys.

Session 4B: **Technology Transformation at the Census Bureau: Building a Data-Centric Ecosystem**

Overview

Michael Thieme

US Census Bureau

The U.S. Census Bureau's mission is to provide quality data on the nation's people and economy. We have historically answered those questions by conducting censuses and surveys and publishing the results. While critical, censuses and surveys alone can no longer answer these questions completely or quickly enough to satisfy the modern appetite for information. At the same time, our society produces vast amounts of data that are directly related to the Bureau's mission from a multitude of sources and much of it in real time. These nontraditional (for official statistics) data sources have great potential to help the Census Bureau improve the information it provides data users on the characteristics and wellbeing of the nation's people and businesses. However, without significant modernization of the Census Bureau's approach, it will not be possible to leverage this unprecedented amount of data and provide the timely, high-quality products our data users need. This session describes the early steps in a new era for the Census Bureau. It describes cutting-edge linking of survey, census, and third-party data; modernized data processing; quality product creation; and innovative dissemination to the public. The session will show how The Census Bureau is moving from fielding surveys and censuses and publishing the results to combining data science with traditional survey methods, diversifying our data products, and placing data at the center of our approach.

Frames

Michael Ratcliffe

US Census Bureau

The Frames Program envisions a growing variety of linked datasets within the EDL. While some of these datasets already exist as standalone entities at the Census Bureau (e.g., Master Address File [MAF], Business Register [BR]), the Frames approach will collocate these and any number of curated datasets and provide an easy and efficient way to link them for purposes both familiar (e.g., providing a tailored survey frame) and unanticipated (e.g., answering a new question about jobs and COVID vaccination rates). Centralization and

”linkability” will increase efficiency, reduce duplicative efforts to maintain and manage data, and greatly expand our capacity to answer critical questions about the population and economy at multiple geographic scales. These linked, augmented, and continuously updated datasets will provide a more comprehensive means for maintaining and updating the inventory of our nation,s addresses, jobs, businesses, people, and other linked data. They will be used as improved collection and sampling frames for our censuses and surveys with augmented information from the linked sources.

Data Ingest and Collection for the Enterprise (DICE)

Greg Hanks

US Census Bureau

Providing a modern platform for both data collection and ingest, DICE will be a key entry point for data into the Census Bureau for subsequent transfer, storage, and use in the EDL. DICE will refresh legacy field, online, and paper data collection technology with updated, flexible capabilities that reinforce the new operations and data ecosystem approach. DICE will also provide much needed functionality to interact with external data ingest, frames, and other modern data processing capabilities. DICE will leverage both operations research and data science techniques to enable more efficient operations and adaptive survey design. DICE will enable flexible scaling to support the diversity of the Census Bureau’s data collection operations, from rapid, lightweight surveys to the decennial census, without the need for costly updates or system rebuilds. Many of the key functions provided by DICE were developed and successfully deployed in the 2020 Census, providing a strong foundation for further development and use by the entire Census Bureau.

Center for Enterprise Dissemination Services and Consumer Innovation

Zach Whitman

US Census Bureau

As the Census Bureau’s primary platform for data dissemination, CEDSCI will provide the gateway to our information for the public. As new data products are produced in the EDL with collected, ingested, and linked data, CEDSCI’s standardized platform will allow the Census Bureau to provide those products quickly. Allowing for discovery of data products and new visualizations and renderings of data, CEDSCI will provide a scalable solution for long-term data dissemination and a better experience for the user. Because CEDSCI has been built to enable easier data discovery for our external data users via data. census.gov and census APIs, reusing and repurposing existing CEDSCI code may also provide a distinct benefit to the integration of the four key initiatives by providing a similar data discovery capability to internal Census Bureau users. These capabilities will be particularly useful in the Frames context - for example, to discover and link diverse datasets within Frames that can quickly enable new survey frames, analysis questions, or innovative data products. We propose using CEDSCI in this way in keeping with a ”build once, use many times” approach.

Session 4C: **Leveraging data science to improve data quality**

Leveraging Paradata to Assess Respondent Behavior and Data Quality in the Consumer Expenditure Online Diary Survey

Graham Jones

Bureau of Labor Statistics

Outside of the flexibility that multiple mode options afford survey respondents and data collectors, online modes offer a unique perspective for assessing data quality by using paradata. This presentation will discuss the analysis of the first year of paradata resulting from the use of an online expenditure diary mode in the Bureau of Labor Statistics’ CE Diary Survey. These paradata contain information about user activity and engagement within the CE online diary, including user logins; devices used; operating systems used;

languages used; and time spent in the diary. Some of these measures (e.g., logins) will be analyzed to identify changes in respondent behaviors over the diary keeping period, while others (e.g., time in diary) will be compared by month to see changes occurring over 2021. Estimates will also be compared to paradata from a previous online diary test in CE. In addition, paradata measures will be analyzed alongside traditional data quality indicators within CE like total expenses and expenditure counts. Along with examining behavior and data quality in 2021, this analysis will establish standard paradata metrics to be regularly monitored and revised as needed.

Improving Consumer Price Index (CPI) Survey Frames Using Administrative Data

Trevor Bergqvist

Bureau of Labor Statistics

The Consumer Price Index (CPI) measures average price change faced by urban consumers in the United States. To track changes over time, the Bureau of Labor Statistics (BLS) collects the prices of goods from a sample of establishments. For most categories of goods and services, the BLS constructs the sample frame from data collected in the Consumer Expenditure Surveys (CE). In recent years, the Bureau of Labor Statistics has increased its use of the BLS business register, the Longitudinal Database (LDB), as an alternative frame source. Maximizing the use of LDB improves data collection efficiency, improves CPI accuracy, and reduces CE respondent burden. However, cleaning and processing business register data sourced from administrative records presents challenges and is not appropriate for all categories. In this presentation, we identify establishment frame data quality concerns (low response and high item-establishment mismatch) and the assessment criteria by which BLS evaluates the LDB as an alternative frame source. When applying the LDB in appropriate circumstances, it can fulfill all of our data criteria and supplement our current sample of the CE surveys for certain industries. Specifically, we will present four case studies (”Funeral Services,” ”Unpowered Boats and Trailers,” Delivery Services,” and ”Food Away From Home”) to demonstrate the careful analysis needed when assessing an alternative sample frame.

Making Sense of Large Volume Transaction Sales

Francis Smart

Censeo Consulting Group Inc

This paper presents a Data Cleanup Plan (DCP) for the handling of large messy data through use of a secondary data source. With over 20 million records, many of them incomplete or suffering from user input error the Transactional Data Repository (TDR) hosted at the General Services Administration (GSA) presents some significant challenges for producing meaningful data products. In this paper we discuss how we constructed a system for matching records across databases through a multiple potential match approach. We discuss common inconsistencies and pitfalls within the TDR database and provide a roadmap that outlines how these problems were addressed. We explore the structure of the TDR database, including its contracts and transaction categories, as well as its usage and growth over time. We then discuss data challenges and the metrics used to detect and exclude data irregularities, such as the Z-score, Interquartile Range, and Price/Mean Price Ratio. Finally, we discuss outlier detection and exclusion rules and their impact on subsequent data products.

Using paradata and metadata to assess effects of SOGI items on postsecondary longitudinal survey data quality

David A. Richards

National Center of Education Statistics

The National Center of Education Statistics included measures of sexual orientation and gender identity (SOGI) in several surveys of postsecondary students: Baccalaureate and Beyond 2016/2017 (B&B:16/17),

Baccalaureate and Beyond 2020/2022 (B&B:20/22), and the 2020 National Postsecondary Student Aid Study (NPSAS:20). Though these potentially sensitive SOGI measures were extensively examined prior to their addition to national survey instruments, paradata and metada were used, after the completion of data collection, to examine whether these items led to data quality concerns. This presentation will share results of these analyses, including investigations of breakoffs, item-level nonresponse, and time spent on item screens.

Creating PSUs with Geography-Based Sampling Units

George Zipf[1], Richard Valliant[2],

[1]DOT, [2]Universities of Michigan

In a survey population of geographic sampling units, it may be required that a sampling frame of primary sampling units (PSUs) be developed where geographic sampling units are combined. This is particularly true in two-stage sample designs, where consolidated geographic units may be chosen for a team of survey researchers to visit and secondary stage sampling units (SSUs), e.g. schools or businesses, are the units of interest. In this presentation, we review how to use the GeoDistPSU and GeoDistMOS functions in the PracTools package to create a sampling frame of geography-based PSUs and show an example. We provide some guidelines in defining PSUs, how to display the United States geography-based PSUs with maps, and also show how the results of geography-based sample frame can be integrated with two-stage sample size calculations.

10:30 am – 10:45 am, April 12

Break

10:45 am – 12:15 pm, April 12

Session 5

Session 5A: **Field Labor Challenges in CAPI Surveys** (Roundtable)

Brad Edwards[1], James T. Christy[2], Grant Benson[3], Sean Coleman[4], Kyle Fennell[5], Jill Carle[1]

[1]Westat, [2]US Census Bureau, [3]University of Michigan, [4]RTI, [5]NORC

The work of in-person field data collectors - CAPI interviewers -- is critical for the success of many large government-sponsored surveys. Despite response rate declines and increasing costs, the mode remains the gold standard for meeting the most rigorous survey requirements. However, technological advances over the last decade have catalyzed a seismic shift in where and how work is done throughout the US with important implications for CAPI surveys with large field staffs. Data collector recruitment and retention are in crisis across the industry. Advancing technology has facilitated new facets of work, increasing competition for part-time labor and providing more opportunities for job flexibility and remote work. While new field technologies have enabled data collection advancements, they have made job tasks more challenging for many field data collectors. The pandemic exacerbated this already looming challenge: part-time workers have more flexible job opportunities; the labor force ages; applicants with the right balance of social, technical, and physical skills for the unique CAPI interviewer role are scarcer. As the pandemic recedes, recruitment and retention have become more difficult, with many survey projects falling far short of field staffing needs. These challenges have led to missed recruitment goals, and increased need for attrition trainings and traveling data collectors. The presenters in this roundtable represent the largest field survey data collection organizations in the U.S. They will identify key challenges in recruitment and retention, and

present recommendations for going forward. Presenters will also consider ways to alleviate CAPI workforce demands (e.g., multimode alternatives; updating value propositions for respondents).

Session 5B: **Improving Response Rates**

Data Collection Initiatives to Face Declining Response Rates at Statistics Canada

Cindy Ubartas, Sylvie Bonhomme

Statistics Canada

Like in many other countries, response rates from social surveys have decreased in the last few years in Canada. Now that the post-pandemic period has started, this presentation will explain how different factors contributed to this trend and will describe initiatives already implemented in collection to address these challenges. It will also summarize the current state of collection (for CAPI or computer assisted in-person interviews in particular) as well as various research projects our organisation would like to conduct to better understand non-response but also determine the best ways to reverse the trend. For example, one of the main initiatives consists in exploring the potential response rate obtained for a voluntary survey by using all collection modes with a reduced sample size.

Investigating Methods to Improve Survey Response for a Multi-Purpose Probability Sample Panel

Ipek Bilgen, David Dutwin, Lindsay Liebert, J. Michael Dennis

NORC at the University of Chicago

Due to increasing survey costs and declining response rates, probability panels have become a major research vehicle for not just private, foundational, and academic research, but also for federally sponsored survey research. Federal research, however, requires high standards of data validity and reliability, as well as achieving high response rates and low nonresponse bias. Accordingly, NORC at the University of Chicago has recently launched AmeriSpeak?? Federal, a probability-based survey panel that offers the highest panel-based response rate in the country, to U.S. government agencies and departments. AmeriSpeak Federal invests substantially in follow-up strategies to convert non-responding households into cooperating households. These strategies also aim to increase the representation of panel studies by recruiting respondents who are more likely to be younger, low-income, non-white, and less educated groups that are typically underrepresented in most surveys. This presentation covers the effectiveness of specific data collection efforts that are incorporated during AmeriSpeak federal data collection studies. Specifically, we will present findings from a field test conducted using the AmeriSpeak Federal panel and will examine different interventions to improve panelist survey response, including the use of advance letter mailings; use of noncontingent cash incentives in advance letter mailings; tailored design of advance letter mailings; and the inclusion/exclusion of instructions on accessing surveys directly using a custom URL or accessing the panel portal log-in, in advance letter mailings. We will also present the impact of each of these treatments on panel representation and panelist survey taking behavior during subsequent surveys.

2020 Nonresponse Followup Proxy Procedures: Successes, Challenges, Future Directions

Sarah Gibb, Christian Garcia

US Census Bureau

The 2020 Nonresponse Followup (NRFU) operation sought to enumerate households that did not respond to the decennial census questionnaire. Field staff, known as enumerators, went door to door for nonresponding addresses seeking to interview a member of that household. When a household member could not be reached, enumerators were prompted by the automated data collection instrument to find a proxy respondent. Proxy respondents were individuals not residing at the census address, but knowledgeable about its inhabitants on Census Day. Common types of proxy respondents include

neighbors, landlords, or utility workers. This presentation reviews the 2020 NRFU proxy contact strategies, explores their successes and challenges, and suggests directions for future research to improve the quality of proxy-provided data and reduce respondent burden.

Making the Most of Priority Reminder Mailings

Alanah Raykovich, Kylie Carpenter, Peter Herman, Jennifer Vanicek

NORC

Priority reminder mailings have been shown to improve survey response, but how can the methods chosen within these mailings, such as the batch size and location of letters sent, be leveraged in the context of a large, national multimode study? In Fall 2022, NORC mailed reminder letters via FedEx to non-respondents from the incoming panel of beneficiaries on the Medicare Current Beneficiary Survey (MCBS). The MCBS serves as the leading source of information on the Medicare program and health care costs for the Medicare population. It is a continuous, multipurpose survey of a nationally representative sample of the Medicare population, with a new panel of beneficiaries recruited every fall. The Centers for Medicare & Medicaid Services (CMS) oversees the Medicare program and contracted NORC at the University of Chicago (NORC) to conduct the MCBS. Letters were shipped in two batches to promote efficient interviewer follow-up. This was further refined by grouping these batches based on geographic clustering and interviewer assignment. Comparative analysis will evaluate this approach relative to mailings sent without geographic clustering in the prior year. This presentation will cover our methods, focusing on geographic clustering strategy in this operational approach. Specific metrics, including the impact of the mailing on inbound hotline calls and interview completion will also be discussed. Finally, qualitative feedback about the speed of interviewer follow-up and field staff sentiment regarding this approach will be captured to contextualize these findings. Analyzing the effectiveness of this approach and sharing lessons learned will inform future respondent mailing strategies.

More is More? The Impact of Doubling Incentive on Food Reporting

Ting Yan[1], Elina Page[2], Tom Krenzke[1], Janice Machado[1]

[1]Westat, [2]USDA Economic Research Service

Food acquisition and purchase data are critical for understanding food expenditures, food consumption, nutrition, food waste and loss, local food environments, and food assistance. The Second National Household Food Acquisition and Purchase Survey (FoodAPS-2) Field Test collected detailed food data by asking survey respondents to report food acquisitions over a period of seven days. To encourage continuous reporting throughout the week, the FoodAPS-2 Field Test implemented an incentive experiment. A random half of households was assigned to the control condition, which offered $5 per person for completing each day's Food Log, yielding a maximum of $35 per person for 7-day reporting. The other half was assigned to the experimental condition, which offered $5 per person per day for completing the first 3 days of the Food Log. The daily incentive amount was increased to $10 starting on the 4th reporting day conditional on respondents having completed three days of the Food Log. The maximum amount of incentive for completing the 7-day Food Log was $55 per person. This paper investigates the impact and implications of doubling the incentive on food reporting. Specifically, we examine whether or not respondents assigned to the experimental condition completed more Food Log days and reported more events and items than those in the control condition.

Session 5C: **Data Science Applications: Machine Learning**

Using Machine Learning Recommendations to Improve Manual Survey Text Coding

Caroline Kery, Emily Hadley, Durk Steed, Ethan Richie, Rob Chew

RTI International

Coding text entries in federal government surveys is often a laborious and frustrating process. Survey coding generally requires one or more individuals to read each text and apply an appropriate code(s) from a classification system. Existing survey coding software can complicate the coding process in time-intensive ways and may require substantial staff training. These systems often have little to no deduplication, forcing coders to label identical text over and over. In addition, changing an assigned label due to a mis-click can be very difficult, and require a software admin to accomplish. SMART, an application developed by RTI in 2018 to help users label text for predictive models, was adapted to address common pain points in survey coding. Specifically, a natural language processing (NLP) model was implemented to recommend labels for text entries interactively during the coding process. Given that survey classification systems often contain hundreds to tens of thousands of nuanced domain-specific codes, this recommendation system helped our coders increase efficiency, reduce labeling burden, and improve the onboarding experience for new staff. In this session, we will discuss the extensions to SMART for survey coding, details of the machine learning approach, and lessons learned throughout development.

The Effects of Translation on Machine Learning Models: A Case Study from the SOII Autocoder

Daniel Todd

Bureau of Labor Statistics

The Bureau of Labor Statistics (BLS) uses Machine Learning (ML) models to classify and autocode cases submitted to the Survey of Occupational Injuries and Illnesses (SOII). With cases collected annually across the United States, thousands of submissions are in Spanish. Given a large majority of cases used to train ML models are in English, BLS investigated the impact of these Spanish cases and potential side effects they may present to the Autocoder. This presentation provides a brief description of the case study performed by BLS, the ML methods used to detect and translate Spanish cases, and an overview of findings regarding performance of ML models using translated and untranslated cases.

When Labels Become Obsolete: Updating the SOII Autocoder to OIICS v3.0

David H. Oh

Bureau of Labor Statistics

Official statistics rely on standardized classification systems, such as the North American Industry Classification System (NAICS) and the Standard Occupational Classification (SOC), to aggregate microdata into meaningful statistics. The Survey of Occupational Injuries and Illnesses (SOII), an establishment survey that collects worker injury and illness information, uses the Occupational Injury and Illness Classification System (OIICS) to classify workplace injury and illness cases into various components, such as the nature of injury or illness, the part of body affected, the event that resulted in the injury or illness, and the source(s) that caused the injury or illness. These standardized classification systems are periodically updated to better reflect the current state of the categories that they are covering. Since 2014, the Bureau of Labor Statistics (BLS) has utilized machine learning models (”autocoders”) to automatically assign OIICS codes. These autocoders are trained on large quantities of labeled SOII cases that use OIICS version 2.01. However, starting in 2024, the SOII autocoder will be required to assign codes using the updated OIICS version 3.0. This poses a challenge for the SOII autocoder since a large proportion of its training data will have labels that are no longer valid. What happens to an autocoder when the labels in its training data become obsolete?

This presentation discusses the method being implemented to enable the SOII autocoder to learn the new labels from OIICS version 3.0.

Applying machine learning language models to link similar text documents together

Monica Puerto, Elizabeth May Nichols, Shaun S Genter, Brian Francis Sadacca

US Census Bureau

In the last five years, researchers and practitioners in Natural Language Processing (NLP) have made significant strides to improve the accuracy and abilities of machine learning models. Large language models using some of the newest deep learning frameworks are trained on large bodies of unstructured text, like books, Wikipedia, and internet forums. These models have more nuanced understanding of language across contexts, enabling the ability to group documents based on their similar content, or to link documents across different groups. In our presentation, we'll share insights into the use of these models in the context of a survey call-center: the 2020 Census Questionnaire Assistance (CQA) operation. The CQA operation assisted respondents over the telephone with responding to and completing the 2020 Census. Through application to calls and call-supporting reference materials (Frequently Asked Questions, or FAQs), we'll review the use of these NLP models and their limitations for both research and operational purposes to better understand and answer survey respondent questions more quickly and efficiently.

12:15 pm – 1:15 pm, April 12

Lunch Break

1:15 pm – 2:45 pm, April 12

Session 6

Session 6A: **Human-Centered Design Approach to Data Dissemination Tools** (Roundtable)

Alda Rivas[1], Kanin L. Reese[1], Jeffrey Stark[1], Bryan Combs[2], Josue de la Rosa[3], Jean Fox[4]

[1]US Census Bureau, [2]National Agricultural Statistics Service, [3]National Center for Education Statistics, [4]Bureau of Labor Statistics

In the United States, different federal agencies collect data and produce statistics about multiple topics and at different levels of geography. These agencies then produce non-partisan reports to inform public policy. Furthermore, these agencies also have the responsibility to make the data available to the general public in a timely manner. However, because the data collected by each agency may span across hundreds of topics, geographies, and years, making these data available in an organized manner is a difficult endeavor. Human-centered design (HCD) involves considering the expectations and the limitations of users when designing products or systems to be used by humans. Designing a data dissemination tool is a difficult challenge not only because of the amount of information that is collected, but also because of the wide range of users to be considered (from policy makers and academic researchers to students and the general public). Adopting an HCD approach can aid in achieving the core goal of the data dissemination tool: making the data available and accessible to the intended users. This round table will be an informal and friendly discussion about the challenges and solutions encountered when designing systems to disseminate the data collected through federal surveys. Employees from different agencies (e.g., US Department of Commerce, US Department of Agriculture) will share their experience and ideas about best practices and difficulties faced in designing, maintaining, or improving a data dissemination tool.

Session 6B: **Data Management**

Improving Web Survey Access through Platform that Enhances Respondent Experience and Protects Systems

Bob Henne, Rebecca Watkins, Mike Price, Matthew Bensen

RTI

The Survey Landing and Information Page System (SLIPS) team at RTI developed a standardized platform for survey projects to use as a landing and login page to enhance security, respondent experience, and create efficiencies. Survey engines provide mechanisms to get respondents into the survey, but the available options to setup or customize a landing/PIN page are often minimal, leaving it to individual projects and companies to create websites to handle this step else settle for a minimal out-of-the-box experience. Further, it may mean using the survey engine itself to validate legitimate users, which can be problematic for high volume projects on shared systems. SLIPS solves these issues by offering an attractive, accessible, and cost-effective base template. These templates can be highly customized for more complex project needs, and the entire system adds a layer of protection between bad actors and the survey systems. In our presentation, we will review the drivers that led us to build SLIPS, discuss how it improves security, and show how it has improved our web survey operations.

Designing a documentation management system for survey data

Weihuang Wong

NORC at the University of Chicago

Comprehensive and clear documentation allows consumers of survey data to understand the provenance of, and make accurate inferences from, the data. Producing high-quality documentation can be a costly and time-consuming effort, spanning multiple teams such as analysts, technical writers, and reviewers. An integrated documentation management system can not only reduce data dissemination costs and shorten the time needed to deliver data to the public, but also improve data reusability, transparency, and reproducibility. In this talk I describe a system that a team at NORC developed to manage and produce documentation for the 2019 National Survey of Early Care and Education (NSECE) datasets. The NSECE is funded by the Office of Planning, Research, and Evaluation (OPRE) in the Administration for Children and Families (ACF). The system consists of a metadata database as well as a user frontend and automated scripts that interact with the database to produce codebooks and other metadata records, such as SAS formats and variable labels. I first discuss the motivation for developing this system and our design principles. I then describe our solution, which integrates a Microsoft Access database with existing SAS-based workflows and a newly developed RMarkdown-based reporting generation process. Finally, I review learnings from the first iteration of our system, focusing on potential improvements to the user experience and process reproducibility.

Addressing data collection management challenges-The role of SOPs in Managing Risk

Craig R Hollingsworth, Christopher Griggs

RTI International

After project award, contractors may find that the Federal agency they are working with may have security requirements that are stricter than those the contractor has implemented in their organization's network. Most contractors work to comply with FISMA and National Institute of Standards and Technology (NIST) 800-53 security controls and in doing so, ensure that data collection instruments are secure and that they meet the standards of most Federal agencies. In some cases, the agency has stricter requirements than the NIST controls and those implemented by the contractor at the network and application level. To address these individual client needs, we have developed a process for discovering deviations, the controls to be

addressed, and for implementing procedures that fulfill the requirements of the stricter control. The process includes examining documentation, working with developers, and developing documentation. The overall process includes (1) Reviewing contract security requirements and identifying any deviations from NIST. (2) Understanding the deviation and working with the project team to devise an implementation/solution. (3) Work with system developers as needed to determine if the control can be implemented at application level. (4) Work with our Chief Information Security Officer to determine if the control should be implemented at the network level. (5) Developing a Standard Operating Procedure (SOP) to ensure the security control is implemented at the level the client requires. In this presentation we will demonstrate how our processes determine control deviations and how we successfully address these deviations with an SOP.

Fostering a community of practice around restricted use surveys for the public good

Benjamin Feder, Jonathan N. Mills, Nathan Barrett

Coleridge Initiative

The expansion of researcher access to federal restricted-use surveys generated by the Foundations for Evidence-Based Policymaking Act of 2018 has generated interest in secure, cloud-based data hosting facilities. Organizations such as the Coleridge Initiative, a national non-profit, are working with federal agencies, including NCES, to create secure computing environments for researchers and agency personnel to better leverage data to inform public policy. Such efforts have only advanced in the wake of the COVID-19 pandemic, which dramatically limited researcher access to physical data facilities. While remote access platforms and increased researcher access can advance the use of data for the public good, additional work is required to ensure that such an environment is not a cloud-based replicant of siloed cold rooms. These advancements do not guarantee a collaborative research environment, where researchers and agency personnel benefit from and improve on shared knowledge. Technology can, however, be used to simultaneously promote data security and foster collaboration across teams and agencies. Such a system requires a technological architecture allowing for the provisioning of isolated project workspaces along with areas for researchers to share their work while not compromising data security. This architecture can be combined with technologies that allow code sharing with version control to guarantee that any advances are attributed to the proper individuals. This presentation will describe the Coleridge Initiative's approaches for establishing a collaborative environment around the use of federal restricted-use surveys, including the National Assessment of Educational Progress (NAEP) and Beginning Postsecondary Students (BPS) surveys, to inform policymaking.

The Transformation of CIDA (Formerly CARI)

Joanne N. Carruba

US Census Bureau

At Census the CARI (Computer Assisted Recorded Interview) system survey operations had been known to be time consuming and costly to perform. In 2020, the CARI Windows/IIS/.net based system was migrated to an Angular and Java/Weblogic/Linux technology stack as the sunset of the Windows Server operating system it once lived on took place. The original CARI Oracle database was left in place and tailored for usage along with new modifications. CARI was then immediately re-branded as the CARI Interactive Data Access System (CIDA) web tool. The new technologies allowed the application to have much more flexibility to maintain and change. Many recent enhancements to CIDA were easily accomplished to give Survey Sponsors more features to save time and provide better analytics. CIDA is generically programmed to allow for sponsor customizations/configurations. Creative ideas for CIDA enhancements came from the collaboration of the CIDA programming team along with key members of the Consumer Expenditure (CE) Survey and Survey of Income and Program Participation (SIPP) CIDA (Census and BLS) community. At this

FedCASIC session we will discuss the innovations of CIDA such as the display of variable text and recorded responses, the tailoring of items to only those needed for evaluation, the export of coding reports to Excel for further manipulation and analysis, as well as current challenges and future plans.

Development of a CARI Research Plan

Brett McBride

Bureau of Labor Statistics

The Consumer Expenditure (CE) Survey contains survey items asking about expenses that range in complexity from ”Do you own this home?” to what percent of property expenses are deducted as farm, rental, or business expenses. Computer audio-recorded interviewing (CARI) is technology that can shed light into item performance in real-world conditions. In an effort to understand the functioning of CE survey items, the Bureau of Labor Statistics (BLS) initiated a project to evaluate the recordings of survey items to understand how they are being administered and answered. This made use of an application developed by the Census Bureau, CARI Interactive Data Access System (CIDA), that facilitated evaluation objectives. In this talk, I will discuss item selection, decisions related to evaluation methods (e.g., qualitative or quantitative through behavior codes), how we arrived at the codes used, some initial lessons learned about recording, and briefly introduce the CIDA app used for evaluation purposes. In addition to learning about interviewer administration of survey items, many household surveys could also benefit from further evaluation of their survey items in the field.

Session 6C: **Data Science Applications: Text Analysis**

A Semi-Automated Nonresponse Detector (SANDS) model for open-response data

Kristen Cibelli Hibben, Zachary Smith, Benjamin Rogers, Valerie Ryan, Travis Hoppe

National Center for Health Statistics

Open-ended survey responses can be valuable because they allow respondents to provide additional information without the constraints of predetermined closed-ended options. However, open-text responses are more prone to item nonresponse and inadequate and irrelevant responses. Time and cost factors associated with processing large sets of qualitative data often further hinder the use of open-text data. To address these challenges, we developed the Semi-Automated Nonresponse Detector (SANDS) model that draws on recent technological advancements in combination with targeted human-coding. The model is based on a Bidirectional Encoder Representation from Transformers model, fine-tuned using Simple Contrastive Sentence Embedding. This powerful approach uses state-of the-art natural language processing, as opposed to previous nonresponse detection approaches that have relied exclusively on rules or regular expressions, or bag-of-words, and tend to perform less well on short pieces of text, typos, or uncommon words. We present our process of training and refining the model and summarize the results of an extensive evaluation. This included the use open-text responses from a series of web probes as case studies, comparing model results against human-coded source of truth or hand-reviewed random samples, and sensitivity and specificity calculations to quantify model performance. Open-text web probe data are from the Research and Development Survey During COVID-19 surveys created at the National Center for Health Statistics. NORC collected the data in multiple survey rounds in 2020 and 2021 using a probability-based panel representative of the US adult English-speaking non-institutionalized population. How to access the model and guidance for best practice are also discussed.

More than Meets the Eye: A Transformer Based NLP Platform for Analyzing Open-Ended Survey Responses

Stewart Jollymore, Mark Schulman

Department of Defense

The past few years have seen remarkable progress in the field of Natural Language Processing (NLP). With Google Lab's BERT in 2018 to OpenAI's GPT3 making headlines today, the Defense Personnel Research Center (DPAC) is bringing these technologies to bear on our vast holdings of open-ended survey responses. Leveraging the DoD's newest cloud computing environment, ADVANA, and DPAC's specific instance, BEACON, we are able to import large language models, fine tune them on domain specific text, and use those models in a production environment to solve Topic Modeling, Sentiment, and Named Entity Recognition problems. In this presentation we focus on the transformer-based NLP tasks of Topic Modeling using the BERTopic pipeline on all-mpnet-base-v2 model to embed the documents and Sentiment using DistilBERT, pre-trained on the Stanford Sentiment Treebank v2 (SST2) and fine-tuned on hand labels from our corpora. Our Sentiment model was compared to baseline models (linear and Naive Bayes) and found to produce better results. Our Topic models were developed using multiple surveys (e.g., Workplace Gender Relations, Status of Forces) addressing various subject matter, and the outputs were compared across these different subjects to ensure cohesion of topics.

Sociodemographic and Methodological Subgroup Correlates of Item Nonresponse to Open-Ended Web Probes

Zachary R. Smith, Kristen Cibelli Hibben, Travis Hoppe, Valerie Ryan, Ben Rogers, Paul Scanlon, Kristen Miller

National Center for Health Statistics

Open-ended web probes are often used to design and evaluate survey questions. While open-ended probes allow respondents to provide unconstrained information and can be useful in identifying patterns of question interpretation, they are also prone to insufficient response that is time-consuming to detect and remove through human coding. To address these challenges, we developed a Semi-Automated Nonresponse Detector for Surveys (SANDS) that, in contrast to existing rule-based approaches, draws on natural language processing to categorize open-ended text data as valid or likely nonresponse. For SANDS's use in question design, however, understanding its performance among subgroups is crucial. If the model performs differentially across sociodemographic or methodological subgroups, its evaluative impact will be lessened and results will need correction. This presentation examines potential model bias through systematic evaluation of sensitivity and specificity within and across subgroups in four web probes on a variety of health- and non-health-related topics. It then applies the findings to a broader assessment of the sociodemographic and methodological subgroup correlates of open-ended web probe item nonresponse. Finally, it concludes with discussion of the implications for the use of open-ended web probes in question design and evaluation.

Exploratory Analysis of Enumerator INFO-COMM Documents

Haley Hunter-Zinck, Arezou Koohi, Patrick Campanello, Kevin Holmes, Melinda Schinstock, Louis Avenilla

US Census Bureau

During the 2020 Decennial, Information Communication forms, or INFO-COMMs, were used as a communication tool between Census field and office staff to record enumeration situations ranging from access to administrative issues. INFO-COMMs are semi-structured with discrete fields for specific information, check boxes to provide annotations, and free-text fields to provide additional details. Overall,

INFO-COMM documents provide information on everything from address errors, misclassified residence types and more details on the reasons for issues such as vacancy, refusals, safety concerns, and inaccessibility. However, while some of the information is contained in structured fields, much remaining information are contained within the free text explanation, making automated extraction more difficult. In this study we look at the Group Quarters Operations. We conduct exploratory analysis of the unique data values and completeness in the form???s fields, including the free text. We then perform matching to the master address file (MAF) to examine the types of residences enumerated as well as to analyze the propagation of any address corrections to the MAF, if applicable. We automate the extraction of relevant concepts from the free text using spaCy named entity recognition (NER) models. Finally, we perform clustering and topic analysis to identify themes of issues. While this analysis is especially relevant for determining themes in documents lacking annotations, we also use the results to explore subthemes within groups of documents relating to a larger issue, such as vacancy or refusals. INFO-COMMs documents provide useful information that can be automatically extracted to update the MAF and detail themes of issues noted by enumerators.

2:45 pm – 3:00 pm, April 12

Break

1:15 pm – 2:45 pm, April 12

Session 7

Session 7A: **Data Collection Challenges and Modernizations on a Longitudinal Household Survey**

Field Interviewer Engagement in a Virtual Environment

Cali Beyer, Elena Navaro, Shannon Nelson, Catherine Haggerty

NORC at the University of Chicago

Collecting Survey of Consumer Finances (SCF) data is challenging because the interview is long, complex, and intrusive. Our field interviewers (FI) are the link between the respondents' data and the central office processing team, so training and keeping our FIs engaged is paramount. Survey trainings, local teambuilding, and motivating incentive programs are given special attention in order to recruit and retain field staff during the pandemic under a period of labor shortages and other economic challenges. Along with so many other surveys, the 2022 SCF had to adapt to a virtual training and data collection environment and rethink strategies to keep interviewers engaged. In this presentation, we will review best practices and lessons learned for field staff engagement in a virtual environment for use moving forward in a hybrid world.

The Challenges of Household Composition and Dynamics in Collecting High Quality Data

Heather Sawyer, Cali Beyer, Abby Rosenbaum, Micah Sjoblom

NORC at the University of Chicago

The "household" is a social construct that has received considerable attention from social scientists. While household arrangements change over time and as a result of shifting cultural patterns, much of survey research relies on a fixed snapshot of a household to organize and standardize data. For example, data collection on the Survey of Consumer Finances relies on the implementation of household screening procedures to identify eligible respondents. This in turn places field interviewers at the forefront of navigating the complexities of household arrangements. This paper explores the implications of diverse and shifting American households on data collection efforts on the Survey of Consumer Finances and presents approaches to sensitize field interviewers to these complexities to ensure high data quality.

Incentive Payments: Thank you Options and Respondent Challenges

Elias Kassa, John Hootman, Kate Bachtell, Frankie Duda, Taifoor Beg, Mariana Patino

NORC at the University of Chicago

The Survey of Consumer Finances (SCF) collects detailed questions about household finances. The survey takes over two hours to administer, on average, and the data we collect are considered private and confidential. While participation in the survey is voluntary and provides families with a unique opportunity to help the Federal Reserve draw an accurate picture of the financial condition of all sorts of households in the United States, most households are reluctant to participate. The SCF has used monetary incentives to thank respondents for their significant contribution to this important research. In the past the incentives were paid either in cash at the end of interviews conducted in-person or by check for those participating via telephone. Dispersing and reconciling cash to interviewers for incentive payments distracted supervisors from their primary activities. Payments by check were not immediate, and in keeping with incentives and reward programs in other contexts, we examined other means to incentivize/reward participation. During the current round of data collection, we decided to offer payments in the form of a bank transfer, a VISA money card, or a variety of electronic store cards. During the use of the new payment forms we discovered barriers for some households to access and use their incentives. This presentation describes respondent incentive preferences, lessons learned using a third-party vendor for incentive payments, and the challenges some respondents experienced.

Identifying Falsifiers: Data Quality and Use of ProofPoint for SCF 2022

Jimmy Herdegen, Kate Bachtell, Frankie Duda, Catherine Haggerty

NORC at the University of Chicago

The Survey of Consumer Finances (SCF) provides the most comprehensive dataset for tracking the personal finances of American households. These data allow policymakers and researchers to make informed monetary and financial policy decisions that influence American households, businesses, and the overall economy. To that end the accuracy of the SCF data is of the utmost importance. While the process of measuring data quality has been enhanced during previous rounds of the SCF, which included validating the first two cases and a randomly selected ten percent of an interviewer's overall caseload, measuring the length of the interview and examining the data for inconsistencies, significant enhancements have been made since the previous SCF data collection period in 2019. This includes the use of ProofPoint, a proprietary quality assurance monitoring system for field data collection. Through SAS Visual Analytics a team of data scientists, IT and survey staff designed a systematic way to incorporate numerous variables to monitor data quality and present these data in a dashboard for ease of use by Field and Central Office staff. In this presentation we will review this new expansion to data quality monitoring and the impacts of adding this new system to the validation process.

Session 7B: **Progress, Challenges, and Opportunities in Implementing a Nationwide, Large-scale Digital Research Platform for Precision Medicine** (Roundtable)

Romuladus E. Azuine[1], Nakia Mack1, Lew Berman[1], Chris Lunt[1], James McClain[1], Izzy Seo[1], Danielle Wilfong[2], Gage Rion[2], Mark Begale[3]

[1]National Institutes of Health, [2]Vanderbilt University, [3]Vibrent Health

An increasing number of biomedical and public health research platforms are establishing or migrating to cloud-based research computing environments (CRCE). Among other benefits, CRCE platforms are cost-effective, they offer enhanced security, and provide researchers with unlimited resources for data computation, storage, and management. CRCE platforms provide equitable access to diverse researchers who may have differential institutional access to big data and computational resources. The NIH All of Us

Research Program is an innovative program that seeks to collect data from at least one million people and make the data available on its CRCE platform for public health and biomedical research discoveries. The program has enrolled 578,000 participants. Data types available from participants include survey data (n=372,000), electronic health records (n=280,000), physical measurements (n=306,000), wearable health data (n=12,800), and genomics data including whole genome sequences (n=98,000). We propose an interactive roundtable discussion session to share the progress, challenges, and opportunities for innovation in implementing a nationwide, large-scale digital research program, managing central data curating, and enabling data access for researchers on an elastic, safe, and secure CRCE. There will be 3 interrelated presentations. 1) Changing the Future of Public Health and Biomedical Research through Cloud-based Computing: Lessons From the NIH All of Us Research Program; 2) Opportunities and Challenges in Implementing a Large-Scale Participant Digital Research Data Collection; and 3) Working with the NIH All of Us Research Program Researcher Workbench: A Guided Tour.

End of Day 2 Program

The schedule is in Eastern Daylight Time

If you have any questions, please contact the 2023 FedCASIC Planning Committee (fedcasic@census.gov)