RESEARCH REPORT SERIES
(Survey Methodology #2020-01)

Issue Paper on Disclosure Review for Information Products with Qualitative Research Findings

Joanne Pascale
Diane K. Willimack
Nancy Bates
Joanna Fane Lineback
Paul C. Beatty

Center for Behavioral Science Methods
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report issued: March, 2020

Disclaimer: This report is released to inform interested parties of research and to encourage discussion of work in progress. Any views expressed are those of the authors and not those of the U.S. Census Bureau.

Abstract: Procedures for disclosure avoidance have long been in place for quantitative data, and are continuing to evolve in response to real-world changes that threaten data security. In particular, advances in computing and access to large datasets combined have the potential, if left unchecked, to facilitate data linkage and identification of research participants. These real-world changes not only pose challenges for quantitative research products; they raise new questions about qualitative research products that were previously considered less vulnerable to such attacks. In June 2018 the issue of disclosure for qualitative findings was discussed at a meeting of the Census Bureau's Data Stewardship Executive Policy (DSEP) Committee. In a separate but related development, in September 2018, the question of disclosure avoidance in qualitative research came to light in response to a legal discovery request that involved release of focus group transcripts. At that time the Census Bureau's Disclosure Review Board (DRB) issued interim guidelines for de-identifying focus group transcript summaries. The Census Bureau then convened a small group of senior survey methodologists who regularly conduct qualitative research. The team's purpose was to draft an issue paper on disclosure avoidance procedures for qualitative research in particular, and to recommend guidelines for implementation. After nearly a year in development the team produced a set of Disclosure Avoidance Guidelines for Qualitative Research that was fully vetted by multiple divisions within the bureau, and ultimately approved by DSEP. This paper documents that process along with the guidelines themselves.

Keywords: disclosure avoidance, qualitative research, de-identification

Suggested Citation: Joanne Pascale, Diane K. Willimack, Nancy Bates, Joanna Fane Lineback, Paul C. Beatty (2020). **Issue Paper on Disclosure Review for Information Products with Qualitative Research Findings.** *Research and Methodology Directorate, Center for Behavioral Science Methods Research Report Series (Survey Methodology #2020-01).* U.S. Census Bureau. Available online at
<http://www.census.gov/content/dam/Census/library/working-papers/2020/adrm/rsm2020-01.pdf>

Acknowledgements: The authors gratefully acknowledge substantive contributions by Census Bureau colleagues William C. Davie, Jr., Amy Newman-Smith and Amy Anderson Riemer in the Economic Programs Directorate. We also acknowledge useful review comments and feedback from members of the Disclosure Review Board, the Methodology and Standards Council, the Center for Enterprise Dissemination – Disclosure Avoidance, and the Data Stewardship Executive Policy Committee – in particular Phil Steele, Laura McKenna, Gary Benedetto, Simson Garfinkel and Rob Sienkiewitz. Finally, we express particular appreciation to John Abowd, Associate Director for Research & Methodology, for championing the development of these guidelines, their supporting documentation, and implementation.

A. BACKGROUND

The Census Bureau's Quality Standards state that all "information products" (working papers, conference presentations, book chapters and journal publications) must be reviewed for "disclosure avoidance" – the practice of protecting respondents' privacy and confidentiality (see Figure 1). Procedures for disclosure avoidance have long been in place for quantitative research, and are continuing to evolve (Abowd, 2018a; Abowd, 2018b). As stated in a recent blog by Census Bureau Chief Scientist John Abowd, "Historical methods cannot completely defend against the threats posed by today's technology. Growth in computing power, advances in mathematics, and easy access to large, public databases pose a significant threat to confidentiality. These forces have made it possible for sophisticated users to ferret out common data points between databases using only our published statistics. If left unchecked, those users might be able to stitch together these common threads to identify the people or businesses behind the statistics." (Abowd, 2018a; Abowd, 2018b). This practice of ferreting out data points to identify the people or businesses behind the statistics is sometimes referred to as "re-identification."

These real-world changes not only pose challenges for quantitative research products; they raise new questions about qualitative research products that were previously considered less vulnerable to such attacks. The question we take up here is: how to develop and adapt disclosure avoidance procedures for information products generated from qualitative research methods.

Qualitative research is commonly used in the social and behavioral sciences to study and understand human social behavior. It is characterized by in-depth interviews and interactions with small, often purposive, samples of human subjects, identified and selected using criteria associated with research goals. Thus statistical inference to a target population is not the aim, nor is it appropriate. Rather, Census Bureau survey research methodologists routinely use qualitative research methods for pretesting and evaluating survey questions, data collection instruments, and related communication materials (such as advance letters), in order to assess and improve the question-answer process to ensure the highest data quality achievable while minimizing respondent burden. Qualitative research methods commonly used at the Census Bureau across its portfolio of surveys include, but are not limited to, the following methods[1:]

- Focus groups
- Cognitive interviews
- Debriefings with interviewers and respondents
- Usability testing
- Exploratory or scoping interviews and consultations

We know of no written guidelines or standards at the Census Bureau with regard to disclosure avoidance for qualitative research specifically. However, at least with regard to cognitive interviews, in October of 2016 the Office of Management and Budget issued an addendum to Directive No. 2 (Office of Management and Budget, 2016) that requires federal agencies to document findings such that results are transparent and replicable (see Figure 2).

[1] These and other pretesting methods are described briefly in an inventory of question and questionnaire evaluation methods prepared by the Federal Committee on Statistical Methodology (Office of Management and Budget, 2016).

Figure 1. Excerpt from US Census Bureau Quality Standards (Reissued Jul 2013)

Statistical Quality Standard E3: Reviewing Information Products (page 98)

Requirement E3-1: All Census Bureau information products must be reviewed before release to ensure that disclosure avoidance techniques necessary to prevent unauthorized release of protected information or administratively restricted information have been implemented completely and correctly. Information protected by federal law (e.g., Title 13, Title 15, and Title 26) and by the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) is covered by this requirement. (Statistical Quality Standard S1, *Protecting Confidentiality*, addresses disclosure avoidance techniques.)

Sub-Requirement E3-1.1: The Census Bureau's Disclosure Review Board (DRB) procedures must be followed for information products that use data protected by Title 13 to prevent unauthorized release of protected information or administratively restricted information, particularly personally identifiable information or business identifiable information. (See the DRB Intranet Web site for further guidance and procedures.)

Source: Thomas L. Mesenbourg, J., Potok, N. A., Jackson, A. A., Vitrano, F. A., Johnson, T. A., Jost, S. J., Wright, T. (Reissued Jul 2013). U.S. Census Bureau. Statistical Quality Standards. Washington: U.S. Census Bureau. Retrieved 2019, from https://www.census.gov/content/dam/Census/about/about-the-bureau/policies_and_notices/quality/statistical-quality-standards/Quality_Standards.pdf

Figure 2. OMB Directive No. 2 Addendum

"Under the *Budget and Accounting Procedures Act* of 1950 (31 U.S.C. 1104 (d)) and the P*aperwork Reduction Act* of 1995 (44 U.S.C. 3504 (e)), the Office of Management and Budget (OMB) is issuing Statistical Policy Directive No. 2 Addendum: *Standards and Guidelines for Statistical Surveys.*

Section A.5 Transparent Analysis

Standard A.5: Analysis of cognitive interviews must be transparent such that study findings can be traced to original data collected in the cognitive interviews.

The analytic process must be transparent so that an outsider can understand and assess the legitimacy of study findings. Each step in the analytic process must be documented in a clear and accessible way, such that the findings can be traced directly back to the raw data. The level of detail at which the analytic process is described must be such that an outside researcher could replicate the analysis.

By making analytic processes transparent, readers can understand, cross-examine and judge the quality of the cognitive interview data as well as the way in which the analysis was conducted. Transparency allows the reader to trust the findings and their reputability.

Source: OPM, (2016). Statistical Policy Directive No. 2: Standards and Guidelines for Statistical Surveys; Addendum: Standards and Guidelines for Cognitive Interviews. Federal Registry, 81, 70586-70587. Washington, DC. Page 70586

In June 2018 the issue of disclosure for qualitative findings was discussed at a meeting of the Data Stewardship Executive Policy (DSEP) Committee. Specifically, the question of "noise infusion" for qualitative reports was raised. The Center for Survey Measurement (now the Center for Behavioral Science Methods, CBSM) submitted comments requesting a waiver to noise infusion procedures for qualitative research. CBSM was then requested to head up an agency-wide effort to craft an "issue paper" on the subject (see Appendix A for meeting notes). In a separate but related development, in October 2018, the question of disclosure avoidance in qualitative research came to light in response to a legal discovery request that involved release of focus group

transcripts. At that time the Census Bureau's Disclosure Review Board (DRB) issued interim guidelines for de-identifying transcript summaries (see Appendix B). Staff in CBSM have combined these two related issues and made it our task to draft an issue paper on disclosure avoidance procedures for qualitative research, and recommend guidelines for implementation.

To put this task into the broader context of disclosure avoidance for all data products at the Census Bureau, we note that we are aware of the current debate on the need to supplement/supplant conventional methods of disclosure avoidance (e.g. data swapping, noise infusion, cell suppression) with more modern techniques (e.g., differential privacy). We suggest that, with the exception of cell collapsing, these traditional methods are not suitable, from a technical standpoint, for qualitative reports involving small numbers of cases. In typical frequency tables in qualitative reports, cell sizes do not meet minimum requirements for applying these disclosure avoidance methods without losing or altering meaning in the published results.

Nevertheless, we fully recognize the need for disclosure avoidance procedures for qualitative research. In keeping with the outcome of the June 2018 DSEP meeting – that guidelines be developed that can be applied across the bureau – our counterparts in the Economic Statistical Methods Division (ESMD) joined this effort early on, to ensure that characteristics of both "household" and "establishment" surveys were accommodated (we define these terms and discuss their unique implications below).

We proceed as follows. In Section B we provide a literature review, which includes a summary of communications with our counterparts at other federal statistical agencies and relevant professional associations regarding any codified standards, guidelines, or best practices for disclosure avoidance procedures for qualitative research. In Section C we begin with more detail about our starting point for our proposed guidelines – specifically, the DRB interim guidelines issued in September 2018. We then discuss how current research practices relate to those interim guidelines, and our proposal for their adaptation, separately for household and establishment surveys. In Section D, we provide a single, consolidated set of disclosure avoidance guidelines for qualitative research to aid the Disclosure Review Board (DRB) and Disclosure Avoidance Officers (DAOs) in their reviews. Section E contains examples of tables before and after these proposed disclosure avoidance guidelines are applied. The document closes with a summary in Section F. After an iterative process of vetting this draft with the DRB and other divisions and branches within the bureau that conduct qualitative research, we hope this document can serve as an industry standard.

One final note with regard to scope. These are the guidelines for decontrolling research results and summary statistics based on confidential qualitative data and declaring them safe for release. Until these disclosure avoidance standards are met and the data has been officially approved for release, raw data (e.g. interview and focus group transcripts, key quotes from respondents) and any interim products created from those data (e.g., frequency tables of respondent characteristics, interview and focus group summaries) carry the confidentiality protections afforded to the data collected under the applicable collection authority (ex. Title 13, CIPSEA, Title 26). This holds true through all of the phases of internal data reduction and analysis produced in the service of final, polished results for the working paper, journal publication, slide presentation, etc. As emphasized in the April 9, 2019 email by Ron Jarmin with the subject line "Handling Pre-Cleared Statistical

Products," (see Appendix C) these raw data and interim data products carry "the same confidentiality protections as the underlying source data and must be handled accordingly." Authors should follow the guidelines outlined in the DS007 - Safeguarding and Managing Information Policy and the associated Data Handling Guidelines with regard to protections, encrypted email, and so on, during the course of producing final information products. The guidelines proposed herein are only meant to apply to the data as represented in the final information product.

B. LITERATURE REVIEW

We first consulted with our counterparts in other federal agencies (including the Bureau of Labor Statistics (BLS), the National Center for Health Statistics, and the USDA National Agricultural Statistics Service). For the most part we found only general guidance, such as removing names and other personally-identifiable information (PII). However, in late April 2019, BLS released a 2-page document from its DRB which hinges on "k-anonymity" methodology, as described in Samarati and Sweeney (1998). We then turned to the National Institutes of Health (NIH), given their vast reach in terms of disciplines and the sheer volume of research conducted and funded under their purview. In November, 2010, the NIH Office of Behavioral and Social Sciences Research (OBSSR) initiated a plan to develop guidelines for conducting "mixed methods" research employing both quantitative and qualitative methods. They convened a working group of 19 individuals – "experienced scientists, research methodologists, and NIH health scientists….representing fields such as public health, medicine, mental health professions, psychology, sociology, anthropology, social work, education, and nursing." (National Institutes of Health, n.d.). Their charge was to develop a set of best practices for "how to rigorously develop and evaluate mixed methods research applications" (Creswell, Klassen, Plano Clark, & Smith, 2011, p. 1). In April 2011 the report was released and includes extensive practical guidelines for how to go about the research, and a statement that the "researcher needs to justify the need for gathering identifying information and put safeguards in place for protecting that information" (Creswell, Klassen, Plano Clark, & Smith, 2011, p. 24). But again, we found nothing in this report that provided specifics on how researchers or reviewers are to evaluate those protections.

Regarding human subjects research in particular, the NIH established a policy on "Certificates of Confidentiality" for NIH-funded and conducted research, which "protect the privacy of subjects by limiting the disclosure of identifiable, sensitive information" (National Institutes of Health (NIH), 2017). On October 1, 2017, the NIH updated its policy for issuing these certificates for all biomedical, behavioral, clinical, or other research. The policy places prohibitions on "disclosure of the names of research participants or any information, documents, or biospecimens that contain identifiable, sensitive information collected or used in research…The term 'identifiable, sensitive information' means information about an individual that is gathered or used during the course of biomedical, behavioral, clinical, or other research, where the following may occur: an individual is identified; or for which there is at least a very small risk, that some combination of the information, a request for the information, and other available data sources could be used to deduce the identity of an individual." The certificates "protect the privacy of subjects by limiting the disclosure of identifiable, sensitive information" (National Institutes of Health (NIH), 2017). So while the NIH is adjusting to real-world changes and putting policies in place to protect its research

subjects against re-identification, we found nothing in NIH resources that provide any guidance to investigators, grantees or reviewers regarding how these protections should be ensured.

Next, we looked toward professional organizations – specifically the American Association of Public Opinion Research (AAPOR), the American Statistical Association, the American Sociological Association and the American Anthropological Association (AAA). In general, their codes of ethics address key themes of informed consent, non-disclosure of PII, secure storage, handling and destruction of PII, and language about avoiding re-identification. However, we found no specific guidance on how to achieve these aims. For example, AAPOR's code of ethics states that members are expected to "act in accordance with all relevant best practices, laws, regulations, and data owner rules governing the handling and storage of such information" by restricting "access to identifiers and destroy them as soon as they are no longer required" and that researchers not "disclose any information that could be used, alone or in combination with other reasonably available information, to identify participants with their data, without participant permission." (American Association of Pubic Opinion Research, 2015, p. 2). The American Statistical Association's 2018 revision of the "Ethical Guidelines for Statistical Practice" states that the "ethical statistician protects the privacy and confidentiality of research subjects and data concerning them, whether obtained from the subjects directly, other persons, or existing records," among other tenets (American Statistical Association, 2018, p. 4). The American Sociological Association published a new code of ethics in 2018, which requires researchers to use "all reasonable precautions to protect the confidentiality rights of research participants" even after the death of participants, when using both primary and secondary data. It even explicitly states that "sociologists do not attempt to re-identify" data from "the collection and analysis of large scale data sets which are generated through technology and internet activities." (American Sociological Association, 2018, p. 10). The AAA Statement on Ethics discusses minimal informed consent, stating that it "must also include establishing expectations regarding anonymity" and the protection of records (American Anthropological Association, 2012). More specifically, it states that researchers:

- have an ethical responsibility to take precautions that raw data and collected materials will not be used for unauthorized ends
- consider and communicate likely or foreseeable uses of collected data and materials
- consult with research participants regarding their views of generation, use and preservation of research records

In other words, the AAA statement offers guidance to researchers on how to manage *expectations* regarding respondent anonymity, but no specific steps on how to avoid respondent re-identification. In sum, across the four associations, we found no specific, technical guidance on how to ensure that the data, findings and other details in any given public-facing document protect respondents' identities.

Finally, we turned to the published literature on the topic of protecting respondent confidentiality in qualitative research and found scant practical, specific guidance. For example, in one article titled, "Protecting Respondent Confidentiality in Qualitative Research," (Kaiser, 2009) the author states:

> "Despite emphasizing the importance of maintaining confidentiality (Grinyer, 2002), the literature on research design and the ethical codes of professional associations offer virtually no specific, practical guidance on disguising respondents' identities and

preventing deductive disclosure in qualitative research." (Giordano et al., 2007; Wiles et al., 2008)"

Other published articles on the subject reiterate much of the above content from professional associations regarding informed consent and reasonable efforts to protect confidentiality, but no practical, specific guidance on how.

Given our fairly extensive due diligence, it appears that we may be in the position of crafting a set of methods for non-disclosure of qualitative research "from scratch." Of course we cannot be sure that our canvassing of other agencies and/or the literature was comprehensive. However, in the interest of advancing the field, we believe it will be a useful exercise to put pen to paper and develop this draft, circulate widely for comment, and hone a set of practical guidelines that can be implemented by a broad range of researchers conducting qualitative research.

C. ADAPTATION OF THE DISCLOSURE REVIEW BOARD INTERIM GUIDELINES

The interim guidelines issued in September 2018, by the Disclosure Review Board consist of one main guiding principle and eight specific directives for how to adhere to that guiding principle, abbreviated here and shown verbatim in Appendix B:

Guiding Principle: Documents must be de-identified so that no individual or establishment can be identified within the text.

Directives:

1. Remove all identifying information, including people names, street numbers, street names, city names, zip codes, place names, or names such as a building or establishment, that identifies a group of people smaller than the population of the smallest US state (i.e. Wyoming)….Any city, county, or minor civil division with population at least the size of the smallest US state may be identified by name.
2. Remove all dates within direct quotes…named holiday…
3. Remove all proper nouns…including sports teams, local schools, business names…
4. Descriptive demographic information about participants such as gender, age, immigration status, rank, titles, and income, can be provided only if it cannot be combined with other information that could uniquely identify participants.
5. Remove all information that could be linked with news reports or publicly available databases, such as accident specifics, drug names and interactions and medical conditions.
6. Remove any photos.
7. Remove any video or voice recordings.
8. Direct quotes are acceptable for release as long as they do not contain uniquely identifiable information that would allow for re-identification.

With regard to qualitative research in general, Census Bureau researchers have historically adhered to the main guiding principle and seven of the eight directives (2 through 8) as a matter of course. Where historical practice and these interim guidelines diverge is mainly in the first directive on geographic identifiers. Most qualitative research aims to capture a certain range and diversity of respondent characteristics, in order to provide context for interpreting and assessing research results. Reports often demonstrate this range and diversity to the reader by indicating the site of the data collection (city, state or region of the country) and by providing a frequency table on the

characteristics of research participants. Due to the small sample sizes, and the methods of identifying and recruiting sample in qualitative research, we propose an alternative that meets the goals of directive #1 but in a different way than the methods used for quantitative research.

Below we discuss the specifics on adapting this first directive given the difference in the nature of quantitative and qualitative sample and methods and their respective disclosure risks. Because the details of qualitative methodology – particularly with regard to the nature of their target populations – varies for household versus establishment surveys, we discuss each of them separately (note that we use the term "household" as a short-hand to apply to demographic surveys more broadly and it includes, for example, surveys of teachers and principals). In household surveys the target population is generally straightforward – (e.g., children under age 5, adults 18 or over) living in households. In establishment surveys, the target population is skewed – that is, a *few* very large businesses account for a substantial proportion of economic activity, while *many* small businesses individually contribute very little economic activity but *collectively* their numbers make up a large proportion of the target population. For our purposes, we adopt the Cox and Chinnappa (1995) definition of "establishment surveys" – surveys where the target population consists of businesses, organizations, institutions, or public sector entities made up of one or more individual establishments, or physical locations – in order to capture the variety of entities they encompass. Additionally, establishment surveys require one or more people (known as "data reporters") to provide data about the establishment on its behalf. We briefly describe the qualitative research methodologies used for each type of survey, and note where each has unique considerations.

Further we note that, for both household and establishment surveys, the guidelines set forth below apply only to qualitative studies. If a given report also includes quantitative data (e.g., estimates of total sales or employment, summary statistics about the companies such as median employment or median sales), then the report must be submitted through the usual established channels for DAO/DRB disclosure avoidance for quantitative reports.

C.1 Household Surveys

For the vast majority of qualitative research conducted for household surveys (e.g., the American Community Survey, the National Health Interview Survey), participants are not drawn from a probability sample, but from a range of other recruiting sources, the most common of which are:

- Advertisements in Craigslist, newspapers and other media;
- Postings on listservs (e.g., NextDoor);
- Flyers, email blasts and other announcements circulated by agencies and institutions with ties to the target population (e.g., senior centers for studies on aging); and
- The CBSM database of respondents, which is a compilation of respondents screened over the years who have contacted the CBSM recruiter via the above outreach methods (note DSMD plans to maintain a similar database).

Potential test participants are generally recruited and screened over the phone to see if they meet the study's particular eligibility criteria.

Given the sources and the method of recruiting and selecting sample, there is no sample frame or universe in a statistical sense. Thus, we will use the term "pool" to describe the group from which

participants are recruited – e.g. a typical pool would be respondents to a Craigslist ad in City X. With regard to disclosure avoidance, we suggest that unlike a true sample frame, this type of pool is not observable to nor replicable by an intruder. Hence the relevant factor to consider in protecting participants' identities is the size of the population from which participants were selected *that match their characteristics (including geographic area disclosed)* – NOT the size of the small set of people who were actually selected. In other words, the absolute number of participants in the study is not what matters; it is the number of participants *relative* to the number of others with their same characteristics in the sample pool. For example, an individual sample member with a common characteristic (e.g., female) drawn from respondents to a Craigslist ad placed in a city of 100,000 people would not be identifiable. However, providing more detail about that same individual sample member who happens to have a less common characteristic (e.g., female, Native Hawaiian), *and* divulging the name of the city could risk disclosing her identity, depending on how many female Native Hawaiians live in that city.

Reports generally include some details about participants in both the methods and the results sections. The methods section typically includes a frequency table showing the basic demographic characteristics of the recruits, such as age range, sex, education level and race to signal to the reader the range and diversity of the participants. In the results section, authors typically discuss participants' reaction to the stimuli in the test (e.g., "Some test subjects interpreted Term XYZ to mean A while others interpreted it to mean B."). Both types of content risk respondent disclosure and we discuss each in turn.

For the frequency table, while there are some exceptions, for most reports a fairly standard set of demographic characteristics is sufficient to convey to the reader the nature of the sample. For this reason, we propose a template approach to creating and evaluating the frequency table. To the extent authors can adhere to displaying only the characteristics in the template (or a subset of characteristics, as needed), without losing any detail that is meaningful for the reader, the disclosure review process can, in theory, be fairly streamlined. Figure 3 displays the full range of characteristics and categories in scope for the template. In the event that characteristics do not appear in Figure 3, then the item must come before the DRB for review. The DRB retains the right to add to this list, as necessary. For the evaluation of cell sizes of the frequency table, we propose guidelines that satisfy the main guiding principle and directive #1 above by relying on principles drawn from current Census Bureau disclosure avoidance practice:

1. Site of data collection can be divulged only if it comprises 600,000 or more people (based on the interim DRB guidelines using the "smallest state" standard noted above).
2. In tables displaying demographic characteristics of respondents, cell counts for the geographic area mentioned in #1 will be evaluated based on the American Community Survey (ACS) 5-year estimates that correspond to the most recent period for the final year of research. In addition, ACS estimates for any given cell of the report should reflect at least 10,000 weighted cases (based on guidelines articulated in "Legacy Techniques and Current Research in Disclosure Avoidance at the US Census Bureau" by Laura McKenna, Matthew Haubach, Caroline Mak, and Christopher Kuang, draft issued September 28, 2018, page 11).
3. Cells should be collapsed until the requirements in #1 and #2 are met, as verified by data.census.gov (formerly American FactFinder or AFF).

Figure 3: Template for Table of Demographic Characteristics of Respondents in Qualitative Studies of Household Surveys

CHARACTERISTIC	CELL SIZE
Sex	
Female	
Male	
Age	
Under 5	
5-17	
18-24	
25-44	
45-54	
55-64	
65-74	
75-84	
85 and older	
Race	
One Race	
White alone	
Black or African American alone	
American Indian and Alaska Native alone	
Asian alone	
Native Hawaiian or Other Pacific Islander alone	
Some other race alone	
Two or more races	
Hispanic or Latino origin (of any race	
Yes	
No	
Average household size	
Number of household members	
1	
2	
3	
4	
5	
6	
7 or more	
Presence of Children in Household	
One or more children under 18	
No children under 18	
Marital Status	
Never married	
Married	
Divorced/separated	
Widowed	
Education Attainment	
Less than high school	
High school graduate	
Some college, no degree	
Bachelor's	
Graduate or professional degree	
Language Spoken at Home	
English only	

CHARACTERISTIC	CELL SIZE
Language other than English	
Speak English less than "very well"	
Nativity	
Native	
U.S. born	
Outside U.S. (e.g., Puerto Rico)	
Foreign born	
Europe	
Asia	
Africa	
Oceania	
Latin America	
Northern America	
Employment Status	
Employed	
Unemployed	
Not in labor force	
Annual Household Income	
Less than $15,000	
$15,000 - $24,999	
$25,000 - $49,999	
$50,000 - $99,999	
$100,000 or more	
Housing Tenure	
Owned	
Rented	
Electronic Device Ownership	
Desktop/laptop	
Smartphone	
Tablet	
Other computer	
Internet Availability	
With subscription	
Without subscription	
No internet	

NOTE: Category break-outs are based on published frequency tables on the American Community Survey as found in data.census.gov.

With regard to the results section, we propose that the level of detail in the text, combined with other data points offered, be evaluated for risk of re-identification on a case-by-case basis, considering the "pool" as discussed above. For example, in the scenarios described above, if the author had submitted a report with the following in the results section: "One participant – female, Native Hawaiian – interpreted Term XYZ to mean A," AND in another part of the report had provided the city, the DRB/DAO would check data.census.gov to determine how many Native Hawaiian females lived in that city. A likely outcome of disclosure review would be to provide the author with two choices regarding the way the sample member is described: drop the specific geographic identifier (city) and provide only the state or region of the country, or maintain the name of the city but drop the Native Hawaiian detail.

C.2 Establishment Surveys

Economic programs at the Census Bureau, for the most part, draw upon sample surveys of business, organizational, or governmental units for developing statistical estimates of economic indicators and other official economic statistics. These statistics are used by a wide range of data users and stakeholders to measure and monitor the health of the U.S. economy, and contribute to the estimation of U.S. Gross Domestic Product (GDP).

Like household surveys, establishment surveys use qualitative research methods for pretesting survey questions, data collection instruments, survey instructions, and associated communication materials. These methods also permit qualitative researchers to take advantage of in-depth interaction with business and organizational respondents to investigate and suggest suitable data collection and communication strategies, along with gauging potential respondent burden.

While many of the qualitative research methods used by establishment survey methodologists are identical to, or adapted from, methods described for household surveys, some research procedures used for establishment surveys differ in important and necessary ways from their household survey counterparts. This is due to the highly skewed nature of establishment survey target populations. Thus, the concept of the "pool" of potential respondents described earlier for household surveys is not appropriate in establishment surveys, because cases selected based on research criteria are, in fact, known to the qualitative researchers.

For establishment surveys, pretesting participants are often recruited from prior establishment survey respondents, and are selected based on criteria specified to meet research needs/goals. Establishments that meet these criteria are identified using information (such as data reporters' contact information) that is already known from other sources:

- The Census Bureau's Business Register, a comprehensive database that contains business identifiable information (BII), such as names, addresses, employer identification numbers, telephone numbers, or email addresses, along with associated data protected by Title 13 (the Census Bureau's authorizing legislation) and Title 26 (IRS data protections);
- Sampling frame or response information from completed surveys, also protected by Titles 13 and 26;
- Outside sources such as company websites, professional networking sites (e.g., LinkedIn), business association lists, publicly available business information web-sites (e.g., Dunn and Bradstreet), or other publicly available lists of businesses or government entities; and
- The Census Bureau's Governments' Master Address File, which contains information for state and local government bodies and other public sector entities (however, these data are, by law, public information, and thus are not protected by Title 13).

Similar to household surveys, potential test participants are generally recruited and screened over the phone to see if they meet the study's particular eligibility criteria. For establishment surveys, often high-level business characteristics, such as industry, size, organization type, and geography, are of interest.

Additionally, establishments cannot "speak" for themselves in order to answer survey questions; instead, a data reporter (usually an employee, owner, or contract agent), must provide data about the establishment on its behalf. It is important to note that these data reporters are rarely, in and of themselves, the subject of a survey inquiry; they are incidental to the establishment survey data collection process. Nevertheless, some of their personal characteristics (e.g., sex, age, education level, position/job title and tenure, and experience with information technology) may be associated with the quality of the reported data, and thus may be of interest to survey researchers using qualitative methods.

Consequently, establishment surveys may require somewhat different attention and precautions with regard to disclosure avoidance. Pretest cases selected from this skewed target population can create challenges in how to strike the balance between informing a reader of the nature of the participants in the study and protecting the establishment's identity, because in some industries very few establishments dominate. Additionally, data reporters' personal characteristics, when combined with characteristics of the establishment, may enable re-identification of either or both the data reporter and the establishment.

Thus, de-identification of establishments in qualitative research reports requires attention to characteristics of both the establishment and the data reporter, and it is usually their combination that may lead to disclosure of businesses' identifiable information. The primary variables that could lead to disclosure in establishment survey qualitative research reports are:

- size (in terms of number of establishments, number of employees, and/or the monetary value of payroll and/or revenue)
- type (as defined by industrial classification)
- organizational structure (businesses with only one location versus those with many locations)
- geography and/or
- PII of the data reporter (race, gender, age, job title, etc.)

In addition, because IRS Federal Tax Information (FTI) is sometimes used to inform sampling, data collection, and statistical estimation procedures, establishment surveys may be subject to provisions of Title 26, U.S. Code. This includes associated regulatory requirements set forth in IRS Pub 1075 (Internal Revenue Service, 2016), which provides direction to governmental entities that acquire and use tax return data for statistical purposes. Restrictions of IRS Pub 1075 impacting disclosure avoidance requirements are included in the specific guidelines below.

Due to these considerations, it is infeasible to provide establishment survey qualitative researchers with a template parallel to Figure 3 that would display the characteristics of the participating establishments. Rather, we advise authors to lessen the potential for re-identification by reducing the reported level of granularity of associated variables, and by avoiding association of data reporters' personal characteristics with specific business identifiable information. Example actions include, but are not limited to, the following:

1. Avoid identifying specific cities and their vicinities, or small, localized geographic areas where research was conducted, particularly for industries that are geographically concentrated. An industry that is geographically concentrated is one where a large

proportion of its total number of establishments are physically located in a single/few small or well-specified geographic areas, or where only one/few enterprises dominate the industry in terms of employment, payroll, and/or the value of goods and services produced.

2. Avoid industry descriptions at a level where specific establishments may be readily identified, particularly in industries that are highly concentrated. Industries may be identified at the lowest level of granularity relevant to study goals, to the degree that individual BII cannot be discerned, separately, or in combination with other variables. Generally, disclosure avoidance should be achievable at the "sector-level" as defined by the NAICS system for "2-digit" industries[2] for economy-wide studies.
3. Avoid associating specific position or job titles of data reporters with BII, such that either or both the data reporter and the establishment may be identified.

For DAOs of establishment surveys, we propose guidelines for qualitative research that, like household surveys, rely on principles drawn from current Census Bureau disclosure avoidance practice:

1. If BII may be discerned by associating specific findings with a given level of granularity for any of the variables noted above, separately or in combination, then such specific values must be removed, or stated in a generalized manner and/or the level of granularity must be raised. Specifically, based on Pub 1075 and the "Rule of 3" (a long-standing practice used for establishment surveys at the Census Bureau) counts cited in the report must meet the following:
   - Counts at the national level must contain at least three units of analysis (e.g. company, establishment, project).
   - Counts at the state level must contain at least 10 units of analysis.
   - Counts at the sub-state level must contain at least 20 units of analysis.
   - If not, then counts must be collapsed to the next largest geographic level or North American Industry Classification System (NAICS) level, or any other subcategory until the required number is achieved.
   - If this additional detail is needed, the more detailed counts could be retained in a separate table in an appendix that would be deemed administratively confidential, labeled accordingly, and thus cannot be publicly released.

2. Personal characteristics of participating data reporters should not be described in a manner or at a level of detail that, when associated with individual or summarized data about the establishment, may reveal their personal or business identifiable information.

D. DISCLOSURE AVOIDANCE GUIDELINES FOR QUALITATIVE RESEARCH

We propose both general principles and specific steps for authors and DAOs to follow for disclosure avoidance. The five general principles are as follows:

1. Authors are to be provided with the disclosure guidelines in Figure 4 (inclusive of the template in Figure 3) so they are fully aware of the kinds of information that could raise concerns for

[2] The NAICS groups industries into a total of 20 sectors (e.g., construction, real estate, retail trade) and assigns each with a 2-digit code. Within each sector the NAICS further categorizes industries into subsector, industry group, etc., and with each level of detail an additional digit is assigned, for a total of 6 digits to categorize industries. For further details see https://www.census.gov/eos/www/naics/.

respondent re-identification, and so that they can exercise due diligence in complying with the guidelines. This should help authors in their deliberations over whether to include certain details in the draft report before it goes to the DAO.

2. If authors feel the integrity of the research would be compromised by strict adherence to the guidelines in Figure 4 (e.g., for household survey reports, the template in Figure 3; for establishment survey reports, the "Rule of 3") they are encouraged to include a note for DAO reviewers as to the rationale. For example, if authors of household survey reports wish to include frequencies of respondent characteristics in addition to those shown in Figure 3 – either as a row in a table or in the text itself – they should note this for the DAO and it will be discussed on a case-by-case basis.
3. For household survey reports, the researcher, DAOs and DRB will be trained to check submitted tables against data.census.gov. If any cell size is insufficient for that geographic area, steps will be taken to collapse the cell. Accountability for compliance with the guidelines is shared among the researcher, the DAOs and the DRB.
4. Authors and reviewers should closely examine data discussed in the *body* of the paper's text – in addition to tables – to assess whether an individual or establishment could be re-identified. If the level of detail in the text, combined with other data points offered, risks re-identification, then it should be redacted. Given the idiosyncrasies of establishment surveys and associated qualitative research, this principle may be even more important in establishment survey reports than for their household counterpart reports.
5. DAOs and authors are urged to collaborate towards workable solutions that adhere to the guidelines, without compromising the meaning of the findings, in order to reduce the need for redaction or reduction of detail in reports. If authors feel the integrity of the research is greatly enhanced by providing more detail than the guidelines advise, the added details would be considered on a case-by-cases basis.

Figure 4 displays our proposed comprehensive guidelines on practical steps to be taken by DAOs responsible for reviewing qualitative research information products. The one guiding principle and directives #5 and #8 are taken, verbatim, from the interim DRB guidelines issued in September 2018. Directive #1 is a more succinct version of the general approach laid out in Section C above. In directive #2 and #3, we changed the term "focus group" to "study" to broaden the scope. In directive #4 we dropped the requirement to create a table and linkage variables, given that procedures for presenting and reviewing tabular data is now articulated in directive #1. In directives #6 and #7 we elaborate on photos and videos and explain they must be removed if they enable identification of an individual. The reason for this detail is that in many reports (usability in particular) it can be important to include images of individuals interacting with the instrument (e.g., on a mobile phone or laptop screen). In these images, the main focus is the question as displayed in the instrument, and sometimes a hand or fingers are part of the image, but no faces or other images sufficient to identify an individual. Finally, we added directive #9 in response to review by DAOs in the Economic Directorate.

Figure 4: Disclosure Avoidance Guidelines for Qualitative Research

It is the responsibility of the DRB to assure that documents are sufficiently redacted such that the confidentiality requirements of Title 13, Title 26, and the Confidential Information Protection and Statistical Efficient Act, ("CIPSEA") are observed. Specifically, such documents must be de-identified so that no individual or establishment can be identified within the text. At minimum, the guidance below should be followed:

- Remove all personal or business identifying information (PII/BII), including people names, street numbers, street names, city names, ZIP codes, place names, names of buildings or establishments, legal names of business/organizational establishments or entities, as well as "doing business as" names, or any specific information that may permit linkage to previously reported data or other data in Census Bureau databases protected by Title 13/26, such as the Business Register. In addition:

For household surveys:

- The site of data collection can be divulged only if it comprises 600,000 or more people
- In tables displaying demographic characteristics of respondents, researchers should use the template displayed in Figure 3. Cell counts for the geographic area mentioned in #1 will be evaluated based on the most recent American Community Survey (ACS) estimates. Specifically, ACS estimates for any given cell of the report should reflect at least 10,000 weighted cases. In cases where the margin of error results in the ACS estimate straddling 10,000, the more conservative lower bound estimate should be used.
- Cells should be collapsed until the requirements in #1 and #2 are met, as verified by data.census.gov (formerly American FactFinder or AFF).

For establishment surveys:

- If BII may be discerned by associating specific findings with a given level of granularity for any of the variables noted above, separately or in combination, then such specific values must be removed, or stated in a generalized manner and/or the level of granularity must be raised. Specifically, based on Pub 1075 (Internal Revenue Service, 2016) and the "Rule of 3," counts cited in the report must meet the following:
- Counts at the national level must contain at least three units of analysis (e.g. company, establishment, project)
- Counts at the state level must contain at least 10 units of analysis
- Counts at the sub-state level must contain at least 20 units of analysis
- If not, then counts must be collapsed to the next largest geographic level or North American Industry Classification System (NAICS) level, or any other subcategory until the required number is achieved.
- If this additional detail is needed, the more detailed counts could be retained in a separate table in an appendix that would be deemed administratively confidential, labeled accordingly, and thus cannot be publicly released.
- Personal characteristics of participating data reporters should not be described in a manner or at a level of detail that, when associated with individual or summarized data about the establishment, may reveal their personal or business identifiable information.

2. Remove all dates within direct quotes, replacing them with [D] or [D:DATE]; named holidays should be replaced with a [D] or [D:HOLIDAY]. Dates that are not within direct quotes can be released as MONTH and YEAR, including the MONTH and YEAR of the study, provided that they cannot be used, alone or in combination with other data, to identify a participant or the location of a study.
3. Remove all proper nouns associated with participants in the study, including sports teams, local schools, business names, and similar information. These should be replaced with a [D].
4. Descriptive demographic information about participants such as gender, age, immigration status, rank, titles, and income, can be provided only if it cannot be combined with other information that could uniquely identify participants.
5. Remove all information that could be linked with news reports or publicly available databases, such as accident specifics, drug names and interactions and medical conditions.
6. Remove any photos that enable an individual to be identified.
7. Remove any voice recordings and any video that enable an individual to be identified.
8. Direct quotes are acceptable for release as long as they do not contain uniquely identifiable information that would allow for re-identification.
9. Excerpts or screenshots of data collection instruments, questionnaires, or associated materials are acceptable, however, only fictional data may be displayed along with a disclaimer to that effect.

E. POTENTIAL RISKS

In terms of risk, for both household and establishment surveys, the risk of re-identification is minimized somewhat by the very nature of the findings, which rarely require exposition of quantitative, numerical response data, specific responses to survey questions, or other information that might be associated with PII or BII. Rather, the key results in most qualitative research reports summarize respondents' verbal statements describing their behaviors, and it is unlikely that this kind of content can be found in existing datasets that could be linked to the participants. For example, typical reports divulge respondents' understanding of the intent of specific questions, their interpretation of key terms and phrases in the question, and their strategies for estimating constructs such as household income (note it is the *strategy for estimating* and NOT the household income itself that is of interest).

Establishment surveys' qualitative research participants are typically recruited using information from the Business Register, and may be respondents to specific surveys. Thus, their frame information is knowable, unlike HH survey's "pool of volunteers." Nevertheless, like HH surveys, disclosure risk to research participants (the businesses) is minimal because of the behavioral and descriptive nature of the qualitative findings. Although establishment surveys themselves typically request factual, often financial, data, and responses may be based on businesses' financial records, the response process is rarely replicated during qualitative interviews, because accessing records is often time-consuming. Instead, researchers obtain verbal descriptions of the steps respondents *would* take (hypothetically) in order to answer our questions. If a respondent were to provide numeric data during an interview, such information would not be quoted or included in draft or final reports. If strategic or proprietary information might be needed in a report in order to adequately provide meaningful findings, then that can be addressed by DAOs on a case-by-case basis.

Finally, we note that toward the end of our research and development efforts we identified a set of guidelines for disclosure avoidance from the health field, specifically with regard to HIPAA (the Health Insurance Portability and Accountability Act passed by Congress in 1996). Those guidelines are entirely consistent with the guidelines we are proposing; they are just not as comprehensive and do not go as far in specifying how to conduct disclosure review. They do, however, call for non-disclosure of the usual types of PII (names, ID numbers), geography smaller than a state, "full face" photos, and biomarkers (including voice and fingerprints). The guidelines also call for disclosure review by an "expert" (akin to our DAO specialists) and state that "The expert must be a person with appropriate knowledge and experience of using generally accepted statistical and scientific principles and methods for removing or altering information to ensure that it is no longer individually identifiable. When those methods and principles have been applied, the expert must determine that the risk of reidentification of an individual is very small." With regard to the term "very small" the guidelines state "…HIPAA does not define the level of risk of re-identification other than to say it should be 'very small'. The expert should define 'very small' in relation to the context of the data set, the specific environment, and the ability of an anticipated recipient to be able to reidentify individuals. Experts may come from a number of different fields and do not require any specific qualifications. What is important is experts have experience of deidentifying data. It is that experience that regulators will look at in the event of an audit, not specific qualifications or certifications." (HIPAA Journal, 2017)

F. EXAMPLES

We demonstrate how these proposed guidelines would be applied with a set of examples all using hypothetical/fictional data. First, on the household side, Table 1a displays hypothetical data meant to represent the demographic characteristics of respondents in a typical report on qualitative testing submitted for DAO review. Columns, rows and cells that are problematic are highlighted, and the problems with each are described beneath the table. Table 1b shows the resulting table after redaction by the DAO, and beneath the table is a description of how and why the problematic columns, rows and cells were modified. On the establishment survey side, Tables 2 and 3 show various levels of geography and disclosure recommendations. Tables 4a and 4b illustrate a typical table of characteristics of establishments in sample before and after redaction.

Table 1a. Demographic Characteristics of Cognitive Interview Respondents BEFORE Redaction

Demographics	All Counties	Mont-gomery	Prince George's	Baltimore	Anne Arundel	Howard	Harford
Total	165	40	35	30	25	20	15
Sex							
Male	84	20	18	15	13	10	8
Female	81	20	17	15	12	10	7
Age							
18-24	12	3	3	2	2	1	1
25-44	27	7	6	5	4	3	2
45-54	45	10	9	8	7	6	5
55-64	42	10	8	8	6	6	4
65-74	27	7	6	5	4	3	2
75-84	8	2	2	1	1	1	1
85+	4	1	1	1	1	0	0
Race							
White alone	100	25	22	14	15	14	10
Black alone	40	10	8	10	4	4	4
AIAN alone	2	0	1	0	1	0	0
Asian alone	10	2	2	2	2	1	1
NHPI alone	3	0	1	1	0	1	0
Some Other Race	2	1	0	1	1	0	0
Two or More Races	8	2	1	3	2	0	0
Hispanic Origin							
Yes	25	8	7	4	3	2	1
No	140	32	28	26	22	18	14
Education							
Less than high school	15	5	4	3	2	1	0
High school graduate	25	8	7	4	3	2	1
Some college	35	8	7	6	5	5	4
Bachelor's	40	9	8	7	6	6	4
Graduate degree	50	10	9	10	9	6	6
Sexual Orientation							
Straight/heterosexual	136	32	29	26	19	18	12
LGBTQ	29	8	6	4	6	2	3

- Columns in dark grey cannot be displayed individually because each of these three counties has a population below 600,000, based on ACS 5-year estimates:
  - Anne Arundel County, total population = 564,600
  - Howard County, total population = 312,495
  - Harford County, total population = 250,132

  However, when combined, the population of the three counties adds up to 1,127,227 which is above the 600,000 threshold, so data from the three counties can be combined.
- Rows in light grey cannot be displayed because this question is not asked in the ACS and hence the characteristic is not in data.census.gov and cell sizes cannot be verified.
- Cells with thick borders cannot be displayed, because the weighted estimates for these subgroups within their respective geographies are less than 10,000, based on ACS 5-year

estimates. (The following numbers are fictional for the sake of the example. In practice, the team would utilize the direct weighted count estimates from the ACS 5-year):

- Montgomery County, total population = 1,039,198
  - AIAN (American Indian/Alaska Native): 2,078
  - NHPI (Native Hawaiian/Other Pacific Islander): 0
- Prince George's County, total population = 905,161
  - Age 85+: 9,052
  - AIAN: 2,715
  - NHPI: 0
- Baltimore County, total pop = 828,637
  - AIAN: 1,657
  - NHPI: 0
  - Some Other Race: 8,286

Table 1b. Demographic Characteristics of Cognitive Interview Respondents AFTER Redaction

Demographics	All Counties	Montgomery	Prince George's	Baltimore	Ann Arundel, Howard and Harford Counties
Total	165	40	35	30	60
Sex					
Male	84	20	18	15	31
Female	81	20	17	15	29
Age					
18-24	12	3	3	2	4
25-44	27	7	6	5	9
45-54	45	10	9	8	18
55-64	42	10	8	8	16
65-74	27	7	6	5	9
75+	12	3	3	2	4
Race					
White alone	100	25	22	14	39
Black alone	40	10	8	10	12
Asian alone	10	2	2	2	4
Any Other Race alone	7	1	2	2	3
Two or More Races	8	2	1	3	2
Hispanic Origin					
Yes	25	8	7	4	6
No	140	32	28	26	54
Education					
Less than high school	15	5	4	3	3
High school graduate	25	8	7	4	6
Some college	35	8	7	6	14
Bachelor's	40	9	8	7	16
Graduate degree	50	10	9	10	21

- The columns in dark grey Table 1 have been combined into a single column in dark grey in Table 2 so that the combined geographies have a population over 600,000. The column has been relabeled accordingly.
- Rows with cells from Table 1 with thick borders have been combined with other categories to the new rows in Table 2 with thick borders so that the weighted estimates for combined subgroups are greater than 10,000
    - New age group row ("75+") combines totals for age groups 75-84 and 85+
    - New race group row ("Any Other Race") combines totals for race groups American Indian/Alaska Native, Native Hawaiian/Other Pacific Islander, and Some Other Race
- The row in light grey in Table 1, for sexual orientation, was taken on a case-basis. Given the existing literature on prevalence of LGBTQ and the size of the counties (even the combined "other county" column) the DAO determined that this population is too small to be included in the table. Authors were advised to instead state in the text of the methodology section that both straight/heterosexual and LGBTQ respondents participated in the study, but not to provide specific numbers of participants.

Table 2. Number of Interviews per State

State	# of interviews
California	9
Texas	3
Michigan	2
Ohio	1
Massachusetts	1
TOTAL	16

Issue with the table: The number of interviews for each state is less than 10.
Disclosure Recommendation: The disclosure recommendation is to omit this table from the publicly released report and, instead, describe the nature of the sample in more general terms with text such as: "Participants for the 16 interviews were from California, Texas, Michigan, Ohio, and Massachusetts."

Table 3. Number of Interviews per Location

Location	# of Interviews
Los Angeles, CA MSA	9
Houston, TX MSA	4
Dallas, TX MSA	3

Issue with the table: The number of interviews is less than 20 for each MSA and less than 10 for each state.

Disclosure Recommendation: The disclosure recommendation is to omit this table from the publicly released report and say something like in the text and, instead, describe the nature of the sample in more general terms with text such as: "Participants for the interviews were from Los Angeles, CA, Houston, TX, and Dallas, TX."

Table 4a. Number of Interviews by Primary NAICS Code of Company BEFORE Redaction

NAICS Code	Description	# of Interviews
325411	Medicinal and Botanical Manufacturing	2
331210	Iron and Steel Pipe and Tube Manufacturing from Purchased Steel	1
333249	Other Industrial Machinery Manufacturing	2
334220	Radio and Television Broadcasting and Wireless Communications Equipment Manufacturing	2
334310	Audio and Video Equipment Manufacturing	1
334413	Semiconductor and Related Device Manufacturing	1
335314	Relay and Industrial Control Manufacturing	2
335921	Fiber Optic Cable Manufacturing	2
335931	Current-Carrying Wiring Device Manufacturing	2
335999	All Other Miscellaneous Electrical Equipment and Component Manufacturing	2
336214	Travel Trailer and Camper Manufacturing	2
336390	Other Motor Vehicle Parts Manufacturing	2

Issue with the table: The number of interviews for each six-digit NAICS is less than three.
Disclosure Recommendations: The disclosure recommendation is to collapse the industries to the 3-digit level and for the cells that still do not have at least 3 cases, group those together. If collapsing does not make sense or does not add any value then omit the table from the publicly released report.

Table 4b. Number of Interviews by Primary NAICS Subsector Code of Company AFTER Redaction

NAICS Code	Description	# of Interviews
334	Computer and Electronic Manufacturing	4
335	Electrical Equipment, Appliance, and Component Manufacturing	8
336	Transportation Equipment Manufacturing	4
325, 331, 333	Chemical, Primary Metal, or Machinery Manufacturing	5

G. VETTING, CLEARANCE AND APPROVALS

At the inception of this effort we reached out to seasoned staff in the area of disclosure avoidance research and methods, and members of the DRB, who provided valuable advice, technical assistance and guidance during the development stage. We also solicited input from Disclosure Avoidance Officers, who have the task of regularly reviewing reports with qualitative findings. Once the guidelines and documentation of their development began to take shape, they were vetted throughout the bureau. Our first formal presentation of the draft guidelines was to the Methodology and Standards Council, which consists of the heads of statistical methodology groups in various program areas throughout the bureau. Their role is to advise on policy and issues affecting research

and methodology, to ensure the use of sound statistical methods and practices, and to facilitate communication and coordination of statistical methodology and research throughout the Census Bureau and the broader statistical community (Thomas L. Mesenbourg, et al., Reissued Jul 2013). Council members distributed the proposed guidelines to their respective staffs with the aim of ensuring that all staff across the bureau who conduct and/or review qualitative research had the opportunity to comment. Staff in several divisions specializing in the areas of decennial statistical studies, demographic statistical methods, statistical research and methods, economic statistical methods, field operations and survey and behavioral science methods were targeted. Comments were received from more than a dozen staff members, whose feedback was incorporated. The full DRB then provided their formal recommendation and finally on October 10, 2019, the guidelines were approved by the DSEP.

Our goal with this project was to bring to light any existing practices or guidelines in the area of disclosure avoidance for qualitative research findings. We canvassed our counterparts at other research agencies, explored materials offered by professional associations, reviewed the published literature and fully vetted the draft guidelines with a broad range of research scientists (including career professionals in the field of disclosure avoidance) and research practitioners who conduct qualitative research on a regular basis at the Census Bureau. Given this due diligence, we believe this set of guidelines may be the first of its kind in terms of specific, written procedures for disclosure avoidance of qualitative research findings. We offer it as a baseline and encourage debate, discussion and refinement.

REFERENCES

Abowd, J. M. (2018a, August 17). Protecting the Confidentiality of America's Statistics: Adopting Modern Disclosure Avoidance Methods at the Census Bureau. *Census Blogs: Research Matters*. Retrieved from Census Blogs: Research Matters: https://www.census.gov/newsroom/blogs/research-matters/2018/08/protecting_the_confi.html

Abowd, J. M. (2018b, September 4). *Protecting the Confidentiality of America's Statistics: Ensuring Confidentiality and Fitness-for-Use.* Retrieved from Census Blogs: Research Matters: https://www.census.gov/newsroom/blogs/research-matters/2018/08/protecting_the_confi0.html

American Anthropological Association. (2012). *AAA Statement on Ethics.* Retrieved from American Anthropological Association: https://www.americananthro.org/LearnAndTeach/Content.aspx?ItemNumber=22869

American Association of Pubic Opinion Research. (2015). *The Code of Professional Ethics and Practices.* Washington, DC: American Association of Public Opinion Research. Retrieved February 13, 2019, from https://www.aapor.org/Standards-Ethics/AAPOR-Code-of-Ethics/AAPOR_Code_Accepted_Version_11302015.aspx

American Sociological Association. (2018, June). Code of Ethics. Washington, DC: American Sociological Association. Retrieved from http://www.asanet.org/sites/default/files/asa_code_of_ethics-june2018.pdf

American Statistical Association. (2018). *Ethical Guidelines for Statistical Practice.* Alexandria, VA: American Statistical Association. Retrieved March 12, 2019, from https://www.amstat.org/asa/files/pdfs/EthicalGuidelines.pdf

Creswell, J. W., Klassen, A. C., Plano Clark, V. L., & Smith, K. C. (2011). *Best Practices for Mixed Methods Research in the Health Sciences.* Office of Behavioral and Social Sciences Research (OBSSR). Washington, DC: National Institutes of Health. Retrieved from https://obssr.od.nih.gov/wp-content/uploads/2016/02/Best_Practices_for_Mixed_Methods_Research.pdf

HIPAA Journal. (2017, October 18). *De-identification of Protected Health Information: How to Anonymize PHI.* Retrieved from https://www.hipaajournal.com/de-identification-protected-health-information/

Kaiser, K. (2009). Protecting Respondent Confidentiality in Qualitative Research. *Qualitative Health Research, 19*(11), 1632–1641. Retrieved from https://doi.org/10.1177/1049732309350879

National Institutes of Health (NIH). (2017, September 7). Notice of Changes to NIH Policy for Issuing Certificates of Confidentiality. *Notice Number: NOT-OD-17-109*. Washington, DC. Retrieved from https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-109.html

National Institutes of Health. (2017, October 2). *Certificates of Confidentiality: Background Information*. Retrieved November 23, 2018, from National Institutes of Health: https://humansubjects.nih.gov/coc/background

National Institutes of Health. (n.d.). *Mixed Methods Research*. Retrieved January 2019, from Office of Behavioral and Social Sciences Research (OBSSR): https://obssr.od.nih.gov/training/online-training-resources/mixed-methods-research/

Office of Management and Budget. (2016, October 12). Statistical Policy Directive No. 2: Standards and Guidelines for Statistical Surveys; Addendum: Standards and Guidelines for Cognitive Interviews. *Federal Registry, 81*, 70586-70587. Washington, DC. Retrieved from https://www.govinfo.gov/content/pkg/FR-2016-10-12/pdf/2016-24607.pdf

Thomas L. Mesenbourg, J., Potok, N. A., Jackson, A. A., Vitrano, F. A., Johnson, T. A., Jost, S. J., . . . Wright, T. (Reissued Jul 2013). *U.S. Census Bureau. Statistical Quality Standards.* Washington: U.S. Census Bureau. Retrieved 2019, from https://www.census.gov/content/dam/Census/about/about-the-bureau/policies_and_notices/quality/statistical-quality-standards/Quality_Standards.pdf

APPENDIX A
Meeting Record of the Data Stewardship Executive Policy Committee
June 7, 2018

Background:

During the winter and spring of 2018, many people in program areas had questions, concerns, and misperceptions about recent DSEP decisions on disclosure avoidance. This was especially the case surrounding the decision to suspend data releases for sub-state geographies if those data releases had not been developed using vetted noise-infusion tools or approved formal privacy methods. Several program areas had questions about progress and timelines on the modernization efforts.

In response, the chair of DSEP sent a memorandum to all division chiefs on April 26, 2018 explaining recent DSEP actions on disclosure avoidance and inviting written comments. The memorandum also explained that DSEP would invite all interested parties to a special meeting where DSEP would respond to comments. Nine divisions sent in comments by the May 3, 2018 deadline, with three of the divisions simply acknowledging receipt or that they had no comments. The Associate Director for Research and Methodology created a draft memorandum to the Chair responding to all comments and invited discussion at the June 7 DSEP meeting.

Discussion:

DSEP discussed the comments and responses largely in the order that DSEP had received the comments.

- Center for Survey Measurement (CSM) CSM
    - CSM requested a waiver from noise-infusion/formal privacy requirements for the publication of summaries of the characteristics of focus group participants and other qualitative evaluation tools used by CSM and other areas.
        - ADRM clarified when such studies may publish the counts of persons in the study by broad demographic characteristics without noise infusion and that noise infusion would be unnecessary on the first set of DRB-approved tabulations made on the underlying confidential data. However, there were outstanding questions on how to handle subsequent retabulations of data from these sources.
        - DSEP asked CSM to engage other programs doing focus group/qualitative studies and prepare an issue paper outlining their dissemination models and questions regarding the applicability of the noise-infusion/formal privacy rules for review by the DRB.

APPENDIX B

INTERIM Guidelines for De-Identifying Transcript Summaries Census Bureau Disclosure Review Board; September 20, 2018

Guidelines

Transcript Summaries (including Transcript Summaries, Transcript Narratives, Audience Summary Reports, and After Action Reports) are internal analytical documents, usually based on focus groups, interviews or other qualitative methods that were not initially intended for release to the public. Such reports are Title 13 Sensitive Controlled Unclassified Information, and may not be released until they have cleared a review by the Disclosure Review Board. Whether the request for public release is part of a Freedom of Information Act (FOIA) request, legal discovery production, communication with a contractor not working in an approved location with Special Sworn Status, or for any other purpose, a DRB clearance release number must be obtained and displayed on the document. It is the responsibility of the DRB to assure that the documents are sufficiently redacted such that the confidentiality requirements of Title 13, Title 26, and the Confidential Information Protection and Statistical Efficient Act, ("CIPSEA") are observed.

Specifically, such documents must be de-identified so that no individual or establishment can be identified within the text. At minimum, the guidance below should be followed:

1. Remove all identifying information, including people names, street numbers, street names, city names, zip codes, place names, or names such as a building or establishment, that identifies a group of people smaller than the population of the smallest US state (i.e. Wyoming). These should be replaced with a [D], with the meaning to be inferred by context. If necessary, generalized information can appear after the D, as shown in the examples below. Any city, county, or minor civil division with population at least the size of the smallest US state may be identified by name.

2. Remove all dates within direct quotes, replacing them with [D] or [D:DATE]; named holidays should be replaced with a [D] or [D:HOLIDAY]. Dates that are not within direct quotes can be released as MONTH and YEAR, including the MONTH and YEAR of the focus group, provided that they cannot be used, alone or in combination with other data, to identify a participant or the location of a focus group.

3. Remove all proper nouns associated with participants in focus groups, including sports teams, local schools, business names, and similar information. These should be replaced with a [D].

4. Descriptive demographic information about participants such as gender, age, immigration status, rank, titles, and income, can be provided only if it cannot be combined with other information that could uniquely identify participants. If it can be, replace this information with a [D]. If this information is important for the data release, it should be explicitly coded in a separate table and linked with a linkage variable. Both the table and the linkage must be explicitly approved by the DRB using rules for the release of microdata.

5. Remove all information that could be linked with news reports or publicly available databases, such as accident specifics, drug names and interactions and medical conditions.

6. Remove any photos.

7. Remove any video or voice recordings.

8. Direct quotes are acceptable for release as long as they do not contain uniquely identifiable information that would allow for re-identification.

These guidelines explicitly do not apply to full transcripts. These guidelines also do not apply to other documents such as reports intended for release. In released reports, it is sometimes necessary to provide details (e.g., additional demographic or descriptive variables) that would be difficult to do using these guidelines. Guidance for such documents that allow a DAO or the DRB to assure that the details provided do not allow unique identification will be forthcoming.

The DRB does not allow the use of automated software for de-identifying transcript summaries; evaluations of automated free format text de-identification tools in 2012 and 2013 found that the systems were not sufficiently reliable to support automated de-identification of free-format medical text for public release (See NISTIR 8053, pp. 30-31). Thus, all redactions must be done manually and verified by a second person.

Example #1: De-identification of a transcript narrative:
Original text: The focus group was held on January 10, 2015, in an office building in Flint, Michigan and consisted of 3 white men, 2 African-American men, 4 white women, and 3 African-American women.
De-Identified Text: The focus group was held in on [JANUARY, 2015], in an office building in [D:MIDWESTERN CITY, POPULATION LESS THAN 100,000] and consisted of 3 white men, 2 African-American men, 4 white women, and 3 African-American women.

Example #2:
Original text: “Due to the politics of this country, I doubt we would be targeted. I'm a white woman in a house of white men. I'm secular. I can’t imagine any kind of ethnic or religious targeting.”
No de-identification is required.

References:

BoB, a best-of-breed automated text de-identification system for VHA clinical documents, developed by the Meystre Lab at the University of Utah School of Medicine. http://meystrelab.org/automated-ehr-text-de-identification/
BoB, a best-of-breed automated text de-identification system for VHA clinical documents. Ferrández O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. J Am Med Inform Assoc. 2013 Jan 1;20(1):77-83. doi: 10.1136/amiajnl-2012-001020. Epub 2012 Sep 4.

APPENDIX C

Handling Pre-Cleared Statistical Products
BOC BROADCAST (CENSUS/DIR)
Tue 4/9/2019, 11:55 AM

To: All HQ Staff
From the Desk of** Ron **Jarmin, Deputy Director and Chief Operating Officer...

As we approach the May 31 deadline to complete the mandatory 2019 Data Stewardship Awareness training, I thought it would be a good time to remind everyone of the U.S. Census Bureau's policy on handling statistical products that have not been cleared by a Disclosure Avoidance Officer (DAO) or the Disclosure Review Board (DRB), and issued a clearance number.

A DAO with bypass authority or the DRB must approve the release of any microdata, tabular data, or other statistical product derived from a confidential source (ex. Title 13, Title 26, CIPSEA, or comingled). Until a product is approved for release, it carries the same confidentiality protections as the underlying source data and must be handled accordingly, including the requirement that every page of the document contain the header or watermark "Title 13 Sensitive Information" (appropriately adjusted if another law covers the confidentiality protection).

Products that still carry the protections of Title 13, 26, or CIPSEA cannot be sent in the body of an email, as an unencrypted attachment, or be stored or transmitted in the Office 365 environment including SharePoint Online (uscensus.sharepoint.gov) and OneDrive. They should only be shared through properly secured network drives, on-premises SharePoint (collab.ecm.census.gov) sites with appropriately restricted permissions, or transmitted through an approved encryption method such as the Department of Commerce Accellion software as an encrypted attachment (How to Use Accellion).

Products that do have a DRB or DAO-bypass clearance number should display that number; otherwise, a BOC CIRT will trigger.

Please direct any questions you have regarding the clearance of data products to your DAO. You can find your DAO and other information on the disclosure review process on the DRB's Intranet Site. If you have any questions about using Accellion, contact the Policy Coordination Office at pco.policy.office@census.gov.