

## Appendix H: Data Accessibility Overview Presentation

### **Expanding Record-Keeping Study Methodology to Assess Structure and Availability of Data in Business Records**

**Melissa A. Cidade**

**Diane K. Willimack**

**Kristin Stettler**

**Demetria V. Hanna**

Sixth International Conference on Establishment Surveys (ICES VI)  
Thursday, June 17, 2021

Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

The U.S. Census Bureau has reviewed this data product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied. (Approval ID: CBDRB-FY21-ESMD001-016)

Shape  
your future  
START HERE >

United States  
**Census**  
2020

Thank you very much to my colleague, Diane Willimack, for organizing this exciting session. I am eager to share with you some of the important work we have been doing at the Census Bureau regarding measurement error and economic surveys, and what a forum to do it in!

I am Melissa Cidade, a survey methodologist in the Census Bureau's Data Collection and Methodology Research Branch in the Economic Statistical Methods Division. I have had the pleasure of working on an extensive redesign of our economic surveys in collaboration with my co-authors: Diane Willimack, Kristin Stettler, and Demi Hanna. During this presentation, I am going to walk you through some of the methodological innovations we have developed over the past two years or so in support of this redesign.



A few years back, the Census Bureau enlisted the National Academies of Sciences – referred to as naz – to systematically review our annual economic surveys. This panel was charged with providing recommendations to improve the “relevance and accuracy of the data, reduce respondent burden, incorporate alternative sources of data where appropriate, and streamline and standardize Census Bureau processes and methods across surveys” (NAS 2016: 6).

On your screen now are the surveys that were in-scope for that review. You can see that for the most part, the Census Bureau has used a sector-driven approach to survey development – note that manufacturing and services have their own set of surveys, trade has a set of surveys, and so on. One of the recommendations from the NAS panel is the implementation of an Annual Business Survey System – which has evolved into the Integrated Annual Survey, a streamlined, cross-sector, harmonized survey instrument designed to lower respondent burden while still achieving high quality, timely data in the service of the American economy.

Driving the development of the Integrated Annual Survey has been a portfolio of research projects to bring together disparate sources of data to one survey instrument. This presentation will provide an overview of the innovative methods we

have used to understand the record-keeping practices of businesses, in order to develop this streamlined instrument.

Citation: National Academies of Sciences, Engineering, and Medicine. (2018). *Reengineering the Census Bureau's Annual Economic Surveys*. Washington, DC: The National Academies Press. doi: <https://doi.org/10.17226/25098>.

## Slide 2

---

**MAC(F1** Reviewers: We recognize that the Integrated Annual Survey may have a name change before the date of this presentation and will update accordingly.

Melissa A Cidade (CENSUS/ESMD FED), 4/28/2021

## Research Questions:

1. **Definitions:** how do businesses define themselves relative to the Census Bureau definitions?
2. **Accessibility:** how accessible are key data points at varying business units?
3. **Burden:** how resource intensive is gathering data at these varying business units?

Throughout the research period, we were guided by a few key research concepts and questions. First, we were interested in how businesses defined themselves, both internally and relative to Census Bureau definitions. This included the business' units of operation, industry, and other key identifiers. We were also driven to understand how accessible data were at differing levels within a company – that is, could respondents get the data to the level of granularity we were asking with minimal effort and maximum accuracy? Finally, as with all of our data collections, we asked about the burden – or resource intensiveness – of pulling these data at various levels within the company.

This research is building on the emergent body of literature referred to as the “unit problem” – the mismatch between the administrative unit, that is, how the business sees itself, and the statistical unit, that is the standardized unit created by statistical agencies for data collection purposes. In order to minimize what van Delden et al call “unit errors” (2018) we must begin by understanding how businesses are keeping their records before we can ask them to map these records to our data requests.

Citation: van Delden, A., Lorenc, B., Struijs, P., & Zhang, L.-C. (2018). Letter to the Editor: On Statistical Unit Errors in Business Statistics. *Journal of Official Statistics*, 34(2), 573–580.

## In-Scope Businesses and Respondents

### Eligibility Criteria:

- Sampled in at least two in-scope surveys
- In at least two industrial sectors
- More than one establishment

	Phase 1	Phase 2
<b>Number of Industries*</b>		
Three or fewer	16	25
Four or more	5	5
<b>Number of establishments*</b>		
30 or fewer	9	19
31 or more	12	11
<i>*Numbers may not sum to total interviews because of missing data.</i>		

4 2020CENSUS.GOV

Shape  
your future  
START HERE >

United States  
Census  
2020

In particular, we focused on so-called “medium” sized firms with some evidence of complexity. As Snijker and Jones point out, medium sized businesses have more diverse response processes for economic surveys because of their various reporting structures and developing and diverse internal infrastructure to handle such requests (2013: 375). Additionally, medium companies had the lowest response rate to the 2017 Economic Census as a group. Because of this diversity in form and capabilities, coupled with lower response rates, understanding the response process for medium companies underpinned much of our research.

(click) On your screen now is the frequency distribution for the businesses we interviewed by each phase and by complexity characteristics. In phase 1, we interviewed 28 companies in the fall of 2019. These interviews were in-person. For phase 2, we interviewed an additional 30 companies in the winter of 2021. These interviews were all done remotely due to federal travel restrictions on account of the COVID-19 global pandemic.

### Citation:

Snijkers, Ger and Jacqui Jones. (2013). “Business Survey Communication.” in *Designing and Conducting Business Surveys*, Snijkers, G., Haraldsen, G., Jones, J., &

Willimack, D. K., eds. (2013). John Wiley & Sons, Inc.



# Phase 1 Interviewing

5

Shape  
your future  
START HERE >

United States<sup>®</sup>  
**Census**  
2020

In this next section, I'll talk about the phase 1 interviewing we conducted.

## Phase 1: The Chart of Accounts

CHART OF ACCOUNTS As of 12/31/20XX					
<b>ASSETS</b>					<b>EXPENSES</b>
100 Checking - Operating Account					500 Gross Payroll
101 Checking - Other					501 Rent
102 Retainer Checking Account					502 Interest
103 Trust Checking Account					505 Taxes - FICA Expense
105 Reserve Account					506 Taxes - DC DOES
115 Petty Cash					507 Taxes - FUTA Unemployment
116 Postage					508 Taxes - DC Personal Property
122 Assets					509 Taxes - DC Professional
130 Cost Advanced					510 Taxes - Other
					515 Repairs
TOTAL ASSETS					516 Employee Benefits
					517 Employee ABRA
<b>LIABILITIES</b>					518 Office Move
200 Accounts Payable					519 Professional Services
202 Retainers on Deposits: Contra - a/c					520 Library & Subscriptions
203 Trust A/C					521 Equipment Rental
220 US & FICA Taxes Withheld					525 Telephone Expense
221 MD/DC Taxes Withheld					526 Supplies
					527 CLIE
					528 Food
<b>INCOME</b>					530 Entertainment
301 Fees					531 Xerox
305 Interest Income					532 Dues
306 Dividend Income					533 Marketing
310 Other Income					534 Parking Expense
					535 Liability & Cas. Insurance
TOTAL INCOME					536 Travel
					537 Postage Expense
<b>EQUITY</b>					538 LEXIS Expense
Partners Equity					540 Miscellaneous
420 Capital					541 Costs Advanced - Uncollectible
					542 Depreciation Expense
					549 Charitable Contributions
					550 Other Non-Deductible Expenses
					TOTAL EXPENSES:
					TOTAL LIABILITIES & STOCKHOLDERS EQUITY

6 2020CENSUS.GOV

Shape  
your future  
START HERE >

United States  
Census  
2020

For Phase 1 interviewing, we built our interviewer protocol around a generic company chart of accounts. A *chart of accounts* (COA) is an index of all the financial *accounts* in the general ledger of a company. It is an organizational tool that provides a breakdown of all the financial transactions that a company conducted during a specific *accounting* period, broken down into subcategories

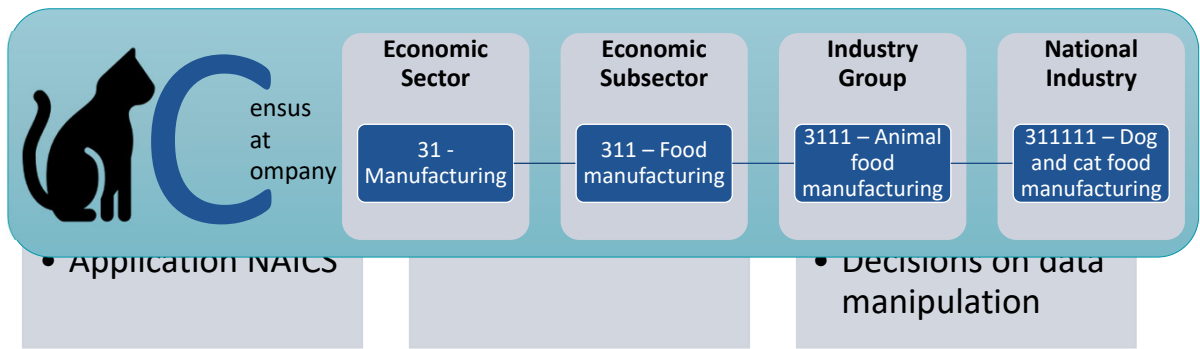
First, we showed respondents the mock chart of accounts on your screen now. We asked respondents to compare and contrast how their business is structured and maintains its records. We probed respondents on their chart of accounts relative to their company's structure, industries in which the company operates, and locations, as well as the types of software used to maintain their chart of accounts.

Once we had a better understanding of the company chart of accounts and record keeping practices, we could then ask follow-up questions about specifics within their chart of accounts. Here, we were really interested in mismatches between our understanding of how records are kept and retrieved and the questions respondents encountered on Census Bureau surveys. We centered these questions around five areas as applicable to the company:

- Business segments by industry (kind of business)

- Sales/receipts/revenues
- Inventory
- Expenses, including payroll and employment
- Capital expenditures.

## Phase 1 Findings



7 2020CENSUS.GOV

Shape  
your future  
START HERE >

United States  
Census  
2020

All companies followed a general chart of accounts with varying levels of detail. It was within these details that we made some interesting findings.

(click) Before I get too far, Let me just take a moment to talk about the North American Industry Classification System – or NAICS. NAICS is a hierarchical taxonomy with nested values – our example today is a fictitious company, the Census Cat Company. This company – at the two-digit NAICS level – is in the Manufacturing Sector. At the three-digit level, we see it is identified as a food manufacturing company, a more specific type of manufacturing. At the four-digit level – the industry group – the Census Cat Company is designated as an animal food manufacturing company. And, at the six-digit level – the national industry code – we see that it is a dog and cat food manufacturing company. Note that NAICS is transnational, used in Canada, Mexico, and the United States down to the fifth digit. The sixth digit is nation-specific so that each country can produce country-specific detail. A complete and valid NAICS code contains six digits.

(click) One of the major findings to come from this interviewing was the mismatch in unit definitions. We noted that at least seven companies may have been misclassified or may not have understood Census Bureau distinctions among classifications, for

example, a 4 digit vs 6 digit NAICS classification. We also noted that the NAICS taxonomy is unnatural for respondents; that is, because NAICS is a standardized classification system, and businesses often need more or different details in their chart of accounts, mapping records to the corresponding NAICS is challenging for some and impossible for others.

(click) The second takeaway from the first round of interviewing is that businesses are using disparate terminology to describe their various operating units. When asked about “establishments,” for example, respondents indicated that their company used a different term – such as region, office, department, line of business, and business segment – or did not track data by individual locations at all.

(click) The third finding from round one interviewing was insight into companies’ response processes. Almost all respondents indicated that completing Census Bureau surveys required more than one person in the company to respond. They also indicated that Census surveys do not match internal reporting, and are uncomfortable making decisions on how to manipulate their data to match our requests.

All three of these findings directly influenced the phase 2 interviewing.

(To access the NAICS manual, click: <https://www.census.gov/naics/>)

## Phase 2 Interviewing

8

Shape  
your future  
START HERE >

United States<sup>®</sup>  
**Census**  
2020

Taking the information we learned in phase 1, we then introduced a novel methodology to assess data accessibility.

# Definitions and Equivalencies

Company	Establishment	Line of Business
<ul style="list-style-type: none"><li>• Highest level</li><li>• Easily understood</li></ul>	<ul style="list-style-type: none"><li>• Location</li><li>• Region</li><li>• Not applicable</li></ul>	<ul style="list-style-type: none"><li>• Revenue stream</li><li>• Kind of Activity</li><li>• Not applicable</li></ul>

2020CENSUS.GOV

Pre-decisional -- Disclosure Prohibited -- Title 13 U.S.C.

Shape  
your future  
START HERE >

United States  
Census  
2020

We started by talking about the unit mismatches. This time, though, we showed respondents these three units – company, establishment, and line of business - one at a time, with corresponding definitions. We then asked them to “map” themselves to these concepts – what is the word or phrase that the business uses to mean the same thing?

<click> First, the word “company.” We meant company to be as broad and encompassing as possible, and respondents obliged. Asked what they would include or exclude from “company”, most respondents said they would not exclude any part of their business – that company is the broadest defined business unit.<click>

<click> Next was “establishment.” Here’s where we started to see things break down a little. Most respondents identified establishment to be equivalent to location. Some, however, thought of establishment as a collection of sites, akin to a region. Still a third group saw the concept of ‘establishment’ as out of scope for them. This mirrors the round 1 findings. <click>

<click> Finally, we asked about Line of Business. In this case, we were thinking that line of business would match more closely with industry. However, after the first few interviews, we noted that it did not align, and switched to asking about line of business and industry as separate concepts.

# General and Specific Industry

## NAICS works well....

"These [NAICS codes] are a good fit. Most of what we do would fall under the first one."

"[I] agree with the [given] NAICS."

"Specific industry: that's a perfect fit. General industry: that works as well, too."

## ...until it doesn't.

"Yes, I am familiar, but I hate them. They have me in warehousing, and say 'You operate the warehouse' but we do not. Not what we do."

"Somewhat familiar with. Do both at many locations. So one NAICS wouldn't apply to one location. We did not have specific NAICS looked up in advance. Mix within locations."

"All other codes are coming from a long time ago, when was non-profit. All other entities dissolved. Focused entirely on [one industry] now. Dissolved these businesses previous to 4 years ago."

2020CENSUS.GOV

Pre-decisional -- Disclosure Prohibited -- Title 13 U.S.C.

Shape  
your future  
START HERE >

United States  
Census  
2020

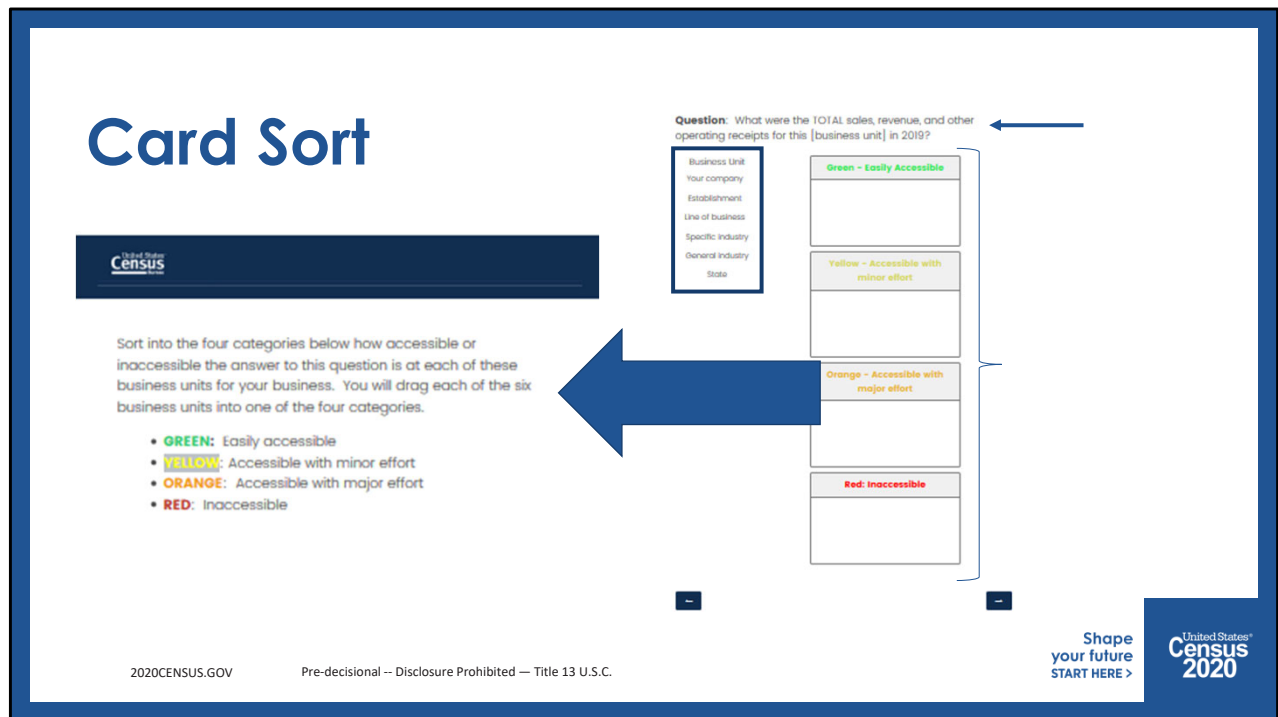
Thinking about the problem of misclassification identified in phase 1, we then asked respondents pointedly about their NAICS Classification. In this case, first, we asked about their six digit NAICS classifications, calling it their 'specific' industry. Remember that six-digits is the most specific principal business activity code we have. We then asked about the four digit NAICS classification – so, less detailed - and called it the 'general industry'. Note that the interviewers walked respondents through each of the six digit NAICS codes we could find for their company, asked for feedback or impressions, and then did the same for the four digit NAICS codes. This part of the interview was time consuming and difficult; we noticed that respondents had trouble understanding their NAICS classification, and then struggled to think of how their business units might related to their NAICS classification. Classifying a business is a critical component to collecting data on that firm, both in terms of directing respondents to the appropriate survey forms based on their classification and in terms of sampling, weighting, imputation, reporting and other important data handling techniques.

It seems that the industry classification either worked or didn't, with few falling in between: We were surprised at how the NAICS data that we had in our records was inconsistent both across and between companies.



<click> Here you can see a few examples of respondents positively reacting to their general and specific industry codes. In these cases, the NAICS we had on file made sense to respondents and fit how respondents saw their company relative to the NAICS categorization scheme.

<click> And, here are a couple of quotes where respondents struggled with the NAICS codes we have assigned them. Remember that this was after discussing the NAICS with the respondent, and having them focus in on their classifications: they still did not agree with or understand their assignments.



Once we had established company-specific working definitions for these various units, we wanted to understand how accessible the company data is at each unit for specific topics. This would help in our efforts in unit harmonization and question harmonization to move toward a streamlined data collection approach.

To do this, we extended a framework put forth by Snijkers and Arentsen (2015), who developed a four-point color coded scale as a reference for respondents when assessing the accessibility of their data at various increments, in terms of both time and organization. We operationalized the Accessibility Scale using card sort methodology by asking respondents to categorize the accessibility of data at each of the different levels of measurement by assigning each level to a color representing accessibility.

Remember that this round of interviewing took place in the winter of 2021, and so, had to be remote. We moved the card sort online - here is a screenshot of what respondents saw for the card sort.

<click> On the left of your screen is the introductory text that was repeated for each topic covered in the card sort. <click>

<click> On the right is the revenue screen – (click) you can see the general question at the top, (click) followed by four color-coded boxes, (click) and six business units nested under the heading “Business Unit.” Respondents were instructed to click on each business unit and move it to the box that corresponded with the accessibility of the requested data at that business unit. We prompted respondents to ‘think aloud’ as they moved the units to the corresponding box so that we could capture their responses and ask follow-up questions about why they categorized the data the way that they did.

Citation:

Snijker, G., Arentsen, K. (2015): Collecting Financial Data from Large Non-Financial Enterprises: a feasibility study. Paper presented at the 4<sup>th</sup> International Workshop on Business Data Collection Methodology in Washington DC September 14-16

## Defining Accessibility

### Green:

"Green means go. Green means info is available."

"Can run a report and get information."

"Green is anything I pull directly off of a financial statement that I'm already producing."

### Yellow:

"I'd probably have to reach out for help."

"I would run a new report for, but not have to do a lot of analysis and digging to find [the data], or I can modify an existing report."

### Orange:

"No one has any idea what we are looking for so they need to dig. If we don't know who to ask for it or know where to get it, but are pretty sure the data exist."

"Orange would take more effort - involving other people or creating additional reporting that we don't normally run."

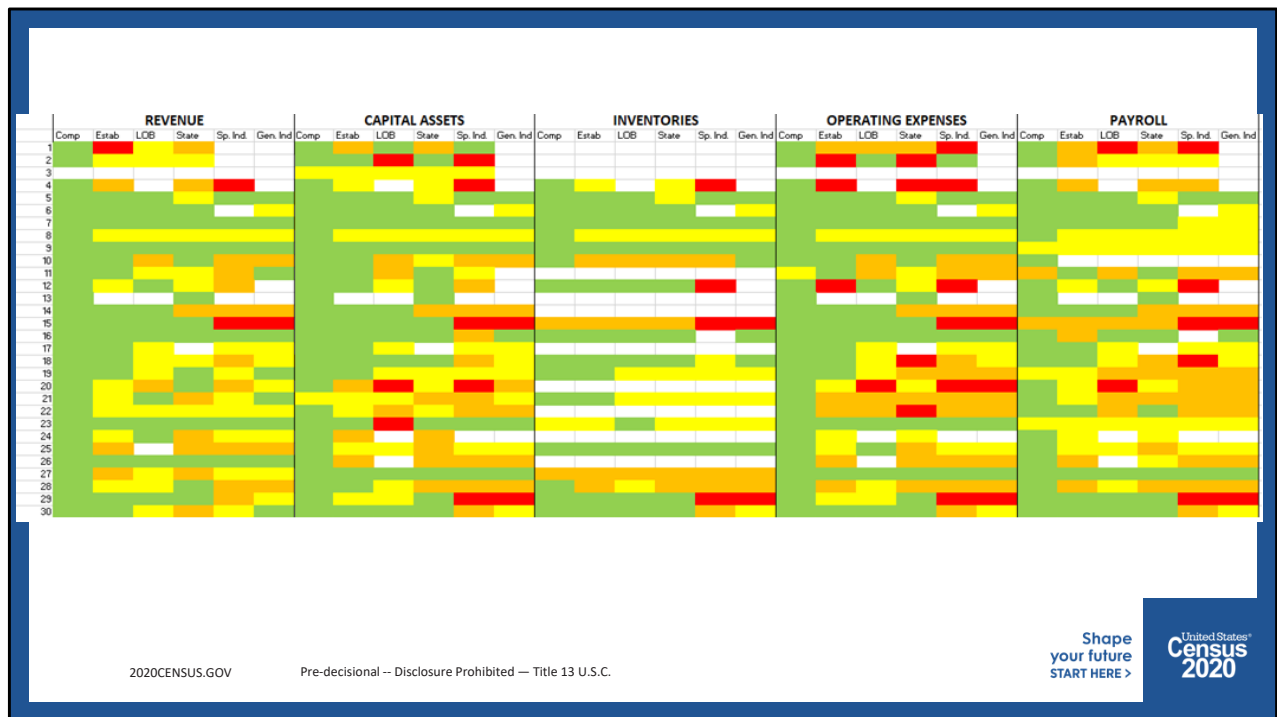
### Red:

"Red is inaccessible; there's no way for me to get that information, and it not tracked or maintained."

"Red is we just can't pull it."

Before we moved to the card sort, we needed to understand how respondents were using the categories of accessibility. On your screen are quotes from respondents setting the parameters around the different colors; we asked specifically about the differences between yellow, orange, and red. For the most part, the difference between yellow and orange is that orange involves additional people or systems, whereas yellow means modifying reporting already in place. Red, universally, means that the data are unavailable or not tracked. We ask respondents' cognition regarding these categories so that we can contextualize their responses to the card sorts.

You can think of this system as sort of like a stop light: <click> we want as many greens as we can get, followed by yellows; at orange and red, things get a little trickier.



And, here is the big reveal from our card sort exercise.

On this slide – each row is a completed interview. Each column is a business unit – company, establishment, line of business, state, specific industry, and general industry. And, each color corresponds to the accessibility of the data at that unit within that topic – grouped across the top: revenue, capital assets, inventories, operating expenses and payroll. Note that blank spaces denote where respondents were either unable to speak to this topic (that is, they were not involved in reporting on this topic), this concept didn't apply to the specific business (that is, this business didn't have inventories, for example), or the interviewer suspected that the respondent wasn't clear on the task at hand (that is, the respondent wasn't understanding the units being used or some other communication challenge).

# Card Sort: Revenue

## Question:

What were the TOTAL sales, revenue, and other operating receipts for this [business unit] in 2019?

REVENUE	GREEN	YELLOW	ORANGE	RED	X	
COMPANY	29	0	0	0	0	1
ESTABLISHMENT	15	9	3	1	1	2
LINE OF BUSINESS	13	10	2	0	0	5
STATE	13	8	7	0	2	2
SPECIFIC INDUSTRY	6	7	9	2	6	6
GENERAL INDUSTRY	9	9	4	1	7	7



2020CENSUS.GOV

Pre-decisional -- Disclosure Prohibited -- Title 13 U.S.C.

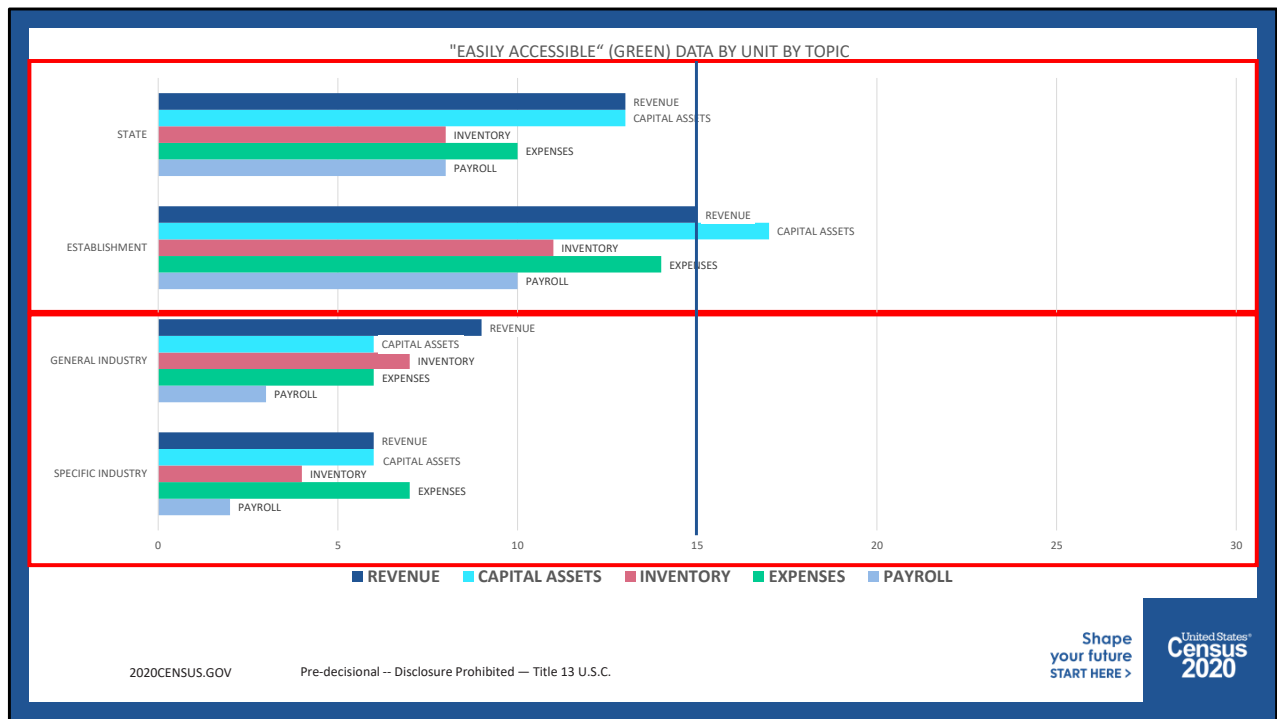
United States  
Census  
2020

One of the benefits to conducting the card sort is that the resultant data can be displayed in compelling ways. On your screen now are the results of just the revenue section.

<click>Within revenue, almost all respondents could provide data by company, the largest unit. Notice, though, as we move through the various units we see less green.<click>

<click>At the Establishment, line of business, and state, about half of respondents could provide revenue data with little to no effort (yellow or green), however, we do start to see some respondents say that it would be a 'major' effort to pull these data, especially at the state level. <click>

<click>Looking at NAICS – remember, we are using the terms 'general' and 'specific' industry - we see a decline in the number of respondents who said that revenue data were easily accessible or accessible with minor effort. This is our first piece of evidence that the data are more accessible at business units that make sense to the company, as opposed to external classifications, like NAICS.



And, because the data can be visualized numerically, we can also compare responses across business units. This chart, for example, displays the “easily accessible” (green) identified data by business unit and topic to try to suss out what is working and where. So far, we really have two groups of business units that appear to be the most accessible for the most firms.

(click) One is geographically based – states and establishments. We can see that establishment shakes out better than state – especially for revenue, capital assets, and expenses. (CLICK)

(Click) On the other hand, when we impose our taxonomy on respondents, we see the accessibility of the data decrease.

## Key Takeaways:

- Using a generic Chart of Accounts during interviewing helps to center respondents to the task at hand.
- Use cognitive methodology to give context to the resultant data.
- Card sorts can be a useful tool in establishment surveys.
- Visualization of qualitative data can have a powerful impact with stakeholders.

16 2020CENSUS.GOV

Shape  
your future  
START HERE >

United States  
Census  
2020

We are just now churning through the rich data that these almost 60 interviews produced. Since this session is focused on methodology and not findings, though, I want to cue you in on three major methodological findings of the work so far.

First, when asking about record keeping practices, we found that providing respondents with a generic Chart of Accounts helps them to understand the task at hand, and to identify and explain differences in the ways that they maintain their records.

Next, we found that leaning on tried and true cognitive testing methods provides a way of assessing content validity – that is, by asking respondents pointedly about their definitions of ‘accessibility’ and of ‘unit,’ we could then provide context for the results of these interviews.

Third, this research is the application of a method not usually used in testing establishment surveys. We found that by using the card sort, respondents were engaged in the interview. The card sort acted as a way of operationalizing the four-point scale measuring accessibility, a complex construct.



Finally, we have found with our stakeholders that the visualizations from the card sort are a captivating way to present complex interview data. We have found that even our most quantitatively-minded colleagues like the display of the qualitative data in a way that is more “rows and columns” than we usually have.

# Thank you!

**Diane K. Willimack**

[diane.k.willimack@census.gov](mailto:diane.k.willimack@census.gov)

1+ (301) 763-3538

**Demetria V. Hanna**

[demetria.v.hanna@census.gov](mailto:demetria.v.hanna@census.gov)

1+ (301) 763-3351

**Melissa A. Cidade**

[melissa.cidade@census.gov](mailto:melissa.cidade@census.gov)

1+ (301) 763-8325

**Kristin Stettler**

[kristin.j.stettler@census.gov](mailto:kristin.j.stettler@census.gov)

1+ (301) 763-7596

We will be chewing through these interviews in the next weeks and months and are excited at the insights to be gained. To that end, if you want to follow up with any of us, our contact information is listed on the screen now. Thank you very much!