# Federal Committee on Statistical Methodology Meeting

## The Role of Data Quality in a Twenty-First Century Federal Statistical System

### Director's Remarks as Prepared for Delivery

**October 25, 2022**

- Good morning, everyone. I'd like to thank the organizing committee for their invitation to address you today.
- I consider it an absolute honor to be here as your director of the U.S. Census Bureau.
- Yes, and I did say your director.
- This first year of my term I've try to say that each time I speak. I see this position very much as one that serves the public… that serves our nation.
- And that's how I'm approaching my role: as your public servant in a leadership position serving everyone—including you.
- And as part of this service, today I'll speak about one of my favorite topics—data quality.
- Specifically, I'll focus on a specific aspect of data quality that twenty-first century federal statistical agencies like the Census Bureau should think about.
- OK, just so you know so I used the conference theme to motivate my keynote title… hey, I want to be relevant.
- And by the way, keep that word *relevant* in mind for later.
- You know, it was easy to use this conference theme to generate my keynote title because it reflects my intense passion for helping people.
- It's both a strength and weakness. Ok, please allow me to explain with a little story.
- When I first came on board after being sworn in, I immediately wanted to do as many media interviews as I could.
- I wanted to help the public understand who this new kid on the block was and how he intends to lead.
- Well, of course our communications folks had something to say about that and offered some media training.
- But ok, I was totally up for it. So we meet, and after giving me a few pointers I go through my first mock interview.
- And let me tell you, I was totally proud of my performance.
- I answered every question and talked about my first days at the Census Bureau… I even answered a tough zinger with aplomb. Or so I thought.
- Well, we met the next day and the trainers showed me the story that could have emerged as a result of my mock interview.
- Maybe you guessed it by now. It was all about the darn zinger question.

- Not a sentence was devoted to my first days as director.
- What was really interesting to me, is that I'd fallen victim to my own true professional passion.
- As I've said many times, there are two things that I'm most passionate about in my career—statistics and helping people.
- In fact, a huge chunk of my career involved providing statistical design consultation.
- And I'd trained myself over the years to instinctively answer someone's question as directly as possible.
- I'd do that by using *their framing* of the question.
- And naturally I'd gently guide them to a clearer research objective so they could achieve more meaningful insights and knowledge.
- That strategy works great for research projects.
- But when it comes to media interviews, you need to be fastidious about the message you want to convey.
- If you don't, you dilute or can even obliterate that message.
- And if I'm not getting out my principal message, I'm not advancing the mission of the Census Bureau.
- Bottom line is that I learned that it takes some additional skills to do an effective interview.
- Sure, occasionally I still fall victim to my passion of answering every question in the framing of how it's posed.
- But hey, I am a work in progress! We all are, for that matter.
- Now, this notion of focusing on principle, but inadvertently straying you're your objective... well that story has relevance to *data quality*.
- And that's what I'll talk about this morning.
- Many folks know that I am particularly enamored with the concept of data quality.
- It can mean so many things to so many people.
- Perhaps it's like music... some are attracted to the beat, others to the melody, still, others find the lyrics captivating.
- And others appreciate the composition—how all the elements fit together to create a true piece of art.
- Being a live music photographer, I've pretty much seen it all: the rappers, hip hop, hard rock and metal, indie, blues and jazz, country, K-pop, folk, classical... You name it.
- And in their own way, they're all magnificent.
- But to really appreciate them, the music needs to speak to you.
- As they say, beauty is in the eye of the beholder. And perhaps that applies to data quality, as well.
- OK, well let me be real: from my perspective, it DOES apply to data quality. Think about it.
- In fact, let's think about using a framework that should be pretty familiar to you.
- Let's consider data quality using the FCSM framework published in 2020 and appropriately titled of all things, "A framework for data quality."
- In that paper, data quality has three domains:
  - Utility,
  - Objectivity, and
  - Integrity.
- *Utility* relates to how well the data address one's needs.
- Obviously, needs can change from one person to the next, so utility varies by use case.

- *Objectivity* reflects data accuracy, reliability, and error structure. Many folks—especially us statisticians—focus on this dimension when thinking about data quality.
- And *integrity* relates to scientific rigor used to generate the data as well as protections against manipulation, influence and, of course, unauthorized access.
- Analysts, statisticians, policymakers and the public seem to favor one of these dimensions over the other.
- And based only experience over the years, not to many folks appreciate the overarching composition… the delicate balance attained among these often-competing dimensions when developing a statistical data product.
- Now let's get back to the framework.
- There are eleven dimensions that span these three dimensions.
- I'll not define each for the sake of time. You can find them in the FCSM paper using a simple Google search on the terms *FCSM data quality.*
- Instead, I'll restrict attention to one of these dimensions. It's called utility.
- I'll circle back to the composition later.
- OK, returning to our FCSM framework, utility features the following dimensions:
    - Relevance
    - Accessibility
    - Timeliness, and
    - Granularity.
- *Relevance* is defined pretty much as the word suggests.
- It's the extent to which data meet a user's needs.
- It's not that different than its parent dimension *utility*, although *utility* more broadly covers important dimensions of timeliness, accessibility, and granularity.
- Although in a real sense, these are all related to relevance.
- Next is *accessibility*. It focuses on the ease of acquiring statistical data and associated documentation, as well as the ability to understand the products themselves.
- For example, our American Community Survey public-use micro data are relatively easy to access.
- But it takes a bit of processing knowledge and statistical training to use the data.
- Then there are data visualization tools like the Community Resilience Estimates, the Census Business Builder, and like My Community Explorer.
- These offer easy access to data using geographic data visualizations that can facilitate community needs assessment and economic development planning.
- The next dimension is *timeliness*. It simply refers to the time lag between the date of observations and the public release of the data.
- The pandemic made us well-aware of this dimension.
- For instance, our nation needed contemporaneous data during the pandemic, and two high-frequency surveys were developed to meet that need—the Household Pulse Survey and Small Business Pulse Survey.
- Next, there's *punctuality*. While related to timeliness, it specifically focuses on the ability to adhere to an official data release schedule.
- For instance, *punctuality* was a key concern in the delay of the apportionment counts and our redistricting data release.
- And finally, *granularity* is all about disaggregation of data items by time, geographic detail, and response characteristics such as socioeconomic variables and the like.

- Granularity competes with disclosure avoidance in the data quality arena.
- The 2020 redistricting data had to strike a balance between the two.
- As you probably know, the Census Bureau uses stakeholder feedback to determine the best balance for its remaining 2020 data products—the Demographic and Housing Characteristics and its more detailed sister products.
- OK as I said earlier, I'd like to focus on just one of those dimensions of data utility—relevance.
- Now, like most things in life, relevance can mean many different things to different people.
- On top of that, relevance is situational to the use case.
- For instance, I just returned from a visit to Anchorage, Alaska, where I addressed the Alaska Federation of Nations, an annual gathering of all tribes in that state.
- Some tribal leaders reported their need for better ACS data for their governance, specifically for their planning in housing, education, and health.
- They spoke about how the employment and earnings questions don't really capture the sustenance work performed in remote villages.
- The absence of such data clearly reflects a cultural relevance gap from tribes' perspectives.
- But for macro-economists, the failure to capture sustenance work in Remote Alaska make no bearing on their analysis.
- Now let's look at this a bit more broadly.
- For about the past decade, I've spoken about a renaissance that is unfolding before us as a result of technology and globalization.
- It has fundamentally changed society and our culture.
- We're now data-driven. We increasingly rely on immediate, easy access to information.
- In fact, we expect to be "catered to" by way of tailoring algorithms based on our internet use.
- This renaissance also features cultural changes.
- With the ever-increasing use of social media, connectivity between families and friends, a new wave of virtual communities have sprung up.
- More and more people have smart phone access, and they use social media apps to connect with each other.
- Technological advances in genealogical research and DNA testing have amplified interest in people knowing who they are racially, ethnically, and culturally.
- Our nation's population is becoming more diverse, including mixed race, mixed-ethnicity, and mixed race-ethnicity.
- We recognize the diversity within racial and ethnic groups.
- Sexual orientation and gender identity is recognized as an important part of who we are as a nation.
- And then there is our economy. As it evolves, so do aspects of work such as telework, remote work, as well as the classic office work and manufacturing work locations.
- New industries are emerging, for instance, the cannabis industry with a slew of new products.
- And even the term "work" has taken a more complex meaning.
- Sure, we still have folks working regular hours, with salaried monthly and biweekly paychecks.
- But a sizeable chunk of our workforce can only make ends meet by working multiple jobs, or doing gig work, or some combination.
- Back in Austin many of my photographer friends held multiple jobs besides passionately pursuing their gig work as a photographer.

- Same is true of musicians and other artists.
- A number of my media interviews this year have been with freelance journalists… gig workers.
- This is all important context when thinking about data quality in our federal statistical system.
- We pride ourselves in gold-standard data collections like the American Community Survey, the Current Population Survey, our decennial census, and the many other data collections we conduct.
- But as society evolves—and does so rapidly—the relevance of data items we collect will be affected over time, and not necessarily in a good way.
- And that will affect their utility, a key dimension of data quality.
- In fact, we've already seen that with our race-ethnicity collection and reporting standards.
- We continue to use the 1997 standards.
- Fortunately, through the leadership of our new Chief Statistician Karin Orvis, OMB is in the process of reviewing and revising those standards.
- There's also the issue of collecting data on those involved in the ever-growing, gig economy.
- Many Americans supplement their employment income with gig work.
- Gigs are sometimes referred to as contingent jobs or alternative work arrangements.
- Both the ACS and CPS capture some data on these, but only when they are the main sources of work.
- So, for people who use a combination of jobs and gig work to make ends meet as well as for conventional workers who use gig work to supplement their income, there is a known data gap.
- Now to be clear, the federal statistical agencies do adapt to our changing society.
- The Contingent Worker Supplement to the CPS has been in place since 1995.
- Questions on computer and internet use first appeared in the ACS in 2013.
- The question on whether a house had a flushing toilet was removed from the ACS in 2016.
- And, of course, the 1997 standards for collecting and reporting race-ethnicity are currently being revised.
- So change does occur. The issue is the extent to which societal change is outpacing revisions in our data collections.
- It seems to me like we are getting further behind.
- There are, however, shining examples of quick adaptation.
- It relates to the data quality dimension of timeliness, although one could argue that it overlaps with relevance.
- Early in the lockdown period of the pandemic, the federal statistical system realized there were no contemporaneous data available to tell the story of how households in our nation were faring.
- The ACS would be of little use due to the 9-month lag between the last data of collection and the publication.
- So a collaboration was formed initially with five other federal statistical agencies to create the Household Pulse Survey.
- It featured quick turnaround national surveys focusing on how the pandemic affected households socially and economically.
- The first survey was launched April 23, 2020, and the survey program continues to this day.
- As an online survey with a low response rate, the quality threshold of the Household Pulse Survey is well below that of our flagship surveys.
- In fact, the product was labeled *experimental*.

- Yet the information still provided a valuable, contemporaneous glimpse into how the nation was dealing with the pandemic.
- Policymakers and the public loved it.
- It was the right data product for the right time in our history.
- It aligned with the public's appetite for contemporaneous data during a profound and vulnerable time in our country's history.
- I'll note, by the way that there's a similar sister survey of establishments called the Small Business Pulse Survey that served the same purpose but for businesses.
- Ok, I've talked about data quality in the context of data utility… specifically the dimension of data relevance maybe with a little timeliness sprinkled in.
- So where does this leave us? What lessons might there be?
- First, let's recognize that data can be collected that are highly accurate, reliable, and demonstrate little bias.
- Yet they can have little relevance, or have relevance that diminishes with time.
- And the risks rise when an ongoing assessment of the content we collect is not not baked into our statistical programs.
- For instance, we always begin planning the next decennial as data collection for the current one ends.
- I suggest that we adopt that model for all the data we collect, be it from surveys or even from administrative data sources.
- If you hadn't yet realized, this is also a matter of data equity.
- Collection of job and income data is easiest for those fortunate enough to enjoy conventional jobs paying a monthly salary or biweekly paycheck.
- But among underserved, vulnerable populations, it's difficult to provide accurate employment and income responses for those involved in episodic work, gig work, and multiple part time jobs.
- The cognitive response burden is considerably higher for these folks.
- Unless we're willing to explore better ways to capture data from these segments of society, an inequity will exist in data quality across various subpopulations, mostly associated with those who are most affected by inequities.
- That's why we need to continuously challenge ourselves to consider all the data elements we currently collect, even if they have recently been revised.
- And a second lesson is that differing data needs lead to different three-way balances between our data quality dimensions—utility, objectivity, and integrity.
- Flagship surveys like the ACS require a heavier reliance on objectivity and integrity and less so on timeliness.
- On the other hand, our high-frequency survey programs—like the Household Pulse Survey—sacrifice objectivity to bolster timeliness.
- Both contribute to our knowledge-base and both benefit the public.
- My final thought relates to accessibility, a dimension I've not addressed so far.
- Note that the Census Bureau now produces both gold standard surveys like the ACS alongside experimental products like the Household Pulse survey.
- And other examples exist. We at the Census Bureau have an obligation to effectively communicate the data quality properties to our data users.
- For instance, the ACS data are used to allocate federal funds and monitor compliance with some federal laws.

- Such uses demand high data quality.
- The Household Pulse Survey seeks to get a sense of how the public is dealing with the pandemic.
- It’s not intended, nor would it be appropriate to use it for federal funding allocation.
- We know this… but the public may not fully understand. And that would be to our peril.
- The Census Bureau and more generally the federal statistical system could benefit from standards for effectively communicating the data quality associated with its products to reduce the risk of inappropriate usage.
- Perhaps we should provide more clear guidance.
- OK, so there you have it… my thoughts on just a subset of the broad topic of data quality.
- Returning to the title of my keynote, I see the role of data quality in a twenty-first century federal statistical system to be rather complex.
- I see us generating new data products from blended data sources.
- Each data source will have its own data quality profile.
- Depending on the mix of data sources the final statistical product may have a high or low level of quality which will determine its fitness for various uses.
- We should capitalize on our nimbleness at the start of the pandemic.
- We cannot become complacent. For if we do, we’ll surely end up collecting less relevant, more expensive data.
- In fact, I believe we have already entered an era of opportunity… our own twenty-first century data renaissance!
- That’s why I’m eager to lead the Census Bureau and help enable our modernization efforts.
- Thank you for your time, everyone. It was an honor to address you.