DATA QUALITY LITERACY

Grace Liu
2024

A Knowledge Brief

CONTENTS

Data Reference Interview

Questions to ask when the researcher has a request related to data.

Start with Three Basic Questions

- **Who** needs the data?—a senior undergraduate, doctoral student, or faculty member? The researcher's experience with data influences the data needed to serve their purposes.
- **What data** is needed? Ask about the researcher's data needs and further explore their research plan, methodologies, bibliographic references, etc.
- **When** is it needed? Ask the scope and deadline of the user's research project.

Understand the Researcher's Plan

- What is the research question? How will it be refined?
- What data was used by articles from the literature review? How widely is it used?
- What is the data collection methodology? What are the limitations and possibilities for the data needed?
- What is the theoretical framework? Does it indicate the need for microdata or statistical summary?

Take More out of the Literature Review

- What data has been collected or used before? Who collected the data?
- Is there data to be collected? Can the existing data be reused?
- How has data been used or interpreted by other researchers?

Get Down to the Specific Data Attributes

- Who or what is the subject of the research? What is the unit of analysis or observation?
- When and where of the data?
  - » Data Currency: How recent does the data need to be?
  - » Data Frequency: Does the data need to be updated monthly, quarterly, annually, or at other intervals?
  - » Geographic Level: Is the research focused on a local, regional, national, or global level?
  - » Cross-sectional vs. Longitudinal: Does the research require data collected at a single point in time, or does it need repeated observations to analyze changes or trends over time?
- What are the concepts that need to be measured? What are the variables?

References

Liu, G., Bordelon, B., Nagar, R., Sarti, J., Nguyen, U., & Boettcher, J. (2024). *Data Quality Literacy: A Guidebook.* Institute of Museum and Library Services (IMLS) Grant Project. https://doi.org/10.31219/osf.io/ruawm

The Knowledge Brief is compiled by Grace Liu, based on the National Forum presentation from Ron Nakao, reviewed by the IMLS Data Quality Literacy project team, and designed by Niko Galioto. This project was made possible in part by the Institute of Museum and Library Services [RE-252357-OLS-22].

Evaluating Data Documentation

Data Documentation

Data documentation provides the contextual information needed to discover, understand, access, and reuse data. Examples include README files, metadata, data dictionaries, codebooks, methodologies, survey instruments, and lab records.

Indicators of Documentation Quality

Generally, data documentation is of higher quality if:

- it is produced for use by researchers;
- it is transparent and well-documented by its creator;
- it is validated via a peer review process; and
- time and expertise are invested to curate it by experts.

Characteristics of Good Data Documentation

Good data documentation tells the prospective user:

- **Why** the data is collected (e.g., project context or descriptions);
- **How** the data is collected, structured, and managed (e.g., methodology, study design, sample, universe/population, questionnaires, restrictions, revision history);
- **Who (or what)** the unit of analysis/observation is (e.g., individual, household, business establishment);
- **Where** is being covered and its granularity (e.g., geographic coverage and smallest geographic unit);
- **When** is covered (e.g., time periods covered, date of collection, or dates for different waves of collection); and
- **What** concepts are being measured (e.g., variable data dictionary)?

Fill Data Documentation Gaps

If you obtain a dataset without documentation or if the documentation is confusing or missing information, consider the following:

- **Who collects the data?** Is it a commercial vendor, government/organization, or individual researcher? You can research and connect with the potential data producers to request the details.
- **Who distributes the data?** Are they libraries or data archives/repositories, commercial/independent organizations, or other researchers? You may check your license, or tap into your network of other libraries or repositories that may purchase or distribute the data.
- **Who uses the data?** Search for published articles, working papers, or unpublished works to find who uses the data. Connect with researchers to work out gaps in documentation. If many people use it, it is often another indicator of good data quality.

Comparing Documentation with Datasets

After evaluating the critical information included in the data documentation, a further step is to compare the documentation with the datasets. This helps identify any undocumented variables, out-of-range codes, unexplained missing data, illogical skip patterns, or other discrepancies between the data and its documentation.

References

Liu, G., Bordelon, B., Nagar, R., Sarti, J., Nguyen, U., & Boettcher, J. (2024). *Data Quality Literacy: A Guidebook.* Institute of Museum and Library Services (IMLS) Grant Project. https://doi.org/10.31219/osf.io/ruawm

The Knowledge Brief is compiled by Grace Liu, based on the National Forum presentation from Ron Nakao and Barbara Esty, reviewed by the IMLS Data Quality Literacy project team, and designed by Niko Galioto. This project was made possible in part by the Institute of Museum and Library Services [RE-252357-OLS-22].

Evaluating Dataset for Research Needs

Data quality depends on the fit between the dataset and the research question. A high-quality dataset allows researchers to answer the question of interest.

General Dataset Features that Indicate Data Quality

Dataset features that contribute to quality include:

- data elements that are relevant to the research question
- sample that allows inferences about the population of interest
- data provenance or documented trace of origin is understood
- data can be cross-checked against other sources
- the sample size is large enough to allow for meaningful statistical testing

Think Critically about Existing Datasets

Thinking critically about existing datasets involves asking:

- Is the dataset accessible to the researcher?
- Is the dataset available within a reasonable timeframe?
- Does the data include measures relevant to the research question?
- Why is the data collected?
- Who does the sample represent, and who is missing from the data? Why are they missing?
- Can I compare this data to data from another source?
- Are the means, minimums, and maximums sensible? Why are some elements missing data?
- How large does the effect need to be to detect it in this sample size?

Questions to Ask When Combining Datasets

When combining datasets, ask a much richer array of questions, including:

- Are there common identifiers between both datasets? The most ideal identifiers are those less likely to change over time such as Employee Identification Numbers (EINs), Central Index Keys (CIKs), or geographic codes such as Federal Information Processing Standards (FIPS) county codes or International Organization for Standardization (ISO) 3166 Codes for the representation of names of countries and their subdivisions. In contrast, a company name or ticker symbol may not be ideal since they may change due to restructurings or name changes. A geographic name may be formatted differently depending on the source.
- What is the smallest unit for which data can be combined and what are the implications for aggregation?
- Who does the combined dataset represent? What are the implications for sample size and representativeness?
- Are there differences in the definition of the same variables (e.g. "Year" can be defined differently in different sources)?

References

Kalinowski, A., & Hines, T. (2020). Eight Things to Know about Business Research Data. *Journal of Business & Finance Librarianship, 25*(3/4), 105–122. https://doi.org/10.1080/08963568.2020.1847548

Liu, G., Bordelon, B., Nagar, R., Sarti, J., Nguyen, U., & Boettcher, J. (2024). *Data Quality Literacy: A Guidebook.* Institute of Museum and Library Services (IMLS) Grant Project. https://doi.org/10.31219/osf.io/ruawm

The Knowledge Brief is compiled by Grace Liu, based on the National Forum presentation from Janet Currie and Barbara Esty, reviewed by the IMLS Data Quality Literacy project team, and designed by Niko Galioto. This project was made possible in part by the Institute of Museum and Library Services [RE-252357-OLS-22].

Using and Evaluating U.S. Federal Statistics

U.S. Federal Statistics and its Quality: The Basics

- The U.S. federal statistical system is a decentralized, interconnected network of 13 principal statistical agencies with about 100 additional federal statistical programs.
- Legislation such as the **Information Quality Act** ensures the quality of information disseminated by federal statistical agencies.
- The Office of Management and Budget issued Information Quality Guidelines and statistical policy directives such as Standards and Guidelines for Statistical Surveys to specify quality requirements regarding:
  - » Utility: the usefulness of the information to the intended users.
  - » Objectivity: the information is accurate, reliable, and unbiased and is presented in an accurate, clear, complete, and unbiased manner.
  - » Integrity: protecting information from unauthorized access or revision and not compromised through corruption or falsification.
- Agencies including the U.S. Census Bureau created their own Statistical Quality Standards to ensure statistical information quality from planning programs, acquiring data, producing estimates, analyzing data, reporting results, and releasing information to documentation.

Evaluating Federal Statistics for Research Needs

- Despite the high quality of federal statistics, researchers still need to assess the fit between the dataset and their research needs (refer to **Data Quality Literacy Series 03: Evaluating Dataset for Research Needs**).
- Researchers also need to be aware of the potential **dataset structure quality** issues in federal statistics:
  - » **Changes in the dataset schema and structure.** Data collection can also be discontinued due to funding cuts or changes in mandates. This change can affect data processing and the comparability of data across time.
  - » **Inconsistent use of variable codes.** Do not assume variable codes are the same between data collections. For example, if age 12 was coded as XYZ in a 2020 survey, it can be coded as XYZ2 in the next release. There can be underlying differences in how data was collected or tabulated.
  - » **Changes in survey questions or variable codes or labels between surveys.** For example, a new survey collection instrument is used in one year but not the next. Then, the trend can become discontinuous.
  - » **Latency or source release schedule changes.** For example, the Internal Revenue Service (IRS) can be late in publishing an updated statistical release.
  - » **Changes in reference materials**. For example, updates to North American Industry Classification System (NAICS) codes every five years can affect data comparability over time, potentially disrupting continuity in series.
  - » **Lack of documentation.** Especially for administrative records, the data documentation may be insufficient for researchers to understand how to use the data.

References

Liu, G., Bordelon, B., Nagar, R., Sarti, J., Nguyen, U., & Boettcher, J. (2024). *Data Quality Literacy: A Guidebook.* Institute of Museum and Library Services (IMLS) Grant Project. https://doi.org/10.31219/osf.io/ruawm

Office of Management and Budget. (2022). *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information.* https://www.federalregister.gov/documents/2002/02/22/R2-59/guidelines-for-ensuring-and-maximizing-the-quality-objectivity-utility-and-integrity-of-information

The Knowledge Brief is compiled by Grace Liu, based on the National Forum presentation from Katherine Wallman and Jill Blaemers, reviewed by the IMLS Data Quality Literacy project team, and designed by Niko Galioto. This project was made possible in part by the Institute of Museum and Library Services [RE-252357-OLS-22].

Understanding Administrative Data

Understanding administrative data, the benefits of using administrative data, and its limitations.

Administrative Data

Administrative data refers to data collected for operational, programmatic, or regulatory purposes rather than statistical or research purposes.

Administrative Data Examples

Federal Administrative Data	State and Local Administrative Data	Commercial Administrative Data
Internal Revenue Service (IRS) data | Social Security Administration (SSA) administrative records | U.S. Patent and Trademark Office patent applications | Center for Medicare and Medicaid Services data	Department of Motor Vehicles Drivers License Data | Supplemental Nutrition Assistance Program and Temporary Assistance for Needy Families (SNAP/ TANF) Data | Unemployment Insurance data	Black Knight (master address data, mortgage data)|Experian (credit bureau header data) | InfoGroup (household member data) | Circana (point of sale scanner data) | J.D. Power (new vehicle transaction data) | D&B (business directory; credit and risk data)

Benefits of Using Administrative Data in Statistics

Administrative data is increasingly used in conjunction with federal statistics products (e.g., the 2020 Census, American Community Survey, USDA Consumer Food Data System). It involves linking the restricted versions of administrative datasets with a survey or another administrative dataset based on common identifiers such as Social Security Number (SSN) or Employer Identification Number (EIN). The administrative data can help:

- Build survey sampling frames.
- Evaluate and enrich survey data, reducing sampling and nonsampling errors.
- Fill in missing information and reduce the questions asked in a survey.
- Form the basis for comparing participant and non-participant outcomes or between communities.
- Save costs for data collection.

Benefits of Using Administrative Data in Evidence-Based Policy Research

- Routinely collected and broadly covered administrative data tend to be inherently longitudinal and more representative.
- The large size allows experiments with more treatment arms and detecting small or heterogeneous effects between groups, without losing statistical power.
- Often more objective, avoiding social desirability or recall biases common in survey data.
- Often more reliable and accurate, particularly for biometric data or geo-tagging data.
- Helping reduce the cost and complexity of research data collection.

Limitations of Using Administrative Data

- When repurposing administrative data, population coverage and sampling biases (e.g., self-selection bias or survivorship bias) may be of particular concern.
- Meanings of particular data values in administrative data are likely to be different from the user's concept of interest and it may not include broader variables of interest such as economic and demographic variables.
- Administrative records alone often cannot be used to address all analysis questions; for example, eligibility data doesn't provide information about nonparticipants.
- Micro-level administrative data is often difficult to access. Privacy and disclosure concerns are major constraints.
- Data cleaning and preparation can be complex, especially if the goal is to link administrative data with other data sources.

To learn more, refer to **Data Quality Literacy Series 06: Evaluating Administrative Data Quality**.

References

Cole, S., Dhaliwal, I., Sautmann, A., & Vilhuber, L. (2020). Using Administrative Data For Research And Evidence-Based Policy: An Introduction. *In Handbook on Using Administrative Data for Research and Evidence-based Policy.* Cambridge, MA: Abdul Latif Jameel Poverty Action Lab.

Liu, G., Bordelon, B., Nagar, R., Sarti, J., Nguyen, U., & Boettcher, J. (2024). *Data Quality Literacy: A Guidebook.* Institute of Museum and Library Services (IMLS) Grant Project. https://doi.org/10.31219/osf.io/ruawm

McNabb, J., Timmons, D., Song, J., & Puckett, C. (2009). *Uses Of Administrative Data At The Social Security Administration.* https://www.ssa.gov/policy/docs/ssb/v69n1/v69n1p75.html

United States Census Bureau. (2023). *Administrative Data*. https://www.census.gov/topics/research/guidance/restricted-use-microdata/administrative-data.html

The Knowledge Brief is compiled by Grace Liu, inspired by the National Forum presentation from Bill Sermons, Jill Blaemers, and Patrick W. McLaughlin, reviewed by the IMLS Data Quality Literacy project team, and designed by Niko Galioto. This project was made possible in part by the Institute of Museum and Library Services [RE-252357-OLS-22].

Evaluating Administrative Data Quality

Administrative Data Quality Dimensions

Administrative data, collected for operational, programmatic, or regulatory purposes, require a distinct quality evaluation approach compared to survey data. Federal statistical agencies broadly define data quality as "fitness for use," recognizing that different users of the same data may have different assessments of its quality. The following **quality dimensions** adapted from the Data Quality Assessment Tool for Administrative Data by the Federal Committee on Statistical Methodology can be utilized to aid the administrative data assessment:

- **Relevance**
  - » Do the content, scope, level of measurement, coverage period, frequency, and timelines meet the user's needs?
- **Accessibility**
  - » What are the administrative restrictions for accessing the data? How will the data be accessed by the user?
  - » Was there any data collected but not included, due to confidentiality or other reasons?
  - » What control methods or modifications were used to protect the confidentiality of the data?
- **Interpretability**
  - » Are data variables and valid values clearly defined? Is the data dictionary available?
  - » What methodology is used to recode original data or create a value for a new variable (e.g., assigning a reported age to an age range)?
  - » Are data architecture and the relationship between key variables explained?
  - » Does the dataset contain records for those who are denied eligibility to administrative programs and can those records be identified?

Does the value reflect only the "latest" version (refer to the concept of "Header Data" in **Data Quality Literacy Series 08: Evaluating Commercial Data Quality**)? Has the data been revised or updated? Have the superseded records been removed at a given time?

- **Coherence**
  - » Are there any classification systems (e.g., race and ethnicity categories or NAICS) used for categorizing or classifying the data? Were there any changes to the system in the extracted data?
  - » Were there changes or differences across geographic areas covered that would cause breaks in consistency such as different questions, revised questions, questions in different languages, or deleted questions?
  - » Were there changes to the instructions that the data was collected or processed (e.g., instructions for completing the application form)?
  - » Were there changes to geographical boundaries?
  - » Were there substantial changes or differences across the geographical areas that influenced who participated in the program (e.g., legislative changes, eligibility changes, program expansions, or natural disasters impacting program participation)?
- **Accuracy**
  - » What percentage of eligible participants are not included in the data file? What is known about their characteristics?
  - » Are there duplicate records or missing values? What are the known sources of error?
  - » What questions are most often misinterpreted?
  - » Are there any revisions to the reported value and why are the changes made?
- **Institutional Environment**
  - » Does the purpose of the administrative program align with the research purpose?
  - » Who is the data collected from and how is source data collected?
  - » Are there quality control standards and processes applied?
  - » Will there be new records or revisions to existing records after data acquisition?

To learn more, refer to **Data Quality Literacy Series 05: Understanding Administrative Data**

References

Liu, G., Bordelon, B., Nagar, R., Sarti, J., Nguyen, U., & Boettcher, J. (2024). *Data Quality Literacy: A Guidebook.* Institute of Museum and Library Services (IMLS) Grant Project. https://doi.org/10.31219/osf.io/ruawm

Iwig, W., Berning, M., Marck, P., & Prell, M. (2013). *Data Quality Assessment Tool for Administrative Data.* https://www.fcsm.gov/assets/files/docs/DataQualityAssessmentTool.pdf

The Knowledge Brief is compiled by Grace Liu, inspired by the National Forum presentation from Bill Sermons, Jill Blaemers, and Patrick W. McLaughlin, reviewed by the IMLS Data Quality Literacy project team, and designed by Niko Galioto. This project was made possible in part by the Institute of Museum and Library Services [RE-252357-OLS-22].

Understanding Commercial Data

Understanding commercial data, the benefits of using commercial data, and its limitations.

Commercial Data, also called private-sector data or third-party data, broadly refers to data created and provided by commercial entities rather than government agencies.

Commercial Data Examples

Structured Commercial Data		
Structured Survey Data	**Structured Administrative Records**	**Other Structured Data**
Media Market Data (e.g., Nielsen) | Market Research Probability Survey or Opt-in Panel data (e.g., Ipsos, Gfk, Kantar, Mintel) | Customer Satisfaction Surveys	Banking and Stock Records (e.g., Bloomberg, S&P, Moody's, Compustat, CRSP) | Commercial Transactions (e.g., Refinitiv) | Point of Sales Data (e.g., IRI/Circana) | Credit Card Records (e.g., Experian) | Housing Data (e.g., Zillow) | Private Employment Data (e.g., ADP) | State and Local Tax Data (e.g., CoreLogic) |Climate and Self-reported Environmental Data (e.g., CDP)	E-commerce Transactions | Mobile Phone Location Sensors | GPS Sensors | Utility Company Sensors | Weather or Pollution Sensors
Semi-Structured Commercial Data	XML or JSON Files; Data from Computer/Online Systems (e.g., web logs); Emails; Articles from Full-text databases Unstructured	
Unstructured Commercial Data	Social Media Data (e.g., Facebook, Twitter, LinkedIn); Internet Searches (e.g., Google), Videos (e.g., YouTube), Traffic Webcams, Satellite Images	

The Benefits of Using Commercial Data

- It can have content or a level of granularity that federal statistics do not provide.
- It can be provided more timely and frequently than federal statistics.
- Commercial data vendors can have business relationships with private firms, so they can acquire and synthesize proprietary data.
- It can be more cost-effective than collecting data on your own and can reduce the response burden.
- It can complement and enhance the analysis of federal statistics.

Limitations of Commercial Data

- Commercial survey data generally have lower response rates than government surveys. Many firms have chosen opt-in Internet panels over probability surveys, which may cause concerns about the representativeness of the sample.
- Administrative data collected for transactional purposes tend to be less stable in data definition and data-generating processes.
- Commercial data are often vulnerable to changes or discontinuation without notice and subject to manipulation for private interest.
- Vendors often provide the latest data at a point in time in a dashboard interface rather than datasets and data file format may not be compatible with statistical software.

To learn more, refer to **Data Quality Literacy Series 08: Evaluating Commercial Data Quality**.

References

Harris-Kojetin, B. A., & Groves, R. M. (Eds.). (2017). I*nnovations in Federal Statistics: Combining Data Sources while Protecting Privacy.* https://nap.nationalacademies.org/catalog/24652/innovations-in-federal-statistics-combining-data-sources-while-protecting-privacy

Harris-Kojetin, B. A., & Groves, R. M. (Eds.). (2018). *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps.* https://nap.nationalacademies.org/catalog/24893/federal-statistics-multiple-data-sources-and-privacy-protection-next-steps

Liu, G., Bordelon, B., Nagar, R., Sarti, J., Nguyen, U., & Boettcher, J. (2024). *Data Quality Literacy: A Guidebook.* Institute of Museum and Library Services (IMLS) Grant Project. https://doi.org/10.31219/osf.io/ruawm

Muth, M., Sweitzer, M., Brown, D., Capogrossi, K., Karns, S., Levin, D., Okrent, A., Siegel, P., and Zhen, C. (2016). Understanding IRI Household-Based and Store-Based Scanner Data. Economic Research Service. *USDA Technical Bulletin 1942.* https://www.ers.usda.gov/publications/pub-details/?pubid=47636

The Knowledge Brief is compiled by Grace Liu, inspired by the National Forum presentation from Bill Sermons, Todd Hines, and Patrick W. McLaughlin, reviewed by the IMLS Data Quality Literacy project team, and designed by Niko Galioto. This project was made possible in part by the Institute of Museum and Library Services [RE-252357-OLS-22].

Evaluating Commercial Data Quality

Commercial data can be evaluated based on its fitness for use and quality dimensions such as relevance, accessibility, interpretability, coherence, accuracy, and institutional environment (refer to **Data Quality Literacy Series 05: Understanding Administrative Data**). Common data quality issues that need attention include the following (for a more detailed explanation of each issue, see (Liu, 2020)):

- **Missing Values:** Missing values can occur due to skipped (or optional) questions in a questionnaire, data suppression for confidentiality, restrictions due to vendor agreements, data not being collected (e.g., not all retail stores collect random-weight data for fruits or vegetables), or other reasons. This can lead to confusion about the existence of the data or incomplete recording. A significant number of missing values can make a dataset unusable.
- **Data Errors** can happen in different forms including simple typos, arithmetic errors, coding errors, date errors, classification errors, etc. One way of identifying data errors is to compare different data sources.
- **Biases** are systematic errors caused by various reasons. Commercial datasets are prone to selection bias such as (1) **sampling bias**, where samples do not represent the population (e.g., opt-in market research panels); (2) **under-coverage bias**, where specific segments (e.g., senior citizens, low-income households, small or independent stores, or low-performing firms) are excluded or less represented in the sample); (3) **survivorship bias**, especially in financial datasets (e.g., Yahoo! Finance), when the dataset concentrates on collecting data about surviving stocks/firms while overlooking the data from delisted firms.
- **Inconsistencies** can happen from variable definitions to value format. For time-series data, consistency can be broken due to changes in questions or instructions in a questionnaire, classification system (e.g., North American Industry Classification System (NAICS)), geographic boundaries, eligibility, etc. Aggregate datasets from multiple private sources are susceptible to inconsistencies.
- **Discrepancies** between databases or datasets often arise from differences in coverage, definitions, coding policies, classifications, or errors. This can lead to the "database effect," where researchers may draw different conclusions based on the database they use.
- **Header Data** refers to data that only reflects the latest available value, which may not always be the most updated. Common examples of header data include company name, ticker symbol, stock exchange, industry code, and headquarters location. Header data in these data points can mislead time-series analyses by attributing data to incorrect or outdated identifiers, and distort cross-sectional studies by misclassifying entities based on incorrect groupings.
- **Standardization** improves data comparability across companies, time, and geography, but it can also result in understating or overstating the original outcome, leading to inaccuracies in certain prediction models.
- **Superseded Data** occurs when a dataset is revised or updated due to error corrections, restatements, or other changes. This can result in data downloaded at different times having different values. Substantial or systematic changes may raise concerns over data integrity.
- **Actual vs. Estimated Data:** Values in a dataset may be estimated based on models rather than exact numbers. This is common in financial data, especially for private companies' revenues or industry sizes in databases such as Dun & Bradstreet, Data Axle, or Bizminer. As a result, data sources often present widely divergent numbers in their estimates.
- **Reporting Time Issues** may occur when the dates that the data becomes available to the public are different from what researchers assume the date to be. Improperly recorded reporting time can lead to look-ahead bias or selection bias.
- **Misuse of Data** may occur when researchers improperly use the data as proxies or measurements, leading to unreliable research results. For example, data from databases that only cover public firms in an industry can be a poor proxy for calculating actual industry concentrations.

- **Lack of Transparency** is prevalent among database vendors. They often view their data collection methods or projection models as proprietary and are reluctant to disclose information about their data collection and management practices, as well as potential data problems and biases. To address this, probing vendors with quality-related questions is important in data acquisition.

To learn more, refer to **Data Quality Literacy Series**
07: Understanding Commercial Data
09: Commercial Data Quality: Conversation with the Vendors
10: Commercial Data Quality: Conversation with the Researchers

References

Liu, G. (2020). Data Quality Problems Troubling Business And Financial Researchers: A Literature Review And Synthetic Analysis. *Journal of Business & Finance Librarianship, 25*(3-4), 315-371.

Liu, G., Bordelon, B., Nagar, R., Sarti, J., Nguyen, U., & Boettcher, J. (2024). *Data Quality Literacy: A Guidebook.* Institute of Museum and Library Services (IMLS) Grant Project. https://doi.org/10.31219/osf.io/ruawm

The Knowledge Brief is compiled by Grace Liu, inspired by the National Forum presentation from Todd Hines and Patrick W. McLaughlin, reviewed by the IMLS Data Quality Literacy project team, and designed by Niko Galioto. This project was made possible in part by the Institute of Museum and Library Services [RE-252357-OLS-22].

Commercial Data Quality: Conversation with the Vendors

Commercial Data Quality: Questions to Ask

Commercial data vendors may be hesitant to share detailed documentation of their data. However, they are often willing to engage in discussions with data users to answer specific questions, provide clarity, or offer additional context regarding their proprietary methods. It is crucial for librarians or researchers to probe this "**black box**" as much as possible. Here are some questions that can help broadly assess the data quality and fitness for use:

- **Content:** What data fields, variables, identifiers, or classifications (e.g., North American Industry Classification System (NAICS)) are included? Check if the desired fields or variables are available. If the dataset needs to merge with other datasets, make sure there are common identifiers.
- **Geographic Coverage:** What is the geographic coverage of the dataset, its granularity, and how complete is the coverage? Be cautious of datasets claiming global coverage that might only include data for 10 countries, or those claiming county-level data that might only have detailed data for 20 states.
- **Temporal Coverage:** What is the temporal coverage of the dataset, its frequency, and how complete is the coverage? Be aware that while the dataset may have monthly data starting in 1995, the data might not be complete before 2000, or may only have annual data available for the earlier years.
- **Data Sources:** How does the vendor gather the data? The data may be obtained through one or a combination of the methods, including (1) collected data: obtained from public sources such as administrative records, surveys, public data, or web scraping; (2) created data: derived or modeled using proprietary algorithms or in-house analysts; (3) third-party data: purchased from other firms or vendors.
- **Methodology:** What is the data collection methodology, and is there any standardization or conversion of the original data? Is the data normalized or indexed to a specific baseline? How is this done? Is there a crosswalk to map data between different coding systems? Ask about the availability of detailed documentation, data dictionaries, file setup documents, or access to vendor experts who can answer questions. At a minimum, the vendor should provide a list of variables and their definitions.
- **Sample Size:** What is the number of records or observations included in the dataset, and how does this compare to the entire population? For example, in a public company dataset, what percentage of all public companies are included? Does the dataset also cover delisted firms, foreign firms listed on US stock exchanges, or US firms listed overseas?
- **Changes in Data:** Are there any changes to the datasets, such as changes in collection methods, classification, or shifts in coverage focus over time? How frequently is the data updated, and what is the time lag between data collection and its inclusion in the database? Is there a process to identify data errors and make corrections? Request documentation on gaps in coverage, change logs, imputations, cleaning, instruments, and any changes in collection methods, variable definitions, or geographic boundaries.
- **Data Product:** Is the data product a database, an interface, or a dataset? Web interfaces usually include the latest data and may not allow extensive data extraction. The cost of extracts or backfiles can be high.
- **Data Format:** What format is the data delivered? Is it structured or unstructured? Can it be delivered in common file formats (e.g., comma-separated values (CSV) file)? Is the format compatible with statistical software, or does it require additional cleanup?
- **Delivery Frequency:** How frequently will the data be delivered? Is it through daily, weekly, or monthly data feeds, or can it be provided as a one-time extract or annual subscription? Can updates be purchased on a scheduled basis or as one-time purchases?
- **Restrictions:** Does the data contain restricted variables? Researchers might need extra permissions to access restricted data such as health, labor, or social surveys that capture personally identifiable information. Does the vendor require

the submission of findings or full papers before publication? Do they retain the right to deny data usage in future publications? Can the data be shared for replication studies? How does the data need to be cited or attributed? Are there limitations on the number of records downloaded or types of analysis allowed? Some vendors prohibit machine learning or artificial intelligence applications on their data.

- **Access and Confidentiality:** Who can access the data, and what are the required processes to access the data? Besides the researcher, is it accessible to all campus affiliates, faculty, a specific department, or co-authors at other institutions? Does the project need an Institutional Review Board (IRB) review, or does the researcher need to sign a Data Use Agreement (DUA) or Non-Disclosure Agreement (NDA)? Consider the data's sensitivity, confidentiality, and required security measures, as well as any restrictions on data storage.
- **Resources Required:** How much time, money, expertise, and storage space are needed to acquire and make the data available to the researcher?
- **Trial or Samples:** Can the vendor provide a specific sample or set up a trial before purchase? Ensure the trial reflects actual database functionality, including download capabilities, exporting, searching, and printing.
- **Support:** What ongoing support is available for researchers regarding data errors, clarification, functionality, and methodology?

Issues Around Third-Party Data

Commercial data providers often provide third-party data on their platforms, but instead of giving direct access to third-party data, it may be integrated into their platform in different ways:

- **Input to their Models:** They may use third-party data as input into their proprietary models.
- **Proprietary Identifier/Rating:** They may append their data with proprietary identifiers or ratings from another company.
- **Convenience Data (Demographics Add-On):** They might enhance their data with third-party convenience data such as demographic data.
- **Aggregator (Tools for Automation):** The data provider may be an aggregator and provide tools for automation and all the data are from third parties.

When dealing with commercial data that includes third-party sources, here are some questions to ask:

- **Third-Party Source:** What is the third-party source, and how does the third party collect or create their data?
- **Licensing Rights:** How long are the licensing rights available to the source provider? Commercial data providers can gain or lose access to other third-party data or change to another data provider, which can alter the resource's methods and affect data quality.
- **Use Limitations:** Are there any limitations in using third-party data? Sometimes, the third-party data is view-only or cannot be easily exported or exported at all, so it would be hard to use for further analysis.
- **Enhancements:** What, if any, enhancements are made? Examples include adding demographic information from the Census, rounding, and modeling or forecasting data until observed. It is important to know what enhancements are made and ensure any alterations to the original data are transparent to the data users.
- **Vendor Consolidation:** Are there instances of data vendor consolidations, mergers, or acquisitions? If so, ask how data will be incorporated into the other product, if all data will come over, the timelines, and any changes to product specialists from the previous provider, methodology, or delivery and access method. Although we may not be able to do anything about a new and more tedious workflow, that is something to keep in mind as we evaluate those products.

Additional Questions on Competitive Products

To better understand the competitive landscape and the vendor's position in the market and open up new options to compare and further negotiate with the vendor, consider asking the following questions:

- What other sources do you compare your products to?
- Who are your competitors in the market?
- How does the data compare to another familiar or well-known source?

Vendors are usually quick to demonstrate how their data is different, highlight where they believe they excel, and are often upfront about areas where their data may not compare as favorably. This can also introduce you to other sources you had not considered or were unaware of.

After gathering this information and requesting a sample or trial, you are ready to investigate and discuss the data quality with the researcher. For further details, refer to **Data Quality Literacy Series 10: Commercial Data Quality: Conversation with the Researchers**.

References

Kalinowski, A., & Hines, T. (2020). Eight things to know about business research data. *Journal of Business & Finance Librarianship*, 25(3/4), 105–122. https://doi.org/10.1080/08963568.2020.1847548

Liu, G., Bordelon, B., Nagar, R., Sarti, J., Nguyen, U., & Boettcher, J. (2024). *Data Quality Literacy: A Guidebook.* Institute of Museum and Library Services (IMLS) Grant Project. https://doi.org/10.31219/osf.io/ruawm

The Knowledge Brief is compiled by Grace Liu, inspired by the National Forum presentation from Stephanie Tully and Barbara Esty, reviewed by the IMLS Data Quality Literacy project team, and designed by Niko Galioto. This project was made possible in part by the Institute of Museum and Library Services [RE-252357-OLS-22].

Commercial Data Quality: Conversation with Researchers

After one or more conversations with the vendors (Refer to **Data Quality Literacy Series 09: Commercial Data Quality: Conversation with the Vendors**), librarians can work with the researcher to examine the data quality more closely through the sample or trial. Here are some further questions to consider to assess its fitness for use:

Coverage

- **Data Fields and Variables:** Does the dataset include all relevant data fields, variables, identifiers, and classifications needed for the research?
- **Geographic and Temporal Scope:** Does the geographic and temporal coverage align with the research needs? Is the level of detail (granularity) sufficient for the analysis?

Clarity

- **Data Sources:** Is the data collected, created, or purchased? Is the methodology for data collection, creation, or aggregation valid and reliable? If the data originates from a survey, questionnaire, or form, are the questions used and detailed methodology available for review?
- **Variable Definitions:** Are all data variables and their values clearly defined and documented? When date or year is a variable, is it clear whether it refers to the date/year the data was collected or reported?
- **Actual vs. Modeled Data:** For variables with assigned values, how are these values calculated? Do the data represent actual values, or are they averaged, estimated, or projected (which is common for variables such as private companies' revenues or industry sizes)? Is the researcher aware that different sources can present widely divergent estimates?
- **Changes to Original Data:** Are there any standardization, conversion, normalization, or indexing made to the original data? Will these changes affect the data's fitness for use?

Completeness

- **Sample Size:** It is rare for a dataset to be 100% complete; it often consists of a sample rather than the entire population. Does the sample size allow for valid inferences about the population of interest? Is the sample large enough to support meaningful statistical testing?
- **Sample Representativeness:** Does the dataset include all necessary subgroups for the intended study? Can the database produce a sample that represents the broader population? Is the data a proper proxy or measurement for the phenomenon under study (e.g., data about public firms can be a poor proxy for calculating actual industry concentration)?
- **Missing Values:** Are there missing values for any variables? Is there a discernible pattern to the missing values, or are they sporadic? Can these missing values be filled through imputation or other methods? Is the proportion of missing values acceptable?
- **Header Data:** Are there data variables, such as company name, ticker symbol, stock exchange, industry code, headquarters location, or other demographic variables that only contain the latest available values (which may not be the most current)? Is the researcher aware that the header data can lead to misclassifications or mismatches in time-series and cross-sectional analyses, potentially distorting trends and comparisons by organizing data into incorrect categories?

Accuracy

- **Errors:** Are there observable data errors, such as typos, arithmetic errors, coding errors, date errors, classification errors, or outliers? Can the data be cross-checked or spot-checked against other sources? Is the researcher aware of the "database effect," where using different datasets can yield different research results?
- **Biases:** Are certain groups excluded, underrepresented, or not timely included in the datasets? Is the researcher aware of potential biases, such as sampling bias, under-coverage bias, survivorship bias, or look-ahead bias? Can procedures be applied to mitigate these biases?

- **Up-to-Dateness:** Are all values in the dataset current and reflective of the most recent information, or is there a mix of up-to-date and outdated values? Is the update process periodic or synchronized?
- **Revisions:** Are there revisions due to error corrections, restatements, or other changes? Are these revisions done on a regular schedule? Is the researcher aware that this can cause data downloaded at different times to have different values, potentially leading to different results?

Consistencies

- **Format:** Is the format of the data variables (e.g. name, address, units of measurement) consistent? If not, is it costly to standardize and clean the data?
- **Classification:** Is a classification system (e.g., NAICS) used to categorize the data? Is the application of classification codes and their level of granularity consistent? When the classification system is updated (e.g., NAICS 2017 to 2022), has reclassification been consistently applied over time? Is the researcher aware of potential issues in comparing, aggregating, or merging data caused by classification inconsistencies?
- **Breaks:** Are there changes in data sources, data availability, collection methods, levels of measurement, geographic boundaries, regulatory policies, or other reasons that affect consistency over time?
- **Multiple Sources:** For data collected or aggregated from multiple sources (e.g., point-of-sales data from different stores) or geographic areas, are there differences in data availability, collection methods, or levels of measurement across these sources and areas?
- **International Data:** What data is available for international coverage? How is the data translated? How are measurement units (e.g., currency, weights, lengths) converted? How is standardization implemented? Is the translation, conversion, and standardization consistent across geographic areas and over time?
- **Duplicate Records:** Are there duplicate records in the dataset? If so, is the data value for these duplicates consistent? Do these duplicates indicate larger issues, such as problems with data entry, dataset merging, or system flaws?

If the librarian and researcher have identified data quality issues, but the database is still the best available option on the market, these quality concerns can be factored into the price negotiation.

References

Liu, G., Bordelon, B., Nagar, R., Sarti, J., Nguyen, U., & Boettcher, J. (2024). *Data Quality Literacy: A Guidebook.* Institute of Museum and Library Services (IMLS) Grant Project. https://doi.org/10.31219/osf.io/ruawm

The Knowledge Brief is compiled by Grace Liu, inspired by the National Forum presentation from Janet Currie and Barbara Esty, reviewed by the IMLS Data Quality Literacy project team, and designed by Niko Galioto. This project was made possible in part by the Institute of Museum and Library Services [RE-252357-OLS-22].

Evaluating International Government Data Quality

International Governmental Data

International government datasets are usually sourced from national statistical authorities. International Governmental Organizations (IGOs) then compile, standardize, and disseminate this data to improve comparability and access. This process is complex and can involve delays in release. The quality and reliability of the data are often dependent on the capabilities and integrity of the governments that provide the data. There are two basic categories of international government data:

Aggregate Data and Statistical Databases: IGOs collect, harmonize, and publish data in centralized systems according to internationally agreed-upon standards and in official languages. Data is typically at the country level and an annual frequency. Users can generate custom tables using a graphical user interface or download entire datasets, which allows users to compare statistics across multiple countries over multiple time periods in the language users are most comfortable with.

Microdata: IGOs conduct their own surveys and may publish the associated microdata (individual level data). This is quite different from the data collected from national governments.

Data Quality Frameworks and Standards

IGOs invest significant effort in developing and refining frameworks, methods, and guidelines through international agreements and collaboration with statisticians and national experts. Notable examples include:

- UN National Quality Assurance Frameworks Manual for Official Statistics (UN NQAF Manual)
- World Bank International Comparison Program (ICP)
- IMF Data Quality Assessment Framework
- European Union Statistical Requirements Compendium
- UNESCO Education Data Quality Assessment Framework

These frameworks are continuously developed and debated, focusing on data quality aspects such as validity, timeliness, completeness, consistency, and integrity.

The Common Data Quality Issues

International government data is extraordinarily useful and convenient for temporal and cross-country comparisons. It is harmonized, standardized, and translated. It is reviewed by national and IGO statistical authorities. With some exceptions, it is largely open access and public. There is often a lack of alternatives or the alternatives can be expensive. However, there are specific data quality issues with compiled international government data that need attention:

- **Missing Geographies and Values:** It is common to have missing geographies and values. Not every country reports every metric.
- **Temporal Limitations:** Harmonizing data across countries is time-consuming, often taking months or years. Challenges include discontinued indicators, regime changes, lack of transparency, and insufficient statistical capacity in some countries.
- **Adjustments:** The adjustments that need to be made sometimes can be inconvenient, and occasionally demoralizing. Converting national currencies to US dollars and switching between constant (real) and current (nominal) prices are common tasks. The metrics may not be done the way it needs to be done for research purposes.
- **Changes in Methodologies:** Governments may revise their statistical methodologies due to changes in the economy or counting methods. New versions may not map well to older versions, and IGOs may not keep older data, which can be disconcerting.
- **Changes in Historical Data:** Not all IGOs have data preservation policies. Data may be updated, revised, deleted, or taken down without explanation.
- **Discrepancies:** International data can be published in different databases (e.g., in both the International Labor Organization statistics and UNData). When data in one system is updated or deleted, it may still exist in another. The

discrepancies in the data retrieved from different systems can lead to confusion and inconsistencies in analysis.

- **Data Manipulation:** Data may be intentionally misreported by countries. Occasionally, data may be manipulated to achieve a goal or a target. For example, the World Bank's Doing Business Report was discontinued after accusations that it gave certain countries preferential treatment in the report's annual country rankings.
- **Concept-Measurement Gap:** There can be a gap between a concept defined in a manual and the capacity of a government or statistical authority to measure it accurately.

When considering international government data, it is crucial to distinguish between data compiled by international organizations and the surveys they conduct. Be aware of high-profile rankings and their political or economic implications. Pay attention to changes in methodologies and be mindful of any campaigns, incentives, or agendas influencing data quality. Recognize the differences and capacities of national statistical agencies. No data source is perfect—do not take the data quality for granted and always maintain a healthy skepticism.

References

Liu, G., Bordelon, B., Nagar, R., Sarti, J., Nguyen, U., & Boettcher, J. (2024). *Data Quality Literacy: A Guidebook.* Institute of Museum and Library Services (IMLS) Grant Project. https://doi.org/10.31219/osf.io/ruawm

The Knowledge Brief is compiled by Grace Liu, based on the National Forum presentation from James Church, reviewed by the IMLS Data Quality Literacy project team, and designed by Niko Galioto. This project was made possible in part by the Institute of Museum and Library Services [RE-252357-OLS-22].

Understanding Survey Data and Public Polls

Survey Research

Survey research as a research methodology (comparable to experimental research, ethnography, grounded theory, case studies, etc.) relies on questionnaires or interviews for data collection. It can employ quantitative research strategies (e.g., using questionnaires with numerically rated items), qualitative research strategies (e.g., using open-ended questions), or both (i.e., mixed methods). Surveys are crucial for gathering data in social sciences, market research, public health, and policy-making. They help researchers understand public opinion, behaviors, and trends, and inform decisions based on empirical evidence.

Key Concepts in Survey Research

Several concepts are critical to understanding survey data and assessing its quality:

- **Target Population** is the entire group of individuals or entities to which the survey results will be generalized. This group should be clearly defined based on the research objectives.

- **Sampling Frame** is the list or database from which the sample will be drawn. It should be as comprehensive and up-to-date as possible to accurately represent the target population. A perfect sampling frame is complete and includes all elements or sample units from the population, with each element listed once and only once, and without any irrelevant or extraneous elements.

- **Sample Size** refers to the number of individuals or units selected from the sample frame that is included in a specific study. A larger sample size generally increases the study's statistical power, reduces the margin of error, and provides more confidence in the findings.

- **Sampling Method** is the technique used to select a sample from a population, and it directly affects the accuracy, reliability, and representativeness of survey results, with probability sampling methods generally providing higher quality data compared to non-probability methods.

  - **Probability Sampling Methods**
    - **Simple Random Sampling:** Every member of the sampling frame has an equal chance of being selected.
    - **Systematic Sampling:** Individuals are selected at regular intervals from the sampling frame.
    - **Stratified Sampling:** The population is divided into subgroups (strata) based on certain characteristics (e.g., age, gender, income level, geographic region), and samples are drawn from each stratum. Stratified sampling expects that the measurement of interest varies between the different subgroups and enhances representativeness by ensuring that key subgroups are proportionately represented.
    - **Cluster Sampling:** The population is divided into clusters, and a random sample of clusters is selected. All individuals within the chosen clusters are surveyed. Cluster sampling is cost-effective and practical for large, dispersed populations, though it may introduce cluster-related bias if clusters are not homogeneous.

  - **Non-Probability Sampling Methods**
    - **Convenience Sampling:** Samples are selected based on ease of access or proximity to the researcher. It is a quick and inexpensive choice but often results in a biased and unrepresentative sample, reducing the generalizability of the survey results.
    - **Quota Sampling:** Samples are selected to ensure certain characteristics are represented in specific proportions. For example, a market researcher decides to survey 750 people over 20 years old and set quotas to ensure that the sample resembles the proportion in the U.S. population: 270 individuals aged 20-39, 256 aged 40-59, and 224 aged 60 and above. Quota sampling

does not use random selection within each subgroup. Instead, the researcher selects readily available individuals who meet the quota criteria. The non-random selection can introduce bias.
  - » **Judgment (or Purposive) Sampling:** Samples are selected based on the researcher's judgment about which units will be most useful or representative. It is a selective and subjective sampling.
  - » **Snowball Sampling:** Samples are selected by asking the participants to nominate subjects known to them. This method is commonly used when investigating hard-to-reach groups. However, snowball sampling is subject to bias due to the lack of control over recruitment and those with more connections are more likely to be included.

- **Data Collection Mode** refers to the method or approach used to collect information from respondents in a survey. Common modes include online surveys, telephone surveys, face-to-face surveys, mail surveys, and mixed-mode surveys.

- **Survey Instruments** refer to specific tools or mediums used to collect information from respondents. Common survey instruments include:
  - » **Paper Mailout-Mailback Instrument:** A traditional paper questionnaire sent to respondents via postal mail, which they complete and mail back.
  - » **Self-Administered Questionnaire (SAQ):** A paper or electronic questionnaire that respondents complete on their own without an interviewer.
  - » **Face-to-Face Interview:** A structured or semi-structured questionnaire administered in person by an interviewer.
  - » **Computer-Assisted Personal Interview (CAPI):** An interview conducted face-to-face using a tablet or laptop, where the interviewer enters responses directly into the computer.
  - » **Computer-Assisted Telephone Interview (CATI):** An interview conducted over the phone with the aid of a computer system that guides the interviewer through the questionnaire and records responses.
  - » **Diary Methods:** Participants keep a diary or log of activities, behaviors, or experiences over a period of time.
- **Nonresponse Followup (NRFU)** is a process used in surveys to address and reduce the impact of nonresponse, where some respondents do not initially participate or complete the survey. It involves reaching out to nonrespondents through various methods, such as additional reminders, alternative contact methods, incentives, and in-person visits.

Types of Surveys

Surveys can be categorized into different types based on their design and data collection approach:

- **Cross-Sectional Survey:** Collects data from a representative sample of respondents at a single point in time, providing a snapshot of the population's characteristics or opinions at that moment.
- **Longitudinal Survey:** Tracks the same individuals or group of respondents over multiple points in time, allowing for the study of changes and developments within the sample over a period.

Survey Data Examples

Survey	Collection Approach	Target Population	Sampling Frame	Sample Size	Sampling Method	Survey Instrument	Data Products
American Community Survey	Cross-sectional survey	The U.S. population	The Census Bureau's Master Address File (updated twice a year with the USPS Delivery Sequence File)	1,980,550 housing units in 2022; 124,846 group quarter people	Multi-stage probability sampling	Internet self-administered questionnaire; paper instrument; CAPI follow-up for a sample of nonrespondents (NRFU).	Census American Community Survey data products
IPSOS Knowledge Panel	Longitudinal panel (by invitation) with an additional online opt-in panel	Adult U.S. population	60,000 random sampled panel members drawn from the USPS Delivery Sequence File frame.	Sample size varies for each survey	Address-based probability sampling	Only online questionnaire (it provides non-internet households with a tablet and mobile data plan)	IPSOS Poll
American Trends Panel (managed by IPSOS)	Longitudinal panel (by invitation)	Adult U.S. population	12,000 adult panel members drawn from the USPS Delivery Sequence File frame	Sample size varies (a subset of panelists)	Address-based probability sampling	Since 2016 online-only panel	Pew Research Center Survey Data
National Consumer Panel (NCP)	Opt-in Panel	U.S. households	Over 120,000 households in the Panel profile (among which, 46-52% is in its Static Panel)	n/a	Non-probability convenience sampling	Scanning equipment, or NCPMobile App to transmit shopping data	Data feeds Circana (formerly IRi) and NielsenIQ
YouGov	Opt-in Panel	Total population	27+ million registered panel members worldwide	1,500+ for each poll	Non-probability convenience sampling (responses weighted to be representative of the full population)	Online questionnaires	Data feeds the New York Times; and CBS News public polls.

To learn more, refer to **Data Quality Literacy Series 13: Evaluating Survey Data Quality**.

References

Census Bureau. (2022). *American Community Survey and Puerto Rico Community Survey Design and Methodology.* https://www2.census.gov/programs-surveys/acs/methodology/design_and_methodology/2022/acs_design_methodology_report_2022.pdf

Ipsos. (n.d.). *KnowledgePanel: A Methodological Overview.* https://www.ipsos.com/sites/default/files/ipsosknowledgepanelmethodology.pdf

Johnson, R. B., & Christensen, L. (2017). Methods of Data Collection in Quantitative, Qualitative and Mixed research. *Educational Research: Quantitative, Qualitative and Mixed Approaches,* 179-206. https://uk.sagepub.com/sites/default/files/upm-assets/106363_book_item_106363.pdf

Liu, G., Bordelon, B., Nagar, R., Sarti, J., Nguyen, U., & Boettcher, J. (2024). *Data Quality Literacy: A Guidebook.* Institute of Museum and Library Services (IMLS) Grant Project. https://doi.org/10.31219/osf.io/ruawm

Pew Research Center. (2014). *Q&A: What the New York Times' Polling Decision Means.* https://www.pewresearch.org/short-reads/2014/07/28/qa-what-the-new-york-times-polling-decision-means/

Pew Research Center. (2024). *American Trend Panel.* https://www.pewresearch.org/the-american-trends-panel/

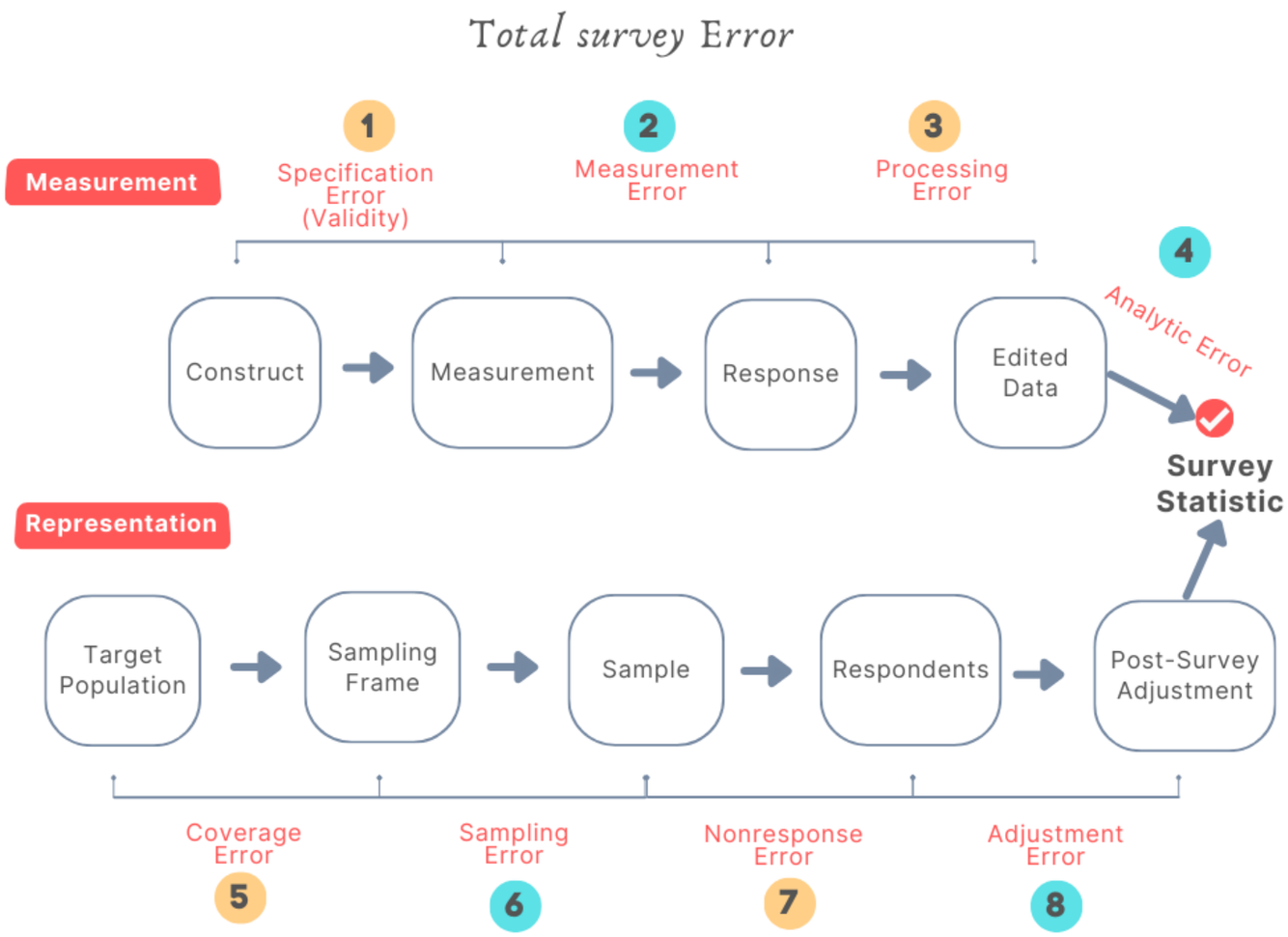
Ponto, J. (2015). Understanding and Evaluating Survey Research. *Journal of the Advanced Practitioner In Oncology, 6*(2), 168. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4601897/

YouGov. (n.d.). *Panel Methodology.* https://today.yougov.com/about/panel-methodology

The Knowledge Brief is compiled by Grace Liu, inspired by the National Forum presentation from John M. Abowd and Kathleen Weldon, reviewed by the IMLS Data Quality Literacy project team, and designed by Niko Galioto. This project was made possible in part by the Institute of Museum and Library Services [RE-252357-OLS-22].

Evaluating Survey Data Quality

The Total Survey Error (TSE) paradigm is a framework widely used to assess and enhance the quality of survey data. It addresses all potential sources of error throughout the survey process, from design and data collection to processing and analysis. The following factors for evaluating survey data quality are adapted from the TSE paradigm (see references).



Measurement Aspect

1. **Specification Error (Validity):** Do the data variables accurately measure the theoretical construct they are intended to measure? For example, are variables such as household income, education, or household wealth valid indicators to measure socioeconomic status?
2. **Measurement Error:** Does the survey instrument accurately measure what it is intended to measure?
   a. **Respondents** may deliberately or unintentionally provide incorrect information. For example, respondents agree or disagree with statements regardless of their content (response style behaviors); use less effort to provide optimal responses (satisficing); are unable to remember information accurately (recall bias); answer questions in a manner that will be viewed favorably by others (social desirability bias); or withhold accurate information on sensitive issues like drug use or sexual behavior.
   b. **Interviewers** may inappropriately influence responses or record the responses incorrectly.
   c. **Questionnaires** may be designed with unclear terms or jargon, ambiguous or leading questions, confusing instructions, or inadequate response options.
   d. **Mode of Data Collection:** There may be differences in responses due to the data collection method (e.g., online, phone, face-to-face) (also called “mode effect”).

3. **Processing Error:** Does the edited data accurately capture the survey responses? Are there any errors caused by data entry, coding, outlier editing, assignment of survey weights, or non-response imputing?
4. **Analytic Errors** arise in the post-processing steps after data has been collected from the field and stored in an analytic dataset. Errors may arise from incorrect merging, attribution of response to the wrong individual, incorrect use of survey weights, design features for estimation and inference, etc.

Representation Aspect

5. **Coverage Error:** Does the sampling frame represent the target population? Are some members of the target population excluded from the sampling frame? Examples include individuals without phone access in phone surveys, those without a permanent address in address-based sampling, or institutionalized populations (e.g., prisoners or dormitory residents), who often face undercoverage. The sampling frame may also contain errors such as omissions, duplicates, incorrect inclusions, and non-up-to-date information.
6. **Sampling Error:** Is the sample selected representative of the target population?
   a. **Sample Scheme:** Is the sample selected using probability-based random sampling methods, or less reliable non-probability methods such as online polls, opt-in panels, convenience sampling, or interactive voice response methods (e.g., robocalls or automated calls)?
   b. **Sample Size:** Is the sample size sufficient to ensure reliable and accurate results?
   c. **Estimator:** Are the statistical methods, formulas, or algorithms used to estimate population parameters robust and unbiased?
7. **Nonresponse Error:** Are there gaps between respondents and the sample?
   a. **Unit nonresponse** occurs when a sample unit (individual, household, or organization) does not respond to any part of the questionnaire;
   b. **Item nonresponse** occurs when the questionnaire is only partially completed and some items are not answered;
   c. **Incomplete response** occurs when the response to an open-ended question is incomplete or very short and inadequate;
   d. **Panel attrition** occurs when a sample unit is lost over the period of a longitudinal study.
8. **Adjustment Error:** Does the adjustment or correction made after survey data is collected reduce its accuracy or introduce new errors? Adjustment errors can happen in the following instances:
   a. **Weighting Adjustments** may be applied to correct for unequal selection probabilities, nonresponse, or to align with known population characteristics;
   b. **Post-Survey Adjustments** such as imputing missing values or making corrections based on identified biases;
   c. **Nonresponse Adjustments** such as methods used to address and account for nonrespondents in the sample.

Total Survey Quality

The TSE focuses on the accuracy dimension of survey quality. Total survey quality is also dependent on other non-statistical quality dimensions:

- **Credibility:** Is the data collection methodology credible and considered trustworthy by the survey community?
- **Comparability:** Are demographic, spatial, and temporal comparisons valid?
- **Usability/Interpretability:** Is the documentation clear, and is the metadata well managed?
- **Relevance:** Does the data satisfy the researcher's needs?
- **Accessibility:** Is access to the data user-friendly?
- **Timeliness and Punctuality:** Does data delivery adhere to the schedule?
- **Completeness:** Are the data rich enough to meet analysis objectives without placing undue burden on respondents?
- **Coherence:** Can estimates from different sources be reliably combined?

For details on assessing survey data's fitness for use and engaging in discussions with data providers and researchers, refer to **Data Quality Literacy Series 09 Commercial Data Quality: Conversation with the Vendors and 10 Commercial Data Quality: Conversation with the Researchers**.

To learn more about survey data, refer to **Data Quality Literacy Series 12: Understanding Survey Data and Public Poll**.

References

Biemer, P. P. (2010). Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly, 74*(5), 817-848. https://academic.oup.com/poq/article/74/5/817/1815551

Groves, R.M., and Lyberg, L. (2010). Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly, 74*(5), 849–879.

ITSEW, Lyberg, L. & Inizio. (2019). *Total Survey Error: Roots and Evolution.* INIZIO Presentation. https://www.niss.org/sites/default/files/ITSEW2019%20Primer%20-%20Lyberg.pdf

Liu, G., Bordelon, B., Nagar, R., Sarti, J., Nguyen, U., & Boettcher, J. (2024). *Data Quality Literacy: A Guidebook.* Institute of Museum and Library Services (IMLS) Grant Project. https://doi.org/10.31219/osf.io/ruawm

Maslovskaya, O. (n.d.). *Data Quality: Total Survey Error (TSE).* National Center for Research Methods Report. https://www.ncrm.ac.uk/resources/online/data_quality_and_survey_error/downloads/slides/data_quality_total_survey_error.pdf

Moya, C. (n.d.). *Inferential Statistics and Complex Surveys: Chapter 5 Total Survey Error Framework.* https://bookdown.org/cristobalmoya/iscs_materials/tse.html

West, B. & Schulz, P. (2018). *Total Survey Error: A Framework for High-quality Survey Design.* https://pdhp.isr.umich.edu/workshops/total-survey-error-a-framework-for-high-quality-survey-design-with-west-and-schulz/

The Knowledge Brief is compiled by Grace Liu, inspired by the National Forum presentation from Kathleen Weldon, reviewed by the IMLS Data Quality Literacy project team, and designed by Niko Galioto. This project was made possible in part by the Institute of Museum and Library Services [RE-252357-OLS-22].