Considering The Use Of Alternative Data In A New National Longitudinal Youth Survey

National Longitudinal Survey of Youth (NLSY)

Alison Aughinbaugh, Keenan Dworak-Fisher and Holly Olson

U.S. Bureau of Labor Statistics

April 5, 2020

National Longitudinal Surveys of Youth (NLSY)

For each cohort of individuals we follow, we select a random sample of youth – as they are entering the labor market or before they enter the labor market

- Sample sizes 9K-15K people
- Nationally representative (residents at time of survey start)
- NLSY79: 14-22 in 1979 (now 56-64) – just finishing R29
  - Additional Survey: Children of NLSY79 mothers
- NLSY97: 12-16 in 1997 (now 36-40) – R20 in fielding
- NLS26: 12-16 in 2026!!

Uses of NLSY Data

- Studies of changes over the life course
  - School to work transitions
  - Job search, Career evolutions
  - Health, Wealth, Human and Social Capital
- Studies of how an event affects future outcomes
  - School / Job training
  - Unemployment spells
  - Family structure, crime, program participation, etc.

Wide Range of Topics Covered → Many Potential Data Sources

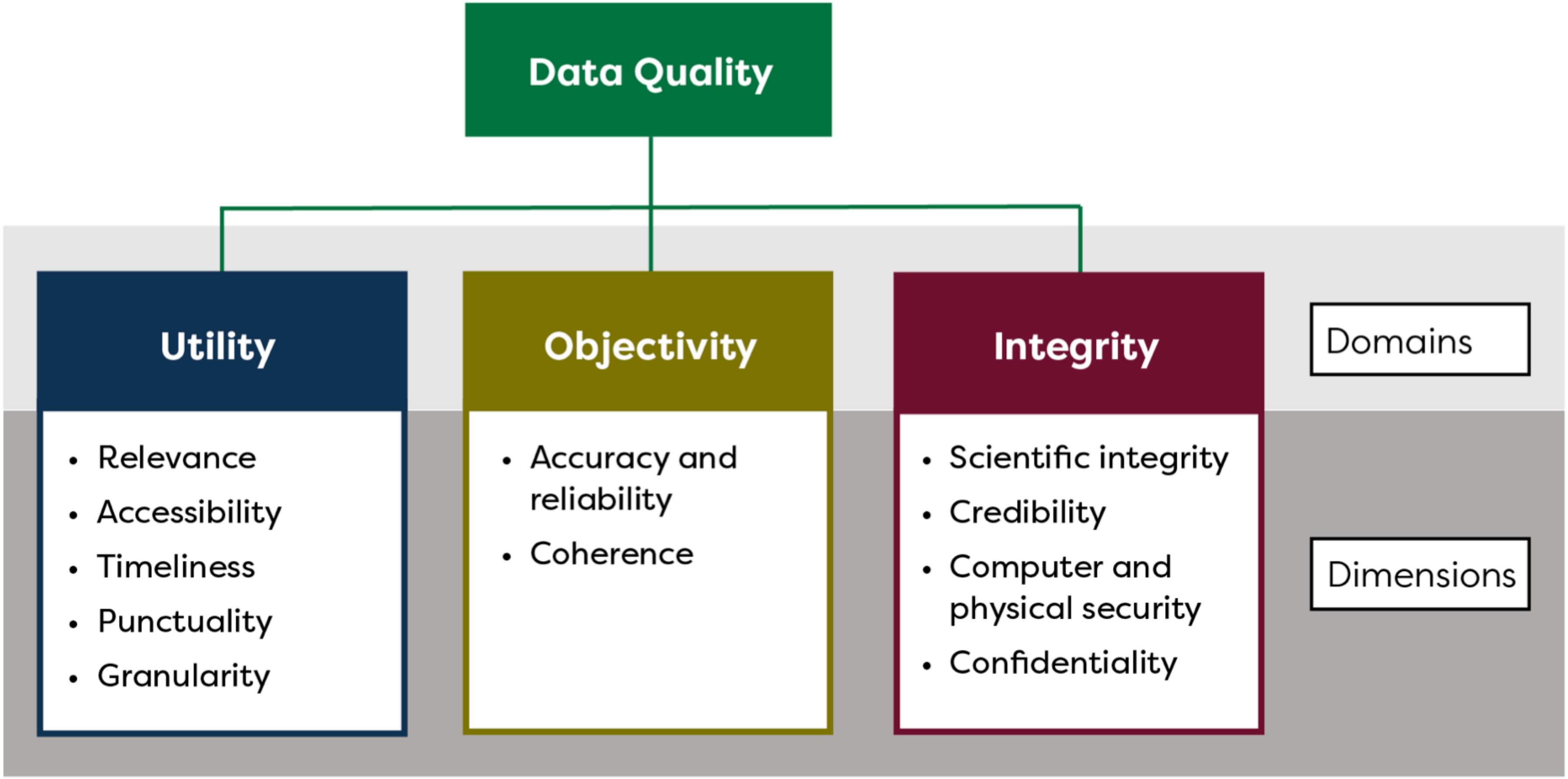
- Employment and Wages
  - UI Wage Records, NDNH, Social Security Records
- Education
  - National Student Clearinghouse, Student Loans
- Program Participation
  - Medicaid, CHIP, SNAP, WIC, Housing Assistance
- Crime
  - Criminal Justice Administrative Records System (CJARS)
- Personal Finance
  - Experian, Equifax

Multi-Dimensional Frameworks

- Murphy and Konny (2017) evaluate sources for Consumer Price Index (CPI) program
- Erhard, McBride and Safir (2021) evaluate sources for Consumer Expenditures (CE)
  - Build on Seeskin *et al* (2018)
- Federal Committee on Statistical Methodology (2020) - Broad View of "Data Quality"

https://nces.ed.gov/fcsm/pdf/FCSM.20.04_A_Framework_for_Data_Quality.pdf

FCSM Data Quality Framework



Data Quality

Utility

- Relevance
- Accessibility
- Timeliness
- Punctuality
- Granularity

Objectivity

- Accuracy and reliability
- Coherence

Integrity

- Scientific integrity
- Credibility
- Computer and physical security
- Confidentiality

Domains

Dimensions

4th Domain: Practical Considerations

Practical Considerations

- Financial Costs
- Respondent Burden
- Continuity / Feasibility

Data Source Evaluation Within the FCSM Data Quality Framework

- We ask a series of questions about the data source to assess how incorporation of the data would impact each quality dimension.
- We answer the questions using the best available information.
  - Answers may be revisited with more information
- Decision rules under development
  - Go / No-Go Structure for first round

Relevance

1. Can you identify potential user(s) of the data?
2. Are the data available through other avenues?
3. Are these data more or less useful to the user(s) than other available data?
4. Are the data complementary with other NLS data?
5. Will the data be readily accepted and used by researchers?

Timeliness and Punctuality

1. What is the time lag between the reference date(s) of the data and its collection by its source?

2. What is the time lag between collection of the data and its transmission to BLS?

3. What is the time lag between acquisition of the data by BLS and its availability to users?

4. Given the potential use(s) of the data, do the combination of these time lags reduce their applicability?

5. Can the data be obtained and processed on a reliable schedule?

Granularity

1. Does the data source provide information at a level of detail that supports researchers' needs?

2. Is the NLS sample large enough to provide meaningful statistics at this level of detail?

3. Will confidentiality concerns limit the level of detail that can be provided to users?

Accessibility

1. Can the data be reliably obtained from its source?

2. Can the data be easily incorporated into the NLS data base (e.g. through record linkage)?

3. Will users be aware of the data and/or be able to discover it readily?

4. How will users be able to access the data (e.g., PUF, RUF, other)?

5. Is the data well-documented (e.g., details of its origination, metadata for users)?

6. Will the data be treated for nondisclosure (e.g., through suppression or perterbation)?

7. Will users incur a cost for using the data?

Accuracy and Reliability

1. Does the target set (frame) of the data source (or an identifiable subset thereof) match the NLS cohort?

2. What proportion of units in the target set provide data?

3. What are the rates of duplication and missingness in the data?

4. What proportion of the data will be made available to NLS?

5. Does the data match the measurement concept of interest?

6. To what extent does the data contain internal inconsstencies or other known collection errors?

Accuracy and Reliability (ctd)

7. What proportion of the data are imputed or modified by the data provider after collection?

8. What other processing is done to the data before it comes to NLS, and are errors accrued?

9. What proportion of the data can be linked to records in the NLS cohort?

10. To what extent are duplicates and false matches present in the linkages?

11. To what extent are linked data consistent with other variables in the NLS cohort?

Scientific Integrity

1. Do the data acceptably or defensibly match with rigorous measurement principles?

2. Are the data subject to interference, manipulation, or misinterpretation by political interests?

3. Are the data prone to obsolescence as new data sources or methods arise?

Credibility

1. Will the data be accepted as factual by users?

2. Do the data conflict with other reliable sources of data?

3. Can the incorporation of the data be supported by transparent documentation of their origin?

Computer and Physical Security

1. Does the source of the data maintain high levels of computer security / integrity?
2. Does the process for transfering the data to BLS pose any issues for potential data corruption or breach?
3. Will the data be maintained by BLS in a way that is suitably secure?

Confidentiality

1. Do the data contain sensitive information for which loss of confidentiality may cause substantial harm?

2. To what extent do the data contain unique values that may be used to re-identify NLS respondents?

3. Are the data matchable to public sources of information that can be used to re-identify NLS respondents?

4. Does BLS have a suitable plan for ameliorating increases in the risk of re-identification that may be incurred by the incorporation of the data?

Financial Costs / Respondent Burden

1. What is the cost to the government of obtaining the data?

2. What costs may be saved by the use of this data (e.g., if it can be used to substitute for another source)?

3. How do the costs compare to the expected usefulness of the data?

4. Are the data supported by external funding sources?

5. How would incorporation of the data affect collection of the NLS survey and associated respondent burden?

Continuity / Feasibility

1. Does use of the data make BLS beholden to a given data source?

2. Can the data be expected to remain available for as many future rounds as are desired?

Illustrative Example 1: SNAP (Food Stamps) Participation

- Other survey programs – notably the Survey of Income and Program Participation (SIPP) have linked administrative records
- Research shows survey responses contain many false negatives
  - Meyer and Mittag (2019)
  - Celhay, Meyer and Mittag (2021)
- Previous NLSY cohorts collect survey response

Illustrative Example: SNAP (Food Stamps) Participation

- Relevance: high value to many users
- Timeliness: 6-month lag is not too long
- Accuracy: high when available
  - + Records are well-maintained
  - + Match rates are very high, esp with SSNs
  - – Coverage may be confined to certain states
- Coherence: high; Credibility: high
- Costs to Government: may be prohibitive

Illustrative Example 2: National Student Clearinghouse

- Student-level data on nearly all enrollments at post-secondary, title IV, degree-granting institutions in the US

Illustrative Example 2: National Student Clearinghouse

- Relevance: high value to many users
- Timeliness: 45-day lag is acceptable
- Accuracy: high
  - Participation, though voluntary, is high
  - Data come from official records (transcripts)
  - Linkages by name, DOB relatively accurate
- Coherence with other NLSY data: high
- Accessibility: Limits may be significant (FERPA)

Contact Information

Keenan Dworak-Fisher

dworak-fisher.keenan@bls.gov