Workshop 3: Transparency of Output Data Quality

Jennifer D. Parker, Ph.D.

Director, Division of Research and Methodology

Federal Committee on Statistical Methodology Research and Policy Conference

March 7, 2018

Workgroup 3 members

- Paul Marck, Census Bureau
- Jennifer Parker, National Center for Health Statistics
- John Popham, Office of Justice Programs
- Richard Reeves, National Center for Education Statistics
- Linda Young, USDA National Agricultural Statistics Service
- John Eltinge (ex-officio), Census Bureau

Disclaimer

- The findings and conclusions in this presentation are mine and do not necessarily represent the official position of the National Center for Health Statistics, Centers for Disease Control and Prevention.

Workshop 3 topics

- Opening: Opening remarks, recap of prior workshops, introduction to Workshop 3
- Session 1: Break in Series
- Session 2: Combining Data from Disparate Sources
- Session 3: Frameworks for Assessing Data Quality
- Summary

Session Overview by Linda Young

- Transparency of Output Data Quality (and next steps)

- Levels of Transparency
  - High Transparency (academics, agency specialists, subject-matter experts)
  - Medium Transparency (professional journalists, students, policy-makers)
  - Low Transparency (general public)

- Topics not covered
  - Sensitivity analysis

Output Data

- Blended estimates
  - Outcome statistics
  - Supporting statistics
- Micro-data files
  - Record-linked data files
  - Variables or other content on data files
- Results from internal and external research studies

Session 1: Break in Series

- Main Speaker: Lynn Langton, Bureau of Justice Statistic
  - Identifying and Addressing a Break (Blip) in Series

- Discussant: John Eltinge, Census Bureau

Langton:
Identifying and Addressing a Break (Blip) in Series

- National Crime Victimization Survey (NCVS)

- Design implemented 2016 incorporated changes in population and increased locations and sample size for state estimates

- Analysis confirmed that estimates had changed between designs more than expected and that there were effects for new field staff and for ‘time in survey’. However, it was not clear that these differences were real effects or artifacts of the design change.

Langton (continued)

- Decision to release unadjusted estimates, though adjustments for research purposes being considered

- Communicating the results to the different stakeholders was critical (e.g. data users, media, academics, DOJ officials)

Eltinge: Discussion of Langton's talk on break in series

- NCVS example not specific to blended data however the issue and the resulting decision processes apply to blended data
- General break in series issues affect the quality of statistical products and have possible impacts for stakeholders
- Statistical adjustments, judgement, and planning can mitigate impact
- Two-way communication is critical; listen to priorities of users

Session 1: Discussion

- Breaks/blips are particularly challenging when a key purpose of the survey is to monitor change

- There are trade-offs between incremental changes and big changes

- “Research series” that are consistent over time can be valuable
  - Statistical agencies are best suited to produce research series that incorporate possible adjustments

Session 2: Combining Data from Disparate Sources

- Main Speaker: Trivellore Raghunathan, University of Michigan
  - Combining Information from Multiple Data Sources: Challenges and Opportunities

- Discussant: William Bell, Census Bureau

Raghunathan:
Combining Information from Multiple Data Sources

- Large study to estimate prevalence rates and trends for multiple disease outcomes, attribute costs to these outcomes, and determine how much change in overall cost over time are due to changes in prevalence or changes in treatment costs

- 7 survey data sources and about 5 non-survey sources are being used in project

- About 120 health outcomes are being estimated

- Propensity and imputation methods used to combine information from each data source (not a record-linked study)

Raghunathan:
Combining Information from Multiple Data Sources

- Issues
  - Types of respondents and sources of information differ
  - There are mode effects, different survey designs, response error properties, question wording, coverage and other measurement issues across data sources

- Exciting opportunities to use Big Data and improve non-probability information using probability sample data

- “**It is dangerous to think that we do not need high quality probability surveys anymore**”

Bell: Discussion of Raghunathan

- Connected Raghu's approach to Small Area Estimation
- Highlighted many assumptions needed for success
    - Relationships between Y and X
    - Good estimates of sampling error are available and used
    - External standard can be used to assess error if it is unbiased or biases are negligible
- Assessments of estimates are optimistic as they assume models are true
- If improvements to estimates are modest, effort may not be worth risks of model failure

Session 3: Frameworks for Assessing Data Quality

- Speaker 1: Paul Biemer, RTI International
  - Assessing and Improving the Accuracy of Estimators from Blended Data

- Speaker 2: John Czajka, Mathematica
  - Transparency in the Reporting of Quality for Integrated Data: International Standards

Biemer: Assessing and Improving the Accuracy of Estimators from Blended Data

- Application of **total error framework** for hybrid estimators where error sources can be identified (and possibly mitigated) at each stage of the hybrid estimation process

- Framework can be applied to decompose intrinsic profiles for unified data and to hybrid estimates by source so that error risks for key components can be identified

- Comparisons of risk profiles between survey and blended estimates (and among survey, administrative, and unified datasets) can inform decisions

Czajka: International Standards

- International standards maybe be useful for our purposes as administrative data systems more developed, more rapid decline in response rates, and international organizations particularly active

- Key documents from the European Union
  - European Statistics Code of Practice for the National and Community Statistical Authorities
  - Quality Assurance Framework for the European Statistical System
  - European Statistical System Handbook for Quality Reports

Czajka: International Standards

- TSE was extended for integrated data by Statistics Norway (Zhang) and applied at Statistics New Zealand (Reid)
  - Zhang renamed TSE concepts to accommodate administrative data and blending in his two phase life-cycle model
    - Phase 1: sources of error for input data
    - Phase 2: sources of error from blending and harmonization processes
  - Extended by Stats NZ
    - Quality indicators for Phase 1 and Phase 2
    - Phase 3: sources of error for assessing estimates from Phase 2 products (no indicators yet)

Summary

- Speaker: Frauke Kreuter, Joint Program in Survey Methodology (JPSM)
  - Discussion of Workshop 3 (and next steps)

Kreuter: Summary

- Issues raised
  - Focus on the quality of the target estimate rather than datasets
  - Think about proxies; no data are perfect
  - Strengthen collaborations to exchange knowledge, build on data combining efforts, and share burdens/costs
    - Interdisciplinary teams broaden expertise and perspective
    - Burden shifting from front end to back end of process
  - Think differently about what we are doing

Workshop 3: General Discussion and Conclusions

- Perspectives are diverse
- Important to consider sources of error throughout the process for all types of outputs
- Frameworks for documenting errors and reporting quality exist for non-blended data that can be (and have been) adopted for blended data outputs
- While the most important components of such frameworks will differ among outputs and will differ among levels of transparency, having general principles for quality reporting will help users decide among data sources and make appropriate inferences

Jennifer Parker
jdparker@cdcgov