Leveraging Auxiliary Frame Data to Improve Survey Quality:

A Total Survey Error Perspective

Paul J. Lavrakas, Ph.D.

Independent Consultant

2023 AAPOR Webinar: June 20

Leveraging Auxiliary Frame Data

I. Intro/Overview/Instructor
II. Creating/Securing auxiliary date for your Frame and Samples
III. Noncoverage usage
IV. Sampling usage
V. Recruiting usage
VI. Nonresponse Bias usage
VII. Missing Data and Imputation usage
VIII. Weighting usage
IX. Discussion
X. Key Takeaways

Intro/Overview

- Auxiliary data, a form of Big Data, are those that can be matched to most, if not all, units/elements on a Frame, Initial Sample, and/or Final Sample
  - These are data that are known/knowable BEFORE the survey is planned and implemented
  - These data vary by the frame, but with some frames they can be extensive in quantity and rich in nature
- Today, almost all the examples I provide will be linked to using an address-based frame, but most of what I say can be, at least partially, adapted to other frames, including…
  - Name frames and samples
  - Telephone number frames and samples
  - Opt-in panel samples

Intro/Overview

- Today, I will be presenting information aimed primarily for those who have some awareness of the fact that many researchers are now routinely using auxiliary data to help improve their survey outcomes, but those of you that do not have experience using auxiliary data
- Such attendees may not realize the myriad ways that auxiliary data can be used to improve survey outcomes

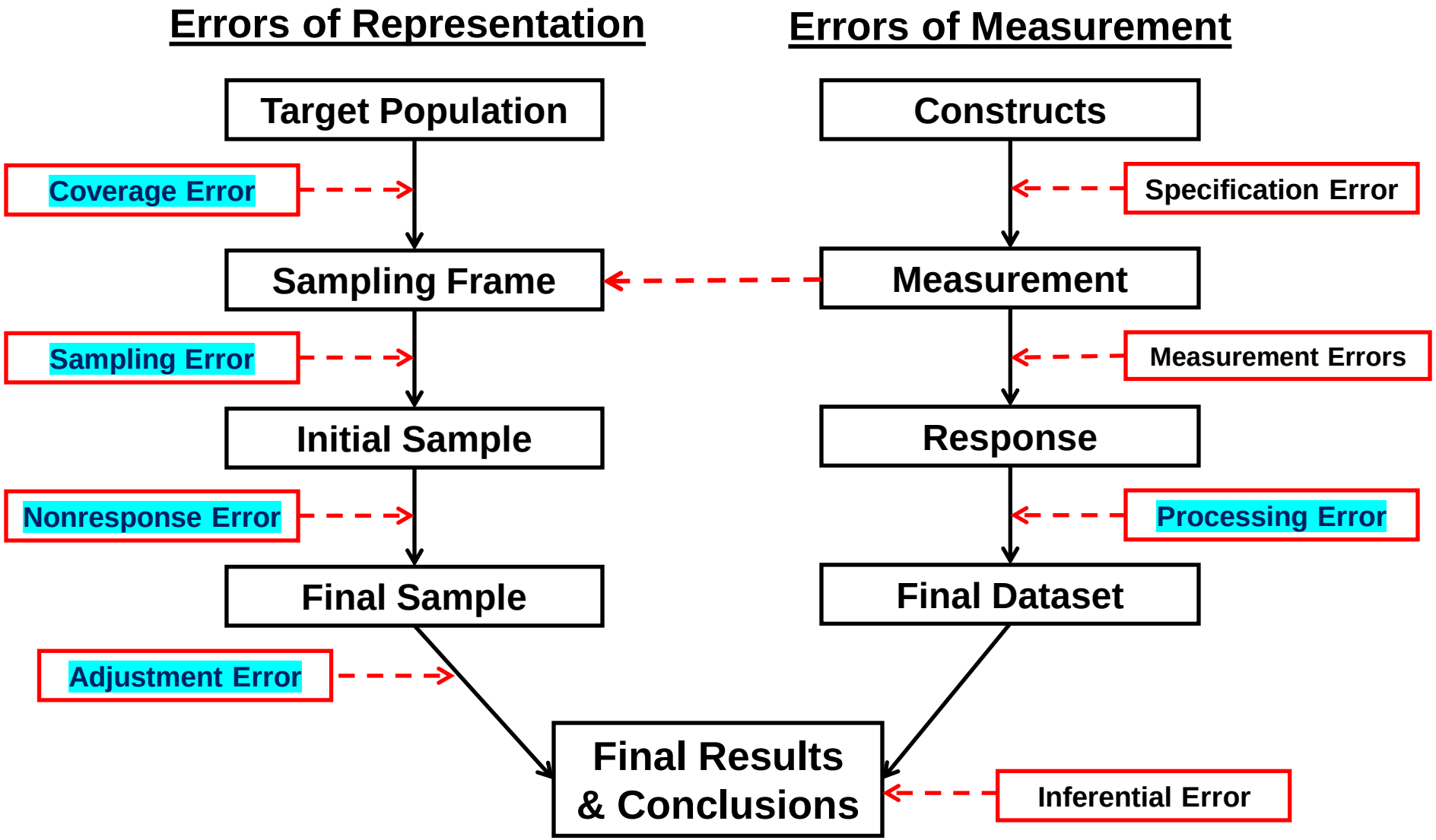
Intro/Overview

- Many survey researchers fail to take advantage of auxiliary data
  - In doing so, they are likely to limit the quality and usefulness of their survey findings
  - They may not use auxiliary data because they…
    - Not aware of the availability of and access to such data
    - Not aware of the value of such data
    - Not aware of how to use such data
    - Think they cannot afford (cost and/or time) to use such data, or that it is impractical for them to do so

Instructor – Why Me?

- I first started thinking about how to use auxiliary data in 1999 as I was preparing my remarks as a discussant at the *International Conference on Survey Nonresponse* in Portland OR in a session on "The Effects of Repeated Call-Backs and Reallocation on Non-Response Bias."
- Since then, I have used auxiliary data for many purposes, but apart from a presentation at the *BIGSURV18* conference in Barcelona, teaching a workshop at the 2021 WAPOR conference, and discussing the topic in the graduate course on survey design and operations that I teach, I've not formally presented anything about the topic prior to today
- **DISCLOSURE**: My goal today is to share my own views about why and how auxiliary data should be routinely used by survey researchers
  - My views are not meant to represent those of anyone else
  - No one supported me to prepare and make this presentation
    - During the past two decades, I have occasionally worked with staff at MSG on research related to what I am speaking about today, including my now-deceased friends, Dale Kulp and Ashley Hyon, but I have never received any compensation from MSG

Total Survey Error Framework



Securing Augmented Data for Your Survey

- Planning for using such data, before actually selecting your frame
  - What auxiliary data will correlated with the important constructs your survey is measuring?
  - What auxiliary data will correlate with Response/Nonresponse at the time of survey recruitment?
  - Where to get the frame and does that source have the auxiliary data that you need/desire?
- Enhancing an existing frame after gaining access to it
  - When no auxiliary data are appended
  - When some auxiliary data are appended
- Having to build your own frame
  - Often is impractical because of labor costs and time
  - But if it is done, then it is possible that a sampling company can match auxiliary data to the elements in your frame, initial sample, and/or final sample

Securing an Augmented Frame/Sample

- Types of Frame
  - Address frames seem to provide the greatest opportunity to append data onto them
  - Online Panel frames, will depend on what contact info is available about each person
  - Telephone frames and Name frames less so, but a lot of data still can be appended
  - Uncertain about Email frames

Securing an Augmented Frame/Sample

- Access to auxiliary data will vary greatly by country
  - EXAMPLE: USA vs Australia
- Basic types of auxiliary data that may be able to be appended
  - **Person level, e.g., head(s) of household**
    - Demographics: sex, age, race, education, etc.
    - Psychographics: hobbies, expenditures, past voting history, etc.
  - **Household level**
    - Demographics: family size, HH income, marital status, presence of children, etc.
    - Psychographics: credit card usage, magazines subscriptions, technology present, etc.
  - **Local area level**
    - Geophysical characteristics: Urban/Suburban/Rural, Hi/Lo Population Density, Physical size of local region, etc.
    - Geopolitical characteristics: % Political Party Affiliations, Type of local government, etc.
    - Socio-Economic characteristics: Pct. HHs in poverty, Pct. HHs use second language, % HHs w/Indoor plumbing, % Occupied dwelling units that are mobile homes, etc.
    - Other Area Demographics: % Females in labor force, % Bachelors degree, % HHs w/Retiree
    - Census Low Response Score (in the USA); Erdman & Bates 2016 *POQ* article

Securing an Augmented Frame/Sample

- Sources of Data to Match/Append, likely available for purchase from a Sample Vendor
  - Official Statistics publicly available from federal agencies; e.g., from a government census
  - Other publicly available data from state and local government agencies; e.g., local crime statistics, property taxes, automobile registration, birth and death statistics, past voting history
  - Proprietary databases assembled by private sector companies; e.g., property taxes, investments, expenditures, consumption behaviors, free-time activities and hobbies, media access, technology in HH, etc.

Securing an Augmented Frame/Sample

- Varied Quality of Data
  - Excellent to High quality with most government data
  - Reasonably Good to Modest quality with most private sector data
  - Missing Auxiliary Data for some elements/units
    - **BUT**, knowing of their "missingness" may still be very useful to survey researchers
      - For example, "missingness" of auxiliary data for some units/element is often predictive of nonresponse in a future survey
- Even if auxiliary data are imperfect, they may still be "good enough" to have value to researchers
  - Especially if the problem is reliability, and not bias

Noncoverage Bias Use of Auxiliary Data

- Investigate Coverage/Noncoverage of the Frame
  - Once your frame is chosen, compare the frequency distributions for local area federal characteristics available for the frame against frequency distributions for those same characteristics that the most recent Census/ACS produced for the geopolitical area your sample is supposed to represent
  - It is up to you to choose those appended characteristics that are available to you – ones that you think are most important for you to represent accurately
    - For example if you are doing a political survey you may decide that age, sex, race, education, and income in the geopolitical area that you are sampling are the most important demographics characteristics that you want well covered/represented by your frame
    - If so, then compare the distribution of Census/ACS parameters for these characteristics against the distribution of the same statistics for your frame

Noncoverage Use of Auxiliary Data

- But, before you do these comparisons, however, you should decide how "close" is "close enough" when comparing a frame statistic for your chosen geopolitical target population area, against its population parameter
  - For example, for one survey the researchers may decide that they need the frame to be within ± 1.0 percentage point of the parameters for the characteristics being compared; yet, another set of researchers may want no more than a ± 0.5 percentage point difference; other researchers may be able to tolerate larger differences
  - The size of the difference that is "acceptable" is up to the individual researchers to select and to be able to justify
  - If the differences between the frame statistics and their population parameters are all within the *a priori* threshold that the researchers selected, then the researchers have shown that their frame appears to have no nonignorable noncoverage errors on those characteristics
  - However, if the differences from the comparisons exceed the *a prior* threshold that the researchers judge to be acceptable, then the researchers basically have two choices:
    1. Accept the fact that your frame has nonignorable noncoverage and take that into account when adjusting and interpreting the findings from the survey, OR
    2. Try to find another frame that covers your target population better

Use with Sampling

- **Stratification of Initial Sample** – using auxiliary data known for essentially all elements on the frame
  - Does the variable correlate "strongly enough" with what will be measured in the survey?
  - If it does, it becomes a candidate for consideration to use in stratifying the selection of the initial sample
- **Oversampling**
  - Decide what subgroups, if any, need to be oversampled when creating the initial sample
    - For example, oversampling Spanish-dominant households in the USA
  - Use auxiliary frame data to best identify those subgroup(s) on the frame and draw proportionally more of those elements for the oversampling
    - For example, in the U.S. use block-level Census data for Hispanicity and Non-English language usage to select elements for a Spanish-Dominant oversample for the initial sample

Use with Sampling

- Investigate the Representativeness of the Initial Sample
  - Once your initial sample is chosen, compare the frequency distributions for local area federal characteristics available for the initial sample against frequency distributions for those same characteristics that the most recent census produced for the geopolitical area your sample is supposed to represent
  - It is up to you to choose those appended characteristics that are available to you, that you think are most important for you to represent accurately
    - For example if you are doing a health survey you may decide that age, sex, race, and education in the geopolitical area you are sampling are the most important demographics characteristics that you want well covered/represented by your initial sample
    - If so, then compare the distribution of Census/ACS parameters for these characteristics against what the distribution of the same statistics are in your initial sample

Use with Sampling

- Before you do these comparisons, however, you should decide how "close" is "close enough" when comparing an initial sample for your chosen geopolitical target population area, against its population parameters
  - For example, for one survey the researchers may decide that they need the initial sample that is drawn from the frame to be within ± 3.0 percentage points of the parameters for the characteristics being compared; yet, another set of researchers may want be able to tolerate as much as a ± 5.0 percentage point difference
  - The size of the difference that is "acceptable" is up to the individual researchers to select and to be able to justify
  - If the differences between the initial sample statistics and their population parameters are all within the *a priori* threshold that the researchers selected, then the researchers have shown that their initial sample appears to represent the target population well enough on those characteristics
  - However, if the differences from the comparisons exceed the *a prior* threshold that the researchers judge to be acceptable, then the researchers basically have two choices:
    1. Accept the fact that the initial sample is not representative to the extent desired, and take that into account when adjusting and interpreting the findings from the survey, OR
    2. Draw a different initial sample from the frame that represents the target population better

Use with Recruitment

- This is a form of Adaptive Design, that is carried out BEFORE the field period begins
- Tailoring heterogenous recruitment strategies (ala Dillman, Smyth & Christian, 2014) to particular types of initially sampled elements/units
  - As opposed to using a homogeneous One-Size-Fits-All (OSFA) recruitment strategy where all elements/units in the initial sample are recruited the same way
  - This adaptive approach to recruitment is highly consistent with Leverage Salience Theory (Groves et al., 2000; *Public Opinion Quarterly* article)
- Response Propensity Modeling (RPM) – possibly based on prior surveying (cf. Lavrakas, Jackson & McPhee, 2019; *Survey Practice* article)
  - 1st stage is to build a **RP model** possibly using similar past surveys to predict the propensity for a initially sampled element/unit to cooperate with the survey request
  - 2nd stage is to apply that RP model to the initial sample on a new survey before the field period begins to create a **RP score** (ranging from 0.00 to 1.00) for each element/unit in the initial sample and use that to form **RP subgroups** that will be recruited differently from each other

Use with Recruitment

- **Example**: If conducting a single cross-sectional survey, with no past "equivalent" survey to model from
  - Use auxiliary data matched to all elements in the initial sample; in the USA, include the Census Bureau's "Low Response Score"
  - Build an RP score variable that is expected to predict Response/ Nonresponse in the forthcoming survey
    - Use auxiliary data to create an RP score that are historically known to correlate with response/nonresponse in your geopolitical area of interest; e.g., in the USA use education, race, age, income, housing type, employment status, urbanicity, local usage of non-English languages, the LRS from Census, etc.
  - Create the RP score variable for all elements in the initial sample, and then create two or more RP groups
  - Devise differential recruitment treatments for each group (see next slide)

Use with Recruitment

- **Example (CONTINUED)**:
  - Within the constraint of the final budget available for recruitment, tailor different recruitment protocols for each RP group so that more effort/resources are expended to gain cooperation from the groups predicted to have lower RP scores and less effort/resources are expended to gain cooperation from groups predicted to have higher RP scores
  - For example, use three subgroups and vary the value of a noncontingent incentive, the value of a contingent incentive, and the use of an advance contact postcard
    - **For the group with lowest RP scores**: $5 noncontingent incentive, $20 contingent incentive, and send an advance postcard
    - **For the group with medium RP scores**: $2 noncontingent incentive, $10 contingent incentive, no postcard sent
    - **For the group with highest RP scores**: $1 noncontingent incentive, $0 contingent incentive, no postcard sent
  - This will differentially affect response rates for the different subgroups with the expectation that it will reduce the possibility of nonresponse bias
    - Response rate for the Highest RP group will be lowered so as to not overrepresent them in the final sample (and thus need to later downweight them) by as much as they would be otherwise if using a OSFA recruitment for all the initial sample; the opposite should happen for the Lowest RP group

Use for Nonresponse Bias Investigations

- Studying nonresponse biases in a survey has become an increasingly common activity in the past 15+ years in the USA
- Ways to do this are explained by Montaquila and Olsen (2012)
  - Google "Practical Tools for Nonresponse Bias Studies" to get the PDF of their slide set
- Directly comparing nonrespondents, using auxiliary data appended to the initial sample with respondents, is one of the important and useful ways to investigate and estimate the size and nature of nonresponse biases
- This can be done by using auxiliary data matched to essentially all elements in the initial sample to make such comparisons after the survey has been conducted, since it then is known which elements in the initial sample became respondents and nonrespondents (excluding those elements/units that were found to be ineligible during the field period)

Use for Nonresponse Bias Investigations

- **Example**
  - Choose variables that are reasoned to correlate both with response/nonresponse in your survey and with what your survey is measuring
    - For example, with a survey on health, choose auxiliary variables such as:
      - **Local Level**: % Minority, % Foreign born, % Adults with no H.S. Degree
      - **Household level**: Engages in physical activities, Retired, Home owner, no matched phone number to their address
      - **Person level**: Householder speak foreign language, age of Householder, employment status of Householder, Householder is a Smoker
  - After survey is completed, identify which variables, if any, are reliably correlated with response/nonresponse and consider the strengths of those correlations
  - Identify those auxiliary variable that are both correlated with (1) response/nonresponse and (2) the key health measures in the survey
  - Decide whether to use the identified auxiliary variables in weighting the survey

Use to Reduce Missing Data

- This issue only concerns a survey's respondents – i.e., those who completed the questionnaire
- The issue is how to use auxiliary data from the frame to help impute missing data
- Steps
  - Choose one variable at a time that has enough missing data to want to use imputation to replace those missing values
  - Identify auxiliary data and other data from the questionnaire that are expected to correlate with the variable that has missing data and which are available for essentially all the respondents
  - Use those auxiliary data and other variables – from the respondents in the survey who provided a substantive answer to the variable being imputed – in multivariate analyses to identify the best set of predictors of the variable that has missing data
  - Use those predictors in the imputation analyses to generate values to replace the missing data in the selected variable

Use to Reduce Missing Data

- **Example:** In a political survey, build a model to impute missing data for the "Horserace" question by using the following types of data that are available on respondents:
  - From the rest of the questionnaire use variables such as sex, race, age, education, political party affiliation, and income
  - From the auxiliary data use variables for local-area characteristics such as
    - % females in the labor force in the locality
    - recent past voting history in the locality
    - racial/ethnic makeup of the locality
    - median single family home value in the locality
    - type of dwelling units in the locality
    - Population density on the locality

Use in Weighting Adjustments

- To Supplement or Compensate for Lack of other Weighting Variables
  - **Coverage:** If there are no reliable population parameters available, use auxiliary frame data which explain noncoverage as "parameters" to weight the final sample's characteristics on these same frame variables against
  - **Sampling**: Use the frame auxiliary data which were used in the sampling design if that led to unequal probabilities of selection of the initial sample
  - **Nonresponse**: If there are no reliable population parameters available, use auxiliary frame data which explain nonresponse as the "parameters" to weight the final sample's characteristics on these same frame variables against
    - Try to use psychographic characteristics that are known to correlate with the key measures of the survey AND are reasoned to correlate with response/nonresponse AND which are available as auxiliary data on the frame
    - But will need to gather these variables in the survey's questionnaire

Use in Weighting Adjustments

- Possibly add the above auxiliary variables to any other survey characteristics of the respondents in the final sample for which there are population parameters available, that the researchers will use in weighting

Discussion

- With finite budgets one will most likely not be able to do everything possible with auxiliary data on a given survey project
- It is still too early in the history of leveraging auxiliary frame data to understand with confidence which use(s) of auxiliary frame data will most likely reduce TSE vs. whether the cost of using them to do so is justified
- Costs to a Survey Organization
  - The cost of appending auxiliary data will vary by whether they are appended to only the initial sample, the final sample, or to all or some part of the frame

Discussion

- Costs within a Survey Organization (CONTINUED)
  - The labor costs of using appended frame data may be expensive depending on how many ways the data will be used
    - If used for all possible purposes, then likely at least one FTE at the survey organization may be need to devote all or a substantial portion of her/his time
      - As this person's expertise and experience using auxiliary data grows, costs should decrease if demand does not grow
      - But as the benefits of using auxiliary data are realized by the organization it would be expected that demand will grow
  - Use of auxiliary frame data may provide a Competitive Advantage
- Longer-term planning
  - Even if a survey organization cannot soon start leveraging auxiliary data, systematic thinking about the infrastructure that will be eventually be needed can (and should) commence

Review – Takeaways

- Many ways to leverage auxiliary frame data
- If a survey organization has not started using auxiliary data then it needs to give very careful thought to creating a systematic plan to introduce their usage into the organization and build the necessary infrastructure
- Once a survey organization begins to use auxiliary data, an R&D program should be established to help the organization identify the most cost-beneficials ways that the organization should use auxiliary data
- Budgeting: Should the organization offer clients an ala carte menu of auxiliary data usages to choose from???
- Special POQ issue in 2023
  - *Augmenting Surveys with Paradata, Administrative Data, and Contextual Data*
- The future for the benefits that should accrue when using auxiliary frame data seems rosy…

Thank you!

pjlavrakas@comcast.net