Imputing Across Interviews: Balancing Time Savings with Data Quality

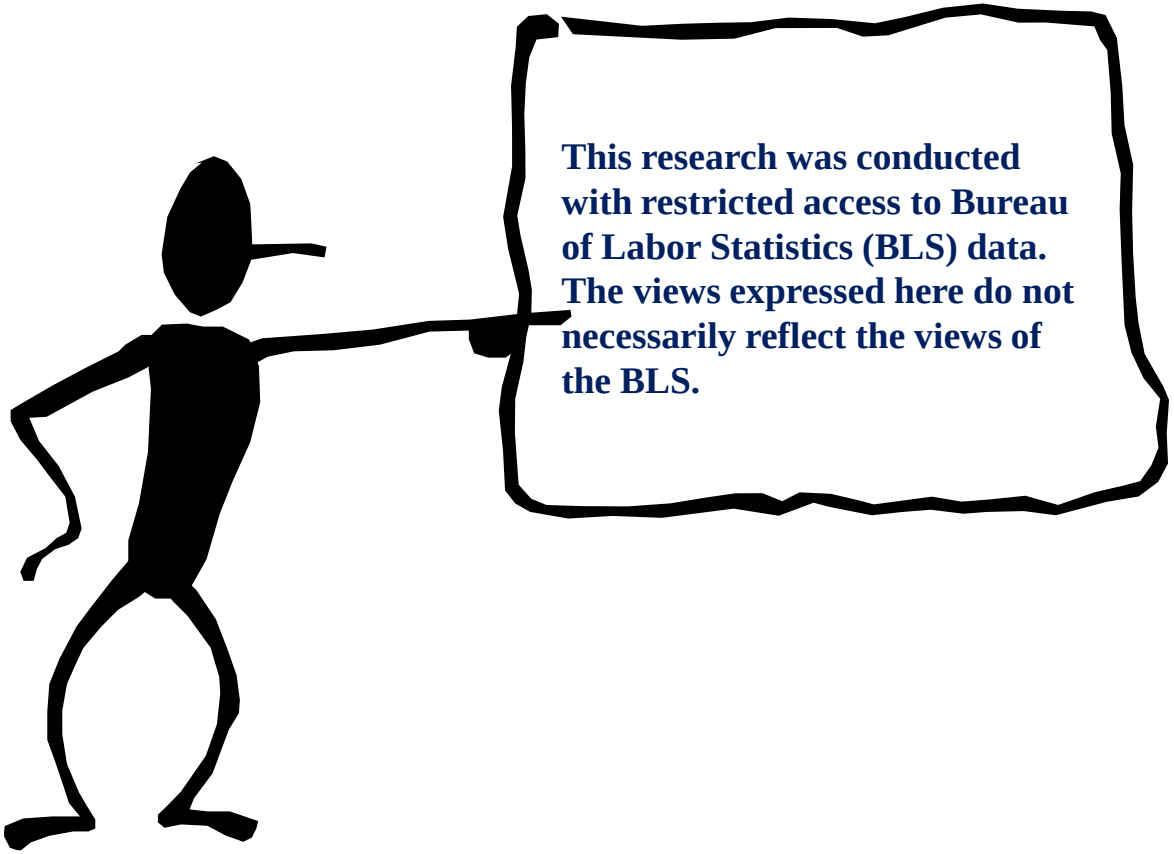
by

Geoffrey Paulin, Ph.D.
Senior Economist
Bureau of Labor Statistics

NISS Workshop on Analyzing Complex Survey Data with Missing Item Values
October 17, 2014
Washington, DC

BUREAU OF LABOR STATISTICS
U.S. DEPARTMENT OF LABOR

Disclaimer:



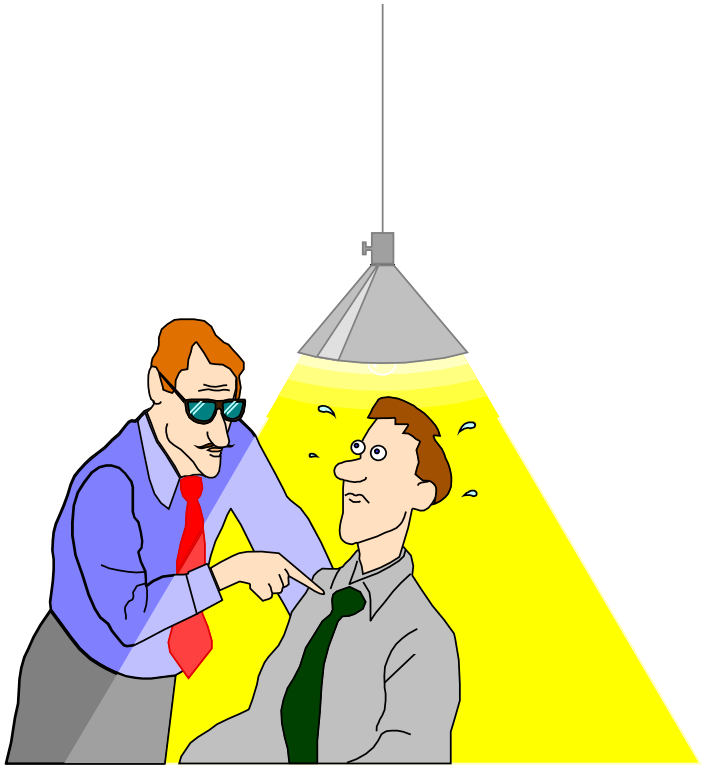
125 YEARS BLS

The Interview component of the Consumer Expenditure Survey (CE) is:

- The most detailed source of expenditures, demographics, income, assets, and liabilities collected directly from consumers by the Federal government.
- Currently collected in five visits over consecutive three-month periods (i.e., quarters).

125 YEARS BLS

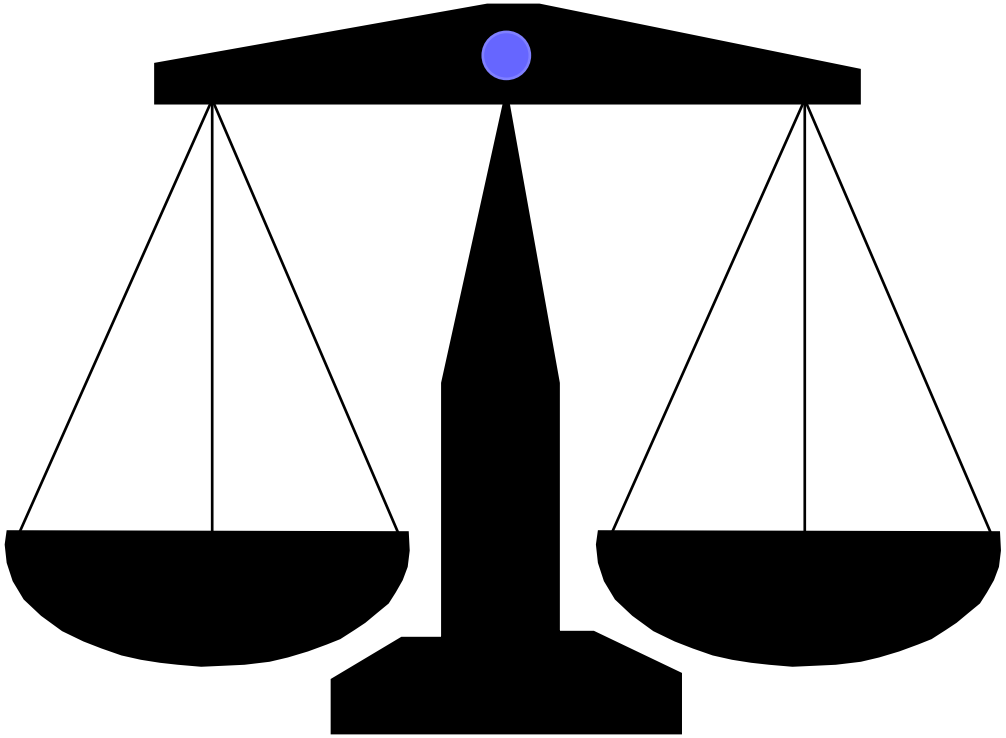
Reducing respondent burden is an important goal…



…In 2011, the average quarterly interview was one hour; 10 percent exceeded 100 minutes.

BLS

However, this must be balanced with maintaining high quality of data.



BLS

Data quality includes:

- Providing accurate estimates of means and variances of expenditures for tabular data
- Preserving correlations among expenditures, demographics, and other variables for microdata users

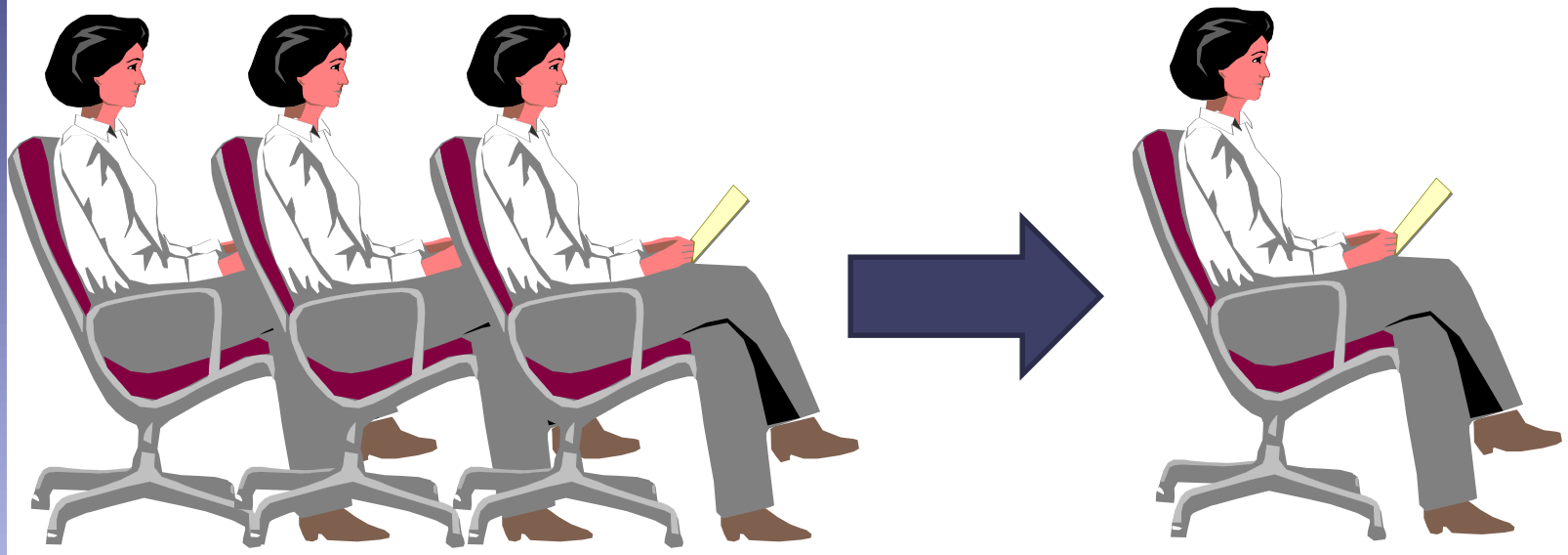
BLS

In 2015, the Bounding Interview will discontinue.

- CONSEQUENCES:
  - Need to add bounding information to current 2nd interview
  - Current 2nd interview time will increase, which was already shown to be a concern
- QUESTION: Can expenditures collected in the (current) 2nd interview be successfully imputed from (current) 3rd, 4th, & 5th interviews to minimize response burden?

BLS

To achieve this, the CE program investigated the feasibility of imputing results from later interviews to the current second interview.



3rd 4th 5th 2nd

This presentations includes:

1. The conceptual framework investigated
2. Problems encountered
3. A request for comments

Two basic categories of expenditure were considered:

1. Food at home
2. Utilities
    a. Electricity
    b. Telephone
        1) Landline
        2) Cell
        3) Phone card/voice over IP

BLS

1. Food at home

- Global question: Respondents are asked about "usual" weekly/monthly expenditures
- Nearly universally reported (almost 99 percent in 2011)
- In 2011, Section 20 is the second most time-consuming expenditure section

BLS

2. Utilities

Reasons for considering:

- Section 4 is the most time consuming
- Expenditures are expected to occur each month, which makes processing easier (no need to decide in which month to place an expenditure; just allocate across the three)
- Expected to be highly correlated with explanatory variables already collected (housing size, types of appliances, region/State/PSU, urban/rural, city size)

BLS

Procedural Concerns and Clarifications:

- "Back Imputation"—that is, using reports from a specific consumer unit's 3rd, 4th, and 5th interviews to impute that consumer unit's 2nd interview is not feasible as it:
  - Causes delays in production (process cannot start until subsequent interviews have been completed;
  - Is still subject to nonresponse. (What happens if the unit participates in the 2nd, but no subsequent, interview?)
- For these reasons, regression using data from ALL consumer units participating in 3rd, 4th, and 5th interviews will be performed. Collection periods will be matched for source data. For example: 3rd, 4th, and 5th interviews from January of a given year will be used to impute 2nd interview values collected in January of that same year.

BLS

As noted, expenditures were estimated by regression analysis.

- **Hot decking was considered but rejected.**
  - Currently, hot decking is used when respondents report that an expenditure occurred, but not the amount. The team investigated the possibility of adopting this approach for the larger project.
  - However, the limitations of hot decking are well-documented (e.g., ability to use few predictor variables; effects on variance).
  - The limitations are less problematic for filling in nonresponse blanks, especially when item nonresponse rates are low. But in this case, all expenditures would be imputed.
  - The inability to properly preserve correlations among expenditures and independent variables would be detrimental to microdata users.
- **Therefore, regression was used.**

BLS

The first item considered was Food at Home.

BLS

Models:

- Included several independent variables
  - Standard demographic characteristics (age, education, etc.);
  - Geographic identifiers (region, PSU);
- Were run for different family types (e.g., single person, husband and wife only, etc.).

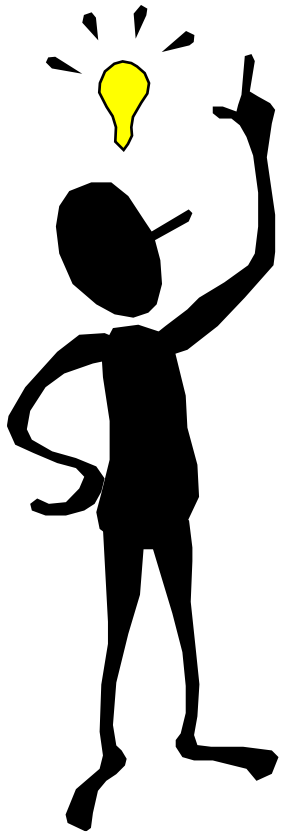
BLS

Results:

- R-squares were small
- Mean absolute deviations were large in percent terms

BLS

The second item considered was electricity.



Models were:

- Large. They included:
  - Standard demographics (age, education, etc.)
  - Special variables such as—
    - Number/type of appliances in household, where known
    - Detailed geographic data as described earlier
    - Type of housing (detached, townhome, highrise, dormitory, mobile home, etc.)
- Complicated. Run separately by Region:
  - Within region, by housing tenure (homeowner or renter)
    - Within housing tenure by family type
      - Single person, husband and wife only, etc.
      - Economies of scale, number of potential users of electricity are related to family type

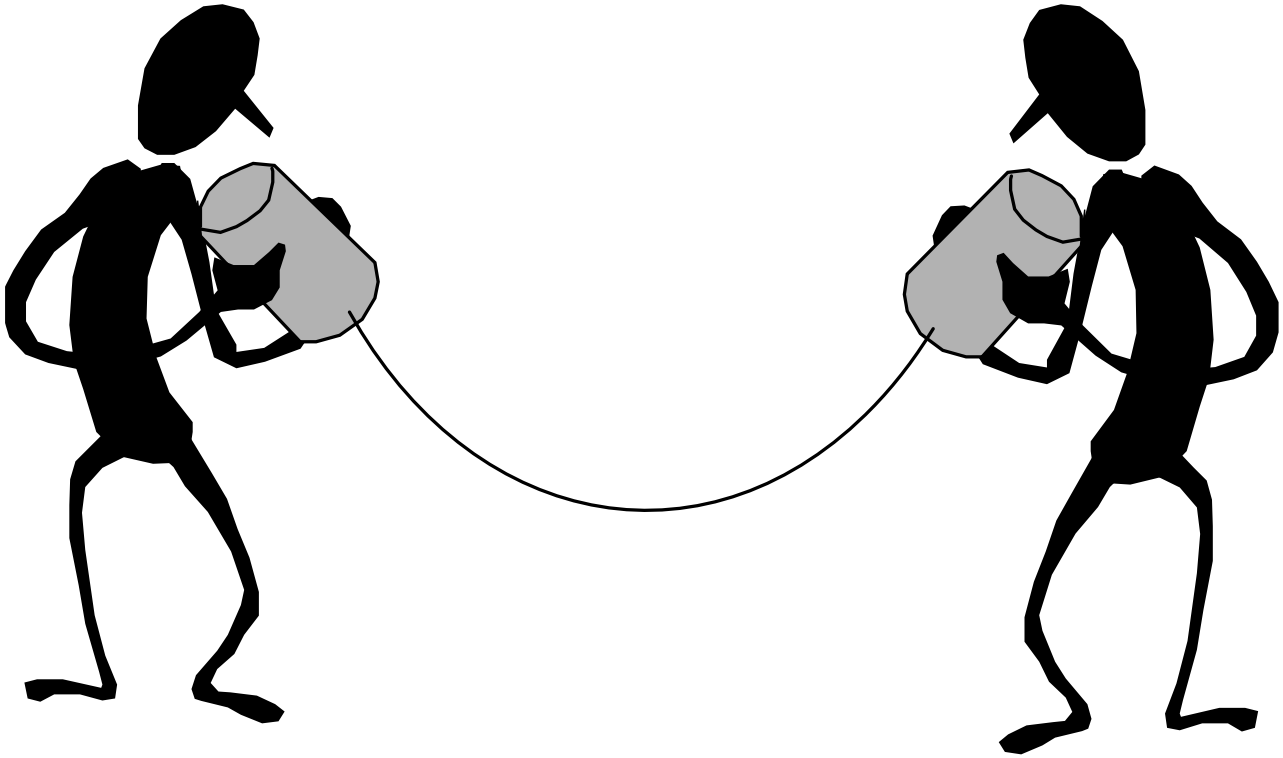
BLS

Results:

- R-squares were larger than for food at home, but still not large
- Mean absolute deviations were still large in percent terms.

BLS

The third item considered was telephone service.



Caveats:

- As with electricity, expenditures can vary substantially by billing location of the consumer.
- Bundling plans (e.g., cable television and internet included with phone service) increase the difficulty of imputing phone service alone.

BLS

Findings:

- As with food at home and electricity, R-square values were low.
- Presumed correlation of telephone expenditures for families with each type (e.g., landline and cell phone) raises additional concerns:
  - Which is better: A simultaneous equations model, or impute one type, and use results to impute the next?
  - In latter case, the second imputation includes an independent variable that is 100 percent imputed, affecting quality of the result.

BLS

In Summary:

- The dropping of the bounding (1st) interview from the Interview Survey in 2015 will necessitate asking new questions in what is currently the 2nd interview.
- To minimize the burden this will add, the CE program investigated the feasibility of using expenditures collected in the current 3rd, 4th, and 5th interviews to impute expenditures, instead of collecting them.

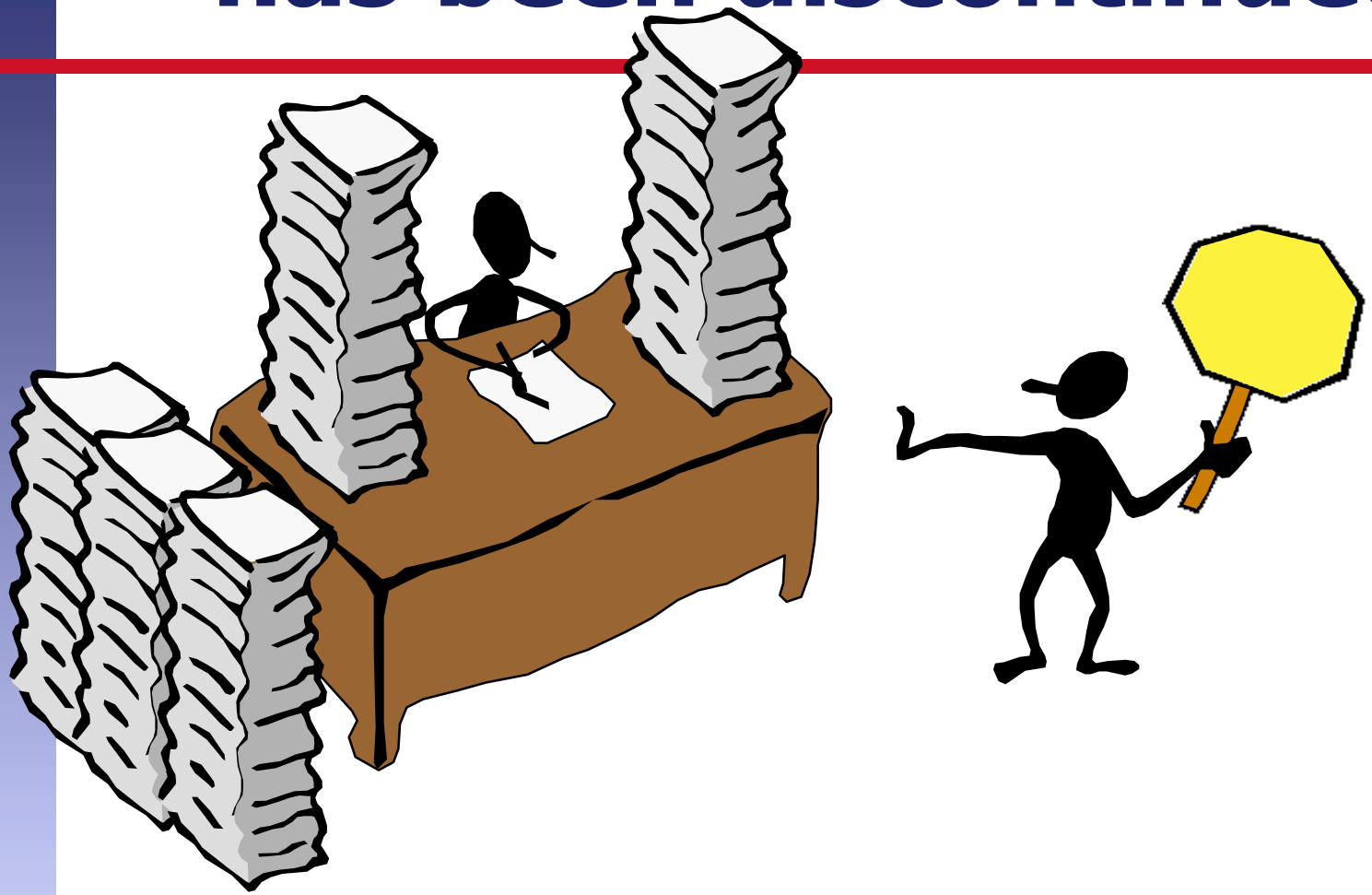
BLS

Findings:

- The quality of the results (low R-square values, large absolute deviations) were insufficient to warrant further investigation.
- The utilities tested (electricity and telephone) require complicated regression models and/or methods making both research and implementation difficult.

BLS

Based on this, the work has been discontinued.

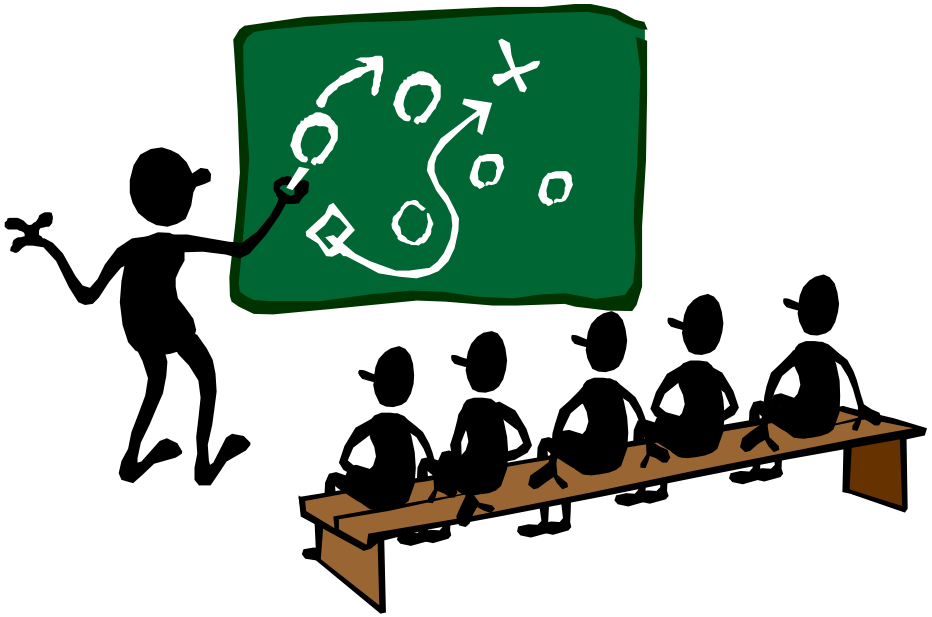


Next Steps:

- The CE program has chartered a new team to investigate imputation of assets and liabilities when values are missing due to nonresponse.
- Methods currently under consideration range from hot deck to multiple imputation.
- Perhaps other methods described today will also prove to be viable options.

BLS

Therefore, if you have any suggestions, comments, or questions of your own…



…The team looks forward to hearing from you.

BLS

Contact Information

Geoffrey Paulin, Ph.D.
Senior Economist
Consumer Expenditure Survey Program
www.bls.gov/cex
202-691-5132
Paulin.Geoffrey@bls.gov

BLS
BUREAU OF LABOR STATISTICS
U.S. DEPARTMENT OF LABOR