Split Questionnaire Designs for Consumer Expenditure Survey

Trivellore Raghunathan (Raghu)
University of Michigan

BLS Workshop December 8-9, 2010

Rationale

- Well Accepted Concept:
  - Not all subjects in the population need to be measured in order to obtain inference about the population
- Extend the same notion
  - Not all variables need to be measured on every sample subjects in order to construct population inferences
- Reducing the burden may increase response rate, improve data quality and hence better inferences
  - Reducing the burden may itself be a goal

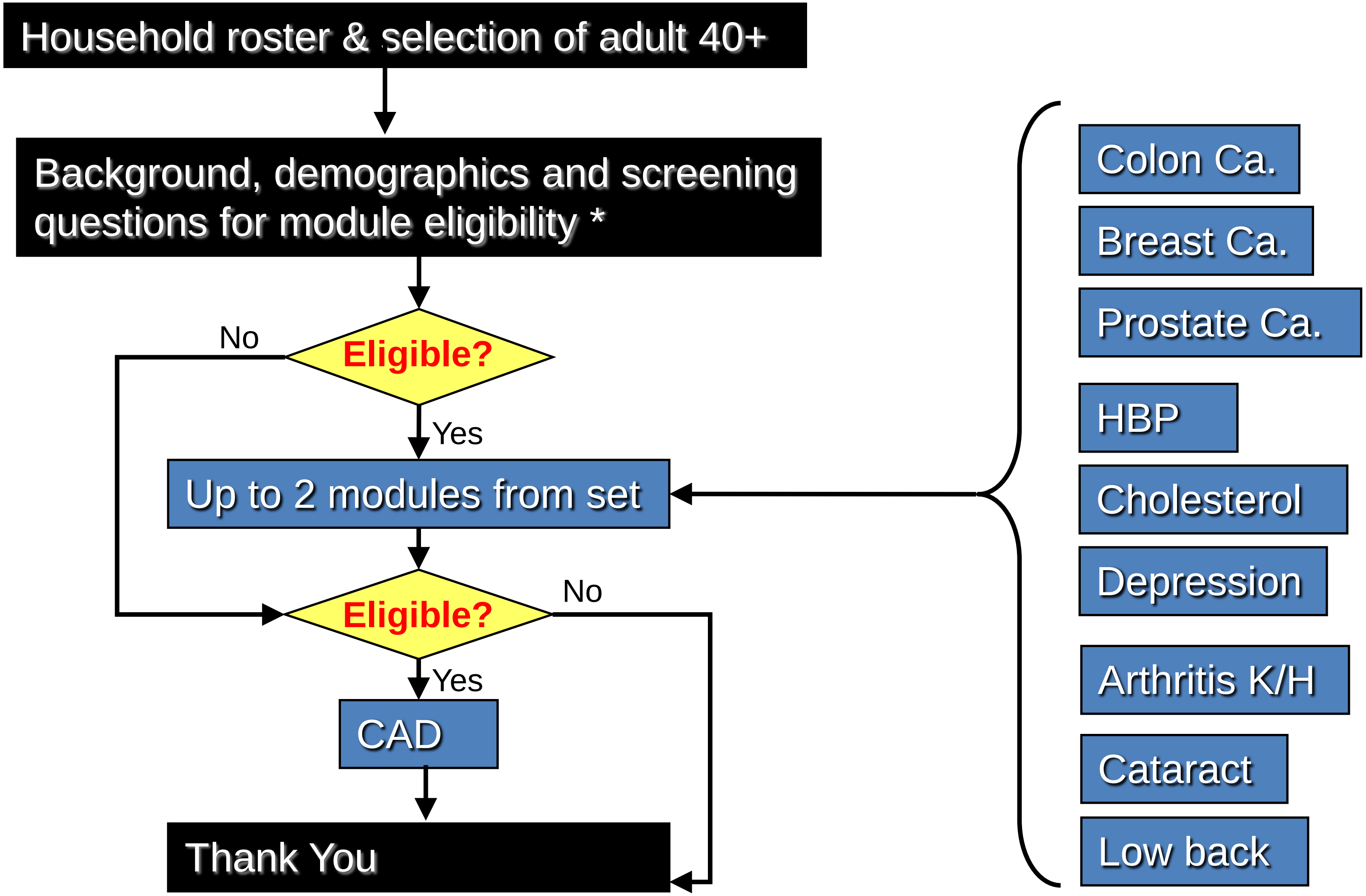
Origins

- Market Research
  - Index constructed from a long list of items (e.g. mean)
  - More missing data or neutral or middle options after a few items
  - Solution: Sample the items for each subject
- Raghunathan and Grizzle (1995, JASA)
  - Designed sampling of items to be able estimate certain key statistics from the observed data
    - Mean and the covariance matrix
    - Up to third order interaction
- National Assessment of Educational Progress (NAEP) uses a similar design (not all students can be tested on all domains)

National Survey of Medical Decision Making

- A large scale survey based on a national probability sample
- Very long questionnaire involving many aspects of decision making
- Wanted to survey respondents about 10 specific medical decisions, but keep the burden of the survey to a minimum (goal of 25 minutes)
- While the relative prevalence of these decisions could be estimated, the marginal or joint prevalence of making one of these medical decisions in the past 2 years is unknown

Instrument Design Structure



Household roster & selection of adult 40+

Background, demographics and screening questions for module eligibility *

Up to 2 modules from set

CAD

Thank You

Colon Ca.

Breast Ca.

Prostate Ca.

HBP

Cholesterol

Depression

Arthritis K/H

Cataract

Low back

Module Selection Algorithm

- Allocation of modules inversely proportional to prevalence rate

  P_i =Prevalence rate for condition i

- Suppose that a subject is eligible for K modules leading to $\binom{K}{2}$ pairs of modules
- Assign pairs to the subject with probability proportional to

$$\pi_{ij} \propto \frac{1}{P_i P_j},\ \sum_{ij} \pi_{ij} = 1$$

Issues

1. What is the impact of breaking up the interview on data quality (e.g., reporting, response rates) and respondent burden?

Two issues:

What are the current issues on item response rates and “patterning of responses” by the order in the questionnaire?

Is it possible to conduct a split-ballot experiment to compare the full and split-questionnaire designs?

2. What are the cognitive aspects of breaking up the interview that CE should consider?

- Reducing the cognitive burden is an important advantage in the split questionnaire design
- Breaking-up the questionnaire needs some thinking about the context for each item
  - “Context integrity” may require certain items to be placed in the same split

3. What features of the current CE (e.g., types of expenditures being collected, panel) would have the greatest influence on design and estimation issues?
    - The details does not have to be collected on all types of expenditures on all subjects
    - Judicious use of stem and leaf questions
        - Option 1: Collect all stems (as a part of core) subsample leaves as a part of split
        - Option 2: Split the stem-and-leaf combinations

- It is not necessary to give the same splits across the panel for all subjects
    - Maximize information by giving the same splits for some subjects and different splits for other subjects

4. What are the primary statistical issues, in addition to the ones cited above, that CE needs to address when investigating the utility of these methods?
   - From the imputation point of view, need to be able to predict the missing component from the observed component (Thomas et al 2006, Survey Methodology)
   - Need a careful study of the existing CEX data to develop split questionnaires and imputation model
   - Administrative data sources and some external data resources needed

- Potential Simulation Study
  - Take the half sample impose split questionnaire on it and analyze
  - Compare with the other half
  - Repeat

5. What are the implications for the primary CE data users (e.g., CPI, published tables, and academic community)?

Realistically, Core and Splits might be most palatable way . Items needing high efficiency and those not well predicted from other variables may have to be in the Core

Need to educate of the user (after educating ourselves)

6. What are the operational challenges associated with implementing these types of designs?

- In the context of CATI and CAPI interviews these can be programmed and does not need to involve interviewers or the field staff
- Mail questionnaire requires a bit more organization by the field staff
- Paper-questionnaire is probably the hardest

7. What should the next steps be to explore and research this issue for a possible change in CE methods?

- Need detailed analysis of the current CEX data to develop potential split questionnaire
- Simulation study to evaluate the estimation properties
- Field split-ballot experiments to compare the split versus full-questionnaire

Conclusions

Current research at BLS shows promise for this approach

- Need more research on refined models for the analysis for descriptive statistics and analytical statistics
- Imputation models
- Need research on developing potential splits informed by the past data and additional external data resources
- Need simulation and Field experiment to compare the split and full questionnaire survey designs