"Challenging Research Issues in Statistics and Survey Methodology at the BLS"

Problem Statement: Application of Sequential Design and Testing Methods and Adaptive Sampling Methods to the Design and Evaluation of Usability Tests

Key words: Adaptive design of experiments; Capture-recapture methodology; Data collection; Data dissemination; User interface.

Contact for further discussion:
John L. Eltinge
Office of Survey Methods Research, PSB 1950
Bureau of Labor Statistics
2 Massachusetts Avenue NE
Washington, DC 20212
Telephone: (202) 691-7404
Fax: (202) 691-7426
E-mail: Eltinge.John@bls.gov

Background:

BLS behavioral scientists often carry out usability tests intended to identify and correct problems with, e.g., questionnaires and other instruments used for data collection, and websites used for data dissemination. These instruments or websites are generically called "interfaces." The standard approach is to ask several potential users to attempt to use the interface and then identify specific problems they encountered during the attempted use. Through sequential identification and correction of these problems, the researcher intends to produce an improved interface.

For some general background on usability testing, see, e.g., Blair and Conrad (2005), Conrad and Blair (2004), Nielsen and Landauer (1993), Nielsen (1994) and references cited therein. For the current discussion, three important questions in this literature are:

A. How many testers do we need to identify most of the usability problems in a given interface?

B. Can we use information from initial tests to estimate the remaining number of usability problems in the interface?

C. What are specific ways in which we should structure the testing to identify and correct as many usability problems as possible in the interface?

Issue: To what extent can methods of sequential and adaptive experimental design, adaptive sampling, capture-recapture sampling methods, or response-surface experimental design shed some light on questions (A)-(C) above?

Questions on the Application of Sequential and Adaptive Statistical Methods:

1. Within the statistical literature, there is a substantial body of work on sequential and adaptive design of experiments, and in the interim and final analyses of such experiments. This literature has arisen primarily in biostatistics and especially in the sequential design of clinical trials, e.g., Müller and Schäfer (2001), Rosenberger (1996, 2002); Wei, Su and Lachin (1990); Yao and Wei (1996) and references cited therein. Much of this literature is focused on, e.g., comparison of two specific medical treatments or other comparative work that is qualitatively different from the usability testing framework encountered at the BLS. However, it appears that much of the underlying mathematical structure developed in the sequential and adaptive literature could potentially be applicable to usability testing.

   In addition, there is some related work specifically in the literature on software testing, e.g., Dalal and Mallows (1988).

   - Are there specific ways in which some results from this literature can be applied to develop improved methods for identification and correction of usability problems?

   - If so, can the resulting methods be applied to practical testing of BLS data collection instruments or BLS data dissemination web pages?

2. In addition, there is a substantial statistical literature on adaptive sampling. See, e.g., Christman and Feng (2001); Seber and Thompson (1996); Schwarz and Seber (1999) and references cited therein. This literature covers several complex topics, but a simple motivating example is the estimation of the total number of members of a species in a given area, when the members of the species tend to move in herds of unequal sizes. Stated in a slightly more abstract form, much of the literature considers estimation of population totals and identification of relationships (or links) among members or groups of members in the presence of population clustering and linking.

   - In some cases, usability issues can be mapped on physical or conceptual spaces, and there are natural links between some points within such spaces. Can the literature on adaptive sampling suggest some specific sampling methods for the identification and correction of usability problems?

   - If so, can the resulting methods be applied to practical testing of BLS data collection instruments or BLS data dissemination web pages?

3. Some usability testing sub-topics have some features similar to those in capture-recapture methodology (e.g., Alho, 1994; Alho et al., 1993; Bunge and Fitzpatrick, 1993; Ding and Fienberg, 1994; Pollock et al., 1994; Wolter, 1990)

and response surface methodology (e.g., Myers and Montgomery, 2002; Khuri, 1996; and references cited therein).

- To what extent can the capture-recapture or response-surface literature provide additional insights into methods that are appropriate for usability testing?

Acknowledgements

The author thanks John Dixon for helpful comments on an earlier draft of this topic statement. The views expressed here are those of the author and do not necessarily represent the policies of the Bureau of Labor Statistics.

References

Alho, J.M. (1994), Analysis of sample based capture-recapture experiments, *Journal of Official Statistics*, **10**, 245-256

Alho, J.M., M.H. Mulry, K. Wurdeman and J. Kim (1993), Estimating heterogeneity in the probabilities of enumeration for dual-system estimation, *Journal of the American Statistical Association*, **88**, 1130-1136

Blair, J. and F. Conrad (2005). The effect of sample size on cognitive interview pretest results. Preliminary report to the Bureau of Labor Statistics, February 10, 2005.

Bunge, J. and M. Fitzpatrick (1993), Estimating the number of species: A review, *Journal of the American Statistical Association*, **88**, 364-373

Christman, M.C. and L. Feng (2001). Inverse adaptive cluster sampling. *Biometrics* **57**, 1096-1105.

Conrad, F. and J. Blair (2004). Aspects of data quality in cognitive interviews: The case of verbal reports. In S. Presser et al. (eds.), *Methods for Testing and Evaluating Survey Questionnaires*, in press.

Dalal, S.R. and C.L. Mallows (1988). When should one stop testing software? *Journal of the American Statistical Association* **83**, 872-879.

Ding, Y. and S.E. Fienberg (1994), Dual system estimation of census undercount in the presence of matching error, *Survey Methodology* **20**, 149-158

Khuri, A.I. (1996), Multiresponse surface methodology. Pp. 377-406 in *Design and Analysis of Experiments (Handbook of Statistics, Volume 13,* S. Ghosh and C.R. Rao, eds., New York: North-Holland).

Müller, H.-H. and H. Schäfer (2001), Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches, *Biometrics* **57**, 886-891

Myers, R.H. and D.C. Montgomery (2002), *Response Surface Methodology: Process and Product Optimization Using Designed Experiments.* New York: Wiley.

Nielsen, J. (1994). Heuristic evaluation. Chapter 2 in J. Nielsen and R.L. Mack (eds.), *Usability Inspection Methods*, New York: Wiley.

Nielsen, J. and T.K. Landauer (1993). A mathematical model of the finding of usability problems. *Interchi 1993*, 206-213.

Pollock, K.H., S.C. Turner and C.A. Brown (1994), Use of capture-recapture techniques to estimate population size and population totals when a complete frame is unavailable, *Survey Methodology*, **20**, 117-124

Rosenberger, W.F. (1996). New directions in adaptive designs. *Statistical Science* **11**, 137-149

Rosenberger, W.F. (2002). Randomized urn models and sequential design (with discussion). *Sequential Analysis*, **21**, 1-41.

Schwarz, C.J. and G.A.F. Seber (1999). Estimating animal abundance: Review III. *Statistical Science* **14**, 427-456

Thompson, S.K. and G.A.F. Seber (1996). *Adaptive Sampling.* New York: Wiley.

Wei, L.J., J.Q. Su and J.M. Lachin (1990). Interim analyses with repeated measurements in a sequential clinical trial. *Biometrika* **77,** 359-364

Wolter, K.M. (1990), Capture-recapture estimation in the presence of a known sex ratio, *Biometrics*, **46**, 157-162

Yao, Q. and L.J. Wei (1996), Play the winner for phase II/III clinical trials (with discussion). *Statistics in Medicine* **15** , 2413-2458