Conceptual and Practical Issues in the Statistical Design and Analysis of Usability Tests

John J. Bosley (Bosley_J@bls.gov), BLS,
John L. Eltinge (Eltinge_J@bls.gov), BLS,
Jean E. Fox (Fox_J@bls.gov), BLS,
Scott S. Fricker (Fricker_S@bls.gov), BLS

Presenter and Contact Author: John J. Bosley, Office of Survey Methods Research, PSB 1950, Bureau of Labor Statistics, 2 Massachusetts Avenue NE, Washington, DC 2021

Key Words: Computer-assisted interviewing (CAI); Data dissemination; Mean and dispersion effects; Predictive distribution.

1. Introduction

In recent years, the Federal statistical community has focused a considerable amount of effort on usability testing for data collection instruments and for data dissemination facilities. Usability testing can help ensure that data are collected easily, efficiently, and accurately. It can also help ensure that those who retrieve Federal statistics, particularly from websites, find and understand what they are looking for. At the Bureau of Labor Statistics (BLS), we have incorporated usability testing into the development process for many data collection instruments, including the Commodities and Services (C&S) survey for the Consumer Price Index, the Consumer Expenditures Survey, and the American Time Use Survey. We have also conducted usability testing for data dissemination facilities such as BLS' own website, LabStat.

The goal of usability testing is to find problems with an application and identify solutions before the application is distributed to end users. In a usability test, typical users perform typical tasks with a prototype of the application. The usability specialists collect objective data, such as number of users who completed a task correctly and the number of screens accessed to complete the task. Usability specialists also collect subjective data such as their own observations and participant comments and ratings, and then evaluate the data to reveal problems and solutions. The evaluation usually includes an analysis of the quantitative measures collected, but the analysis tends to be very simple. Usability specialists look for cases where the values seem inappropriately high or low. For example, they may look for unusually high task times or error rates or unusually low satisfaction ratings. One reason for the simple analyses is that usability tests often do not meet the assumptions necessary for more rigorous statistical analyses. Dumas and Redish (1993) list a variety of reasons that usability test results may not warrant more detailed statistical analysis. According to these authors, sample sizes generally are small (5-10 participants per user population), and sample selection and assignment of treatments to experimental units may not include randomization. Further, many of the measures are nominal or ordinal, but not interval variables, so they violate the assumptions of many statistical tests typically used in psychological research. In addition, to the extent that statistical methods are used in the analysis of usability data, previous authors have tended to focus on formal hypothesis tests (e.g., t, chi-square or F tests) for equality of treatment means.

As interface development becomes more refined, one encounters important questions that are not readily addressed by the simple analyses described above. For example, non-usability specialists, e.g., managers of software development projects, may not readily accept expert conclusions based on a casual analytic approach. Similarly, knowledgeable clients may raise informed questions regarding trade-offs in performance as measured in multiple dimensions. Consequently, this paper notes some ways in which statistical methods in experimental design, formal inference and exploratory analysis may shed additional light on the abovementioned usability testing issues. Section 2 outlines four issues that may be of special interest in the statistical analysis of usability data. Section 3 introduces an example involving testing of alternative approaches to a topical-link function for federal statistics; and presents some related exploratory data analyses. Section 4 suggests some areas for additional work.

2. Four Statistical Issues in Usability Testing

2.1. Exploratory Analyses Linked with the Practical Implications of Extreme Outcomes

In usability testing, the primary practical goal generally is to identify interfaces that may entail unreasonably high risks of unusually high or low values of one or more outcomes Y. Within this framework, the concepts of "unreasonably high risks" and "unusually high or low values" often are very context dependent. For example, in some applications there may be an expectation that almost all users should be able to complete a task in less than a specified number of minutes, and any interface that did not meet this criterion would be deemed unacceptable. On the other hand, in other applications there may be a greater tolerance for the fact that some users may take a much longer amount of time to complete a task with a given interface, or may not be able to complete the task at all. Also, even within a given context, the terms "unreasonably high risk" and "unusually high or low values" often are not precisely defined *a priori;* and the relative importance of different outcome measures (e.g., time to completion or reported frustration) are not prespecified. Consequently, agencies may obtain better insights into the usability of an interface through exploratory analysis of the outcome data, rather than through formal hypothesis testing procedures. For these reasons, it may be advisable to structure analyses of usability data in forms that allow relatively direct examination of the full predictive distribution of outcome measures Y that are expected to arise with the particular user group(s) of interest.

2.2 Dispersion Effects

In some cases, one may address the issues of Section 2.1 through statistical analysis of mean and dispersion effects. For example, if for each interface i, the vector of outcomes $(Y_{i1},...,Y_{ir})'$ follows a normal distribution with mean m_i and variance-covariance matrix Σ_i, then the prevalence rate of extreme values Y depends entirely on the parameters m_i and Σ_i. Thus, for this case, it is appropriate to focus principal analytic effort on identification of interfaces i that have exceptionally large means or variances. To date, mean effects have received principal attention in the statistical analysis of usability data, and tend to follow relatively standard approaches to formal hypothesis testing for the equality of means, as covered, e.g., in Johnson and Wichern (1998). However, in some cases different treatments may have relatively similar means, but have marked differences in their variances. For these cases, a treatment that leads to a higher variance may in turn lead to an unacceptably large proportion of users having exceptionally high (or low) outcomes Y. In such cases, it is worthwhile to consider detailed modeling of the underlying variances as functions of the treatments and relevant covariates. A large body of literature on variance function modeling has been developed in the literature for biostatistics and engineering statistics (e.g., Carroll and Ruppert, 1988, Chapter 3; Box and Meyer, 1986; and Davidian, 1990), and could be applied to analysis of dispersion effects in usability data.

2.3. Nonparametric Comparison of Predictive Distributions

In cases for which we have relatively large sample sizes and for which the outcome vectors do not follow an approximate normal distribution, it is of interest to carry out a nonparametric comparison of the predictive distributions of the outcomes Y across different interfaces. For some outcome variables (e.g., user preferences recorded on a five-point Likert scale) this nonparametric comparison is relatively simple. For other outcome variables (e.g., the time or number of "mouse clicks" required to complete a given task) the nonparametric comparisons may be somewhat more complex. One option involves display of prediction intervals for new observations Y, or associated tolerance intervals. Another option involves display of side-by-side quantile plots or boxplots of observations Y.

2.4 Distinctions Among User Subpopulations

In principle, a given interface may be used by members of different subpopulations defined by differences in such characteristics as levels of formal training, experience, self-perceived ability to use computers, or motivation. For example, in work with usability testing in computer-aided interviewing, the principal subpopulations of interest may be defined by levels of training and experience within the interviewer corps. In this example, three subpopulations might be field supervisors, experienced interviewers and relatively new interviewers. In the data dissemination context, Marchionini and Hert have attempted to define various qualitatively distinct user types, with a view toward designing and providing differently-configured interfaces for each different type. (Hert and Marchionini, 1997; Marchionini, 2000.) In general, these researchers sought to make role-wide distinctions, e.g., "teacher," "student," "economic research professional," "business owner," and so on. They did focus on attempting to define a task-based classification scheme on occasion, but this was not the primary approach. This has two practical implications for statistical work with usability data. First, it is important for the design of the usability study to select persons appropriately from the user subpopulation(s) of interest; and for results to be interpreted accordingly. Second, if the usability study involves more than one subpopulation, then there may be special interest in evaluation of interactions between interface effects and subpopulation effects.

3. An Example: Usability Testing of Alternative Approaches to a Topical-Link Search Function for Federal Statistics

One commonly-used technique to enable users to find information on a large, information-rich website is to create a content index to the site and then use that index as a list of hypertext links that will take the user to the topic that the link label describes. The cross-agency "portal" site, Fedstats (http://www.fedstats.gov) is one of several major federal statistical websites for data dissemination that uses a topical index to support searches. This site contains links to data that reside at more than 100 different U.S. federal statistical agencies with a wide range of missions. The Fedstats home page presents a variety of structured search functions as alternatives to the search engine. Among these is a "Topics A-Z" index list of hyperlinks, along with a search by agency and a search by geography ("Mapstats").

Prior usability tests (Marchionini and Hert, 1997) of the Fedstats interface had raised concerns about the effectiveness of the "Topics A-Z" index for certain tasks and for certain types of users. To address these concerns, Ceaparu and Shneiderman (2003) conducted three empirical studies of alternate organizational schemes for FedStats "Topics A-Z." After each test, the FedStats site was modified to incorporate the results. Each study had 15 participants, all of them graduate students from a wide variety of disciplines. The gender composition of the three samples was reasonably equivalent. In each test, all participants performed three "scripted" search tasks ("scenarios") per version. These three tasks were always presented in the same order. The first scenario required construction of an answer to a complex question involving urban development in formerly rural areas. The second scenario had a much more specific goal--locating data about sales trends for organic products in a metropolitan area. The third scenario asks the participant to find data that would allow comparison of two U.S. cities for "livability" by user-defined criteria. Of the three, the last appears to be the least well-defined or structured, with the second the most clearly-defined task and the first somewhere between them.

Ceaparu and Shneiderman (2003) recorded task completion, the total elapsed time for a given task, an overall measure of reported user frustration (on a scale of 1-10), and seven measures of specific types of user frustration (each also measured on a scale of 1-10). For the current paper, we have carried out additional analyses of these data; some relatively simple results are displayed in Figures 1 through 7. More detailed analyses will be reported in a longer version of this paper. Figure 1 presents boxplots of the "Overall Frustration" scores for subjects who received treatment 1, with separate plots for each scenario. For a given scenario, the top and bottom of the grey box represent the seventy-fifth and twenty-fifth percentiles of the observed values; the middle line represents the median; and the solid dot represents the mean. In addition, the upper and lower "whiskers" represent the extreme upper and lower values of the observed scores. Figures 2 and 3 present related boxplots for treatments 2 and 3, respectively. Comparisons of Figures 1 through 3 indicate that the degree of variability in the "overall frustration" scores (as reflected in the interquartile range, equal to the difference between the seventy-fifth and twenty-fifth percentiles) varies considerably across scenarios and across treatments. For example, the interquartile range for Scenario 3 is substantially larger for Treatment 2 than it is for Treatments 1 or 3. In other analyses, not reported in detail here, we observed similar (and in some cases, more pronounced) patterns of heterogeneity of variance in task completion time and other scores. Figures 4 through 6 present a corresponding set of boxplots for the "Confusing Search" component of the frustration score. These plots again display some indication of heterogeneity of variance (e.g., for Scenario 1 in Treatments 1 and 2). In addition, note that for Treatment 3, the distribution of the "Confusing Search" scores was also influenced by a "floor effect" since most of the scores were relatively close to zero.

Finally, Figure 7 presents a scatterplot of frustration scores against completion times, with different plotting symbols used for each treatment. The frustration scores and completion times are "jittered" in this plot to avoid problems with overstrikes. The plot indicates a moderate degree of positive association between the (subjective) frustration scores and the (objective) completion times; a related simple linear regression of frustration score on completion time had an R^2 value equal to 0.115. Subjects corresponding to points in the upper left and lower right corners of the scatterplot may be of special interest for follow-up analyses. For example, subjects in the upper left corner completed the assigned task relatively quickly, but nonetheless reported relatively high levels of frustration. It would be of interest (but beyond the scope of the current work) to explore the extent to which these phenomena may be associated with subpopulation membership (which in turn may be linked with expectations regarding completion time); or associated with specific components of the frustration score that are relatively independent of completion time.

4. Discussion

This paper has noted some areas in which statistical analysis methods may offer some insights into usability data. A more detailed examination of this area would need to consider several additional topics, e.g., sample size selection and power analysis; nonresponse; measurement error; Hawthorne and other intervention effects; and use of formal factorial or fractional factorial experimental designs to identify important factors in mean and dispersion effect models, and to construct correspondingly improved interfaces. In considering these issues, it would be important to identify specific cases in which the costs of increased sample sizes, complexity of study design, and analytic burden can reasonably be justified by the resulting additional insights into human-computer interaction that would not have been obtained through simpler approaches. (See also, for example, Nielsen, 1993).

Acknowledgements

The authors wish to thank Ms. Irina Ceaparu and Prof. Ben Shneiderman of the University of Maryland's Human-Computer Interaction Laboratory for kindly permitting use of their data in this paper. We also thank Dr. Cornell Juliano of the Xerox Corporation for a conversation that helped motivate writing this paper, and for obtaining corporate permission to use some of his data in our work.

References

Box, G.E.P. and Meyer, R.D. (1986). Dispersion Effects from Fractional Designs *Technometrics*, 28, 19-27.

Carroll, R.J. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* New York: Chapman and Hall.

Ceaparu, I., and Shneiderman.B. (2003). Finding Government Statistical Data on the Web: Three Empirical Studies of the FedStats Topics Page. University of Maryland HCIL Report HCIL 2003-31, May 28, 2003. Online: ftp://ftp.cs.umd.edu/pub/hcil/Reports-Abstracts-Bibliography/2003-31html/2003-31.htm,

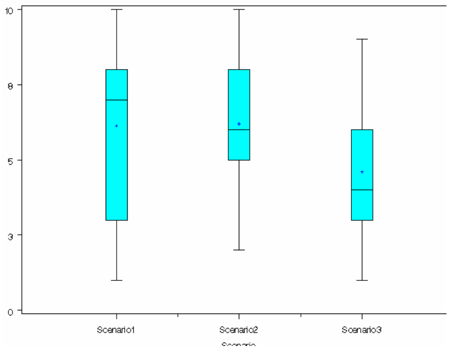

Figure 1: Distribution of Overall Frustration Ratings by Scenario for Treatment 1

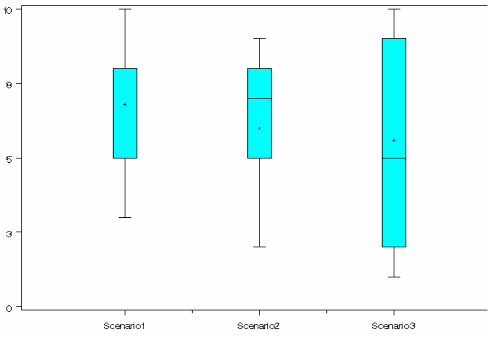

Figure 2: Distribution of Overall Frustration Ratings by Scenario for Treatment 2

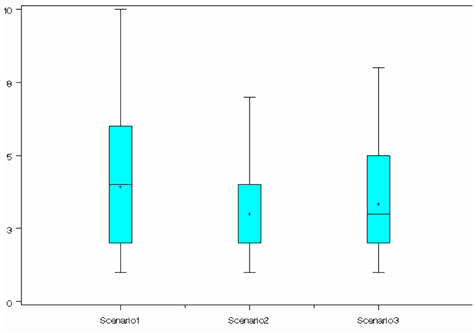

Figure 3: Distribution of Overall Frustration Ratings by Scenario for Treatment 3

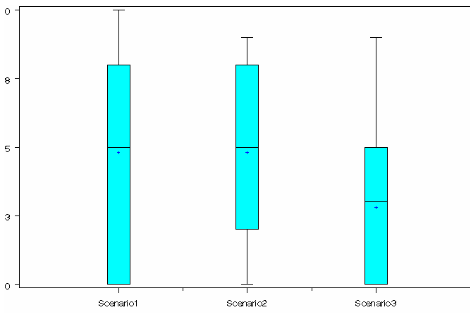

Figure 4: Distribution of "Confusing Search" Ratings by Scenario for Treatment 1

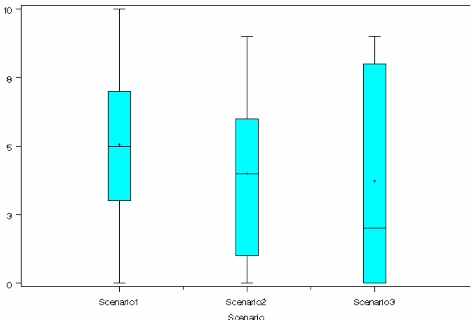

Figure 5: Distribution of "Confusing Search" Ratings by Scenario for Treatment 2

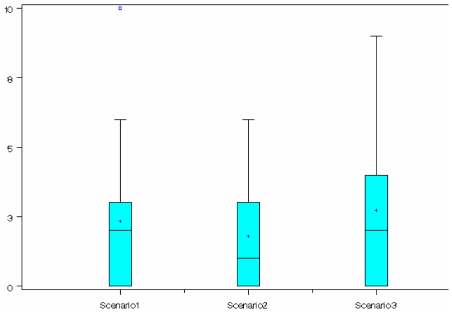

Figure 6: Distribution of "Confusing Search" Ratings by Scenario for Treatment 3

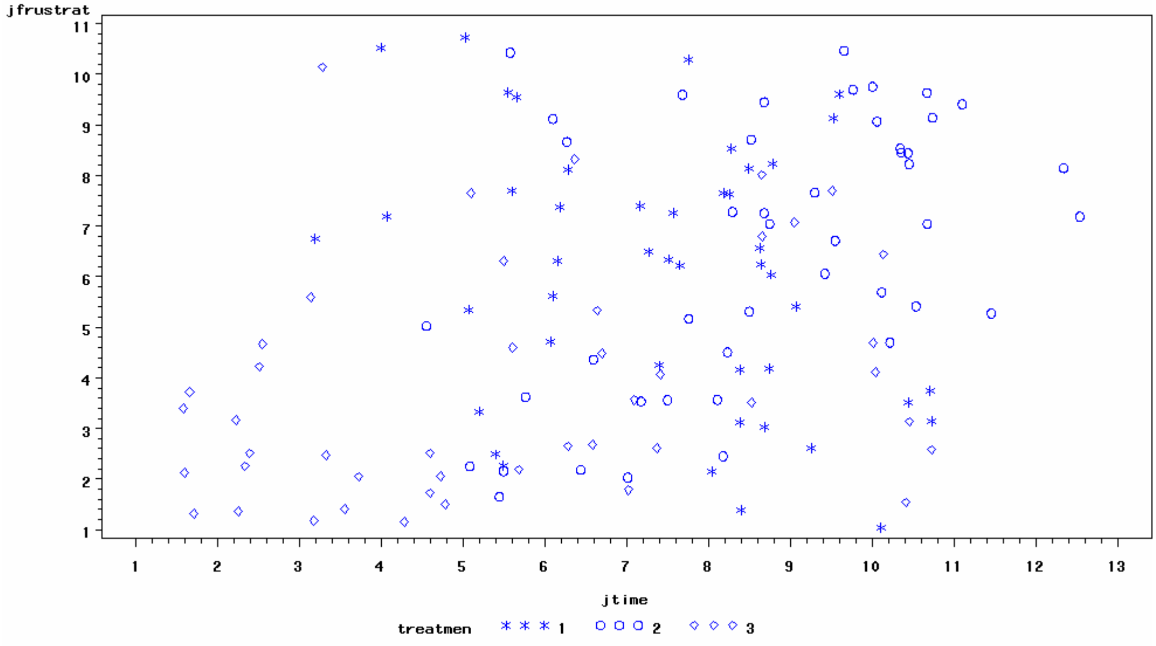

Figure 7: Scatterplot of Overall Frustration Score Against Elapsed Time, With Separate Plotting Symbols for Treatments 1 Th rough 3 (Frustration Scores and Times are Jittered)

Davidian, M. (1990). Estimation of Variance Functions in Assays with Possibly Unequal Replication and Nonnormal Data. *Biometrika*, 77, 43-54

Dumas, J.S. and Redish, J.C. (1993). *A Practical Guide to Usability Testing*. Norwood, NJ: Ablex Publishing.

Hert, C.H. & Marchionini. (1997). Seeking Statistical Information on Federal Websites: Users, Tasks, Strategies, and Design Recommendations. Final Report to the Bureau of Labor Statistics, July 18, 1997. Online: http://www.ils.unc.edu/~march/blsreport/mainbls.html

Johnson, R.A. and Wichern, D.W. (1998). *Applied Multivariate Statistical Analysis*. Englewood Cliffs, NJ: Prentice-Hall Inc

Lazar, J., Bessiere, K., Ceaparu, I., Robinson, J., and Shneiderman, B. (2003). Help! I'm Lost: User Frustration in Web Navigation, *IT&Society*, 1(3), Winter 2003, pp. 18-26. Online: http://www.stanford.edu/group/siqss/itandsociety/v01i03.html

Marchionini, G. (2000). Interfaces to Support Customized Views and Manipulation of Statistical Data, Paper presented at the Second International Conferences on Establishment Surveys, Buffalo, NY, 2000. Online: http://www.ils.unc.edu/~march/ICES_paper.pdf

Nielsen, J. (1993). *Usability Engineering*. New York: AP Professional.

Tourangeau, R., Couper, M.P. and Conrad, F. (2003). The Impact of the Visible: Images, Spacing, and Other Visual Cues in Web Surveys. Paper presented at the WSS/FCSM Seminar on the Funding Opportunity in Survey Methodology, May 22, 2003.