DISCUSSION

Martin David, University of Wisconsin

Disproportionate emphasis will be given to comments on the three papers on the Survey of Consumer Expenditure (CES); the continuing magnitude of our likely expenditures on such surveys and the almost complete inattention that they have received amongst academic statisticians justifies that emphasis. We need a complete, scientific, statistically adequate evaluation of the whole CES design. My comments can be summarized under five headings: No memory, No model, No comment, No dice, and some zip.

No memory. Past work of several of the authors is extremely germane and the CES has been discussed at the 1971 and 1975 ASA meetings. I urge readers to look at that material.

We know from past work that consumer expenditures and savings can not be reconciled with incomes reported. We know that there is differential reporting of information in different categories -- vice and casual expenditures being particularly badly reported. We know that the consumer unit is an artifact of the Bureau of the Census, and that most people can only report expenditure behavior accurately in the areas over which they have control. Pearl remembered these past discoveries and structured his discussion accordingly. It would have been extremely pertinent to the evaluations presented by Dippo to do the same. These considerations imply that we are dealing with a problem in measurement that includes both sampling error and response bias. The conceptually desirable procedure for evaluating the results of the CES would be to appeal to a minimum mean square error criterion (MSE).

Neither of the papers appeal to MSE as a choice criterion. The reason is that the design of the CES anihilates many of the comparisons that one might like: a) Some items are excluded from the diary and included in the interview and conversely; b) even where the same items are measured, the period of measurement may be different, with the result that comparable estimates can not be generated, given the well-known decay curves for recall information.[1] As a result this whole exercise of evaluation of individual items appears to be at sea without a rudder or a paddle -- a combination of Hawthorne effects, telescoping, and respondent fatigue make it unclear whether the diary estimates contain more or less bias than the survey interview estimates.

We do obtain one useful clue to this problem from the Dippo paper. First-day of diary-keeping appears to be biassed upwards by telescoping, and it would appear desirable to incorporate that finding into the estimation procedure used to obtain expenditure aggregates. It is not clear whether that has been done for CPI revision.

However, the Dippo finding is marred by the fact that we learn nothing about the treatment of non-response (including the 10 percent missing data diaries in which interviewer treatment of the first day is not known) in her calculations. How much of an effect does weighting the data have on Table 7?

No model. Both the Pearl and Dippo papers proceed as if we were in a state of ignorance about the nature of response effects and an appropriate psychological model to use for predicting poor performance. Work by Locander, Sudman, and Bradburn,[2] and by Cannell, Oksenberg, and Vinokur[3] give some clues on where and why to expect bias in the use of alternative data collection instruments. Failure to obtain relevant information can either be due to lack of motivation or perceived threats to the respondent from giving the information requested. It would be highly desirable to integrate findings from the evaluation of the CES within this theoretical structure.

The lack of a model, and the lack of emphasis on mean square error minimization imply that Dippo and Pearl reach conflicting conclusions on which data source to use. In large part this is due to the fact that Dippo et al. find significant differences between the diary and the quarterly interview where Pearl reports none. I am astonished to find two users reaching different conclusions or so basic a question. However, it is also the case that Pearl appears to base his choice of the two data collection methods on consistency with the national aggregates (which may themselves not be correct) whereas Dippo et al. use a criterion based on the coefficient of variation (CV). Looking at Pearl's Table 1 does not convince me that the ratio is a compelling criterion for choice -- Is .73 for "clothing" good? Is 1.11 for "food away from home" good? How does this compare to 1961 CES? The nature and logic for a CV choice is also not clear:

a. In the first place a multi-variate procedure would appear desirable for choosing the data collection technique, grouping classes of items together that could be expected to have similar problems in terms of threat, motivation, or recall.

b. Choice of the diary as a preferred source of data when the mean is larger than that for the interview and the CV is not, appears to imply that both numbers are subject only to under-reporting. Therefore larger means represent more complete data, not telescoping, misclassification in the diary or other errors. This assumption should be examined carefully.

The third criticism that I levy under the heading of no model is that both evaluations proceed without reference to the statistical problem for which the CES data were generated, namely revising the weights in the CPI index. It is in the nature of the price index problem that revision is required to maintain a focus on

the quantities that figure importantly in the consumer budget as new products are introduced and relative price changes shift. The design of the SCE must be evaluated by answering the following questions:

a. How does increased disaggregation of products in the weights contribute to the validity of the CPI? Increased disaggregation implies biases due to response effects and burden on the respondent that I feel are unlikely to be compensated by improved validity in the index numbers generated.

b. How does the CES assist in the timely revision of weights? The elapsed time between what is really happening in the world and the capacity of the BLS-Census to update the index makes it hard to believe that the design being evaluated today is reasonable and a cost-effective use of the nation's statistical resources. This is an echo of Pearl's comment on Jacob's paper at the 1975 meeting.

c. Finally, the evaluation of the CES must answer the question -- how do the data collected enhance the capacity of the government to move towards a utility-based cost-of-living price index that reduces the need for revisions of expenditure weights? My own interpretation of the CES is that it does not move us very far in the direction of being able to estimate the systems of structural demand equations that are required for a utility-based index, precisely because the data collection design did not adequately anticipate how to integrate information from the diary and the interview.

No comment. I have briefly mentioned the need for memory. Let me remind Bob Pearl that when he introduced the design for the CES to this association in 1971, he asserted that the design was novel and important because of eight features. His evaluation touches on three of the eight -- quarterly interview vs. diary data collection and the inventory method. Dippo et al. tell us something about the diary-keeping procedure. But several of the features embedded in the design are not touched on in their talks today:

a) Have we learned something about the last payment technique?
b) Has the scheduling of the sample as a time subsample of months and weeks been helpful?
c) How has the awkward problem of migrant families affected the data quality?

In the same meetings in 1971 Lester Frankel commented on the ingenious blending of different samples in the SCE design. 1- and 2-week samples for frequent items of purchase, monthly, quarterly, or annual samples for other items. This complex sample design and the related pattern of recall periods has only been indirectly discussed today, and I feel the profession deserves a report on its strengths and weaknesses. I hope we will see comments on these features in the evaluation reports now being prepared.

Fortunately, we do have some answers today on another feature of the design -- compensation incentives. Cowan's paper gives a clear and admirably documented report on the CES compensation experiment. His paper focuses on reduction in response bias, with the implicit model being a model of respondent omission. His conclusion that compensation is not an important technique for improving response quality must be qualified. The data do point to the fact that increased numbers of responses and response amounts attributable to compensation are a very small fraction of overall variance. What his Eta values do not display is: a. The possible increase in overall response rates that may be associated with compensation. b. Moreover they do not reflect the importance of additional reporting in relation to a measure of mean square error that appears to me to be the appropriate criterion.

Sudman's study also gives us some insight into compensation, and he should be urged to look beyond cooperation rates to the kind of item response analysis that occupied Cowan.

What is interesting about both studies is the light that they shed on the question of respondent motivation. Cannell and his coworkers have found that making the reporting task relevant to the respondent and educating him as to what constitutes a good job is crucial to the complete reporting of health events. Sudman's data demonstrate this effect in the lower cooperation levels of those who have few health events to discuss. The same framework suggests that money should be a more significant motivator to those for whom the task is relevant (i.e., health events to report and for whom incomes are low. This appears to be borne out by Sudman's Table 3 for the mail returned Diary.

Cowan's Table 3 has the potential for giving similar insights, when we see the direction of the interaction effects, which ought to be included. The fact that the interaction effects are strong for urbanicity, race and education offers the possibility that the sensible use of compensation is not to offer compensation to all respondents but to adopt a selective strategy. Identify those in the population for whom money is a good motivator and who are lacking in motivation; then concentrate payments on those individuals. It would seem quite feasible to concentrate compensation, say on urban blacks, if the interaction effects suggest that response in the sub-group could be substantially improved.

No dice. I said at the outset that my penultimate comment was no dice. I refer to the continuing consumer expenditure survey. The inconclusive character of the evaluation of survey interview versus diary data that is reflected in the papers here today stems from a fundamental lack of integration between the

theory of price indices and the measurement processes. This lack of integration was compounded in the CES by the use of independent diary and interview samples. Design of future collections should proceed with a more integrated approach based on the theory of utility-based cost of living price indexes. That will require that diaries and expenditure data be collected from the same sample, together with price statistics so that the behavioral response of consumers to a changing environment can be modelled. Such a design could be carried out with the resources that are required for a CCES. Pearl's recommendations to retain analytical opportunities are particularly wise when viewed from this perspective.

I strongly urge BLS-Census to involve academic statisticians in the design of CCES so that the product can be more useful than the piecemeal CES.

Some zip. My last comment is that Juster's proposal has some zip. The practitioners in the field of expenditure measurement appear to have forgotten that we live in a space age in which technology can be used to assist in data collection. The profession should not be devising ways of burdening the respondent with more forms paperwork or hours of interview -- we should be devising ways of automatically recording behavior as it occurs. The Neilson ratings do this. A pocket electronic memory could be devised that might substantially increase the coverage and accuracy of diary methods.

Juster's suggestion that we look at checkbook records is another way of automating data collection. It probably ought to be supplemented with data from credit card statements, and I am sure that a checkbook study ought at the beginning to be done on a very limited time scale such as the 2-week diaries we have heard about today. I also would caution that the reconcilliation of records with the social scientist's conceptual structure of income and savings is extremely difficult. Tax records are not economic records. Finally Juster implies that we must know the inventory or cash equivalents which are among the most difficult data to get completely reported. Beyond that I can only say good luck!

---

1. Norman Bradburn and Seymour Sudman, Response Effects in Surveys (1974).
2. ASA Proceedings: Social Statistics Section 1975.
3. Journal of Marketing (1977).