Discussion

Walt R. Simmons, National Center for Health Statistics

1. *Introduction*

My discussion will focus mainly on the paper by Dr. Greenberg, et. al. At the close, I shall make two brief comments on the Mallows and Williams paper.

I must warn anyone in the audience who is not already aware of it that on the subject of randomized response I am a prejudiced witness. Ever since I first heard of the original Warner proposal, I have been intrigued by the notion. I have explored a number of theoretical variations of the original model, have talked with quite a few people—including the authors of the present paper—written about advantages and weaknesses of the method, and have had a role in several of the field trials that have been undertaken. Out of this experience, there are three principal conclusions which I should like to pass on:

First: "Randomized Response" is not a single technique, but rather under its umbrella there can be assembled an *extensive family of procedures,* characterized by the fact that they can provide essentially unbiased estimators even though no person other than the respondent himself knows or possibly *can* know how the individual respondent replied to a specified question.

Second: At least some of these procedures are operationally feasible, and are efficient devices for securing intelligence on sensitive issues.

Third: The efficiency of one randomized response model can vary greatly from that of another. Nor is it always initially obvious which of two procedures is better, as was discovered when we first explored the weaker-appearing unrelated question version, only to find it could have smaller variance than certain competitors. The total process, being a function of at least several parameters, is sufficiently complex that it merits and repays the expenditure of considerable effort to locate optimum formulations.

My bias clearly, then, favors randomized response methods, and I am grateful to the authors for their work in advancing the technique. Please remember that overall judgment as I use most of my time to call attention to difficulties, weaknesses, or problems associated with the scheme described in today's paper, or to the method more broadly.

2. *New Contributions*

The principal contributions of the present paper are two in number.

On the theoretical side the approach has been extended for the first time to quantitative variables—e.g., amount of income of head of household—whereas previous work has been restricted to binomial, yes-no, true-false situations. This is obviously a significant extension. The theoretical work follows largely the pattern of earlier treatment of the binomial. The audience may feel that the presentation in this paper lacks some of the satisfying features of the earlier work—especially algorithms for optimum determination of design parameters: choice of the y-statistic given a likely range for the A-statistic; apportionment of n into n_1 and n_2; and choice of P_1 and P_2. Apparently the actual design work did follow the guidelines of the earlier binomial analysis.

The authors recognize in the empirical data evidence counter to the assumption that the A-statistic and the y-statistic have the same probability density function. This is not important in calculation of the key estimator, since that computation is distribution-free. But it can have an impact on the climate of the interview, and on the optimum determination of design parameters. Sensitivity of design analysis to operational deviations from the assumption of equivalent distributions should be investigated further. Remember that we rarely know precisely the density function of the A-variable, and may very well not know that of the y-variable.

The second contribution of the paper is of course the presentation of data from field trials. This is a very important part of any theoretically contrived procedure such as randomized response. One must accumulate experience to determine if the method is feasible. The authors report that the operation was a feasible one, and argue that the results are plausible.

3. *Respondent Cooperation*

Use of randomized response technique is designed to minimize non-response and evasive or incorrect response. Consequently, the charac-

ter of the approach to the respondent is important. Presentation time today was limited, but most serious audiences would like to know precisely how the game was explained to the respondent, and how the interviewer went about convincing the respondent to tell the truth.

One very important aspect of this matter is that of assuring that the respondent truly understands the procedure and the specific questions. The procedure *is* unusual, must in some degree cause diversion of attention of the respondent while he tries to discover what the trick may be. One wishes to forestall that line of thinking if possible.

Question construction can be critical. In an earlier survey, one question read, "There was a baby born in this household after January 1, 1965, to an unmarried woman who was living here." This is a bit more involved question, than one would prefer. There is a suspicion that some respondents missed the *un* in "unmarried," and replied as though the question had asked about children of married women—a more normal idea perhaps.

In the first of the two studies reported in today's paper, the sensitive issue was abortions. The two questions were:

(1) How many abortions have you had during your lifetime?

(2) If a woman has to work full time to make a living, how many children do you think she should have?

Dr. Greenberg has said they "goofed" in not making it clear that the second question related only to married women. I guess that is so, but the unrelated question seems to me otherwise to be particularly well-chosen. Did the framers consciously select a statement which had propaganda value in order to encourage the respondent on the acceptability of abortion, and therefore to answer question 1 truthfully?

4. *Realized Values of P_i*

The unbiased estimator of the sensitive statistic is

$$\bar{A} = \frac{(1 - P_2)\,\bar{z}_1 + (1 - P_1)\,\bar{z}_2}{P_1 - P_2}$$

The $\bar{z}_i$ value is the observed mean of responses in the i^{th} sample, and P_i is the designed proportion of instances in which the sensitive question is asked in the i^{th} sample.

If we start with the values

$$\mu_A = 0.5$$
$$\mu_y = 0.9$$
$$P_1 = 0.7$$
$$P_2 = 0.3$$

following the paper, we get an expected sample estimate for μ_A of $\bar{A} = 0.5$ as we should hope. Suppose, however, that the *realized* values of P_i, secured from faulty construction or handling of the randomizing device, are $P_1' = 0.5$ and $P_2' = 0.1$. Then we should expect an estimate $\bar{A}' = 0.58$, a substantial error. The error from this source tends to be less when μ_A and μ_y are similar, and greater when they are dissimilar. The possible failure of the realized proportion of persons answering question 1 to match the designed probability P_1 represents one of the important inherent technical—or perhaps I should say technological—frailties of the system. Perhaps pilot tests of the randomizing procedure should be made until a non-biased process can be reasonably guaranteed. Perhaps some new technique should be invented which will assure that the realized and designed values P_i are closely similar.

5. *Empirical Evaluation*

Some sturdier empirical tests or evaluations would be useful. We shall agree readily with the authors that in some respects, including directions of observed differentials, results of both of the studies are plausible. Although, uninformed of research in the area and recognizing that a majority of respondents are below the mean income, I'm not sure I would concur with the authors in expecting that both white and nonwhite respondents would report that the average head of the household of the same size had higher income than their own household.

The procedure estimates $\bar{y}$ as well as $\bar{A}$, the sensitive measure. In certain situations it should be feasible to choose a y-statistic for which the mean is known. In this way one can secure an external check on the survey. In fact, there may exist decent external statistics on income of persons in the counties in which this survey was

conducted. In a hasty review I discovered some IRS data for North Carolina which suggest that the report figures on income are at least in the right ball-park.

6. *Sampling Variances*

The paper states that the samples were stratified cluster designs, but that sampling variances were calculated on the assumption of simple random sampling of z-values. I would suspect that there was a more-than-trivial intracluster correlation on the income variables and perhaps on abortion. If so, I suspect the authors will agree that the variances shown in the report are understatements.

7. *The Future*

I hope other variants of the general technique will be developed, and other trials made of the procedures already described. These can take several different paths. I'd like to suggest one effort: Namely, a device which will adapt randomized response to a mail survey. One very important proviso goes with this. That is that it shall indeed continue to be true that it is *impossible* for the surveying agency to know whether or how the respondent has replied to a specified question.

8. *The Mallows and Williams Paper*

There are two remarks I should like to make on this report. Basically, the major part of the paper is a rather elaborate algebraic description of a model which demonstrates that biased measurements will result from a survey unless non-response rates for relevant subcategories are equivalent. We will all agree with this finding. It is why most of us attempt to impute for non-response within subclasses which are as nearly homogeneous as we can make them.

The Mallows and Williams paper offers one technique among a number of possible alternatives for adjustment of raw data. I am certainly among those who believe that we should do everything we can to minimize the impact of non-response. That task will be difficult unless the extent of non-response is kept small. The old adage applies: It's better to do analysis with data than without.