A FIELD-VALIDATION OF A QUANTITATIVE RANDOMIZED RESPONSE MODEL

James Alan Fox, Northeastern University*
Paul E. Tracy, University of Pennsylvania

Randomized response is a survey technique for reducing response bias arising from respondent concern over revealing sensitive information. The randomized response method utilizes indeterminate questions (i.e., the question answered by the respondent is unknown to the researcher) and, thus, maintains the anonymity of the responses. In other words, not even the interviewer knows what question the respondent is actually answering; the interviewer merely records the response to a random question. Based on various stochastic relations between the questions and the observed responses, it is possible to obtain estimates of parameters in the aggregate. Because only aggregate estimates are possible, not only are respondents protected, but many ethical concerns surrounding the solicitation of sensitive information are nullified. [1]

After years of enthusiasm concerning the randomized response approach, there recently has been a negative reaction among some statisticians toward the technique. The contention is that the reduction in bias earned through randomized response may not be sufficient to outweigh the severe loss in efficiency. For example, if a coin is used as a randomizing device ($p = 0.5$), the variance of the estimator would be four times greater than that derived using traditional direct questions. This argument is indeed valid for many applications involving mildly sensitive information where the potential bias with direct questions might be relatively small. Nevertheless, not until research actually determines the size of the reduction in bias achieved through the randomized response technique can this statistical question of bias versus efficiency be answered definitively.

PREVIOUS VALIDATION RESEARCH

The value of the randomized response approach depends on its ability to reduce response bias. Yet, despite a host of substantive applications, only a few validations of the method have been attempted. Also, not only are these studies contradictory in their findings, but various methodological problems render them inconclusive.

Using a sample of persons who had been arrested for driving under the influence (DUI), Folsom (1974) found that the randomized response approach failed to outperform a self-administered questionnaire in eliciting an admission of their DUI arrests. However, the randomizing device used in this study produced response combinations that were exceedingly revealing. Despite the fact that the purpose of the randomized response approach is to safeguard subjects by destigmatizing their answers, in Folsom's particular design, respondents were sufficiently jeopardized so as to destroy this protection. In fact, Folsom admits that, "Too much emphasis may have been placed on improving the precision of the method at the expense of its credibility" (1974:35) and further recommends "moves to deoptimize the [randomizing] device like setting p, the probability of selecting the sensitive question, to one-half [which could] significantly increase cooperation" (1974:54).

In a comparative validation of several methods of survey data collection, Locander, Sudman, and Bradburn found the randomized response technique to outperform, although not significantly so, personal interviews, telephone interviews, and self-administered questionnaires in eliciting truthful responses to sensitive questions (Locander, 1974; Locander et al., 1976; Bradburn and Sudman, 1979). Although this finding is promising, it is far from definitive. The sizes of the validation samples were extremely small (for example, validation checks for the most sensitive randomized response question, DUI arrests, were available for just 23 subjects), and, thus, it would be unlikely that differences would be statistically significant.

Moreover, the design of the study resulted in excessive jeopardy to respondents. Each sensitive question was paired with a nonsensitive alternative regarding month of birth. Unfortunately, after a series of twelve sensitive questions each paired with a different month, a respondent giving an affirmative response to two question sets (i.e., one sensitive question paired with a month of birth question equals one set) would necessarily be implicated in at least one of the sensitive behaviors. In terms of the subject's perception of the instrument, only a few iterations are necessary before a respondent would suspect that the truth concerning the sensitive information is mathematically discernable. Certainly, when designing a series of randomized response questions, it must be possible for a respondent to answer more than one nonsensitive question in the affirmative. Failure to observe this requirement prevents such questions from representing meaningful alternatives.

In sum, neither of the two validation efforts described above has produced definitive indications of the value of the randomized response approach. More validation research is clearly warranted in order to evaluate the extent to which randomized response actually can reduce the response bias so evident in more traditional data-gathering approaches.

VALIDATION STUDY DESIGN

Interview Design

In order to attempt an effective comparison of the response bias exhibited by the randomized response and the traditional direct question methods, it was necessary to select a known criterion that was sensitive to respondents. Both Wyner (1976) and Bridges (1979) have found that respondents systematically distort their actual number of official arrests. Thus, arrest history (i.e., number of adult arrests) constitutes an area of inquiry which appears sufficiently sensitive to the respondent to allow the validation effort.[2]

*This research was funded by U.S. Department of Justice grant #78-AX-0123. Order of authors' names determined by lottery.

Two interview schedules (one for the randomized response condition and the other for the direct question condition) were designed for administration to arrestees to be chosen from the Philadelphia police files. Both schedules solicited such background information as age, education, family income, occupation, and employment status. Both schedules included the Crowne and Marlowe (1964) Need for Approval Scale and directed the interviewer to note race and sex by observation. Also, responses to five sensitive questions were elicited, one of which concerned the number of adult arrests in Philadelphia. Because it was essential that all respondents clearly understand the arrest question, respondents were instructed to report all adult arrests (i.e., after the age of eighteen) by the Philadelphia Police involving a formal booking (i.e., fingerprinting, etc.) in a Philadelphia police station. The sensitive questions were asked directly for one interview condition, and for the other a randomized response procedure was used.

The randomizing device consisted of a lucite container with a sealed neck holding 25 red and 25 white balls.[3] Each white ball had a number printed upon it, and these numbers ranged from zero through eight with a prespecified distribution. The respondent was instructed to write down on a slip of paper (which was never seen by the interviewer) his/her numerical answer to a sensitive question. The respondent was then instructed to shake the container vigorously and then to manipulate the container so that a ball lodged in the neck. If a red ball appeared, the respondent was to report the number he/she had previously written on the slip of paper (i.e., the response to the sensitive question); if a white ball appeared, the respondent was simply to read the number from the ball. Of course, the interviewer did not know if the verbal response given was an answer to the sensitive question or merely a number read from a white ball.

The value of the selection probability p was set at 0.5 (i.e., half of the balls in the container were red) in order to maximize respondent cooperation with the randomized response procedure. Although this value of p is quite inefficient, it was chosen in order to be convincing --some respondents think that any probability other than one-half is not random.

The randomized response approach is designed to permit respondents to answer sensitive questions in a nonthreatening way. When a subject is giving a socially undesirable response to a sensitive question, it should not be apparent from the response that it pertains to the sensitive question; in other words, the respondent's answer should not be jeopardizing. On the other hand, a subject who is "innocent" on the sensitive question should not have to give an answer for the nonsensitive question that implicates him or her in the sensitive behavior; in other words, the respondent's answer should not risk suspicion. These two respondent hazards tend to vary inversely with the design of the procedure. The levels of these hazards depend on the distributional properties of the sensitive and nonsensitive responses. To attain an optimal design (i.e., to minimize total respondent hazard), the two probability distributions should be as similar as possible so that a respondent's answer does not too clearly indicate (correctly or incorrectly) from which distribution the response emanates. For example, if the sensitive question were to elicit small numerical responses and the white balls were to be preprinted with mostly large numerals, then the response given by the respondent might imply the source of that response.

To optimize our randomizing device, we matched the distribution of the numerals on the white balls to that which we expected for the arrest criterion in the eventual sample. (In particular, for these numerals the mean was set at 2.2 and the variance was set at 5.04; see Figure 1.) The anticipated distribution of the arrest criterion was used to structure the device because, of the five sensitive questions, only the arrest question was of interest. The four other sensitive questions were included to obfuscate the purpose of the survey as well as to familiarize the respondents with the randomized response technique prior to the arrest question.

FIGURE 1

Distribution of White Ball Numerals

Numeral	Frequency
0	6
1	7
2	4
3	2
4	2
5	1
6	1
7	1
8	1

Sampling Procedure

In survey research it is customary to select a representative probability sample so as to obtain unbiased and efficient estimates of distributional parameters. However, because the interest in this research involved validation rather than estimation, the necessity of acquiring a sample having known values on the criterion measure dictated that the reverse of the usual procedure be adopted. Consequently, respondents were selected from the adult arrest files of the Philadelphia Police Department, stratifying on race and sex (demonstrated correlates of response error). Also, cases were reviewed to provide diversity in offense type and number or arrests. This purposive selection ensured sufficient variability in factors of interest (e.g., age, education, occupational prestige, and income).

Because the sample did not need to be representative, any chosen arrestee who could not be located could be replaced by another selection

from the arrest files without jeopardizing the validity of the research. Similarly, any chosen arrestee who was unavailable or refused to be interviewed could also be replaced.

Initially, a sample of approximately 1300 listings was drawn from the arrest files and was supplied to the subcontractee for potential interviews. Even though most arrestees experience only one arrest during their criminal careers, multiple arrestees were overrepresented in the listings so that both the effect of arrest frequency on response error could be ascertained and diversity in offender type could be ensured. Unfortunately, this form of oversampling prevented the sample quotas from being achieved because multiple arrestees were unusually difficult to locate (i.e., most were transient, many had been incarcerated, and a few had died). When the original sample of listings was depleted, a second phase sample was chosen involving 700 additional listings. Almost all of these listings were for one-time arrestees because most of the proportionately few multiple arrestees that existed in our population of arrest files were included in the first phase sample.[4]

A total of 600 interviews was targeted. The respondents were to be equally divided among whites and nonwhites, and 480 of the respondents were to be men and 120 were to be women. Also, 480 of the interviews were scheduled for the randomized response condition and 120 for the direct question condition. This four-to-one ratio for the interview condition assignment was desired in order that the estimates obtained in the two conditions would have approximately equal precision.[5]

The targeted number of interviews was not quite achieved. Moreover, after the deletion of cases were information was unclear and/or suspect and after balance was restored to the stratification factors, 530 cases remained available for the analysis.

RESULTS

Comparative Response Error

The first objective of the analysis is a comparison of the magnitude of response error arising in the two conditions (randomized response and direct question). We shall denote x_i as a respondent's admitted number or arrests and μ_i as the actual number of arrests derived from the respondent's rap sheet.

For the direct question condition the respondent's answer to the frequency of arrests question constitutes x_i, the reported number of arrests. However, for the randomized response condition a respondent's score on x_i must be estimated from his/her answer to the randomized response design. Because the probability is 0.5 that the subject's response (R_i) arises from the arrest question and the same probability holds that the response comes from a white ball (y),

$$E(R_i) = \tfrac{1}{2}x_i + \tfrac{1}{2}E(y).$$

Since R_i is observed and E(y), the mean of the numerals on the white balls, is known to be 2.2, the reported frequency of arrests for a subject in the randomized response condition can be estimated by:

$$x_i = 2R_i - 2.2$$

For both interview conditions response error is then defined as

$$\text{Error}_i = x_i - \mu_i$$

where again μ_i, the actual frequency of arrests, is known from the official arrest records.

The top section of Table 1 presents mean reported arrests, official arrests, and response error for the two conditions. On the average, the randomized response subjects underreported .63 arrests per person, while the direct question respondents underreported .72 arrests, thus reflecting a 15 percent reduction in mean response error gained through randomized response.

The sizable difference in mean official arrests between the two interview conditions is reflective of the disproportionate allocation of the two sample phases to the interview conditions. In particular, the direct question interviews, unfortunately, were completed with an overutilization of the phase one sample listings in which multiple arrestees were overrepresented. Moreover, stratifying the comparisons in terms of frequency of official arrests (see bottom of Table 1) not only documents the uneven allocation of arrestee types to the interview conditions, but highlights the substantial effects of respondent hazards (i.e., jeopardy and risk of suspicion).[6]

Recall that a respondent is said to be in jeopardy when he/she is compelled to report very sensitive information. Clearly, this hazard was operative for the portion of the direct question sample having two or more arrests as is reflected in the substantial mean response error for this group of -1.7 arrests. On the other hand, respondent jeopardy is substantially reduced with the randomized response approach because high numerical responses do not definitely concern frequency of arrest. The small mean response error of -.28 for the multiple arrest group under this condition (which is an 83 percent reduction in mean response error over the direct question condition) indicates the substantial reduction in respondent jeopardy earned through this technique.

Corollary to protecting respondents from having to reveal sensitive information is to permit a respondent who is relatively innocent to appear as such. In other words, compelling a respondent to give an answer that provokes undue suspicion should be avoided. In this regard, the direct question approach protects the respondent in that a respondent can directly report his/her innocence (or relative innocence) to a sensitive inquiry. In contrast, in the randomized response condition this respondent, if choosing a high numbered white ball, is compelled to report a number that makes him/her suspicious. This excessive risk of suspicion is apparent for the one arrest sample in the randomized response condition which yielded a mean response error of

-.70 arrests. In comparison, the direct question condition produced for comparable respondents a mean response error of -.24, i.e., a 66 percent reduction.

The disproportionate allocation of arrestee types (one/multiple arrestees) can easily be remedied by weighting the cases in these two subgroups. In particular, the randomized response interviews can be stratified by arrestee type and weighted in order to match the proportions present in the direct question condition (i.e., one-third multiple arrestees). The recomputed mean response errors are shown in column (b) of Table 2. After weighting, randomized response exhibits a mean response error of -.56 while that of direct question is -.72. The difference (.1618) in mean response error represents a nonsignificant 22 percent reduction in error achieved by randomized response.

Recall that in order to counterbalance the two respondent hazards, the distributions of reported arrests triggered by red ball selections and numerals prompted by white ball selections should be as similar as possible so that a respondent's answer does not too clearly suggest (correctly or incorrectly) from which distribution the response arises. Thus, the distribution of the numerals on the white balls in the randomizing devices was constrained to a mean of 2.2, matching the targeted value for mean official arrests for the entire sample. However, difficulties in locating subjects with more than one arrest produced actual means that fell short of the desired 2.2 value, creating an excessive degree of respondent risk of suspicion.

Therefore, in order to approximate the optimal conditions to which the randomized response procedure was tailored, the cases in both conditions were stratified and weighted to achieve statistically a mean of 2.2 official arrests. Not only does this equalize, to some extent, the official arrest distributions for the two interview conditions, but the effects of the competing respondent hazards are counterbalanced. This optimal contrast is given in column (c) of Table 2. The mean response error for randomized response reduces to -.47 and that for direct question increases to -.98, reflecting a significant 52% reduction in response error provided by randomized response.

In sum, all the comparisons constructed in Table 2 yield results that are favorable to the randomized response approach. Given this support for randomized resopnse under both non-optimal and optimal conditions, it would be advisable at this point to maintain a conservative posture. Thus, in the further analysis we utilize the unweighted cases. It is emphasized, however, that the results achieved in the analysis to follow are fairly invariant to choice of weighting scheme.

'lean Squared Error

While one observer may be overly concerned about the additional costs (i.e., inefficiency) of randomized response procedures, another observer may be swayed by the reduction in response bias achieved through this approach. However, the purist might select between randomized response and the traditional direct question procedure on the basis of their respective means squared error. For any application the size of the mean squared error (defined as the variance of an estimator plus its squared bias) depends on the values assumed by its components. However, using the estimates provided here for the response variance and mean response error, a comparison of mean squared error specific to these estimates can be achieved.

The mean squared error operative in the two measurement conditions for varying sample sizes are shown in Table 3. Clearly, for small samples the inefficiency of the randomized response approach makes it an inadvisable alternative (i.e., mean squared error is high). In other words, if only a few observations are possible, they shoud be used in the most efficient manner. However, not much in terms of affordable sample size is required before the potential for bias reduction through randomized response outweighs the concern regarding this method's relative inefficiency. In other words, for sample sizes that are large, excessive cases can be used for other purposes than just reducing sampling variance.[7]

CONCLUSION

The present study has demonstrated that the randomized response approach reduces one of the major methodological limitations inherent in traditional measurement approaches: response bias resulting from survey respondent concern over revealing information of a sensitive nature. When compared to the direct question interview method, the quantitative randomized response design used in this research achieved a substantial reduction in response bias. Moreover, based on the results of a comparison of mean square error, some of the criticisms concerning the inefficiency of the randomized response approach appear to be overstated. It is true that when only small samples can be taken the luxury of randomized response can not safely be afforded; however, it does not take much in terms of available sample size before that which can be gained in terms of bias in estimation by adopting randomized response outweighs that which is lost in efficiency of estimation.

NOTES

1. For a complete review of the randomized response approach, with suggestions for its application, see, Fox and Tracy (1980).

2. The specific criterion used in this research was the number of official adult arrests recorded in the files of the Philadelphia Police Department. Had data on arrests in the entire U.S. been used as the criterion (i.e., from FBI rap sheets), the generalizability of the results would have been maximized but only at the expense of internal validity resulting from measurement error (e.g., interjurisdictional inconsistencies in the definition of arrest). However, because this research involved a validation of measurement methods rather than an attempt at estimating official offensivity, the criterion (arrests) need not be universally

complete but must be as reliable as possible. By limiting the scope of official offensivity to Philadelphia arrests exclusively, the definition of what constituted an arrest was held constant, and the problem of measurement error in the criterion was thereby avoided.

3. We acknowledge that the device described here derives from the work of Liu and Chow (1976). Actually, there are numerous varieties of devices that could be used; the range of models is only limited by one's ingenuity in manipulating probabilistic relations.

4. The discrepancy between the two samples when combined with a deviation from specified interview allocation procedures that occurred in the field created a problem which is discussed and statistically remedied in the next section.

5. In particular, the variances of the estimates given by the respective interview conditions are:

Direct question: $\text{Var}(\hat{\mu}_x) = \sigma^2/n$

Randomized Response: $\text{Var}(\hat{\mu}_x) = \sigma^2/np^2$

Thus, relative efficiency equal to unity is attained by sampling according to the ratio of $1/p^2$ to one (and here p, the selection probability of the sensitive question, was to be $\frac{1}{2}$).

6. Not only does the one-time/multiple arrest division provide a close (although imperfect) approximation to the important distinction between the occasional and habitual offender, but this division leaves sufficient cases in both levels of the dichotomy to permit reliable results.

7. The mean squared error comparisons are based on conservative estimates of mean response bias (i.e., unweighted data). When the comparisons are made using the weighted data, the findings become even more favorable to the randomized response approach.

REFERENCES

Bradburn, N. M. and S. Sudman
1979 Improving Interview Method and Questionnaire Design. San Francisco: Jossey-Bass.

Bridges, G. S.
1979 Levels of Effects of Response Error in Self-Reports of Crime and Delinquency. Unpublished Ph.D. Dissertation, University of Pennsylvania.

Crown, D. and D. Marlowe
1964 The Approval Motive. New York: Wiley.

Folsom, R. E.
1974 "A Randomized Response Validation Study: Comparison of Direct and Randomized Reporting in DUI Arrests." Research Triangle Institute report No. 254-807.

Fox, J. A. and P. E. Tracy
1980 "The Randomized Response Approach: Applicability to Criminal Justice Research and Evaluation." Evaluation Review 4.

Liu, P. T. and L. P. Chow
1976 "A New Discrete Quantitative Randomized Response Model." Journal of the American Statistical Association 71:72-72.

Locander, W. B.
1974 "An Investigation of Interview Method, Threat, and Response Distortion." Unpublished Ph.D. Dissertation, University of Illinois.

Locander, W. B., S. Sudman, and N. M. Bradburn
1976 "An Investigation of Interview Method, Threat, and Response Distortion." Journal of the American Statistical Association 71:269-75.

Wyner, G. A.
1976 "Sources of Response Error in Self-Reports of Behavior." Unpublished Ph.D. Dissertation, University of Pennsylvania.

TABLE 1

Summary Statistics by Official Arrests and Interview Method

RESPONDENT CATEGORY	N	Mean Reported Arrests	Mean Official Arrests	Mean Response Error
Total Sample				
Interview Method				
Randomized Response	410	.8341	1.4512	-.6171
Direct Question	120	1.0583	1.7833	-.7250
One Arrest				
Interview Method				
Randomized Response	326	.2969	1.0000	-.7031
Direct Question	80	.7625	1.0000	-.2375
Two or More Arrests				
Interview Method				
Randomized Response	84	2.9190	3.2024	-.2833
Direct Question	40	1.6500	3.3500	-1.7000

TABLE 2

Response Error by Interview Method

INTERVIEW METHOD	Unweighted (a)	Weighted* (b)	Weighted** (c)
Randomized Response			
N	410	410	410
Mean	-.6171	-.5632	-.4744
Variance	13.4417	15.1616	17.9801
Standard Error	.1811	.1923	.2094
Direct Question			
N	120	120	120
Mean	-.7250	-.7250	-.9843
Variance	2.6682	2.6682	3.6381
Standard Error	.1497	.1497	.1741
Mean Difference in Error	.1079	.1618	.5099
t-value	.46	.66	1.87
One-tail p-value	.323	.254	.031
% Reduction in Error	14.89	22.32	51.81

*The randomized response file was weighted so that it would reflect the same proportion of multiple arrestees (one-third) as in the direct question file.

**The cases in both files were stratified and weighted to achieve a mean of 2.2 official arrests.

TABLE 3

Mean Squared Error Comparisons for Varying Sample Sizes

Sample Size	Mean Squared Error: Randomized Response	Mean Squared Error: Direct Question	Ratio
5	3.3390	0.9458	3.5267
10	1.8599	0.7362	2.5264
15	1.3669	0.6660	2.0523
20	1.1204	0.6309	1.7758
25	0.9725	0.6099	1.5946
30	0.8738	0.5958	1.4666
35	0.8034	0.5858	1.3715
40	0.7506	0.5783	1.2980
45	0.7095	0.5724	1.2395
50	0.6766	0.5677	1.1918
55	0.6497	0.5639	1.1522
60	0.6273	0.5607	1.1188
65	0.6084	0.5580	1.0902
70	0.5921	0.5557	1.0655
75	0.5780	0.5537	1.0439
80	0.5657	0.5519	1.0249
85	0.5548	0.5504	1.0080
90	0.5452	0.5490	0.9930
95	0.5365	0.5478	0.9794
100	0.5287	0.5467	0.9671
105	0.5217	0.5457	0.9560
110	0.5153	0.5448	0.9459
115	0.5094	0.5439	0.9366
120	0.5041	0.5432	0.9280
125	0.4991	0.5425	0.9201