THE EFFICIENCY OF ESTIMATING CORRELATION WITH TWO FORMS OF THE RANDOMIZED RESPONSE TECHNIQUE

Stephen E. Edgell, Samuel Himmelfarb, and Darrell J. Cira, University of Louisville

Warner's (1965) randomized response technique (RRT) was designed to eliminate evasiveness in response to questions of a sensitive, possibly embarrassing or stigmatizing nature that require a dichotomous response. In Warner's technique, the respondent is presented with two questions, the sensitive question (e.g., Have you ever had an abortion?) and its logical complement (e.g., Have you never had an abortion?). Through the aid of a randomizing device, the respondent is directed to answer the sensitive question with probability P and to answer its logical complement with probability 1 - P. Since no one but the respondent knows to which question the answer pertains, the technique provides response confidentiality and should increase respondents' willingness to answer questions of a sensitive nature.

Despite the uncertainty created by the RRT, Warner (1965) showed that estimates of population parameters could be derived by the application of elementary probability theory. Warner's technique is statistically highly inefficient and considerable efforts have been expended to develop other RRT models that would have greater efficiency (see reviews by Horvitz, Greenberg, and Abernathy, 1976; Fox and Tracy, 1980).

In addition to RRT models for dichotomous response questions, a number of models have been developed for questions that require a quantitative or numerical response. In the quantitative version of the Unrelated Question RRT (Greenberg, Abul-Ela, Simmons, and Horvitz, 1969) the respondent is directed by the randomizing device to give a numerical response to the sensitive question with probability P and to give a numerical response to a totally unrelated and innocuous question with probability 1 - P. Another RRT model that can be applied to quantitative responses is the linear RRT (Warner, 1972; Pollock and Bek, 1976; Himmelfarb and Edgell, 1980). In Himmelfarb and Edgell's (1980) Additive Constants RRT the respondent is directed to add a numerical constant, K_i, $(i = 1, c)$ to the true or correct response with probability P_i $(\sum P_i = 1)$. While some authors have considered more general forms such as adding a continuous random variable or multiplying the true value by a probabilistically determined quantity, the present authors conjecture that these forms are not practical with survey applications.

Although the main focus of RRT models has been on the derivation of estimators of the mean or proportion in the population, the availability of quantitative response models suggests the possibility of using the RRT to estimate the correlation between two variables. Indeed, the ability to estimate the relationship between two variables is crucial if the RRT is to have serious scientific value. Recently, formulae have been provided by Kraemer (1980) for estimating the Pearson product-moment correlation coefficient between quantitative responses to two questions each measured by the Unrelated Question RRT and by Himmelfarb and Edgell (1982) for two questions each measured by the Additive Constants RRT model. However, to be useful for researchers, these estimators of the correlation coefficient require further investigation. The purpose of this paper is to report the results of a large number of Monte Carlo simulations directed at answering questions about the possible bias of these estimators, their efficiency relative to the standard bivariate correlation case, and the relative efficiency of the two estimators vis-a-vis each other.

METHOD

The RRT Models

Unrelated Question Model. In this RRT model, the sensitive question is paired with an unrelated and nonsensitive question. The respondent is directed to answer the sensitive question with probability P and is directed to answer the unrelated question with probability 1 - P. To perform the simulation, a number of assumptions about the parameters of the questions were made. First, it was assumed that the unrelated question had the same population mean and variance as the sensitive question. This assumption, if implemented in practice, is extremely useful in providing the respondents with confidentiality since responses to the sensitive and unrelated questions would be indistinguishable. Secondly, it was assumed that the population mean and variance of the unrelated question were known. This assumption represents an extremely idealized case, but has the practical advantage of not requiring a second sample for estimation of the parameters (i.e., mean and variance) of both questions. Finally, it was assumed that each unrelated question was different and uncorrelated with the sensitive question with which it was paired, with the other sensitive question, and with the other unrelated question. This was an unstated assumption of the equation presented by Kraemer (1980) that was used to estimate the correlation between the two sensitive questions.

The above assumptions were made for both convenience and because they give the Unrelated Question RRT its maximum possible efficiency. While this ideal case may not be attainable very often in practice, it provides upper limits for the efficiency of the Unrelated Question RRT, and it is useful to compare these to the efficiency of the Additive Constants RRT. In estimating the population mean of the sensitive question, previous comparisons (Pollock and Bek, 1976; Himmelfarb and Edgell, 1980) have shown that this idealized case of the Unrelated Question RRT is less efficient than the Additive Constants RRT for reasonable choices of constants. It should be noted that while the above assumptions facilitated the execution of the simulation for the Unrelated Question RRT, the implementation of many or all of these assumptions in practice is highly unlikely. Besides the assumptions about known and equal parameters, the requirement of uncorrelated unrelated questions is a matter of some practical concern, and formulae for estimating correlation when this is not satisfied need to be developed.

Additive Constants Model. Himmelfarb and Edgell (1980) present three special cases of this RRT Model. Their Case I is applicable in many situations and thus was chosen for study and comparison. In Case I, the respondents are directed to answer the sensitive question directly (i.e., to add a constant of zero) with probability P. If the randomizing device does not direct them to answer the question directly, then with equal probability (i.e., (1 - P)/4), they are directed to add one of four constants, +K, +2K, -K, -2K. It should be noted that the parameter P in the Additive Constants RRT is equivalent to the same parameter in the Unrelated Question RRT. The estimated correlation coefficient between the two sensitive questions each measured by the Additive Constants RRT was obtained by the equation given by Himmelfarb and Edgell (1980).

Parameter Values. A probability (P) of .8 of being directed to answer directly the sensitive question was chosen for most simulation runs. This probability is likely to be the highest value of P one could choose and still hope to enlist the cooperation of the respondents. To obtain an indication of the effect of lowering P, several cases of P = .6 are also reported.

The parameter K must also be set for the Additive Constants RRT. Since a smaller value of K makes the model more efficient but should produce less cooperation, a reasonably large value of K was selected. The value chosen was K = .75σ. Thus, the respondent could be directed to add to or subtract from their true value .75σ or 1.5σ; these are quite reasonable masks for the true value. Himmelfarb and Edgell (1980) give a formula for the largest value K could take on and have the efficiency of estimating the mean by this special case of the Additive Constants RRT be no less than that for the idealized case of the Unrelated Question RRT considered here. For P = .8, that value of K is 1.06σ. The value is somewhat extreme but useful for comparison of efficiency. For P = .6, that maximum value of K is 1.33σ.

The Simulation

Monte Carlo simulation techniques were used to generate sampling distributions of the correlation coefficient between responses to two sensitive questions in which the true population correlation coefficient, ρ, was equal to 0, .2, .5, and .8. Respondents from a standardized bivariate normal distribution were generated with the required population correlation using an algorithm for generating random normal deviates by Bell (1968). Uniform random deviates used by the Bell algorithm were generated by the DEC-10 FORTRAN RAN function previously tested and found to be acceptable (Edgell, 1979). In applying an RRT to a respondent's score, it was assumed that the respondent followed the RRT honestly and without error. The calculated expected values (means) and standard deviations (standard errors) of the sampling distributions of the correlation coefficients were based on 100,000 samples of size n, where n is the sample size studied. Correlation coefficients were corrected by the appropriate formulae referenced above before being used in calculating the sampling distributions.

Two checks on the accuracy of the simulation results were performed. The first involved generating correlation estimates without the use of an RRT and comparing the obtained means and standard errors against tabled values (Soper, Young, Cave, Lee, and Pearson, 1917). This was done for the twelve combinations of four sample sizes (n = 10, 20, 50, and 100) and for three values of population correlations (ρ = 0, .2, .5). In calculating the expected values better than two decimal place accuracy was obtained in all cases, and better than three decimal place accuracy was obtained in all but two of the cases. All standard deviations were accurate to better than three decimal places. The average error in the standard deviation was around 0.1% with the largest less than 0.4%. The second check involved two runs of the same parameter values on the Additive Constants RRT (n = 100, ρ = .5, K = 1.06). The obtained expected values differed by .000671, and the obtained standard deviations differed by .000385. Certainly, it can be concluded that the simulation is more than sufficiently accurate for the present requirements. It is reasonable from these accuracy checks to assume that the findings below are accurate to plus or minus .001.

RESULTS AND DISCUSSION

The results of the simulations are given in several tables. Table 1 gives the expected value and standard deviation of the standard bivariate correlation coefficient for population correlation values of 0, .2, .5, and .8 and sample sizes of 30, 50, 100, and 200. The values in Table 1 are those expected if the relationship between the two sensitive questions had been measured directly without an RRT procedure and all respondents had answered without bias. This table is provided so that comparisons can be made with the values obtained under RRT conditions. The values in Table 1 for sample sizes 50 and 100 were obtained from standard tables (Soper, et al, 1917), while the values for samples sizes 30 and 200 were calculated using the present simulation.

TABLE 1

Expected Values and Standard Deviations for the Correlation Coefficient

Expected Value

ρ \ n	30	50	100	200
0	.000	.000	.000	.000
.2	.197	.198	.199	.199
.5	.493	.496	.498	.499
.8	.795	.797	.799	.799

Standard Deviation

ρ \ n	30	50	100	200
0	.186	.143	.101	.071
.2	.179	.137	.097	.068
.5	.143	.109	.076	.053
.8	.071	.053	.037	.026

TABLE 2

Expected Values and Standard Deviations for Estimates of Correlation Using the Alternative Question RRT

		Expected Value				Standard Deviation			
	ρ \ n	30	50	100	200	30	50	100	200
P = .8	0	.001	.000	.000	.000	.301	.228	.159	.111
	.2	.201	.201	.200	.200	.299	.226	.157	.110
	.5	.506	.504	.502	.501	.284	.215	.149	.105
	.8	.817	.810	.805	.802	.272	.205	.143	.100

	ρ \ n	Expected Value 100	Expected Value 200	Standard Deviation 100	Standard Deviation 200
P = .6	.5	.511	.506	.294	.203

TABLE 3

Expected Values and Standard Deviations for Estimates of Correlation Using the Additive Constants RRT

		Expected Value				Standard Deviation			
	ρ \ n	30	50	100	200	30	50	100	200
P = .8, K = .75σ	0	.000	.000	.000	.000	.246	.186	.130	.091
	.2	.202	.201	.200	.200	.239	.182	.126	.089
	.5	.505	.502	.502	.500	.207	.155	.108	.075
	.8	.815	.808	.804	.802	.151	.111	.077	.053

		Expected Value			Standard Deviation		
	ρ \ n	50	100	200	50	100	200
P = .8, K = 1.06σ	0	.000	.000	-.001	.237	.161	.112
	.2	.207	.203	.201	.234	.158	.110
	.5	.519	.509	.504	.216	.145	.100
	.8	.835	.815	.807	.198	.127	.087

	ρ \ n	Expected Value 100	Expected Value 200	Standard Deviation 100	Standard Deviation 200
P = .6, K = .75σ	.5	.504	.502	.139	.096
P = .6, K = 1.06σ	.5	.520	.509	.218	.144
P = .6, K = 1.33σ	.5		.523		.214

Table 2 gives the expected values and standard deviations of the correlation coefficients obtained by using the Unrelated Question RRT under the assumptions stated above. The expected values are reasonably close to the population values. It is interesting to note that (for other than ρ = 0.) the Unrelated Question RRT produces a positive rather than a negative bias which occurs in the standard bivariate correlation situation. Also, the bias is somewhat larger using the RRT than for the standard correlation. As would be expected, the bias is larger for P = .6 than for P = .8. The standard deviations are considerably larger for the RRT. This is especially true when P = .6, where they are about twice that for P = .8.

Table 3 gives the expected values and standard deviations of the correlation coefficients obtained by Case I of the Additive Constants RRT. Again, a positive bias is obtained in the expected values (for cases other than ρ = 0.). For the reasonable case of K = .75σ the bias is about the same size as for the Unrelated Question RRT. However, for large K and especially for large values of ρ and small sample sizes, the bias is somewhat larger. Further comparison of Tables 2 and 3 reveals that the standard deviations for the Additive Constants RRT under maximum K (K = 1.06σ for P = .8 and K = 1.33σ for P = .6) are about the same as the standard deviations for the Unrelated Question RRT. This generalizes the Himmelfarb and Edgell (1980) limits on K for equal efficiency in estimating means to the estimation of correlation coefficients.

In executing the simulations for the RRTs a problem was encountered. It is possible for the estimate of the variance of a sample, which is needed to calculate the correlation, to be negative. Obviously this is an unreasonable result and one from which one cannot proceed. In the Additive Constants RRT for P = .8, K = 1.06σ, and n = 30 about 0.3% of the samples yielded negative variances. Thus, no values are given in Table 3 for these cases. A few (less than 10 out of 100,000) samples for n = 50 also yielded negative variances. This frequency of occurrence is so small as to be negligible both in this simulation and in practice. Thus these few samples were just discarded and replacement samples were drawn and used for the results given in Table 3. The same is true for n = 30 in the Unrelated Question RRT and the Additive Constants RRT (P = .8, K = .75σ). If, in the Additive Constants RRT, n is lowered to 20 for P = .8, K = .75σ, the frequency of negative variances stays at or below 0.04%. Although no values for these conditions are presented in the tables, in practice one could use this RRT in a survey using P = .8 and K = .75σ with an n as small as 20 with little chance of obtaining a negative and hence unusable estimate of the sample variance. This problem worsens with lower values of P. In the Additive Constants RRT for P = .6 and K = 1.33σ, the case of n = 100 generated about a 0.7% negative variance rate; again, no data for this condition are presented in the table. Inasmuch as the problem of negative variances only occurs with any frequency for extreme values of K, it need not concern the researcher who wishes to use the Additive Constants RRT in practice. The Unrelated Question RRT gave no negative variances for the P = .6 cases. As the Unrelated Question RRT seems to have less problem with negative variances, the researcher also need not be concerned with this potential problem with this RRT.

Table 4 gives the usual ratio of variances measure of efficiency for the Unrelated Question RRT against the standard bivariate correlation. This table makes more salient the loss of efficiency due to the use of this RRT. The steep decline in relative efficiency as the population correlation increases is of interest. Also of interest in Table 4 is the finding that the relative efficiency for the Unrelated Question RRT is mostly constant across different sample sizes.

TABLE 4

Relative Efficiency of Alternative Question RRT to Direct Correlation

P = .8

ρ \ n	30	50	100	200
0	.38	.39	.40	.41
.2	.36	.37	.38	.38
.5	.25	.26	.26	.26
.8	.07	.07	.07	.07

P = .6

ρ \ n	100	200
.5	.07	.07

In Table 5 are the relative efficiencies for the Additive Constants RRT compared to the standard bivariate correlation. Again the steep drop in relative efficiency as the population correlation increases is seen. Also, the constancy of relative efficiency over sample size is found for the Additive Constants RRT. The relative efficiencies for the more reasonable values of K are not that small for this RRT. This size of loss in efficiency using the Additive Constants RRT may be a very reasonable price to pay if this RRT does in fact overcome respondent evasiveness.

Table 6 gives the direct comparison of the relative efficiency of the Additive Constants

TABLE 5

Relative Efficiency of Additive Constants RRT to Direct Correlation

P = .8 K = .75σ

ρ \ n	30	50	100	200
0	.57	.59	.60	.61
.2	.56	.57	.59	.58
.5	.48	.49	.50	.50
.8	.22	.23	.23	.24

P = .8 K = 1.06σ

ρ \ n	50	100	200
0	.36	.39	.40
.2	.34	.38	.38
.5	.25	.27	.28
.8	.07	.08	.09

P = .6 K = .75σ

ρ \ n	100	200
.5	.30	.30

P = .6 K = 1.06σ

ρ \ n	100	200
.5	.12	.14

P = .6 K = 1.33σ

ρ \ n	200
.5	.06

TABLE 6

Relative Efficiency of Additive Constants RRT to Alternative Question RRT

P = .8 K = .75σ

ρ \ n	30	50	100	200
0	1.50	1.50	1.50	1.49
.2	1.57	1.54	1.55	1.53
.5	1.88	1.92	1.90	1.96
.8	3.24	3.41	3.45	3.56

P = .8 K = 1.06σ

ρ \ n	50	100	200
0	.93	.98	.98
.2	.93	.99	1.00
.5	.99	1.06	1.10
.8	1.07	1.27	1.32

P = .6 K = .75σ

ρ \ n	100	200
.5	4.47	4.47

P = .6 K = 1.06σ

ρ \ n	100	200
.5	1.82	1.99

P = .6 K = 1.33σ

ρ \ n	200
.5	.90

RRT to the Unrelated Question RRT in estimating the correlation. A value greater than 1.0 favors the Additive Constants RRT. As can be seen in Table 6, the Additive Constants RRT is more efficient for values of K less than the limiting values given by Himmelfarb and Edgell (1980) and is about equal in efficiency for the limiting values of K. Since values smaller than the limiting values of K are quite reasonable in practice, it is clear that the Additive Constants RRT is statistically preferred over the Unrelated Question Model for applied work.

On the basis of bias and problems with negative variances, there is no preference between the two RRTs considered here. Indeed both are quite usable based on these two criteria. However, on the basis of efficiency, the Additive Constants RRT is preferred. Of course, the efficiency of an RRT procedure is not the sole criterion. If it were, then direct questioning would always be preferred. Empirical work on how well these two RRTs achieve their primary task of eliciting truthful cooperation from respondents is needed. This recommendation cannot be taken lightly as no amount of statistical efficiency in a technique can overcome respondent evasiveness. Indeed, while all RRTs have face validity, there is some recent empirical evidence that at least one form of the RRT still seriously suffers from one kind of respondent evasiveness even under near optimal conditions (Edgell, Himmelfarb, and Duchan, 1982). These authors found considerable evasiveness to an Unrelated Question Technique where the unrelated question was in fact a randomizing device. While they studied a discrete variable case rather than continuous, this form of the Unrelated Question RRT is basically the one studied here, as using a randomizing device is one method (and maybe the only one) of knowing the population mean and variance of the unrelated question.

REFERENCES

Bell, J.R. Algorithm for generating normal random deviates. Communications of the ACM, 1968, 11, 498.

Edgell, S.E. A statistical check of the DEC System-10 FORTRAN pseudorandom number generator. Behavior Research Methods and Instrumentation, 1979, 11, 529-530.

Edgell, S.E., Himmelfarb, S., and Duchan, K.L. The validity of forced responses in a randomized response model. Sociological Methods and Research, 1982, 11, 89-100.

Fox, J.A. and Tracy, P.E. The randomized response approach: Applicability to criminal justice research and evaluation. Evaluation Review, 1980, 4, 601-622.

Greenberg, B.G., Abul-Ela, A.L.A., Simmons, W.R., and Horvitz, D.G. The unrelated question randomized response model: Theoretical framework. Journal of the American Statistical Association, 1969, 64, 520-539.

Himmelfarb, S. and Edgell, S.E. Additive constants model: A randomized response technique for eliminating evasiveness to quantitative response questions. Psychological Bulletin, 1980, 87, 525-530.

Himmelfarb, S. and Edgell, S.E. Note on the randomized response approach. Evaluation Review, 1982, 6, 279-284.

Horvitz, D.G., Greenberg, B.G., and Abernathy, J.R. Randomized response: A data gathering device for sensitive questions. International Statistical Review, 1976, 44, 181-196.

Kraemer, H.C. Estimation and testing of bivariate association using data generated by the randomized response technique. Psychological Bulletin, 1980, 87, 304-308.

Pollock, K.H. and Bek, Y. A comparison of three randomized response models for quantitative data. Journal of the American Statistical Association, 1976, 71, 884-886.

Soper, H.E., Young, A.W., Cave, B.M., Lee, A., and Pearson, K. On the distribution of the correlation coefficient in small samples. Appendix II to the papers of Student and R.A. Fisher. Biometrika, 1917, 11, 328-413.

Warner, S.L. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 1965, 60, 63-69.

Warner, S.L. The linear randomized response model. Journal of the American Statistical Association, 1971, 66, 884-888.