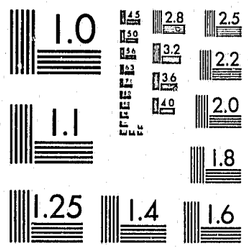

MEASURING ASSOCIATIONS WITH RANDOMIZED RESPONSE*

James Alan Fox

College of Criminal Justice
Northeastern University

and

Paul E. Tracy

Center for Studies in Criminology and Criminal Law
University of Pennsylvania

(April, 1982)

*Order of authors names determined by lottery. Direct communications to James Alan Fox, College of Criminal Justice, Northeastern University, Boston, Massachusetts 02115. The authors gratefully acknowledge Wesley Skogan of Northwestern University whose inquiry stimulated this note.

MEASURING ASSOCIATIONS WITH RANDOMIZED RESPONSE

Abstract

The view of many in the social science research community, it appears, underestimates the scope and potential of the randomized response technique, particularly with regard to its analytic capabilities. That is, there seems to be some concern that this data collection strategy is, by its nature, analytically restrictive. However, in this note we demonstrate how the quantitative, unrelated question model of randomized response can be reformulated into a measurement error model, which in turn allows a wide range of multivariate approaches. Last, we illustrate through a simulated correlation experiement the type of adjustments that need only be made to manipulate randomized response data more powerfully than has been the practice in the past.

Since its introduction by Warner (1965) the randomized response approach has been the focus of considerable theoretical literature but has not been readily welcomed by survey researchers. The reluctance to utilize randomized response is perhaps rooted in the fact that the decision to employ the technique is irreversible. That is, unlike analytic innovations which can be experimented with and then discarded by the researcher without altering the quantity or quality of survey response data, the researcher cannot recapture information that is lost or altered by using a randomized response data collection strategy. It is easily understood, therefore, why randomized response is considered by some to be more of a validation check than a primary method of collecting data (see, e.g., Penick and Owens, 1976).

Skepticism over the randomized response technique appears to surround two principal issues: whether the reduction in bias earned through randomized response outweighs the inefficiency of the technique and the extent to which the technique limits analytic capabilities. The former issue has been the topic of a number of validation efforts (see, e.g., Folsom, 1974; Locander, 1974; Locander et. al., 1976; Bradburn and Sudman, 1979; and Tracy and Fox, 1981) which seem to suggest that the randomized response technique has value in reducing systematic response error. The second concern, more particularly, that multivariate statistics are not possible with data gathered through randomized response methods, is still an issue around which there is considerable confusion. The application of univariate summary measures (e.g., proportions and means) and the

comparison of these measures across subgroups are obvious. However, the richer multivariate approaches (e.g., correlations, ANOVA, logit analysis) require individual level data which are seemingly unavailable with the randomized response technique. Sudman and Bradburn's (1982:81) recent comment reflects this view:

> By using this procedure, you can estimate the undesirable behavior of a group; and, at the same time, the respondent's anonymity is fully protected. With this method, however, you cannot relate individual characteristics of respondents to individual behavior. That is, standard regression procedures are not possible at an individual level. If you have a very large sample, group characteristics can be related to the estimates obtained from randomized response. For example, you could look at all the answers of young women and compare them to all the answers of men and older age groups. On the whole, however, much information is lost when randomized response is used.

In this note we show how unknown individual level data can be estimated for use in multivariate analyses without loss of information, explicating mathematically an approach recommended conceptually by Boruch (1972:410) (see also Boruch and Cecil, 1979:142). More precisely, we demonstrate that randomized response observations can be treated as individual level scores that are contaminated by random measurement error and, consequently, that multivariate measures need only be corrected for or purged of the effects of this error.

As is true of most of the theoretical and applied literature concerning randomized response, efforts to develop bivariate and multivariate approaches for randomized response data have focused mainly on dichotomous (or polytomous) data in multiway contingency tables (see Boruch, 1971; Barksdale, 1971; Drane, 1976; Clicker and Iglewicz, 1976; Chen, 1979; Kraemer, 1980; and Tamhane, 1981) with only limited attention to quantitative responses (see Rosenberg, 1979; Kraemer, 1980; and Himmelfarb and Edgell, 1981). The "errors in variables" approach to quantitative randomized response data presented here, like Chen's (1979) misclassification perspective in the qualitative case, is sufficiently general to apply to a variety of data collection strategies and analytic techniques.

Distribution of Measurement Error in Randomized Responses

Consider the quantitative, unrelated question model employing an alternative question having a known distribution (see Greenberg et al., 1971). More particularly, a survey respondent is instructed to manipulate a randomizing device following a bernoulli distribution with known parameter p, and then either to answer a sensitive question having unknown mean and variance, μ_x and σ_x^2, or to report innocuously a random digit (or nonsensitive response) having known location and scale parameters, μ_y and σ_y^2, depending on the outcome of the randomizing device.

Let z_i represent the verbal (observed) response given by the ith respondent, and let x_i and y_i be the underlying (unobserved) sensitive and nonsensitive scores. As dictated by the randomizing device with selection probability p,

$$z_i = \begin{cases} x_i & \text{with probability } p \\ y_i & \text{with probability } 1-p \end{cases} . \qquad (1)$$

Conditional on fixed x_i,

$$E(z_i|x_i) = px_i + (1-p)\,\mu_y . \qquad (2)$$

Thus, although individual scores on the sensitive question (x) are unknown, we can estimate

$$\hat{x}_i = [z_i - (1-p)\,\mu_y]/p . \qquad (3)$$

We can define, further, the measurement error introduced by randomized response as

$$u_i = \hat{x}_i - x_i \quad . \tag{4}$$

Alternatively, the estimated individual scores derived from Equation (3) can be viewed as the actual score plus a disturbance term, i.e.,

$$\hat{x}_i = x_i + u_i \quad . \tag{5}$$

Considering next properties of the disturbance u_i, we substitute Equation (3) into Equation (4),

$$u_i = [z_i - (1-p)\ \mu_y]/p - x_i \quad .$$

From Equation (1)

$$u_i = \begin{cases} [x_i - (1-p)\ \mu_y]/p - x_i & \text{w.p. } p \\ [y_i - (1-p)\ \mu_y]/p - x_i & \text{w.p. } 1-p \end{cases} \tag{6}$$

For simplicity, set

$$r_i = [x_i - (1-p)\ \mu_y]/p - x_i \ ,$$
$$= \frac{(1-p)}{p}\ (x_i - \mu_y) \ ;$$

and

$$s_i = [y_i - (1-p)\ \mu_y]/p - x_i \quad .$$

Thus,

$$\mu_r = \frac{(1-p)}{p}\ (\mu_x - \mu_y) \ ,$$

$$\sigma_r^2 = \frac{(1-p)^2}{p^2}\ \sigma_x^2 \ ;$$

and

$$\mu_s = \mu_y - \mu_x \ ,$$

$$\sigma_s^2 = \sigma_y^2/p^2 + \sigma_x^2 \quad .$$

Substituting r and s back into Equation (6), it can be shown that

$$\mu_u = p\,\mu_r + (1-p)\ \mu_s = 0 \ , \tag{7}$$

and

$$\sigma_u^2 = p\left[\sigma_r^2 + (\mu_r - \mu_u)^2\right] + (1-p)\ \left[\sigma_s^2 + (\mu_s - \mu_u)^2\right]$$
$$= p\left[\frac{(1-p)^2}{p^2}\ \sigma_x^2 + \frac{(1-p)^2}{p^2}\ (\mu_x - \mu_y)^2\right]$$
$$+ (1-p)\ \left[\sigma_y^2/p^2 + \sigma_x^2 + (\mu_y - \mu_x)^2\right]$$
$$= \frac{(1-p)}{p}\ \left[\sigma_x^2 + \sigma_y^2/p + (\mu_x - \mu_y)^2\right] \ , \tag{8}$$

producing the desired expressions for the mean and variance of the disturbance term..

Next, the covariance between x_i and u_i can be obtained in a similar fashion. We have

$$\sigma_{xu} = E(xu) - E(x)\,E(u) = E(xu)$$

because $E(u) = 0$. Multiplying Equation (6) by x_i, and collecting x_i^2 terms,

$$x_i u_i = \begin{cases} (1-p)\ (x_i^2 - x_i\ \mu_y)/p & \text{w.p. } p \\ [x_i\ y_i - (1-p)\ x_i\ \mu_y]/p - x_i^2 & \text{w.p. } 1-p \end{cases}$$

Next, we take expected values and collect covariance terms,

$$E(xu) = \begin{cases} \frac{(1-p)}{p}\ (\sigma_x^2 + \mu_x^2 - \mu_x\ \mu_y) & \text{w.p. } p \\ (\sigma_{xy} + p\ \mu_x\ \mu_y)/p - (\sigma_x^2 + \mu_x^2) & \text{w.p. } 1-p \end{cases}$$

Thus,

$$\sigma_{xu} = E(xu) = \frac{(1-p)}{p}\,\sigma_{xy} = 0 \qquad (9)$$

since, by design of the randomized response procedure, x_i and its alternative y_i are independent.[1]

In sum, the randomized response procedure in effect contaminates a response by a random disturbance term having a zero mean, a variance given in Equation (8), and, most importantly, a zero covariance with the variable under investigation. Therefore, since the distributional properties of this type of measurement error are known and are quite tractable, any multivariate approach using the estimated scores in (3) are

possible so long as the summary statistics (e.g., correlations) are corrected for attenuation produced by the randomized response procedure. That is, the randomized response estimated scores are unbiased and, although containing measurement error, the effects of this unreliability can be corrected, as we demonstrate in the next section.

Correcting Summary Statistics for Measurement Error from Randomized Response

In order to demonstrate the correction procedure, consider, for example, estimating the correlation between two sensitive variables x_1 and x_2 which are surveyed under alternative question randomized response designs. The estimated individual scores, given by

$$\hat{x}_{i1} = [z_{i1} - (1-p_1)\ \mu_{y_1}]/p_1$$

$$\hat{x}_{i2} = [z_{i2} - (1-p_2)\ \mu_{y_2}]/p_2 \qquad (10)$$

can, as before, be re-expressed as

$$\hat{x}_{i1} = x_{i1} + u_{i1}$$

$$\hat{x}_{i2} = x_{i2} + u_{i2}$$

where u_{i1} and u_{i2} are uncorrelated with their respective "true" scores as well as with each other.

Because of this strict independence of the error terms, $\sigma_{\hat{x}_1\hat{x}_2} = \sigma_{x_1x_2}$, and the correlation between the estimated scores becomes

$$\rho_{\hat{x}_1\hat{x}_2} = \sigma_{x_1x_2}/\sigma_{\hat{x}_1}\ \sigma_{\hat{x}_2}$$

$$= \frac{\rho_{x_1x_2}\ \sigma_{x_1}\ \sigma_{x_2}}{\left[(\sigma^2_{x_1} + \sigma^2_{u_1})(\sigma^2_{x_2} + \sigma^2_{u_2})\right]^{1/2}}$$

Thus, the correlation between the unobserved true scores is

$$\rho_{x_1x_2} = \rho_{\hat{x}_1\hat{x}_2}\left[(1 + \sigma^2_{u_1}/\sigma^2_{x_1})(1 + \sigma^2_{u_2}/\sigma^2_{x_2})\right]^{1/2} \qquad (11)$$

Since the sample correlation of $\hat{x}_1$ and $\hat{x}_2$ is available directly, we need only construct estimates of the ratios $\sigma^2_{u_1}/\sigma^2_{x_1}$ and $\sigma^2_{u_2}/\sigma^2_{x_2}$ in order to achieve an estimate of $\rho_{x_1x_2}$. Further, because $\sigma^2_{y_1}$ and $\sigma^2_{y_2}$ are known and $\hat{\sigma}^2_{z_1}$ and $\hat{\sigma}^2_{z_2}$ are observable, we derive below estimates of $\sigma^2_{x_1}$ and $\sigma^2_{x_2}$ and then, using Equation (8), estimates of $\sigma^2_{u_1}$ and $\sigma^2_{u_2}$ follow.

Generalizing Equation (1), the observed scores

$$z_{ij} = \begin{cases} x_{ij} & \text{w.p. } p_j \\ y_{ij} & \text{w.p. } 1-p_j \end{cases}$$

where $j = 1, 2$. The mean and variance of z_j are easily shown to be

$$\mu_{z_j} = p_j\ \mu_{x_j} + (1-p_j)\ \mu_{y_j}\ , \qquad (12)$$

and

$$\sigma^2_{z_j} = p_j\left[\sigma^2_{x_j} + (\mu_{x_j} - \mu_{z_j})^2\right] + (1 - p_j)\left[\sigma^2_{y_j} + (\mu_{y_j} - \mu_{z_j})^2\right]\ . \qquad (13)$$

Substituting (12) into (13) and simplifying algebraically,

$$\sigma^2_{z_j} = p_j\ \sigma^2_{x_j} + (1 - p_j)\ \sigma^2_{y_j} + p_j\ (1 - p_j)(\mu_{x_j} - \mu_{y_j})^2\ . \qquad (14)$$

Solving for $\sigma^2_{x_j}$,

$$\sigma^2_{x_j} = \left[\sigma^2_{z_j} - p_j(1-p_j)(\mu_{x_j} - \mu_{y_j})^2 - (1-p_j)\sigma^2_{y_j}\right]/p_j \quad . \tag{15}$$

We have, as a result, sample estimates

$$\hat{\sigma}^2_{x_j} = \left[\hat{\sigma}^2_{z_j} - p_j(1-p_j)(\hat{\mu}_{x_j} - \mu_{y_j})^2 - (1-p_j)\sigma^2_{y_j}\right]/p_j \tag{16}$$

where p_j, μ_{y_j}, and $\sigma^2_{y_j}$ are known; $\hat{\sigma}^2_{z_j} = \Sigma(z_{ij} - \bar{z}_j)^2/(n-1)$; and

$$\hat{\mu}_{x_j} = \left[\bar{z}_j - (1-p_j)\mu_{y_j}\right]/p_j \quad .$$

Also, given $\hat{\sigma}^2_{x_j}$ from Equation (16), we can generalize from Equation (8),

$$\hat{\sigma}^2_{u_j} = \frac{(1-p_j)}{p_j}\left[\hat{\sigma}^2_{x_j} + \sigma^2_{y_j}/p_j + (\hat{\mu}_{x_j} - \mu_{y_j})^2\right] \tag{17}$$

In sum, we arrive at a corrected estimate of $\rho_{x_1x_2}$, the correlation of two randomized response measures, by expressing Equation (11) in sample form

$$\hat{\rho}_{x_1x_2} = \hat{\rho}_{\hat{x}_1\hat{x}_2}\left[(1 + \hat{\sigma}^2_{u_1}/\hat{\sigma}^2_{x_1})(1 + \hat{\sigma}^2_{u_2}/\hat{\sigma}^2_{x_2})\right]^{1/2} \tag{18}$$

where $\hat{\sigma}^2_{x_j}$ and $\hat{\sigma}^2_{u_j}$ are provided respectively in (16) and (17), and $\hat{\rho}_{\hat{x}_1\hat{x}_2}$ is the sample correlation of the scores estimated in Equation (10).

To illustrate this correction procedure, we have simulated randomized response data for estimating the correlation between two randomized response measures as well as between a randomized response and a direct question measure. For both cases the true correlation between the two "sensitive" measures x_1 and x_2 was set at .6. Both measures as well as their respective alternative responses, y_1 and y_2, were drawn from normal distributions. The location and scale parameters in each randomized response pair were purposely set close but unequal (i.e., $\mu_{x_1} = 20$, $\sigma^2_{x_1} = 9$; $\mu_{y_1} = 18$, $\sigma^2_{y_1} = 10$; $\mu_{x_2} = 50$, $\sigma^2_{x_2} = 100$; $\mu_{y_2} = 55$, $\sigma^2_{y_2} = 105$).

In the first simulation the probabilities of selecting the sensitive question, $p_1 = .6$ and $p_2 = .7$. The observed (attenuated) correlation and the corrected correlation computed from Equation (18) are shown in Table 1 for varying sample sizes. As expected the correlations, after correction for substantial attenuation, hover around the true value of .6 and, more importantly, the standard errors of the corrected coefficients, although inflated, grow acceptably small in moderately sized samples.

Next, we considered the more usual case where one of the variables is measured directly (e.g., age), that is, where $p_1 = 1.0$. Further, we reduced the selection probability, p_2, to .5 (which increases respondent protection) as we would recommend in a highly sensitive inquiry.[3] The simulated correlations under these conditions are given in Table 2, again revealing that these correlations can be satisfactorily corrected for contamination produced by the randomized response procedure. As before, the corrected correlations are unbiased and have standard errors which become reasonably small for most applications.

TABLE 1*

Simulated Correlation of Two Randomized Responses

	Attenuated Correlation		Corrected Correlation	
Sample Size	Mean	S.E.	Mean	S.E.
100	.2297	.0974	.6195	.2768
250	.2322	.0611	.6203	.1697
500	.2301	.0450	.6119	.1194
750	.2292	.0347	.6063	.0901
1000	.2310	.0302	.6120	.0818
1500	.2310	.0274	.6099	.0722
2000	.2290	.0220	.6049	.0586
2500	.2264	.0212	.5969	.0545

*Selection probabilities $p_1 = .6$ and $p_2 = .7$; true correlation $\rho = .6$. Simulation includes 100 trials.

TABLE 2*

Simulated Correlations of a Randomized and a Direct Response

	Attenuated Correlation		Corrected Correlation	
Sample Size	Mean	S.E.	Mean	S.E.
100	.2751	.1058	.6002	.2512
250	.2793	.0654	.5923	.1374
500	.2880	.0436	.6033	.0968
750	.2907	.0355	.6069	.0734
1000	.2914	.0319	.6074	.0683
1500	.2842	.0299	.5942	.0652
2000	.2912	.0230	.6109	.0508
2500	.2889	.0207	.6030	.0445

*Selection probability $p = .5$; true correlation $\rho = .60$. Simulation includes 100 trials.

Conclusion

We have demonstrated how estimation with the unrelated question randomized response model can be treated as simply a problem of measurement error. As a result, the sizable literature on analysis of variables with error can be invoked to fashion analytic techniques for randomized response data. Also, the measurement error model derived here is structurally equivalent to Warner's (1971) additive randomized response model in which respondents verbally report the sum of their sensitive score and a random number. Although we feel that Warner's additive contamination model is problematic as a data collection strategy (see, Fox and Tracy, 1980), Rosenberg's (1979) full exposition of multivariate methods for the additive randomized response model can be extended directly to data collected through the unrelated question approach as well.[4] Therefore, although we have illustrated here only the use of correlations for unrelated question, randomized response data, a full range of multivariate techniques, both categorical and quantitative, can be adapted in similar fashion.

Notes

1. On occasions, the condition that the sensitive scores and their associated disturbances are uncorrelated, which is central to this derivation and is assured here by the use of random digits for nonsensitive responses, may be restricting. For example, one may wish to employ, rather than random digits, a substantively meaningful, yet innocuous, alternative question (see Greenberg, et al., 1971), perhaps even one that elicits socially desirable responses in order to neutralize the embarrassing nature of the sensitive question (Zdep and Rhodes, 1976; also see Miller, 1981; and Fox and Tracy, 1981). Because the sensitive and nonsensitive responses would no longer be necessarily uncorrelated, the derived disturbance term would be correlated with the true sensitive score. Nevertheless, more complex interview designs, involving multiple samples and multiple alternative questions (e.g., see Folsom et al., 1973), will permit an estimate of the covariance of x_i and u_i so that the measurement model can be identified entirely.

2. There are certain statistical advantages to using a distribution for the alternative response that is the same as that expected for the sensitive response. However, we tried here to approximate the usual case where these expectations are fallible.

3. For discussions of choosing selection probabilities as well as other design parameters, see Fox and Tracy (1980), and Greenberg et al. (1977).

4. Unlike the unrelated question design (which we prefer for data collection) the additive randomized response model, because it can be

designed with random digits that are distributed normally, produces estimates that enjoy certain desirable statistical properties (e.g., maximum likelihood). Nonetheless, both plans permit usual hypothesis tests as a consequence of the Central Limit Theorem (see Rosenberg, 1979: 91).

References

Barksdale, W. B.

1971 "New randomized response techniques for control of nonsampling errors in surveys." Ph.D. Dissertation, Department of Biostatistics, University of North Carolina, Chapel Hill.

Chen, T. T.

1979 "Analysis of randomized response as purposively misclassified data." Proceedings of the American Statistical Association, Section on Survey Research Methods: 158-163.

Clicker, R. P. and B. Iglewicz

1976 "Warner's randomized response technique: the two sensitive questions case." Proceedings of the American Statistical Association, Social Statistics Section: 260-263.

Drane, W.

1976 "On the theory of randomized responses to two sensitive questions." Communications in Statistics - Theory and Methods A5:565-574.

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.

Folsom R. E., B. G. Greenberg, D. G. Horvitz, and J. R. Abernathy

1973 "The two alternative questions randomized response model for human surveys." Journal of the American Statistical Association 68:525-30.

Fox, J. A. and P. E. Tracy

1980 "The randomized response approach: applicability to criminal justice research and evaluation." Evaluation Review 4:601-22.

1981 "Reaffirming the viability of the randomized response approach." American Sociological Review 46:930-933.

Greenberg, B. G., R. R. Kuebler, J. R. Abernathy and D. G. Horvitz

1971 "Application of the randomized response technique in obtaining quantitative data." Journal of the American Statistical Association 66:243-50.

1977 "Respondent hazards in the unrelated question randomized response model." Journal of Statistical Planning and Inference 1:53-60.

Himmelfarb, S. H. and S. E. Edgell

1981 "Association of variables measured by the additive constants model of the randomized response technique." Department of Psychology, University of Louisville, mimeo.

Kraemer, H. C.

1980 "Estimation and testing of bivariate association using data generated by the randomized response technique." Psychological Bulletin 87:304-308.

Locander, W. B.

1974 "An investigation of interview method, threat and response distortion." Ph.D. Dissertation, Department of Business Administration, University of Illinois.

Locander, W. B., S. Sudman and N. N. Bradburn

1976 "An investigation of the interview method, threat and response distortion." Journal of the American Statistical Association 71:269-75

Miller, J.D.

1981 "Complexities of the randomized response solution." American Sociological Review 46:928-930.

Penick, B. K. E. and M. E. B. Owens

1976 Surveying Crime. Washington, D.C.: National Academy of Sciences.

Rosenberg, M. J.

1979 "Multivariable analysis by a randomized response technique for disclosure control." Ph.D. Dissertation, Department of Biostatistics, University of Michigan.

Tamhane, A. C.

1981 "Randomized response techniques for multiple sensitive attributes." Journal of the American Statistical Association 76:916-923.

Tracy, P. E. and J. A. Fox

1981 The validity of randomized response for sensitive measurements." American Sociological Review 46:187-200.

Warner, S. L.

1965 "Randomized response: a survey technique for eliminating evasive answer bias." Journal of the American Statistical Association 60:63-69.

1971 "The linear randomized response model." Journal of the American Statistical Association 66:884-888.

Zdep, S. M. and I. N. Rhodes

1976-77 "Making the randomized response technique work." Public Opinion Quarterly 40:531-7.

END