STRUCTURAL EQUATION MODELING, *11*(4), 615–620

Randomized Response Analysis in M*plus*

Joop Hox and Gerty Lensvelt-Mulders
Department of Methodology and Statistics
Utrecht University

This article describes a technique to analyze randomized response data using available structural equation modeling (SEM) software. The randomized response technique was developed to obtain estimates that are more valid when studying sensitive topics. The basic feature of all randomized response methods is that the data are deliberately contaminated with error. This makes it difficult to relate randomized responses to explanatory variables. In this tutorial, we present an approach to this problem, in which the analysis of randomized response data is viewed as a latent class problem, with different latent classes for the random and the truthful responses. To illustrate this technique, an example is presented using the program M*plus.*

When a survey studies socially sensitive topics, this can cause substantial nonresponse, including item nonresponse for the sensitive questions, or result in socially desirable answers (Lee, 1993). One way to deal with these problems is to use Warner's (1965) randomized response technique. In Warner's original format the respondent had to answer one of two statements, for example, statement A "I took hard drugs last year" with known probability *p,* or statement B, the complementary statement, "I did not take hard drugs last year," with known probability $1 - p$. A randomizing device (most often dice) is used to decide whether statement A or B has to be answered. Because the researcher does not know the number that the respondent has thrown, the respondent's privacy is fully protected, but on the aggregate level the probability of hard drug use can be estimated.

After Warner, other randomized response designs have been developed. One of the more statistically and psychologically efficient designs is the forced response technique developed by Boruch (1971). In this technique, respondents are con-

Requests for reprints should be sent to Joop Hox, Department of Methodology and Statistics, Utrecht University, PO Box 80140, NL–3508 TC, Utrecht, The Netherlands. E-mail: j.hox@fss.uu.nl

fronted with only one sensitive question and asked to reply dependent on the outcome of two dice. When the dice produce 2, 3, or 4, the respondent is forced to answer yes, regardless of his or her own true answer. The probability of giving a forced yes answer is given by θ (= 1/6). When the dice produce 5 to 10 the respondent is required to answer the question truthfully with probability p_{true} (= 3/4). When the dice produce 11 or 12, the respondent is forced to answer no, again regardless of his or her own true answer, with probability $1 - p_{true} - \theta$ (= 1/12). The psychological advantage of using two dice is that respondents tend to underestimate the probability of throwing 5 to 10 and, therefore, feel more protected than they objectively are (Fox & Tracy, 1986).

Although for individual respondents the true state is unknown, the prevalence of the sensitive behavior in a population ($\hat{\pi}$) can be estimated as

$$\hat{\pi} = \frac{\hat{\lambda} - \theta}{p_{true}} \tag{1}$$

where $\hat{\lambda}$ is the observed proportion of *yes* answers in the sample, with sampling variance

$$\hat{\text{var}}(\hat{\pi}) = \frac{1}{p_{true}^{\ 2}} + \frac{\hat{\lambda}(1-\hat{\lambda})}{n-1} \tag{2}$$

RELATING RANDOMIZED RESPONSE DATA TO EXPLANATORY VARIABLES

A drawback of the randomized response technique (RRT) is that it is difficult to relate the population estimates of the sensitive behavior to explanatory variables. Special logistic regression techniques have been used for this purpose (Maddala, 1983; Scheers & Dayton, 1988; Van der Heijden, van Gils, Bouts, & Hox, 2000). However, these techniques are not generally accessible, for instance because they use proprietary software.

In this tutorial, we present a way to analyze the relations between randomized response estimates and explanatory variables using standard structural equation modeling (SEM) software. This becomes possible when the analysis of randomized response data is viewed as a latent class problem. Latent class analysis assumes that each individual in the sample belongs to class g of G classes with probability p_g, with $\Sigma p_g = 1$. Van den Hout and van der Heiden (2004) described the RRT as a latent class problem using the concept of misclassification, because misclassification and randomized response data have in common that they all deal with a finite mixture of distributions. They showed that a transition matrix with

conditional misclassification can describe the perturbation due to the RRT and that with the use of an Expectation = Maximization (EM) algorithm loglinear models can be estimated.

Our approach is a little different; we employ latent class SEM. Latent class SEM assumes that a different structural equation model characterizes each latent class *g.* A latent categorical variable is used to indicate the latent classes. If latent class analysis is used to model randomized response data, it is in fact known that the sample contains two classes of respondents, with unknown membership. Those respondents that were required to provide a yes or no answer based on the dice roll form one latent class, with random responses coded 1 and 0. Those respondents who were required to provide an honest answer to the real question form the second latent class, with responses that reflect their true state of affairs and are, therefore, potentially related to explanatory variables.

Although randomized response questions can use continuous answer categories (usually asking about the frequency of specific deviant behavior), they are almost always using a dichotomous yes or no question format.

Latent class problems can be analyzed using M*plus* (Muthén & Muthén, 1998). In M*plus*, categorical, ordered, or continuous outcome variables can be modeled by specifying a dichotomous latent class variable as the outcome and by using the observed dichotomous randomized response variable as a latent class indicator. In the latent class for the random responses, the structural model is empty. In the latent class for the truthful responses, the structural part is a regression model that regresses the latent categorical variable on the explanatory variables.

EXAMPLE WITH SIMULATED DATA

To gauge the performance of M*plus* with randomized response data, we generated a data set with randomized response data. We generated 300 cases giving random dichotomous responses Y with $p(Y = 1) = 0.5$ ($\theta = .15$) and 700 cases giving truthful dichotomous answers with $p(Y = 1) = 0.2$ ($p_{true} = .7$). For all cases, a standard normal explanatory variable Z was generated. In the random response class, Z is uncorrelated with the outcome Y, but in the true response class Z correlates 0.5 with the continuous latent variable underlying the observed dichotomous responses Y. Therefore, this data set represents a situation in which the sensitive behavior has a prevalence of .2 and it is related to an explanatory variable Z with $r = .5$.

For the mixture model logistic regression analysis, we specify two classes. In the truthful response class, a logistic regression model is estimated. In the random response class, the regression parameters are constrained to the true population values of a zero intercept (the logit of the known yes proportion of 0.5), and a zero slope for Z. The Appendix presents the M*plus* commands used to analyze

TABLE 1
Mixture Model Logistic Regression Artificial Data

Class	*Predictor*	*B*	*SE*	*Odds Ratio*
1	Threshold	1.83	0.53	
	Z	1.56	0.46	4.76
2	Threshold	0	—	
	Z	0	—	

TABLE 2
Interpretation of the M*plus* Analysis

pY when Z < +1SD	*PY*	*pY when Z > –1SD*
.01	.20	.65

this model. M*plus* estimates the class sizes as 756 in the truthful response class and 243 in the random response class, which is reasonably close to the true values of 700 and 300. The estimated parameters for each class are presented in Table 1.

The predictor variable *Z* is standard normal. The threshold parameter can be interpreted as the estimated proportion of zero in the population on the logit scale; transformed to proportions, it equals 0.86. Using the standard error for the threshold we can determine the 95% confidence interval (CI) on the logit scale; transformed to proportions, we find a 95% CI that ranges from 0.69 to 0.91. The known true value of negative answers is 0.80, which is in the middle of the 95% CI.

How is the estimate for the sensitive behavior *Y* related to the explanatory variable *Z*? *Z* is standard normal distributed with a mean of zero and a standard deviation of one. Respondents with a score on *Z* that is 1 *SD* above this mean have a probability of .65 to be engaged in the sensitive behavior *Y*. Respondents that score 1 *SD* lower than the mean of *Z* have a probability that is close to zero (.01) to be engaged in the sensitive behavior *Y* (Table 2).

DISCUSSION

Structural equation mixture modeling using M*plus* is not the only available approach to analyzing randomized response data. However, SEM in M*plus* makes it possible to embed the logistic regression part in a larger structural model. Thus, one can simultaneously analyze the responses to more than one randomized response question or impose constraints across latent or manifest groups.

In randomized response research, a known randomizing process, such as the dice mentioned previously, decides class membership. Therefore, the fraction of the population in each class is known. M*plus* currently does not allow imposing constraints on the size of the latent classes, but the proportions of the sample estimated for each latent class can be checked against the known class probabilities. In addition, in the mixture model, class membership for individuals is not known but can be inferred from the data. For each individual in the data set, the probability of belonging to each of the latent classes can be computed. This could be useful for diagnostic purposes. Other uses of this model feature should in our view be restricted for ethical reasons; firm attempts to identify the individuals' class membership conflicts with the privacy guarantee assured in the randomized response method.

ACKNOWLEDGMENTS

We thank Peter van der Heijden, Ardo van der Hout, and Bengt Muthén for their comments on an earlier version of this article.

REFERENCES

Boruch, R. F. (1971). Assuring confidentiality of responses in social research: A note on strategies. *The American Sociologist, 6,* 308–311.

Fox, J. A., & Tracy, P. E. (1986). *Randomized response: A method for sensitive surveys.* Beverly Hills, CA: Sage.

Lee, R. M. (1993). *Doing research on sensitive topics.* London: Sage.

Maddala, G. S. (1983). *Limited dependent and qualitative variables in econometrics.* Cambridge University Press: Cambridge, England.

Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide.* Los Angeles: Muthén & Muthén.

Scheers, N. J., & Dayton, M. C. (1988). Covariate randomized response models. *Journal of the American Statistical Association, 83,* 969–974.

Van den Hout, A., & van der Heijden, P. G. M. (2002). Randomized response, statistical disclosure and misclassification: A review. *International Statistical Review, 70,* 269–288.

Van der Heijden, P. G. M., & Van Gils, G. (1996). Some logistic regression models for randomized response data. In A. Forcina, G. M. Marchetti, R. Hatzinger, & G. Galmatti (Eds.), *Statistical modelling. Proceedings of the 11th International Workshop on Statistical Modelling* (pp. 341–348). Perugia, Italy: University of Perugia..

Van der Heijden, P. G. M., & Van Gils, G., Bouts J., & Hox, J. J. (2000). A comparison of randomized response, computer assisted self interview, and face-to-face direct questioning: Eliciting sensitive information in the context of welfare and unemployment benefit. *Sociological Methods & Research, 28,* 505–537.

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

APPENDIX

M*plus* commands for the simulated data

```
TITLE:

DATA:
  FILE IS "d:\simu1000.dat";

VARIABLE:
  NAMES ARE idnr pred response;
  USEVARIABLES ARE pred response;
  CATEGORICAL ARE response;
  CLASSES = c(2);

ANALYSIS:
  TYPE IS MIXTURE;
  LOGHIGH = +15;
  LOGLOW = -15;
  UCELLSIZE = 0.01;
  ESTIMATOR IS MLR;
  LOGCRITERION = 0.0000001;
  ITERATIONS = 100000;
  CONVERGENCE = 0.000001;
  MITERATIONS = 50000;
  MCONVERGENCE = 0.000001;
  MIXC = ITERATIONS;
  MCITERATIONS = 2;
  MIXU = ITERATIONS;
  MUITERATIONS = 2;

OUTPUT: SAMPSTAT;

MODEL:
  %OVERALL%
  [response$1*1.5];
  response on pred*1;
  %c#2%
  response on pred@0;
  [response$1@0];
```