ASSESSING QUALITY OF RANDOMIZED RESPONSE: WERE INSTRUCTIONS FOLLOWED?

James Schmeidler, N.Y. State Division of Substance Abuse Services*
55 West 125 Street, New York, NY 10027

1. INTRODUCTION AND SUMMARY

In a pilot study, the value of innovative questions is assessed by measuring their reliability and, if feasible, their validity. This is particularly true for a complex procedure such as randomized response. However, in randomized response (Warner, 1965), unreliability is intentionally introduced, so the criterion of reliability is not applicable. A substitute criterion is to determine whether the randomized response instructions were followed. Some models for systematic disregard for instructions are presented.

Estimation of model parameters from randomized response data often necessitates iterative computations. This difficulty can be avoided by use of the unrelated question model (Horvitz, Shah, and Simmons, 1967; Greenberg, Abul-Ela, Simmons, and Horvitz, 1969) with equal probabilities for the unequal question. This method for testing whether instructions were followed is illustrated for a pilot telephone study of randomized response questions on substance abuse.

2. VALIDITY AND RELIABILITY

In general statistics, the variance of an estimate may be decomposed into the variance of the attribute being measured, the variance of the estimator, and the square of the bias of the estimator. In psychometrics, these concepts are formulated differently. Unreliability is the proportion of the total variance due to estimator variance; its complement is reliability. Validity is the proportion of the total variance due to the variance of the attribute; its complement, invalidity, is composed of the unreliability and the contribution of the squared bias. Informally, reliability measures the consistency of the response, and validity measures its accuracy. These are different--a consistent but inaccurate response--if there is bias, a systematic inaccuracy.

Reliability is usually estimated relatively easily, by comparing different estimators; their correlation depends only on the respective reliabilities. If alternative forms of a question, or the same question asked at different times, give answers that agree, this establishes high reliability. However, validity is often much more difficult to estimate, since this requires an unbiased estimator for comparison. It may be impossible to verify assertions if a study is being conducted anonymously. Reliability is often used as a surrogate for validity, even though it must be an overestimate, if there is no reason to assume an estimator is severely biased.

A major difference between unreliability and bias as components of invalidity is that replication reduces unreliability but it does not reduce bias. Thus increased sample size cannot increase validity beyond the bound imposed by bias. When asking about sensitive topics, such as substance abuse, where admission may be embarrassing, bias may be a serious problem. To increase validity, bias must be reduced; how can giving stigmatizing answers be made less threatening to the respondent?

3. RANDOMIZED RESPONSE

Randomized response (Warner, 1965; Greenberg, Abul-Ela, Simmons, and Horvitz, 1969) is a technique of asking questions in which the meaning of every answer is intentionally ambiguous. The respondent is instructed to respond to one of two forms of a question, depending on the result of a random device (such as a spinner) seen by the respondent but not by the interviewer. Since the interviewer does not know which question was answered, it is impossible to determine what any particular response means. One question is made more likely than the other, by controlling the probabilities of the random device. Then inferences can be made on a collective basis about a sample, since most answers were based on one question form rather than the other.

In Warner's (1965) original version of randomized response, the questions were antithetical, for example eliciting a true/false response to either "I am a narcotics addict" or "I am not a narcotics addict." A more efficient model of randomized response (Horvitz, Shah, and Simmons, 1967; Greenberg, Abul-Ela, Simmons, and Horvitz, 1969) has two questions with the same set of responses, but one refers to the sensitive topic and the other refers to a neutral unrelated topic, such as place of birth. The probabilities of these neutral answers must be used in the randomized response estimation procedure. If they are not known a priori, they require empirical estimation. In a variation of the unrelated question (Greenberg, Abul-Ela, Simmons,

* The author gratefully acknowledges the critical comments of Bruce W. Hall, University of South Florida. Points of view or opinions in this paper do not necessarily represent the official positions or policies of the New York State Division of Substance Abuse Services.

and Horvitz, 1969), the randomizing device directs the respondent to a specific answer if the truth is not elicited, thus controlling the probability distribution of the "unrelated" answer.

If "I am a narcotics addict" is the more likely form, the response "true" is likely to be indicative of the stigmatizing fact of being a addict. However, t he response "true" may have the opposite or a neutral meaning (depending of the version of randomized response). Thus every respondent who says "true" is shielded by the fact that noone else knows for certain which question was actually answered by that response. An honest admission of a stigmatizing attribute may be given more freely than in direct questioning, since the response might have the nonstigmatizing meaning.

In most surveys, the objective is to draw conclusions about the sample, for inference concerning a population, not about individual respondents. Thus the unavailability of information about individual respondents is not a drawback. If responses to direct questions were biased, such information would not be available without randomized response, either. More accurate information about individuals can be obtained by replicating randomized response for each respondent (Horvitz, Shah, and Simmons, 1967; Liu and Chow, 1976).

4. QUALITY OF RANDOMIZED RESPONSE

Validity and reliability must be reinterpreted to make sense in the context of randomized response. Without randomized response, replications of a perfectly reliable estimator would yield the identical value. However, if randomized response instructions are followed, the result should vary according to the outcome of the random device. Similarly, randomized response replications of a perfectly valid estimator should not always yield the true value. Inclusion of variance due to the randomizing device in the unreliability and the invalidity is a positive feature of randomized response. This contrasts with other sources of unreliability and invalidity, which are undesirable.

If a respondent disregarded randomized response instructions and always answered the sensitive question truthfully, reliability and validity, as usually defined, would be increased. However, inference from such data, using methods adjusting for randomized response, would be incorrect. Thus it is more suitable to use the term quality, rather than reliability, to refer to appropriateness of randomized response. A proper response to a randomized response question follows instructions, rather than responding to the sensitive question.

Randomized response decreases reliability since the response varies due to the result of the random device, in addition to sources of unreliability otherwise present. If this technique induces the respondent to answer more honestly, bias is reduced. For inference about a sample, such a trade can be worthwhile (Warner, 1965). Using a large sample size can counteract the increased unreliability, since the contribution of unreliability to invalidity varies inversely as the square root of the sample size. Then the net effect is increased validity, due to lowered bias. This could not have been accomplished by simply using a large sample size without employing randomized response. However, another cost that must be considered is increased interview time explaining randomized response and performing the randomizations.

In a pilot study, it is important to determine whether the instructions were followed. This is particularly true for a complex procedure such as randomized response. Respondents may balk at responding with forced answers on some occasions, rather than telling the truth at all times. If a respondent pretended to participate in the randomized response procedure, but actually told the truth at all times, this strategy would produce an "improper" response for the randomized response task (although reliable and valid by the usual definitions). Estimates from such responses, using procedures designed to corect for randomized response, would not be accurate.

Another response strategy in which instructions would not be followed would be to deny the stigmatizing characteristic (such as drug addiction) regardless of the randomizing device and the truth. If the respondent did not have the stigmatizing characteristic, this accurate denial would be the improper response described above. However, if he had it, unvarying denial of abuse by an addict would be a biased response--because of its falsity. Such unvarying denial would also be improper--because an admission would never be forced by randomization. These two strategies would not be distinguishable unless the respondent had the stigmatizing characteristic. There would be no way to distinguish between a nonaddict always telling the truth and a nonaddict always denying addiction.

In a study with external validation, it would be easy to distinguish between these two strategies by considering only responses for which it was known that the respondent had the stigmatizing characteristic. However, external validation is difficult in any study of sensitive characteristics, and impossible in a genuinely anonymous survey. In a study without external validation, it is not possible to distinguish between responses that are only improper or also biased. Each of these strategies includes improper response. Although bias cannot be assessed without external validation, it

is possible to test for improper response. This assesment of whether the specifically randomized response aspect of the instructions was followed, as distinguished from any bias, is analogous to the usual assessment of reliability. As in that situation, replication of questions can provide the necessary information.

5. DISCREPANT RESPONSES TO REPLICATED QUESTIONS

Consider an unrelated question randomized response model with probability p of asking the sensitive question, and with known probability r that the unrelated question will have the response "yes." Randomized response questions can be replicated, as is done in one form of reliablity assessment. There are four possible outcomes of two questions: yes/yes and no/no, which are consistent; and yes/no and no/yes, which are discrepant answers. In investigating impropriety of response, the critial outcomes are the discrepant ones, rather than the consistent ones.

If the respondent has the sensitive attribute and follows instructions, the probability of a "no" response is

$$(1 - p)(1 - r).$$

Then the probability that an answer and the replication will be discrepant, in either order, is

$$2 [(1 - p)(1 - r)] \cdot [1 - (1 - p)(1 - r)]. \quad (1)$$

If the respondent does not have the sensitive attribute, and follows instructions, the probability of a "yes" is

$$(1 - p)r,$$

so the probability of a discrepancy is

$$2 [(1 - p)r] [1 - (1 - p)r]. \quad (2)$$

If the respondent always tells the truth, or always answers "no," the probability of a discrepancy is zero.

The probability of a discrepant response depends upon the prevalence of the sensitive attribute, π, and the extent to which instructions are followed. If everyone follows instructions, the overall probability of a discrepant response in a population is a weighted average of (1) and (2),

$$\pi \{ 2 [(1 - p)(1 - r)] \cdot [1 - (1 - p)(1 - r)] \} + (1 - \pi) \{ 2 [(1 - p)r] \cdot [1 - (1 - p)r] \}, \quad (3)$$

with only one parameter, π. This is the null hypothesis when testing for improper response. Models for the alternative hypothesis include improper response and perhaps also bias.

A second model, with only improper response, is that some proportion, ϕ, of all respondents, both those with and without the sensitive attribute, follow instructions. However, the rest fail to follow instructions by always responding truthfully, so they never give a discrepant response. In this model, the probability of a discrepant response is only ϕ times (3),

$$\phi\pi \{ 2 [(1 - p)(1 - r)] \cdot [1 - (1 - p)(1 - r)] \} + \phi(1 - n) \{ 2 [(1 - p)r] \cdot [1 - (1 - p)r] \}. \quad (4)$$

Here there are two parameters, π and ϕ, determining the coefficients of (1) and (2) in (4).

A third model is that some proportion, ψ, of those with the sensitive attribute follow instructions. The rest always deny use, so they never respond discrepantly. As noted above, this model includes both bias and improper response. Here the probability of a discrepant response is a different weighted sum of (1) and (2),

$$\psi\pi \{ 2 [(1 - p)(1 - r)] \cdot [1 - (1 - p)(1 - r)] \} + (1 - \pi) \{ 2 [(1 - p)r] \cdot [1 - (1 - p)r] \}. \quad (5)$$

As in the previous model, there are two parameters, π and ψ, determining the coefficients of (1) and (2) in (5). This model yields probabilities indistiguishable from the second model, when suitable parameter values are substituted into (4):

$$1 - \pi + \psi\pi$$

replacing ϕ and

$$\psi\pi / (1 - \pi + \psi\pi)$$

replacing π. There is little point to considering even more general models, including different rates of truth telling and denial. Regardless of the profusion of parameters, they are still only weighted sums of (1) and (2) and zero, and thus indistiguishable from the two parameter second model with suitable parameter values in (4).

The motivation for the use of randomized response is to reduce the probability that respondents will give biased responses by denying stigmatizing behavior. However, this comparison of models confirms that purely improper response is indistinguishable from biased improper response. Thus if randomized responses are shown to be improper, there can be no way--in the absence of external validation--to demonstrate that they are only improper and not also biased. In any event, whether there was only improper response or also bias, the randomized response results would not be valid.

Since the models of improper response considered above give lower probability of discrepancy than proper response, a one sided test could be considered. However, another possibility is that a respondent might have increased anxiety about giving a second stigmatizing response after responding positively to an earlier question. Instead, a discrepant response might be given to the second question, rather than confirming the earlier stigmatization when directed to do so (whether the second response was the truth or directed by the randomizing

device). This could produce more discrepant responses, specifically "yes" followed by "no."

6. THE NULL HYPOTHESIS WHEN r = .5

When the unrelated question is answered "yes" and "no" with equal probability, the value r = .5 has a fortuitous effect on the assesment of quality of response. Substituting this value into (1) and (2) gives the same value,

$$(1 - p^2) / 2 \qquad (6)$$

for the probability of a discrepancy in both cases. Since (3) is a weighted average of (1) and (2), its value must be (6) regardless of π, the prevalence of the sensitive attribute.

In general, estimation of parameters for replicated randomized response is an arduous process requiring iterative methods (Horvitz, Shah, and Simmons, 1967; Liu and Chow, 1976). However, for the telephone pilot study discussed below, the general solution--estimating all the parameters for the model--is not required. To assess quality of response, it is sufficient to test the null hypothesis that randomized response directons were followed, against the alternative hypothesis of improper (including possibly biased) response. According to this null hypothesis, the number of discrepant responses follows a binomial distribution with parameter (6). Improper response, with or without bias, constitutes the alternative hypothesis. Other sources of unreliability will also affect the number of discrepancies; such models were considered by Horvitz, Shah, and Simmons (1967).

7. RANDOMIZED RESPONSE OVER THE TELEPHONE

Almost all randomized response research has been performed in the context of face to face interviewing. This permits the interviewer to supply a physical device, such as a spinner (Warner, 1965), a deck of cards with questions printed on them (Horvitz, Shah, and Simmons, 1967), or a flask that presents a ball at random (Liu and Chow, 1976). In a telephone survey, or a mail survey in which a spinner or other device is not sent, the repondent must be instructed how to make a randomizing device from materials at hand. Stem and Steinhorst (1985) reported several methods that have been used.

In one method, used by the New York State Division of Substance Abuse Services in a pilot telephone study for substance use questions (Weissman, 1981), the respondent was told to toss three coins. This procedure is described in detail, since results from this pilot study are analysed in the next section. If all three coins were heads, the respondent was to say "yes;" if all tails, "no;" and if there was a mixture of heads and tails, the respondent was instructed to tell the truth. Thus p = .75, asking the sensitive question three quarters of the time. When a response was forced, which was equivalent to asking an unrelated question, r = .5, since forced responses were half "yes" and half "no." This format was used for questions about lifetime use of four drugs: cocaine, PCP, LSD, and heroin, and for a variety of other questions.

This choice of r =.5 was constrained by the simple random device used and the need for simple instructions. This is not the most effective choice of r (Greenberg, Abul-Ela, Simmons, and Horvitz, 1969). On the contrary, when the true proportion of the sensitive attribute is believed to be on one side of .5, as in the case of drugs rarely abused, it is best to choose r as far from .5 as practical, but on the same side as the unknown proportion. Stem and Steinhorst (1984) discuss a more flexible technique for telephone randomized response, based on telephone book "random" numbers.

Another randomized response format was used for one set of questions about these same four drugs: "How recently did you use (name of drug)?" The possible answers were "within the past thirty days," "more than thirty days ago but in the past six months," "more than six months ago," and "never." For these questions, if there were three tails, the response was forced to be "never," but if there were three heads, the response was to refer to the length of time since the respondent's birthday (or that of the respondent's father, mother, or another relative, for other drugs).

The most important practical result of this pilot study was that 55 of 115 respondents randomly assigned to the randomized response procedure insisted they preferred to tell the truth rather than tossing coins and responding accordingly. For three of the four drugs, lifetime rates of use estimated by randomized response were higher than rates for respondents randomly assigned to be asked direct questions (and also higher for three of four drugs than those who objected to participating in randomized response), but none of these comparisons were statistically significant. Small sample sizes and low population rates for these attributes made this an inconclusive test of the merit of randomized response.

For four drugs, randomized response questions were asked both concerning lifetime use and concerning the most recent use. If the time of the most recent use is disregarded, the recency question collapses into a replication of the question on lifetime use. This permits a test using the method of Section 6.

8. EVALUATION OF THE QUALITY OF PILOT STUDY RESPONSES

In this telephone pilot study, randomized response was administered to 60 respondents. Since p = .75, the probability of a discrepancy, if instructions were followed, was .21875 by (6). Thus under the null hypothesis that instructions were followed, the expected number of discrepant responses on the pair of questions about each of the four drugs was 13.125 . The results were 18, 10, 17, and 12, which were clearly not significantly different from the expected value, based on a binomial distribution. If anything, these results showed a tendency toward more discrepancies than the null hypothesis predicted, contrary to several of the alternatives discussed in detail above.

Another possible alternative to the null hypothesis was that the responses to the two forms of question might differ, due to a difference in the randomized response technique, or to a reluctance to confirm a stigmatizing response. Of the 57 discrepant responses, 22 were "yes" to the simple lifetime use question, and 35 were "yes" to the more complex question on recency of use, which was asked later. This was not significantly different from an even division. This analysis pooled results from four questions; since a respondent might have contributed several results, such pooling provides an anti-conservative test of significance.

The conclusion was not to reject the null hypothesis that the randomized response instructions were followed in this pilot study. The respondents who agreed to answer the coin tossing randomized response questions seem to have done so properly. However, it must be remembered that this sample was small and self-selected, in that the many respondents who resisted using randomized response were not included.

REFERENCES

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

Horvitz, D. G., Shah, B. V., and Simmons, W. R. (1967) "The unrelated question Randomized response model," in Proceedings of the Social Statistics Section, American Statistical Association, 65-72.

Liu, P. T., and Chow, L. P. (1976) "The efficiency of the multiple trial randomized response technique," Biometrics, 32, 607-618.

Stem, D. E., Jr., and Steinhorst, R. K. (1984) "Telephone interview and mail questionnaire applications of the randomized response Model," Journal of the American Statistical Association, 79, 555-564.

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

Weissman, A. N. (1981) "Randomized response versus direct questioning: Two methods for asking sensitive questions over the telephone," paper presented at the American Educational Research Association meeting, Los Angeles.