Meta-Analysis of Randomized Response Research

Thirty-Five Years of Validation

GERTY J. L. M. LENSVELT-MULDERS
JOOP J. HOX
PETER G. M. VAN DER HEIJDEN
CORA J. M. MAAS
Utrecht University, the Netherlands

This article discusses two meta-analyses on randomized response technique (RRT) studies, the first on 6 individual validation studies and the second on 32 comparative studies. The meta-analyses focus on the performance of RRTs compared to conventional question-and-answer methods. The authors use the percentage of incorrect answers as effect size for the individual validation studies and the standardized difference score (d-probit) as effect size for the comparative studies. Results indicate that compared to other methods, randomized response designs result in more valid data. For the individual validation studies, the mean percentage of incorrect answers for the RRT condition is .38; for the other conditions, it is .49. The more sensitive the topic under investigation, the higher the validity of RRT results. However, both meta-analyses have unexplained residual variances across studies, which indicates that RRTs are not completely under the control of the researcher.

Keywords: *Randomized response; meta-analysis; multilevel; sensitive topics*

This article describes the outcomes of a meta-analysis of 38 randomized response validation studies. Randomized response designs are especially developed to obtain more valid estimates when studying sensitive topics, that is, topics perceived as threatening to respondents (Lee 1993). Such threats can be either extrinsic or intrinsic. A threat is extrinsic if certain responses carry the risk of sanctions (e.g., if the questions are about illegal or deviant behavior) and intrinsic if the questions concern subjects that are very personal or stressful

AUTHORS' NOTE: *The authors would like to thank the anonymous reviewers for the time and effort they put into the manuscript and for their helpful comments.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 33, No. 3, November 2004 1-30
DOI: XXXXXXXXXXXXXXXXXXXXXXXx

to the respondents, or certain responses imply a negative adjustment in their self-image. It is due to these threats that researchers studying sensitive topics are more often confronted with higher nonresponse rates and obtain more socially desirable answers than those studying neutral topics. The disturbances can lead to underreporting sensitive topics, thus making the data less valid (Lee 1993; Rasinski et al. 1999).

In pursuing a survey method that guarantees the most valid results, nowadays the focus is mainly on improving computer-assisted survey techniques such as computer-assisted interviewing (CAI; Baker and Bradburn 1992; de Leeuw, Hox, and Snijkers 1995). This is so because computers have no social context cues and increase the respondent's perception of privacy protection, which can lead to greater self-disclosure (Supple, Aquilino, and Wright 1999). Regrettably, a meta-analysis on earlier CAI studies shows that using a computer does not have a consistent positive effect on data distortion. The distortion seems largely based on moderating factors, such as whether respondents are tested alone or with others (Richman et al. 1999). In the future, the relative advantage of using computer-assisted self-interviews will be smaller as respondents become more computer literate, and negative media publicity (e.g., about organizations merging files) may lead to less trust in CAI. Weisband and Kiesler (1996) note what they call a *year effect*; the positive effect of using CAI to measure sensitive topics decreases with time. If this trend continues, CAI will become less indicated for the study of sensitive topics.

This is why it is interesting to study alternative data collection techniques that can be used to study sensitive questions validly. The randomized response technique (RRT) is one of these techniques. Although it has never been a leading survey research technique, it has been used in many scientific studies, with variable results. In this article, we carry out a meta-analysis to uncover consistencies and account for variations across studies (Cooper and Hedges 1994).

In the next section, we explain how RRT works. The steps needed to conduct a meta-analysis are described (Halvorsen 1994), as is the literature search. Then we describe how potentially eligible papers are retrieved and evaluated. The coding rules are given, technical details of the meta-analysis are explained, and the results are presented and interpreted.

RANDOMIZED RESPONSE TECHNIQUE

The RRT is an interview method that guarantees privacy and may well overcome respondents' reluctance to reveal sensitive or probably harmful information (Chaudhuri and Mukerjee 1988; Fox and Tracy 1986). By inserting an element of chance (e.g., using cards or dice) in the question-and-response process, the respondents' privacy is fully guaranteed. As a result, respondents are more inclined to cooperate and give honest answers to sensitive questions. We describe Warner's (1965) randomization technique as an example of randomized response methods in general.

WARNER'S RANDOMIZED RESPONSE METHOD

Using a randomization device (colored marbles in a box, coins, or dice), respondents are directed toward one out of two statements—for example, the following:

- A: I am for capital punishment (A: selected with probability p, $p \neq .5$).
- B: I am against capital punishment (not-A: selected with probability $1 - p$).

Without revealing to the interviewer which statement is selected by the dice, the respondent answers *true* or *not true* according to his or her attitude to capital punishment. Elementary probability theory can be used to get a bias-free estimate ($\hat{\pi}$) of the population probability of A (for capital punishment) by

$$\hat{\pi} = (\hat{\lambda} + p - 1)/(2p - 1), \tag{1}$$

where $\hat{\lambda}$ is the observed sample proportion of yes answers. The sampling variance of $\hat{\pi}$ is given by

$$\hat{\text{var}}(\hat{\pi}) = \left[\hat{\pi}(1 - \hat{\pi})/n\right] + \left[p(1 - p)/n(2p - 1)^2\right]. \tag{2}$$

In equation (2), $\hat{\pi}(1 - \hat{\pi})/n$ is the standard for the sampling variance of a proportion, and $p(1 - p)/n(2p - 1)^2$ represents the variance added by the randomized response technique. Equation (2) shows that this added variance decreases if p is further from 0.5.

Since Warner (1965) published his first paper on randomized response, many researchers have improved and further developed this

technique. Efforts have been made to improve the efficiency of the technique by reducing the variance and thus the confidence intervals; other efforts have tried to improve the psychological features of the randomized response technique and enhance the respondents' trust in the technique so they are more inclined to open up. For a good overview of RRT techniques, see Chaudhuri and Mukerjee (1988).

PROFITS AND COSTS OF USING RANDOMIZED RESPONSE TECHNIQUES

The most important RRT claim is that it yields more valid point estimates of sensitive behavior. It is important to note that even though sensitive behavior is measured using RRT, it is still possible to link sensitive behavior to explanatory variables with a specially adapted version of the logistic regression technique (Maddala 1983; Scheers and Dayton 1988; van der Heijden and van Gils 1996). The explanatory variables can be dichotomous, such as gender; ordinal, such as educational level; or continuous, such as attitudes to sensitive behavior.

However, using an RRT entails extra costs (Lensvelt-Mulders, Hox, and van der Heijden forthcoming). Equation (2) clearly shows that Warner's RRT is less efficient than conventional collection methods. RRTs produce larger sampling variances, which leads to reduced power and thus necessitates larger samples. Extra costs are also associated with the increased complexity of RRT questions as compared to more conventional forms of data collection. Survey methodologists employ a question-and-response model developed by Tourangeau and Rasinski (1988). Four steps characterize an optimal question-answering process: (1) understanding the question, (2) retrieving the relevant information from memory, (3) integrating this information into a summarized judgment, and (4) reporting this judgment correctly. Using an RRT adds extra steps to this process since respondents also have to understand and follow the RRT instructions. Clearly, this increases the cognitive load of RRTs as compared to the conventional question-and-answer process. It also allows for new sources of error, such as misunderstanding the RRT procedures or cheating on the procedures (Boeije and Lensvelt-Mulders 2002).

The advantage of using RRT—more valid population estimates—only outweighs the extra costs if population estimates are substantially more valid than estimates from straightforward question-and-answer

designs. Two sorts of studies have been carried out to address this question: individual validation studies and comparative studies.

Individual Validation Studies

Individual validation studies are the standard for testing the value of a method. A study is defined as an individual validation study if we know the true status of each individual on the sensitive issue involved. This enables us to compare the population estimates to the true mean. Individual validation studies have a high internal validity, but they are rare since they require sensitive information at an individual level, and access is thus needed to databases on sensitive topics such as police or medical files. Six individual validation studies have been retrieved, and the results vary greatly across these studies.

Comparative Studies

A study is comparative if RRTs are compared to conventional data collection methods (self-administered questionnaires, telephone interviews, face-to-face interviews and CAI), without the option of individual validation of the results against a known criterion. The results of comparative studies are interpreted according to the *more is better* assumption: A higher population estimate is interpreted as a more valid population estimate (Umesh and Peterson 1991). As in individual validation studies, the results vary greatly across the comparative studies. On the question "Have you ever taken anything from a shop worth over 50 dollars?" Beldt, Daniel, and Garcha (1982) find no indication that the RRT provides better estimates than direct questioning, but Wimbush and Dalton (1997) find positive results for the same question, and in their case, an RRT performs better than face-to-face interviews and self-administered questionnaires.

A thorough look at the literature on RRTs reveals that 35 years of research have not led to a consensus or a description of best practices. Many statistical improvements have enhanced the method's efficiency and reliability, and numerous varieties of randomized response procedures have been developed. However, individual validation and comparative studies exhibit ample diversity in the research outcomes. This is why we have decided to do a formal meta-analysis to better understand the merits of RRTs for producing more valid estimates than the conventional collection methods.

PURPOSE OF THE META-ANALYSIS

The meta-analysis addresses the following questions:

1. *Do RRTs produce more valid population estimates for sensitive topics than conventional question-and-answer designs such as face-to-face interviews, self-administered questionnaires, telephone interviews, and computer-assisted interviews?*

   In this study, we are interested in the merits of the RRT as compared to conventional data collection designs for providing more valid data. A meta-analysis provides an integrated summary of the outcomes of various studies. Insight into the summarized results of these studies can help researchers design better studies for examining sensitive topics.
2. *Why are RRT results so variable across studies?*

   If the differences between studies cannot be explained by sampling variance alone, the results are considered heterogeneous, which means the differences are not due to chance. They can be due to differences in the quality of the studies, the sensitivity of the topics, the samples, or the implementation of the RRT.
3. *Do comparative studies provide the same information as individual validation studies?*

   We include individual validation as well as comparative studies in our research. Due to effect measure differences between individual validation and comparative studies, which are discussed later, they are analyzed separately in two meta-analyses. Individual validation studies have a stronger internal validity than comparative ones but are harder to carry out, for instance, because it is difficult to obtain true sensitive information about the individuals in the sample. As a result, more comparative studies have been retrieved than individual validation ones. Although comparative studies are easier to carry out, they are only informative if the outcomes are comparable to those of individual validation studies.

LITERATURE RETRIEVAL AND CODING

COMPILING A BIBLIOGRAPHY

A bibliography on randomized response studies by Nathan (1988) has served as the point of departure for our search for randomized response literature. The bibliography covers the period from 1965 to

1987 and contains more than 250 theses, research reports, published papers, and books. To supplement this bibliography and expand it from 1987 to 2000, a literature search has been carried out following the instructions given by Cooper and Hedges (1994) and the Campbell Collaboration for evidence-based research in the social sciences. An online search has been conducted in the following computer databases: PsychInfo, Sociofile, Eric, Medline, Sage Publications CD-ROM, the SRM-database of social research methods, the Current Index to Statistics (eighth release), the SSCI (Social Sciences Citation Index), and JSTOR. For the subject search, we use the terms *randomized response* and, in the next step, *sensitive questions*. Search results have been compared with the Nathan bibliography; extra articles (1965-1987) were added, as was more recent literature. In addition, the reference list is supplemented by studies resulting from inspections of reference sections of previously located studies, and a call for unpublished studies was sent to the SMRS (Survey Methods, Research, and Statistics) Internet mailing list.

This search strategy has produced a bibliography of randomized response studies conducted between 1965 and 2000. Most of the studies address statistical issues such as how to deal with sampling problems, the statistical efficiency of the design, and detecting lying in the RRT analysis. To limit the publications to individual validation and comparative studies, in addition to *randomized response* and *sensitive questions*, we have added the terms *compare*, *comparative*, *evaluate*, *validation*, *direct questioning*, *telephone surveys*, *mail*, and *CAI*. Inspection of the abstracts leaves us with only 70 potentially useful papers. This confirms the conclusion of Umesh and Peterson (1991), who note that there have been very few substantive applications of RTTs and that most papers are published to test a variant or illustrate a statistical problem. We have retrieved as many studies as possible, 68 in all, by conducting a library search using online full-text contents (OMEGA) and contacting authors and institutions.

CRITERIA FOR INDIVIDUAL VALIDATION STUDIES

To include an individual validation study in the meta-analysis, the report should provide sufficient information to derive an effect score for differences between the RRT outcome and the known

TABLE 1: Mean Results of Individual Validation Studies (1975–2000)

Study	*Method/Condition*	*Percentage Wrong Answers*
Horvitz, Shah, and Simmons (1967)	RRT	17
Lamb and Stem (1978)	RRT	11
	Face-to-face	25
Locander, Sudman, and Bradburn (1976)	RRT	20
	Telephone	32
	Questionnaire	35
	Face-to-face	32
van der Heijden et al. (1998, 2000)	RRT (Kuk)	54
	RTT (forced response)	51
	CAI	81
	Face-to-face	75
Tracy and Fox (1980-1981)	RRT	43
	Face-to-face	40
Kulka, Weeks, and Folsom (1981)	RRT	28
	Face-to-face	16

NOTE: RRT = randomized response technique; CAI = computer-assisted interviewing.

population value together with its sampling variance (Lipsey and Wilson 2000; Hox and de Leeuw 2003). Six individual validation studies have been retrieved: Horvitz, Shah, and Simmons (1967); Kulka, Weeks, and Folsom (1981); Lamb and Stem (1978); Locander, Sudman, and Bradburn (1976); Tracy and Fox (1980, 1981); and van der Heijden et al. (1998, 2000). If a study's results are described in more than one publication, they are still considered one study (van der Heijden et al. 1998, 2000; Tracy and Fox 1980-1981). The results of the individual validation studies are summarized in Table 1.

CRITERIA FOR COMPARATIVE STUDIES

In the second meta-analysis, we include all the studies that compare a randomized response method with one or more conventional data collection methods. Only 32 of the 68 studies we have been able to retrieve are included. The most important reason why certain studies are not included in the meta-analysis is because they failed to meet the inclusion criteria. Furthermore, studies limited to a comparison of new randomized response results with results from earlier studies or results obtained from literature reviews are not included: This design

makes it impossible to compute an effect score because the studies lack sufficient information on the results of previous studies. Studies using a within-group design with respondents in more than one condition are also excluded because this design can result in biased effect scores due to order effects. To be included in the study, population estimates and their estimated standard errors (or sufficient statistics to calculate them) are needed for all the conditions; many studies provide too little of this vital information to be included. Thirty-two comparative studies meet all criteria and are included in the second meta-analysis.

CODING[1]

For each study, the following variables are encoded:

1. Dependent variables. For individual validation studies, the effect size is calculated as the percentage of incorrect answers, which is the difference between the known population probability of 1 and the point estimate. For instance, the effect size in the RRT condition of the Lamb and Stem (1978) study is 11 percent, and the effect size of the face-to-face condition is 25 percent (Table 1). All effect sizes are underestimations of the true population score of 100 percent. The individual validation studies thus only deal with false negatives, whereas comparative studies also deal with false positives (respondents who are not part of the sensitive group incorrectly state that they are).

The interpretation of the results of comparative validation studies is different from that of individual validation studies because a different effect measure is used. Individual validation studies compare observed estimates to a known true score, resulting in a straightforward effect measure. Comparative studies do not have a known true score. The effect of the randomized response condition is compared to that of other data collection conditions, and higher estimates for the sensitive characteristics are interpreted as more valid estimates. The standardized difference score for proportions (i.e., the difference between the cumulative standard normal value for the proportions p_{rr} in the RRT group and p_{control} in the control conditions),

$$d'_{\text{probit}} = Z_{rr} - Z_{\text{control}}, \tag{3}$$

has been selected as effect measure for the comparative studies.

This standardized difference d'_{probit} is the difference between p_{rr} and p_{control} transformed to a probit scale. Rosenthal (1994) gives the corresponding sampling variance estimates as

$$\hat{VAR}\left(d'_{\text{probit}}\right) = \frac{2\pi p_{rr}(1-p_{rr})e^{z^2_{rr}}}{n_{rr}} + \frac{2\pi p_{\text{control}}(1-p_{\text{control}})e^{z^2_{\text{control}}}}{n_{\text{control}}}. \quad (4)$$

Equation (4) does not take into account the extra variance added by the RRT condition; the variance is simply computed as $p_{rr}(1-p_{rr})/n$, so it underestimates the sampling variance. This is why we replace it with

$$\hat{VAR}\left(d'_{\text{probit}}\right) = 2\pi(se_{rr})^2 e^{z^2_{rr}} + 2\pi(se_{\text{control}})^2 e^{z^2_{\text{control}}}, \quad (5)$$

where $(se_{rr})^2$ equals the variance related to the randomized response technique, and $(se_{\text{control}})^2$ equals the variance for the direct question conditions $(p_{\text{control}}(1-p_{\text{control}}))$.

Although directly surveying sensitive issues is generally thought to lead to an underestimation of the true population score, some sensitive items lead to boasting and thus to an overestimation of the true population score (Brewer 1981; Zdep et al. 1979). If boasting is expected in a specific study, a negative d'_{probit} is recoded into a positive outcome.[2] As a consequence of this recoding, a significant positive d'_{probit} unambiguously means that the randomized response condition provides the more valid population estimate as compared to the non-RRT approach.

2. Data collection methods. We code the data collection methods as conditions within a study: (1) RRT, (2) telephone surveys, (3) self-administered questionnaires, (4) face-to-face interviews, (5) computer-assisted interviews, (6) scenario designs, and (7) unmatched count technique.

3. Forms of the RRT. To gain insight into the differences between various approaches to randomized response, we coded the forms of randomized response techniques: (1) Warner's technique, (2) forced-response technique, (3) unrelated-question technique, (4) Takahasi's technique, (5) two-stage sampling method, and (6) Kuk's card method (Kuk 1990).

4. Variables dependent on the RRT design. For greater insight into the influence that minor features of the randomized response technique can have on respondents' behavior, we code the randomization device that is used: (1) cards, (2) dice, (3) coins, (4) dollar bill, (5) telephone number, (6) spinner, (7) social security number, (8) colored marbles, and (9) other—the magnitude of the chance a respondent has to answer the sensitive question (p-true), and the number of respondents in different conditions (N).

5. Topic of the study. Because of the importance of the topic in sensitive research, the issues used in the studies are also coded: (1) abortion, (2) sexual behavior, (3) drug use, (4) alcohol abuse, (5) criminal behavior, (6) ethical problems, (7) charity, (8) academic cheating, (9) environmental behavior, (10) medicine compliance, and (11) other.

6. Sensitivity of the topic. A measure of the social sensitivity of the topics is also coded following Himmelfarb and Lickteig (1982), who instruct respondents that

> some people may not answer questions truthfully and the study is an effort to find out what sort of questions should be answered truthfully and untruthfully under anonymous conditions. The subjects should not answer the question but indicate whether they think a typical respondent should answer the question truthfully or not. (p. 715)

To code the sensitivity of the topic, we use the same introduction as Himmelfarb and Lickteig (1982). Because of the subjective judgment involved, we use four independent raters (Department of Social Sciences staff at Utrecht University, the Netherlands) to code the sensitivity of all the research issues, including the ones used by Himmelfarb and Lickteig. They score all the items on a social desirability scale, rated from 0 (*no inclination toward social desirable answering should be expected*) to 4 (*the researcher can hardly expect an honest answer to this question*). In the 32 comparative studies, 226 items are coded, resulting in 104 questions. The mean correlation between the raters on the 104 questions is 0.73. The mean ratings across all four raters are coded as a measure for the social sensitivity of the topic. This measure is again validated against the corresponding scores from the Himmelfarb and Lickteig study. The overall correlation is 0.65 (on 55 overlapping questions). In 2002, the

raters are somewhat more lenient on topics such as sexual behavior or cheating on an exam, and in 1982, they are more lenient toward alcohol-related questions. All the raters (2002 and 1982) agree completely on the expected direction of the answer distortion (i.e., boasting or underreporting).

7. Data quality. Differences in methodological strictness among the studies may cause differences in results. To control for the effect of differences in the quality of studies, we compute a measure for study quality. Four indicators of methodological strictness are combined into one measure.

1. Are the sample sizes adjusted for unequal power in RRT and control conditions (RRT n at least twice control $n = 1$, no adjustment $= 0$)?
2. Are the results published in an international, peer-reviewed journal (yes $= 1$, no $= 0$)?
3. Is the sample a convenience sample (yes $= 0$, no $= 1$)?
4. Are all the coded variables retrievable from the publication (yes $= 1$, no $= 0$)?

The sum of the scores on these four indicators is used as a measure for the quality of the study ($4 =$ *high quality* and $0 =$ *very low quality*).

INTERRATER RELIABILITY

It is standard procedure in meta-analysis to assess the reliability of the coding by calculating the intercoder reliability. Two raters independently code a random sample of five comparative publications (Orwin 1994). The interrater reliability for nominal variables is indicated using Cohen's kappa (Cohen 1960) and for scale variables as the coefficient alpha reliability (Nunnally 1967). Table 2 shows the results of the interrater analysis. All the kappas and correlations are high, indicating that the coding is sufficiently reliable.

META ANALYSIS

DESIGN

Meta-analysis is the statistical synthesis of the results of several studies. Due to the lack of data at the respondent level, the aggregated

TABLE 2: Interrater Reliability

Measure	*Kappa*
Method	1.00
Journal	1.00
Sample	0.89
Format randomized response	1.00
Device	0.92
Standard error RRT	1.00
Scale Measures	*Reliability*
Number of respondents in RRT condition	1.00
Number of respondents in control condition	1.00
Population estimate, RRT condition	0.94
Population estimate, control condition	0.95

NOTE: RRT = randomized response technique.

statistical results are analyzed (Hedges and Olkin 1985). This straightforward approach cannot be used in this study because we have to accommodate the special data structure: Every study contains one or more conditions (RRT and conventional data collection methods), and within these conditions, various sensitive items are coded. The data matrix for the individual validation studies consists of 6 studies, 15 conditions, and 34 items. Coding the comparative studies results in a data matrix consisting of 32 studies, 74 conditions, and 226 items. Because of the hierarchical structure of the data matrix, a multilevel approach to the meta-analysis is used for the analysis (Hox 2002; Kalaian and Raudenbush 1996; Raudenbush 1994). Three models are tested in sequence. The precise model equations are presented below for the two meta-analyses separately. We use a three-level weighted regression model with 6 studies, 16 data collection methods, and 35 items for the individual validation studies and a three-level weighted regression model with 32 studies at the highest level, the 74 data collection conditions in studies at the second level, and the 226 effect sizes (one for each item) at the lowest level for the comparative studies.

All the parameters are estimated by restricted maximum likelihood (RIGLS) using MLwiN (Goldstein et al. 1998). The significance of the regression coefficients is determined by the Wald test. For the variance components, we report the standard errors, but their significance

is assessed using a likelihood ratio chi-square test (Hox 2002; Raudenbush and Bryk 2002).[3]

META ANALYSIS OF THE INDIVIDUAL VALIDATION STUDIES

NULL OR INTERCEPT-ONLY MODEL (M0)

The null or intercept-only model (M0) is given by equation (6):

$$Y_{ijk} = b_0 + v_{0k} + u_{0jk} + e_{ijk}, \tag{6}$$

where Y_{ijk} is the effect size i of condition j in study k; here, Y_{ijk} is the percentage of incorrect answers. The null model estimates the mean effect size across all the studies, conditions, and items (b_0), plus residual error terms at the study level (v_{0k}) and at the condition-within-studies level (u_{jk}). The item level (e_{ijk}) variance at the lowest level, indicated by σ^2_{error}, is the known sampling error for each item calculated from the study publication and entered directly into the analysis as data input. If the residual variances (σ^2_{study} and $\sigma^2_{\text{condition}}$) are not significantly greater than zero, all the observed differences between the effect sizes are considered the result of sampling error. In this case, the analysis stops, and we have to conclude that there are no indications for differences between studies and conditions.

MODEL INCLUDING DATA COLLECTION METHODS (M1)

If σ^2_{study} and/or $\sigma^2_{\text{condition}}$ are not equal to zero, the results are considered heterogeneous across the conditions and/or studies, which means that there are systematic differences between the conditions and/or studies. If this is the case, a full set of dummy variables representing the various data collection methods is added as explanatory variables at the condition level (M1). Since a full set of dummies is used, the intercept is removed from the model. M1 is given by regression equation (7):

$$\begin{aligned} Y_{ijk} = {} & b_1 X_{1jk} + b_2 X_{2jk} + b_3 X_{3jk} + b_4 X_{4jk} \\ & + b_5 X_{5jk} + v_{0k} + u_{0jk} + e_{0ijk}, \end{aligned} \tag{7}$$

where X_1 is the dummy for the randomized response method, X_2 is for telephone surveys, X_3 is for self-administered questionnaires, X_4 is for computer-assisted interviewing, and X_5 is for face-to-face interviews.

MODEL INCLUDING SENSITIVITY (M2)

If, after including the various data collection methods, the residual error variances at the condition and/or study level are still significant, we add sensitivity of the research topic as an explanatory variable at the condition level (M2). The sensitivity of the topic can contribute to the respondents' willingness to answer a question. Differences in sensitivity can thus cause differences across studies. Adding sensitivity leads to regression equation (8):

$$\begin{aligned} Y_{ijk} = {} & b_1 X_{1jk} + b_2 X_{2jk} + b_3 X_{3jk} + b_4 X_{4jk} + b_5 X_{5jk} \\ & + b_6 X_{6jk} + v_{0k} + u_{0jk} + e_{0ijk}, \end{aligned} \tag{8}$$

where X_6 is the measure for sensitivity.

META-ANALYSIS OF COMPARATIVE STUDIES

NULL OR INTERCEPT-ONLY MODEL (M0)

The null or intercept-only model (M0) equals equation (6), with the only difference being the meaning of Y_{ijk}, which is expressed in the comparative studies as d'_{probit}.

MODEL INCLUDING DATA COLLECTION METHODS (M1)

If the analysis of the first model, the null model, leads to the conclusion that the data are heterogeneous across studies and/or conditions within studies, the data collection methods are added to model (9):

$$\begin{aligned} Y_{ijk} = {} & b_1 X_{1jk} + b_2 X_{2jk} + b_3 X_{3jk} + b_4 X_{4jk} \\ & + b_5 X_{5jk} + v_{0k} + u_{0jk} + e_{0ijk}, \end{aligned} \tag{9}$$

where X_1 stands for the difference between RRT and face-to-face interviews, X_2 is for the difference between RRT and scenario conditions, X_3 is for the difference between RRT and telephone

surveys, X_4 is for the difference between RRT and self-administered questionnaires, and X_5 is for the difference between RRT and unmatched count techniques.

MODEL INCLUDING SENSITIVITY (M2)

If, after including the various data collection methods, the residual error variances at the condition and/or study level are still significant, again we add research topic sensitivity as an explanatory variable at the condition level (M2). Adding sensitivity will lead to regression equation (10):

$$Y_{ijk} = b_1 X_{1jk} + b_2 X_{2jk} + b_3 X_{3jk} + b_4 X_{4jk} + b_5 X_{5jk} + b_6 X_{6jk} + v_{0k} + u_{ojk} + e_{0ijk}, \quad (10)$$

where X_6 is the measure for sensitivity.

RESULTS

GENERAL RESULTS ACROSS STUDIES

We start with an overview of the general characteristics of the studies. The 38 studies are from 21 different international journals, 2 working papers of the Research Triangle Institute (RTI, North Carolina), and 6 studies from unpublished literature databases. A total of 226 sensitive questions have been recorded. The questions cover 10 sensitive topics: abortion (5.3 percent), sexual behavior (19.1 percent), drugs (9.5 percent), alcohol (5.3 percent), criminal offenses (17.4 percent), ethical problems/attitudes (16.2 percent), charity (3.8 percent), cheating on exams (19.6 percent), the environment (2.9 percent), and miscellaneous (0.9 percent). The studies have been conducted in the United States (26), the Netherlands (6), Great Britain (2), Scandinavia (2), Canada (1), and Turkey (1). The data collection methods used in the studies are telephone interviews (22), self-administered questionnaires (SAQ, 13), computer-assisted self-administered interviews (CASI, 2), scenario methods (1), the unmatched count technique (2), face-to-face interviews (22), and randomized response methods (all). The randomized response technique can be subdivided into the forced-response method (22), the two

unrelated-questions method (12), the two-stage sampling method (1), Kuk's card method (2), the Takahasi and Sakasegawa design (1), and Warner's original technique (1). The following randomization devices are used: cards (3.4 percent), colored marbles (11.1 percent), dice (24.9 percent), coins (head or tails, 36.4 percent), a banknote (2.7 percent), color lists (3.6 percent), a spinner (8.4 percent), a telephone book (8.9 percent), and Social Security numbers (0.6 percent).

Thirty-four percent of the research is based on population-based samples; 66 percent of the studies use convenience samples such as psychology students. The probability of having to answer the sensitive question (p-true) in the randomized response conditions varies between 0.33 (Kerkvliet 1994) and 0.84 (Shotland and Yankovski 1982), with a mean of 0.67 and a median of 0.7. In the comparative studies, 226 sensitive items are coded. The mean d-probit across the items is .205 (standard deviation [SD] = .52). For 168 of the 226 questions (73.6 percent), RRT results in more valid outcomes than the non-RRT methods.

RESULTS OF THE INDIVIDUAL VALIDATION STUDIES

Six individual validation studies have been retrieved. In these studies, 16 conditions and 34 items are coded. The mean number of respondents per study is 157 (range$= 47 - 239$). Since the multilevel analysis takes the number of respondents into account, the power of this analysis to detect real differences between RRT and conventional data collection methods is sufficient. The generalizability of the results beyond these studies is not very high since only six studies have been retrieved. For the individual validation studies, the effect size is the difference between the known population probability of 1 and the observed estimate. All effect sizes are underestimates of the true score. Small outcome values indicate a small discrepancy and hence a better quality of the data collection method. First, the null hypothesis is tested that results are homogeneous across all studies (i.e., residual variance $\sigma^2_{\text{study}} = 0$) and across all data collection conditions within studies (i.e., residual variance $\sigma^2_{\text{condition}} = 0$). The results are in Table 3 under M0.

The significant intercept value of 0.42 in M0 indicates that in general, there is a significant discrepancy of 42 percent between

TABLE 3: Results of the Meta-Analysis for Individual Validation Studies

Step	*n*	*M0 (Intercept Only)*	*M1 (Conditions Added)*	*M2 (Sensitivity Added)*
Intercept		.42 (.09)*		
RRT	7		.38 (.099)**	.04 (.130)
Telephone	1		.46 (.138)**	.13 (.151)**
Questionnaire	1		.47 (.140)**	.15 (.150)**
CASI	1		.62 (.191)**	.26 (.141)*
Face-to-face	5		.42 (.099)**	.09 (.127)*
Sensitivity				.12 (.036)**
σ^2_{study}		.042 (.028)*	.042 (.029)	.025 (.018)
$\sigma^2_{condition}$		.023 (.010)**	.018 (.008)**	.013 (.005)**

NOTE: n = number of data collection conditions; standard error in parentheses. RRT = randomized response technique; CASI = computer-assisted self-administered interviews.
*$p \leq .05$. **$p \leq .01$.

the known population value and the observed percentages. The variance of the residual errors across studies is given by σ^2_{study} (.042) and the variance across data collection conditions within studies by $\sigma^2_{\text{condition}}$ (.023). The chi-square test on the variances indicates that the effects are heterogeneous across conditions within studies ($\chi^2_1 = 246.69$, $p = .000$) as well as across studies ($\chi^2_1 = 9.39$, $p = .001$). Significant residual variances imply that differences in results across conditions and studies cannot be explained by sampling variation alone. There are systematic differences in outcomes at both levels of the model.

The first question in our meta-analysis is whether the differences can be explained by the various data collection methods—in other words, RRT versus non-RRT approaches. To test this hypothesis, a full set of five dummy variables has been created for the data collection method variable. The results of this analysis are given in Table 3 under M1. All the explanatory variables have a significant effect, which means that all the data collection methods produce a significant underestimation of the known population value. Using RRT produces the smallest difference between the observed outcome and the known population score of 100 percent, with a mean underestimation of 38 percent across all the studies. The other data collection methods all deviate further from the true score, which means the RRT achieves the most valid results. Using a self-administered

questionnaire or a telephone interview to gather data produces approximately the same underestimation of 46 to 47 percent. Face-to-face interviews show a mean underestimation of 42 percent, and computer-assisted self-interviews produce the largest discrepancy, with an underestimation of 62 percent. Differences in data collection methods explain 22 percent of the variance at the condition level $((.023 - .018)/.023)$.

Adding data collection methods to the model does not reduce the residual variance at the study or the condition-within-studies level enough to make it nonsignificant (σ^2_{study}: $\chi^2_1 = 6.62$, $p = .01$; $\sigma^2_{\text{condition}}$: $\chi^2_1 = 82.44$, $p = .000$), and thus the third model, which also includes the effect of topic sensitivity (M2), has been tested. Sensitivity makes a significant contribution to the effect size ($b = .12$, $Z = 2.50$, $p = .006$, Wald test). Differences in the validity of the results of the data collection methods and the topic sensitivity jointly explain 43 percent of the variance at the condition level. The effect of the topic sensitivity can be interpreted as follows: If the sensitivity of a topic increases by 1 point, the effect size (the discrepancy between what respondents report and their true score) increases by 12 percent. There is also a large shift in the distribution of the regression coefficients of the data collection methods. If we control for topic sensitivity, the difference between the estimated and known prevalence becomes nonsignificant for RRT and telephone interviewing. Again the RRT condition results in the smallest difference between the observed outcome and population value, with a mean difference of 4 percent across studies. CASI still has the largest difference between the true score and estimated prevalence (26 percent). The underestimations of the other designs vary between 9 and 15 percent if we control for sensitivity.

Adding additional explanatory variables such as indices for the quality of the study, topic sensitivity, or the characteristics of the randomized response technique (randomizer, *p*-true) do not improve the model significantly, which is why these extended models are not included in Table 3.

RESULTS OF THE COMPARATIVE STUDIES

For the comparative studies, the effect variable is the d'_{probit} for the comparison of the outcome of the RRT and one of the other data collection techniques. Positive outcomes indicate that in a specific

TABLE 4: Results of the Multilevel Analysis for Comparative Randomized Response Studies

Step	*n*	*M0 (Intercept Only)*	*M1 (Conditions Added)*	*M2 (Sensitivity Added)*
Intercept		.28 (.077)**		
Telephone interview	3		.23 (.455)	.03 (.449)
Questionnaire	13		.24 (.136)*	.05 (.144)
Face-to-face	23		.39 (.106)**	.21 (.119)**
Scenario	1		−.13 (.224)	−.31 (.230)*
Unmatched count	2		−.08 (.170)	−.24 (.177)
Social sensitivity				.07 (.023)**
σ^2 study		.072 (.029)**	.17 (.05)**	.15 (.049)**
σ^2 condition in study		.031 (.009)**	.02 (.01)*	.03 (.008)*

NOTE: n = number of data collection conditions; standard error in parentheses. The five dummies represent the differences between the randomized response technique (RRT) and the more conventional data collection techniques; for instance, *telephone interview* is the dummy for the difference between the RRT and the telephone interview results.
*$p \leq .05$. **$p \leq .01$.

comparison, the RRT yields a more valid estimate, and negative outcomes indicate a less valid estimate for the RRT condition.

First, we test the data for homogeneity across studies (model M0). The significant intercept of .28 indicates an overall positive effect of randomized response methods as compared to non-RRT approaches (intercept = .28, $Z = 3.68$, $p < .001$). The fact that the residual error variances are significant at both levels (Table 4, M0: [$\chi_1^2 = 69.72$, $p < .001$] for σ^2_{study} and [$\chi_1^2 = 54.12$, $p < .001$] for $\sigma^2_{\text{condition}}$) enables us to conclude that the differences in results across the studies or conditions within the studies cannot be attributed to mere sampling error. There are systematic differences between the results of various data collection methods, and the magnitude of the differences varies across studies.

To test for differences in the effects of RRTs compared to various data collection methods, we created a full set of five dummy variables. The dummies are RRT compared to face-to-face interviews, RRT compared to self-administered questionnaires, RRT compared to telephone interviews, RRT compared to unmatched count techniques, and RRT compared to scenario methods. The dummies have been added to the regression equation (Table 4, model M1). Randomized response techniques produce significantly better population estimates

TABLE 5: Relation Between Topic Sensitivity and Effectiveness of Randomized Response Techniques as Compared to Conventional Collection Methods (*d*-Probit)

Sensitivity Rating	*Mean d-Probit Across Topics and Methods*
0	0.0062
1	0.1980
2	0.3254
3	0.3063
4	0.4037

than face-to-face interviews ($b = .39, Z = 3.67, p < .001$) or self-administered questionnaires ($b = .24, Z = 1.76, p = .04$). A comparison between the RRT and the other three data collection methods does not result in significant differences. The application of the scenario method and the two applications of the unmatched count technique do better than the randomized response techniques, but the differences are not significant (scenario: $b = -.14, Z = -.08, p = .53$; UMT: $b = -.08, Z = 1.36, p = .68$). There is, however, still a significant residual variance at the study level (σ^2_{study}: $\chi^2_1 = 33.26$, $p < .001$) and the condition level ($\sigma^2_{\text{condition}}$: $\chi^2_1 = 77.73$, $p < .001$), so differences in data collection methods do not satisfactorily explain the variability of the outcomes.

In the third step (Table 4, M2), topic sensitivity is added to the equation. The influence of sensitivity on the effect sizes is positive and significant ($b = .071, Z = 3.09, p = .002$). This means that across all the comparative studies, the greater the sensitivity of the research topic, the more valid the results yielded by the RRT. Table 5 demonstrates this result at a more basic level by presenting the relationship between sensitivity and the mean *d*-probit across the studies and methods. It is clear that *d*-probit increases with increasing sensitivity, which means the difference between the RRT results and the results from other conditions increases with increasing sensitivity. The very small positive *d*-probit for nonsensitive questions (*d*-probit = .0062, social sensitivity is 0) is the result of extrapolation since all the topics in the analysis are, by definition, sensitive.

Model M2 shows significant unexplained residual variances at the study level and the condition-within-studies level (Table 4, M2: [σ^2_{study}:

$\chi_1^2 = 71.00$, $p < .001$] and [$\sigma^2_{\text{condition}}$:$\chi_1^2 = 31.35$, $p < .001$]). Adding data quality, topic content, and characteristics of the RRT procedures (randomizer, p-true) to the regression equation does not, however, improve the model significantly, which is why these models are not included in Table 4.

CONCLUSIONS AND DISCUSSION

The results of the individual validation studies show that all the data collection methods present some degree of discrepancy between the observed characteristic and the known population value, with the RRT exhibiting the smallest discrepancy. We thus conclude that RRT results are more valid than the results of conventional data collection methods, although there is room for improvement. The results of the comparative studies partly corroborate those of the individual validation studies. RRTs produce significantly more valid population estimates than the conventional data collection methods of face-to-face interviewing, self-administered questionnaires, and, be it not significantly**[PLS. CLARIFY]**, telephone interviewing. This answers our first research question: The RRT does yield more valid results than the more conventional data collection methods (i.e., face-to-face interviews, self-administered questionnaires, telephone interviews, and some forms of computer-assisted self-interviews).

A significant proportion of variance is left at the study level and the condition-within-studies level. We make an effort to explain this residual variance by adding topic sensitivity to the model. As regards the comparative studies, there are indications that as the topic sensitivity increases, so does the positive effect of using RRT (Table 5). The difference between the outcomes of RRT conditions and the conventional data collection methods increases, which means that the results of RRT studies are more valid. Even if a topic is very sensitive, the RRTs still yields relatively valid estimators. If the sensitivity of a research topic is extremely great, the advantage of using an RRT may well outweigh its disadvantages (i.e., the need for larger samples and increased cognitive load for respondents).

Are the results of comparative validation studies comparable to those of individual validation studies? It follows from the outcome

of the analysis that the two research designs result in comparable outcomes if RRTs are compared to face-to-face or telephone interviews and self-administered questionnaires. Although individual validation studies are doubtlessly the gold standard from a methodological point of view, it is obvious that they are more difficult to carry out since access to sensitive information at an individual level is needed. Due to reasons of privacy, it can be very difficult to obtain this kind of information. This sometimes makes it necessary to use a comparative instead of a validation design. We can conclude from our results that although this design is not as strong as the validation design, it nevertheless appears to result in comparable outcomes and thus in informative results.

There are two results of this study that we would like to address more thoroughly in the remainder of this section: (1) the fact that we still have a large amount of unexplained error variance at the study level and the condition-within-studies level and (2) the dearth of studies that compare CAI to RRT.

EXPLAINING RESIDUAL VARIANCE

The meta-analysis literature points out the potential effect of data quality, but adding data quality to the model does not significantly lower the residual variance at the study level. This can be the result of the small sample of studies, especially in the individual validation studies. As a result of this small sample of studies, explanatory variables such as the overall methodological quality have very little power at the study level. Adding explanatory variables at the condition level, such as the type of RRT or its features, does not significantly lower the residual variance at the condition level. Part of the problem may be our inability to code for some known influences in our meta-analysis due to insufficient detail in the original publications. For instance, the respondents' understanding of the RRT is known to enhance trust and cooperation (Landsheer, van der Heijden, and van Gils 1999), but it is seldom described in research reports. We conclude that it is impossible to explain all the residual variance with the coded variables. Translated into the issue of data collection on sensitive topics, we interpret this as an indication that RRTs are not yet under adequate researcher control. Responses to sensitive questions are known to be susceptible to small variations in the actual data collection process (Jobe et al.

1997; Rasinski 1997). The interview situation in an RRT is complex, and not much is known about the cognitions and motivations that it triggers in respondents. This means that the researcher needs more intensive control over the quality aspects of the actual data collection process when using RRT (Boeije and Lensvelt-Mulders 2002).

THE RELATION BETWEEN RRT AND CAI

In individual validation studies, computer-assisted methods do not work as well as they are often reputed to (cf. Richman et al. 1999). However, this conclusion is the result of only one study, so there is a problem with the external validity. Although CAI is in the mainstream of survey research, we have found only one study that directly compares the RRT to CAI with direct questions at the level of individual validation. One reason for this omission can be that RRTs themselves can be implemented in face-to-face interviews, paper-and-pencil questionnaires, or in a CAI environment. In the RRT conditions, the mode of response is not always clear, and the number of cases that explicitly report that the RRT was used in a CAI environment is too small to make them a special group. This is why no distinction is drawn between the various response modes in this study.

Nowadays, Audio-CASI, a special form of CAI, is generating a great deal of research in the area of sensitive topics and is viewed as a very promising way to study sensitive topics in the future. If Audio-CASI is used, the respondent does not read the questions from the screen but listens to them as they are presented via a headphone. The results of studies on the quality of Audio-CASI data are extremely promising (Lessler and O'Reilly 2001). One recent comparative study compares Audio-CASI to RRT (Lara et al. 2004). The results of this study confirm our findings. Using RRT resulted in the highest estimates of the sensitive behavior (in this study, "induced abortion") as compared to the SAQ, face-to-face interviewing, and Audio-CASI. But again, this is the result of only one study; therefore, more studies comparing RRT, CASI with direct questions, and Audio-CASI are called for to answer this question.

The results of this meta-analysis across 35 years of RRT validation studies demonstrate the superiority of RRTs to data collection methods using direct questions in face-to-face interviews, self-administered questionnaires, telephone interviews, and some

forms of computer-assisted interviews. We can thus conclude that using randomized response questions in surveys on sensitive topics significantly improves the data quality of the surveys. Since using RRTs also increases the cost of data collection, does the increase in data quality justify the extra cost? The results of the meta-analysis show that with increasing topic sensitivity, the benefits of using RRTs also increase (Table 5). We conclude that with increasing topic sensitivity, the advantage of RRTs can counterbalance their costs. Currently available research has not demonstrated the superiority of any data collection method to RRT. This makes the RRT a viable method for the assessment of sensitive topics.

NOTES

1. The meta-analysis files can be found on our homepage (www.fss.uu.nl/ms/RRT). They contain the following variables: name of the author, year of the study, conditions within the study, randomized response technique (RRT), topic, items, study quality, randomizer, p-true, N per condition, population estimates from the study, and the d-probit and its standard error.

2. Although reversing a code is always arbitrary, we feel that in this case, it can be done, supported by the 100 percent concordance between our own ratings on boasting and the Himmelfarb and Lickteig (1982) data. The reversals are made for the topics *charity* (visit elderly people, do volunteer work, donate blood, and collect money for a good cause), *the environment* (walking instead of taking a car to help preserve the environment), and *opinions* (attitude to halfway houses for criminals and protected living for disabled people).

3. The Wald test assumes normality, so for variances, the likelihood ratio test is preferred (Raudenbush and Bryk 2002; Hox 2002). Since the null hypothesis is on the boundary of the parameter space (variances cannot be negative), the usual p value is divided by 2 (Hox 2002).

REFERENCES

Armacost, Robert L., Jamshid C. Hosseini, Sara A. Morris, and Kathleen A. Rehbein. 1991. "An Empirical Comparison of Direct Questioning, Scenario and Randomized Response Methods for Obtaining Sensitive Business." *Decision Science* 22:1073-87.

Baker, Robert P. and N. M. Bradburn. 1992. "CAPI: Impacts on Data Quality and Survey Costs." *Presented at the Public Health Conference on Records and Statistics,* **MONTH, CITY?**.

Barth, Jeremy T. and Howard M. Sandler. 1976. "Evaluation of the Randomized Response Technique in a Drinking Survey." *Journal of Studies in Alcoholism* 37:690-3.

Begin, Guy and Michel Boivin. 1980. "Comparison of Data Gathered on Sensitive Questions Via Direct Questioning, Randomized Response Technique and a Projective Method." *Psychological Reports* 47:743-50.

Begin, Guy, Michel Boivin, and J. Bellerose. 1979. "Sensitive Data Collection Through the Randomized Response Technique: Some Improvements." *Journal of Psychology* 101:53-65.

Beldt, Sandra F., Wayne W. Daniel, and Bikramjit S. Garcha. 1982. "The Takahasi-Sakasegawa Randomized Response Technique: A Field Test." *Sociological Methods & Research* 11:101-11.

Boeije, Hennie and Gerty J. L. M. Lensvelt-Mulders. 2002. "Honest by Chance: A Qualitative Interview Study to Clarify Respondents' (Non-)Compliance With Computer-Assisted Randomized Response." *Bulletin de Methodologie Sociologique* 75:24-39.

Brewer, K. R. W. 1981. "Estimating Marihuana Usage Using Randomized Response: Some Paradoxical Findings." *Australian Journal of Statistics* 23:139-48.

Buchman, Thomas A. and John A. Tracy. 1982. "Obtaining Responses to Sensitive Questions: Conventional Questionnaire Versus Randomized Response Technique." *Journal of Accounting Research* 20:263-71.

Chaudhuri, Arijit and Rahul Mukerjee. 1988. *Randomized Response: Theory and Techniques*. New York: Marcel Dekker.

Chi, I-Cheng, L. P. Chow, and Rowland V. Rider. 1972. "The Randomized Response Techniques as Used in the Taiwan Outcome and Pregnancy Study." *Studies in Family Planning* 3:265-9.

Cohen, Jacob. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20:37-46.

Cooper, H. and L. V. Hedges. 1994. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Dalton, Dan R., James C. Wimbush, and Catherine M. Daily. 1994. "Using the Unmatched Count Technique (UCT) to Estimate Base Rates for Sensitive Behavior." *Personnel Psychology* 47:817-28.

Danermark, Berth and Bengt Swensson. 1978. "Measuring Drug Use Among Swedish Adolescents." *Journal of Official Statistics* 3:439-48.

de Leeuw, Edith D., Joop J. Hox, and Ger Snijkers. 1995. "The Effect of Computer-Assisted Interviewing on Data Quality: A Review." *Journal of the Market Research Society* 37:325-44.

Duffy, John C. and Jennifer J. Waterton. 1988. "Randomized Response Versus Direct Questioning: Estimating the Prevalence of Alcohol-Related Problems in a Field Survey." *Australian Journal of Statistics* 30:1-14.

Edgell, Stephen E., Karen L. Duchan, and Samuel Himmelfarb. 1992. "An Empirical Test of the Unrelated Question Randomized Response Technique." *Bulletin of the Psychonomic Society* 30:153-6.

Edgell, Stephen E., Samuel Himmelfarb, and Karen L. Duchan. 1982. "Validity of Forced Responses in a Randomized Response Model." *Sociological Methods & Research* 11:89-100.

Fidler, Dorothy S. and Richard E. Kleinknecht. 1977. "Randomized Response Versus Direct Questioning: Two Data Collection Methods for Sensitive Information." *Psychological Bulletin* 84:1045-9.

Fisher, Martin, Linda B. Kupferman, and Martin Lesser. 1992. "Substance Use in a School-Based Clinic Population: Use of the Randomized Response Technique to Estimate Prevalence." *Journal of Adolescent Health* 13:281-5.

Fox, James A. and Paul E. Tracy. 1986. *Randomized Response: A Method for Sensitive Surveys*. Beverly Hills, CA: Sage.

Goldstein, Harvey, Jon Rasbash, Ian Plewis, David Draper, William Browne, Min Yang, Geoff Woodhouse, and Michael Healy. 1998. *A User's Guide to MlwiN*. London: Multilevel Models Project, Institute of Education, University of London.

Goodstadt, Michael S. and Valerie Gruson. 1975. "The Randomized Response Technique: A Test on Drug Use." *Journal of the American Statistical Association* 70:814-8.

Halvorsen, Katherine T. 1994. "The Reporting Format." Pp. 425-38 in *The Handbook of Research Synthesis*, edited by H. Cooper and L. V. Hedges. New York: Russell Sage Foundation.

Hedges, L. V. and I. Olkin. 1985. *Statistical Methods for Meta-Analysis*. New York: Academic Press.

Himmelfarb, Samuel and Carl Lickteig. 1982. "Social Desirability and Randomized Response Technique." *Journal of Personality and Social Psychology* 43:710-7.

Horvitz, D. G., B. V. Shah, and Walt R. Simmons. 1967. "The Unrelated Question Randomized Response Model." Pp. 65-72 in *Proceedings in the Social Statistics Section, American Statistical Association*. Baltimore: American Statistical Association.

Hox, Joop J. 2002. *Multilevel Analysis*. Mahwah, NJ: Lawrence Erlbaum.

Hox, Joop J. and Edith D. de Leeuw. 2003. "Multilevel Models for Meta-Analysis." Pp. 90-111 in *Multilevel Modeling: Methodological Advances, Issues and Applications*, edited by N. Duan and S. Reise. Mahwah, NJ: Lawrence Erlbaum.

Jobe, Jared B., William F. Pratt, Roger Tourangeau, Allison K. Baldwin, and Kenneth A. Rasinski. 1997. "The Effects of Interview Mode on Sensitive Questions in a Fertility Survey." Pp. 311-30 in *Survey Measurement and Process Quality*, edited by L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwartz, and D. Trewin. New York: John Wiley.

Kalaian, Hripsine A. and Stephen W. Raudenbush. 1996. "A Multivariate Mixed Linear Model for Meta-Analysis." *Psychological Methods* 1:227-35.

Kerkvliet, Jan. 1994. "Cheating by Economics Students: A Comparison of Survey Results." *Journal of Economic Education* 25:121-33.

Krotki, K. J. and B. Fox. 1974. "The Randomized Response Technique, the Interview and the Self-administered Questionnaire: An Empirical Comparison on Fertility Reports." pp. 367-71 in *Proceedings of the Social Statistics Section, American Statistical Association*. Baltimore: American Statistical Association.

Kuk, Anthony Y. C. 1990. "Asking Sensitive Questions Indirectly." *Biometrika* 77:436-8.

Kulka, Richard A., Michael F. Weeks, and Robert E. Folsom. 1981. "A Comparison of the Randomized Response Approach and Direct Questioning Approach to Asking Sensitive Survey Questions." Working paper, Research Triangle Institute, NC.

Lamb, Charles W. and Donald E. Stem. 1978. "An Empirical Validation of the Randomized Response Technique." *Journal of Marketing Research* 15:616-21.

Landsheer, Hans A., Peter van der Heijden, and Ger van Gils. 1999. "Trust and Understanding, Two Psychological Aspects of Randomized Response." *Quality and Quantity* 33:1-12.

Lara, Diana, Jennifer Strickler, Claudia D. Olavarrieta, and Charlotte Ellertson. 2004. "Measuring Induced Abortion in Mexico: A Comparison of Four Methodologies." *Sociological Methods & Research* 32:529-58.

Lee, Raymond M. 1993. *Doing Research on Sensitive Topics*. London: Sage.

Lensvelt-Mulders, Gerty, Joop Hox, and Peter van der Heijden. Forthcoming. "How to Improve the Efficiency of Randomized Response Designs." *Quality and Quantity*.

Lessler, J. and J. M. O'Reilly. 2001. "Mode of Interview and Reporting of Sensitive Issues: Design and Implementation of Audio Computer-Assisted Self Interviewing." In *The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates*, edited by L. Larrison and A. Hughes. Rockville, MD: National Institute of Drug Abuse.

Lipsey, Mark W. and David B. Wilson 2001. *Practical Meta-Analysis*. Thousand Oaks, CA: Sage.

Locander, William, Seymour Sudman, and Norman Bradburn. 1976. "An Investigation of Interview Method, Threat and Response Distortion." *Journal of American Statistics* 71:269-75.

Maddala, G. S. 1983. *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge, UK: Cambridge University Press.

Miller, J. D. 1984. *A New Survey Technique for Studying Deviant Behavior*. Washington, DC: George Washington University.

Nathan, Gad. 1988. "A Bibliography of Randomized Response: 1965-1987." *Survey Methodology* 14:331-46.

Nordlund, Sturla, Ingar Holme, and Steinar Tamsfoss. 1994. "Randomized Response Estimates for the Purchase of Smuggled Liquor in Norway." *Addiction* 4:401-5.

Nunnally, Jum. 1967. *Psychometric Theory*. New York: McGraw-Hill.

Orwin, Robert G. 1994. "Evaluating Coding Decisions." Pp. 139-62 in *The Handbook of Research Synthesis*, edited by H. Cooper and L. V. Hedges. New York: Russell Sage Foundation.

Prinsen, H. M. and Ron A. Visser. 2000. Eindrapport en advies verbetering naleving taxiregels [Final Report and Recommendation on Improving Adherence to Taxi Rules]. The Hague, the Netherlands: Department of Justice, Expertise Centre for Law Enforcement (in Dutch).

Rasinski, Kenneth A. 1997. "The Effects of Interview Mode on Sensitive Questions in a Fertility Survey." Pp. 311-30 in *Survey Measurement and Process Quality*, edited by L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin. New York: John Wiley.

Rasinski, Kenneth A., G. B. Willis, A. K. Baldwin, W. Yeh, and L. Lee. 1999. "Methods of Data Collection, Perceptions of Risks and Losses and Motivation to Give Truthful Answers to Sensitive Survey Questions." *Applied Cognitive Psychology* 13:465-84.

Raudenbush, Stephen W. 1994. "Random Effects Models." Pp. 301-22 in *The Handbook of Research Synthesis*, edited by H. Cooper and L. V. Hedges. New York: Russell Sage Foundation.

Raudenbush, Stephen W. and Anthony S. Bryk. 2002. *Hierarchical Linear Models*. Thousand Oaks, CA: Sage.

Richman, Wendy L., Sara Kiesler, Suzanne Weisband, and Fritz Drasgow. 1999. "A Meta-Analytic Study of Social Desirability Distortion in Computer-Administered Questionnaires, Traditional Questionnaires and Interviews." *Journal of Applied Psychology* 84:754-75.

Rosenthal, Robert. 1994. "Parametric Measures of Effect Size." Pp. 231-44 in *The Handbook of Research Synthesis*, edited by H. Cooper and L. V. Hedges. New York: Russell Sage Foundation.

Scheers, N. J. and C. Mitchell Dayton. 1987. "Improved Estimation of Academic Cheating Behaviour Using the Randomized Response Technique." *Research in Higher Education* 26:61-9.

———. 1988. "Covariate Randomized Response Models." *Journal of the American Statistical Association* 83:969-74.

Shotland, Lance R. and Lynn David Yankovski. 1982. "The Randomized Response Method: A Valid and Ethical Indicator of the Truth in Reactive Situations." *Personality and Social Psychology Bulletin* 8:174-9.

Smith, Linda L., Walter T. Federer, and Damaraju Raghavarao. 1974. "A Comparison of Three Techniques for Eliciting Truthful Answers to Sensitive Questions." pp. 447-52 in *Proceedings of the Social Statistics Section, American Statistical Association*. Baltimore: American Statistical Association.

Soeken, Karen L. and Shirley P. Damrosch. 1986. "Randomized Response Technique: Applications to Research on Rape." *Psychology of Women Quarterly* 10:119-26.

Stem, Donald E. and Kirk R. Steinhorst. 1984. "Telephone Interview and Mail Questionnaire Applications of the Randomized Response Model." *Journal of the American Statistical Association* 79:555-64.

Supple, A. J., W. S. Aquilino, and D. L. Wright. 1999. "Collecting Sensitive Self-Report Data With Laptop Computers: Impact on the Response Tendencies of Adolescents in a Home Interview." *Journal of Research on Adolescence* 9:467-88.

Tezcan, Sabahat and Abdel R. Omran. 1981. "Prevalence and Reporting of Induced Abortion in Turkey: Two Survey Techniques." *Studies in Family Planning* 12:262-71.

Tourangeau, Robert and Kenneth A. Rasinski. 1988. "Cognitive Processes Underlying Context Effects in Attitude Measurement." *Psychological Bulletin* 103:299-314.

Tracy, Paul E. and James A. Fox. 1980-1981. "The Validity of Randomized Response for Sensitive Measurements." *American Sociological Review* 46:187-200.

Umesh, U. N. and Robert A. Peterson. 1991. "A Critical Evaluation of the Randomized Response Method: Applications, Validation and Research Agenda." *Sociological Methods & Research* 20:104-38.

van der Heijden, Peter G. M. and Ger van Gils. 1996. "Some Logistic Regression Models for Randomized Response Data." Pp. 341-8 in *Statistical Modelling: Proceedings of the 11th International Workshop on Statistical Modelling*, edited by A. Forcina, G. M. Marchetti, R. Hatzinger, and G. Galmatti. Orvieto, Italy: **PUBLISHER?**.

van der Heijden, Peter G. M., Ger van Gils, Jan Bouts, and Joop J. Hox. 1998. "A Comparison of Randomized Response, CASAQ and Direct Questioning: Eliciting Sensitive Information in the Context of Social Security Fraud." *Kwantitatieve Methoden* 19:15-34.

———. 2000. "A Comparison of Randomized Response, CASI and Face-to-Face Direct Questioning: Eliciting Sensitive Information in the Context of Welfare and Unemployment Benefit." *Sociological Methods & Research* 28:505-37.

Volicer, Beverly J., Mary H. Cahill, Evelyn Neuburger, and Gretchen Arntz. 1983. "Randomized Response Estimates of Problem Use of Alcohol Among Employed Females." *Alcoholism: Clinical and Experimental Research* 7:321-6.

Volicer, Beverly J. and L. Volicer. 1982. "Randomized Response Technique for Estimating Alcohol Use and Non-Compliance in Hypertensives." *Journal of Studies on Alcohol* 43:739-50.

Warner, Stanley L. 1965. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association* 60:63-9.

Weisband, Suzanne and Sara Kiesler. 1996. *Self-Disclosure on Computer Forms: Meta-Analysis and Implications*. Tucson: University of Arizona.

Weissman, Arlene N., Robert A. Steer, and Douglas S. Lipton. 1986. "Estimating Illicit Drug Use Through Telephone Interviews and the Randomized Response Technique." *Drug and Alcohol Dependency* 18:225-33.

Williams, Bryan L. and Hoi Suen. 1994. "A Methodological Comparison of Survey Techniques in Obtaining Self-Reports of Condom-Related Behaviors." *Psychological Reports* 7:1531-7.

Wimbush, Dan C. and Donald R. Dalton. 1997. "Base Rate for Employee Theft: Convergence of Multiple Methods." *Journal of Applied Psychology* 82:756-63.

Wyatt, Gail E., Lawrence A. Vodounon, and M. R. Mickey. 1992. "The Wyatt Sex History Questionnaire: A Structured Interview for Female Sexual History Taking." *Journal of Child Sexual Abuse* 1:51-68.

Zdep, S. M. and Isabelle N. Rhodes. 1976. "Making the Randomized Response Technique Work." *Public Opinion Quarterly* 40:531-7.
Zdep, S. M., Isabelle N. Rhodes, R. M. Schwarz, and Mary J. Kilkenny. 1979. "The Validity of the Randomized Response Technique." *Public Opinion Quarterly* 43:544-9.

Gerty J. L. M. Lensvelt-Mulders is an assistant professor of methods and statistics in the Faculty of Social Sciences at Utrecht University, the Netherlands. Her research focuses on research methods for sensitive topics in surveys as well as experimental settings and nonresponse problems in surveys.

Joop J. Hox is a professor of social science methodology in the Faculty of Social Sciences at Utrecht University, the Netherlands. His main research interests are survey methodology, data quality, and the analysis of complex data with multilevel models and structural equation modeling. Recent publications are on interviewer effects on non-response and meta-analysis, and he is the author of a handbook on multilevel analysis.

Peter G. M. van der Heijden is a professor of statistics in the Faculty of Social Sciences at Utrecht University, the Netherlands. He is interested in the estimation of population sizes in cases of sensitive topics and categorical data analysis.

Cora J. M. Maas is an assistant professor of methods and statistics in the Faculty of Social Sciences at Utrecht University, the Netherlands. She is an expert on multilevel analysis and has published on sample size and robustness issues in multilevel analysis.