Meta-analysis of Randomized Response Research: 
35 Years of Validation

Gerty J. L. M. Lensvelt-Mulders
Department of Methodology and Statistics, Utrecht University
P.O.B. 80140, NL-3508 TC, Utrecht, the Netherlands
Tel: +31 30 2535857 /fax: +31 30 2535797
private tel: +31 13 5431948

g.lensvelt@fss.uu.nl

Joop J. Hox
Department of Methodology and Statistics, Utrecht University
P.O.B. 80140, NL-3508 TC, Utrecht, the Netherlands
Tel: +31302539236 /fax: +31302535797
j.hox@fss.uu.nl

Peter, G. M. van der Heijden
Department of Methodology and Statistics, Utrecht University
P.O.B. 80140, NL-3508 TC, Utrecht, the Netherlands
Tel: +31302534688 /fax: +31302535797
p.vanderheijden@fss.uu.nl
Abstract

The Randomized Response Technique (RRT) is a survey method especially developed to improve the accuracy of answers to questions on threatening topics. This paper reports the results of two meta-analyses on RRT studies, the first on the results of six individual validation studies, and the second on 31 comparative studies. The focus of these meta-analyses was on the performance of RRTs compared to conventional question-answer methods. A multilevel analysis with three levels (studies, conditions, and sensitive items) was used to compare the effects of data collection methods across different studies. As effect size for the individual validation studies we used the percentage of incorrect answers and as effect size for the comparative studies we used the standardized difference. Results indicated an overall positive effect for RRT across studies compared to the other methods. This positive effect is stronger as the topic under investigation becomes more socially sensitive by nature. This makes it opportune to use RRT when sensitive topics are studied. However, both meta-analyses showed a large variance component between studies. We interpret this as an indication that RRT is not under complete control of the researchers using it, and advise careful pretesting of the chosen RRT procedure.
INTRODUCTION

Socially sensitive questions are thought to be threatening to respondents. (Lee 1993). This threat is characterized as extrinsic when certain responses carry the risk of sanctions. If the questions are about illegal behavior, these sanctions could be penal prosecution; if the questions are about deviant behavior, these sanctions could be the result of peer group pressure or social control. The threat is intrinsic when the questions concern subjects that are very personal or stressful to respondents, or when certain responses imply a negative adjustment in one’s self-image. Both types of threat may occur at the same time. For instance, one issue is the respondents’ distrust of the researcher’s confidentiality (Landsheer, van der Heijden, and van Gils 1999). This relates to both threats: it is embarrassing to tell an unfamiliar interviewer that one has engaged in unacceptable or deviant behavior, and it is even more difficult when one has doubts about the protection of one’s privacy. For these reasons respondents will be more prone to distort their answers, which may leave the researcher with incorrect data (Rasinski, Willis, Baldwin, Yeh and Lee 1999).

An explanation for incorrect answering as related to intrinsic threat is the effect of social desirability (Sudman and Bradburn 1982). Respondents wish to avoid social embarrassment and strive to project a positive self-image. As a result topics with a negative connotation, like (illegal) abortion, child abuse, rape, or use of hard drugs often encounter underreporting (Finkelhor and Lewis 1988; Soeken and Damrosch 1986). Overestimation is found for behaviors with a positive connotation, like pro-environmental behavior or wearing seatbelts (Himmelfarb and Lickteig 1982; Stem and Steinhorst 1984). Both over- and underreporting can be related to specific sub-populations. For marijuana use overreporting is found in high school populations and underreporting in adult populations (Brewer 1981; Fisher, Kupferman and Lesser 1992). The social sensitivity of a research topic clearly has an effect on answering but this effect is not always straightforward.
THE RANDOMIZED RESPONSE TECHNIQUE

To tackle the problem, Warner (1965) introduced the Randomized Response Technique (RRT). This is an interview method that guarantees privacy, and therefore has the potential to overcome the reluctance of subjects to reveal sensitive or probably harmful information about themselves (Chaudhuri and Mukerjee 1988; Fox and Tracy 1986). By entering an element of chance (for instance using cards or dice) in the question-response process, the privacy of a respondent is fully guaranteed. As a result, the respondent will be more inclined to answer honestly to a sensitive question. We describe Warner’s (1965) randomization technique here as an example of randomized response methods in general. Other RRTs and some of their properties are described in the Appendix.

WARNER’S RANDOMIZED RESPONSE METHOD

With the aid of a randomization device (i.e. colored marbles in a box, coins or dice), respondents are directed towards one out of two statements, for example:

A: I am pro capital punishment (A: selected with probability $p$)

B: I am against capital punishment (not-A: selected with probability $1-p$)

Without revealing to the interviewer which statement was selected by the dice, the respondent answers ‘true’ or ‘not true’ according to his attitude towards capital punishment. Elementary probability theory can be used to get a bias free estimate ($\hat{\pi}$) of the population probability of A (being in favor of capital punishment) ($\pi$) by:

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

(1)

where $p$ is the dice-controlled probability that the sensitive statement A has to be answered ($p \neq .5$) and $\hat{\lambda}$ is the observed sample proportion of yes-answers. The sampling variance of
\hat{\pi} is given by:

\[ \text{var}(\hat{\pi}) = \left[ \frac{\pi(1-\pi)}{n} \right] + \left[ \frac{p(1-p)}{n} \right] (2p-1)^2, \]

(2)

In equation 2, \( \left[ \frac{\pi(1-\pi)}{n} \right] \) is the standard formula for the sampling variance of a proportion, and \( \left[ \frac{p(1-p)}{n} \right] (2p-1)^2 \) represents the variance added by the randomized response technique. Equation 2 shows that this added variance increases when \( n \) is smaller and \( p \) (the dice-controlled probability of replying to statement A) is closer to 0.5.

PROFITS AND COSTS OF USING RANDOMIZED RESPONSE TECHNIQUES

Equation (2) clearly shows that Warner’s RRT is less efficient than standard data collection methods. It suffers from a larger sampling variance, leading to reduced power that makes it necessary to use larger samples. For example, when the true population estimate of some sensitive behavior is 0.2, and the probabilities of having to answer question A and B are 0.7 and 0.3 (for instance if there are 70 red beads and 30 blue beads in a box), Warner’s RRT method needs as many as 10 times the number of respondents to be just as powerful as the straightforward question-answer designs. After Warner’s initial papers, different varieties of the RRT have been developed that are more efficient, for an excellent overview see Umesh and Peterson (1991).

A second less recognized disadvantage of the RRT compared to direct questioning is the increased complexity of the question. Survey methodologists employ a standard question response model developed by Tourangeau and Rasinski (1988). Four steps characterize an optimal question answering process: (1) understanding the question being asked, (2) retrieving the relevant information from memory, (3) integrating this information into a summarized judgment, and (4) reporting this judgment correctly. Using RRT adds an extra step to this process because respondents also have to understand and follow the RRT instructions.
Clearly, this makes the cognitive load larger for RRTs than for more standard data collection methods. It also opens the doors for a new sources of error, namely misunderstanding of the RRT procedures, and /or cheating on this procedure (Boeije and Lensvelt-Mulders 2002). As a result of the extra cognitive load, valid reporting in an RRT environment depends heavily on respondent’s trust in and understanding of the method. This is illustrated in a study by Landsheer et al.(1999), which finds that respondents who are better equipped to understand the rationale behind the way the RRT protected their privacy also developed more trust.

The advantages of RRTs outweigh their extra costs only when the RRTs’ population estimates are much better than estimates from straightforward question-answer designs. Two sorts of studies have been carried out to address this question: individual validation studies and comparative studies. A study is defined as an individual validation study when we know the true status of all individuals on the sensitive issue. This makes it possible to compare the population estimates to the true mean, derived from the known true status of all individuals. Individual validation studies have a high internal validity, but they are rare, since they require information on the individual respondent level, i.e. access to police files or medical reports. We only found seven individual validation studies, of which six could be retrieved. The results of the individual validation studies, summarized in Table 1, vary highly.

A study is defined as a comparative study when RRT’s are compared to one or more standard data collection methods (self administered questionnaires, telephone interviews, face-to-face interviews and CAI), without the possibility of individual validation of the results against a known criterion. The results of these comparative studies are interpreted according to the ‘more is better’-adage, a higher reporting of the sensitive behavior is interpreted as an
indication of higher validity. In comparative studies the superiority of the RRT is as often confirmed as it is not (Umesh and Peterson 1991). For instance, on the question ‘Have you ever taken something from a shop that was worth over 50 dollar?’ Beldt, Daniel, and Garcha (1982) found no indication that the RRT provides better estimates than direct questioning, whereas Wimbush and Dalton (1997) found positive results for the same item question: in their case RRT performed better than both face to face interviewing and self administered questionnaires.

Looking at the literature on RRT we can see that 35 years of research have not led to consensus in the field, or to the description of best practices. Many statistical improvements have been made to enhance the efficiency and reliability of the method, and many different varieties of randomized response procedures have been developed. However, both individual validation studies and comparative studies show a large diversity in research outcomes. Therefore, we decided to use meta-analysis to analyze the outcomes of the individual validation and the comparative studies.

PURPOSE OF THE META-ANALYSIS

The meta-analysis is carried out to answer the following questions:

1. Do RRTs in general produce more valid population estimates for sensitive topics than conventional question–answer designs like face to face interviews, self-administered questionnaires, telephone interviewing and computer assisted interviewing (CASI)?

This is an important question because the results of the randomized response manipulation appear to differ highly across studies. Meta-analysis provides us with an integrated summary of the studies’ outcomes. In addition, it lets us decompose the between-study variance into a sampling variance component and a systematic component. Insight in the summarized results
across studies can help researchers to decide whether or not to use a randomized response technique when doing sensitive research. When the meta-analysis results in a large and statistically significant systematic variance, we come to our second question:

2. Why are the RRT results so variable across studies?

If the meta-analysis shows that the variance between studies is significantly larger than can be explained by sampling variance, the results are considered as heterogeneous, which means that variations between study results are real. This leaves us with the question why in some studies the randomized response technique produces better estimates than in other studies.

We included individual validation studies as well as comparative studies in the meta-analysis. Due to differences in the effect measures between individual validation studies and comparative studies, which will be discussed later, we analyze them separately in two different meta-analyses. Individual validation studies have a much stronger internal validity than comparative studies, but they are more expensive and harder to carry out. As a result, there are many more comparative studies in the literature than individual validation studies. Although these comparative studies are easier and cheaper to perform, they are informative only if their outcomes are as relevant as the outcomes of the individual validation studies. This brings us to the third research question:

3. Do comparative studies provide the same information as individual validation studies?

**LITERATURE RETRIEVAL AND CODING**

**COMPILING A BIBLIOGRAPHY**

Point of departure for our search for randomized response literature has been a bibliography on randomized response studies by Nathan (1988). This bibliography covers the period 1965-1987, and contains over 250 theses, research reports, published papers, and books. To add to
this bibliography and to extend it from 1987 to 2000, an on-line search was carried out in the following computer databases: PsychInfo, Sociofile, Eric, Medline, Sage Publications CD-ROM, the SRM-database of social research methods, the CIS Current index to statistics (8th release), and the SSCI (Social Sciences Citation Index). For the subject search, we used the expressions randomized response and sensitive questions. Search results were compared with the Nathan bibliography, further articles (1965-1987) were added, and the bibliography was extended with more recent literature. In addition, the reference list was supplemented with studies located from reference sections of studies located earlier, and a call for unpublished studies was sent to the SMRS-Internet mailing list (Survey Methods, Research and Statistics).

This search strategy produced a bibliography of randomized response studies performed between 1965-2000. In addition to the seven individual validation studies discussed earlier, we searched for comparative validation studies, using the search terms compare, comparative, evaluate, validation, direct questioning, telephone surveys, mail, and CASI. This strategy produced 70 studies. We have retrieved as many of these studies as possible, by library search, online full text contents (JSTOR) and by contacting authors and institutions. We were able to retrieve 66 studies. Almost half of the retrieved reports did not describe the kind of comparative experiment that we searched for, and were for that reason not relevant for our meta-analysis. At the end we were left with 37 relevant and codable studies.

INCLUSION CRITERIA

To include an individual validation study in the meta-analysis, the published report should at least provide sufficient information to derive an effect estimate of the difference between the RRT outcome and the known population value, together with its sampling variance (Lipsey and Wilson 2000; Hox and de Leeuw 2002). Not all reports contained sufficient information
meta-analysis of randomized response research

to calculate the effect size and sampling variance. This resulted in 6 individual validation studies that were included in the first meta-analysis, i.e. Horvitz, Shah and Simmons (1967), Kulka, Weeks and Folsom (1981), Lamb and Stem (1978), Locander, Sudman and Bradburn (1976), Tracy and Fox (1981), and van der Heijden, van Gils, Bouts, and Hox (1998, 2000).

In the second meta-analysis we included all studies that compared a randomized response condition with one or more standard data collection condition(s). Studies that are limited to a comparison of new randomized response results with results from earlier studies, or results obtained from literature reviews were not included. Also studies that used a within group design with respondents in more than one condition, were excluded.

The population estimates and their standard errors, or sufficient statistics to calculate these, must be given for all conditions in a study. If necessary we calculated the standard errors ourselves, according to the randomized response method that was used (Chaudurhi and Mukerjee 1988, see Appendix). For the non-randomized response data collection methods, standard errors were calculated using the standard formula \( (pq/n)^{1/2} \).

After screening all the retrieved papers 37 studies turned out to meet our inclusion criteria and were included: 6 individual validation studies and 31 comparative validation studies. The results of two studies were described in more than one publication, and both are coded as one study (van der Heijden et al. 1998-2000, Tracy and Fox 1980-1981).

CODING

For each study, the following variables are encoded:

A. To arrive at the dependent variables we coded the population estimates in a study together with their standard errors. With these estimates we computed the appropriate effect sizes and corresponding standard errors to be used as dependent variable in the meta-analysis.
B. We coded the different data collection methods to be used as conditions within a study.

C. We coded all different forms of the randomized response techniques that were used, to get insight in the differences between different approaches to randomized response.

D. To obtain more insight in the influence that minor features of the randomized response technique can have on respondents behavior, we coded which randomization device was used, the magnitude of the chance that a respondent had to answer the sensitive question (p-true), and the number of respondents in different conditions.

E. Because of the importance of the topic in sensitive research the content of the different issues that were used in the studies was also coded.

F. A measure of the social sensitivity of the topics was also coded, following Himmelfarb and Lickteig (1982). They instructed respondents that: ‘…some people may not answer questions truthfully and that the study was an attempt 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 thought that a typical respondent should answer the question truthfully or not’ (Himmelfarb and Lickteig 1982, p.715). To code the sensitivity of the topic we used the same introduction that was used by Himmelfarb and Lickteig. Because of the rather subjective judgment involved, we used four independent raters (coworkers of the Department of Social Sciences from Utrecht University, The Netherlands) to code the sensitivity of all research issues, including the issues used by Himmelfarb and Lickteig. They were asked to score all items on a social desirability scale, rating from 0: no inclination towards social desirable answering has to be expected to 4: the researcher can hardly expect an honest answer to this question. The mean correlation between the raters on the total number of 104 coded topics in this study is 0.73. The mean ratings across all four raters are coded as a measure for the social sensitivity of the topic. The aggregated sensitivity scores of our raters on the Himmelfarb-
issues are validated against the corresponding scores from the Himmelfarb and Lickteig -
study. The overall correlation is 0.65 (on 55 overlapping issues). In 2002 the raters are
somewhat more lenient on topics as ‘sexual behavior’ and ‘cheating on exam’, and in
1982 the raters were more lenient towards alcohol related questions. All raters (2002 and
1982) agreed completely on the expected direction of the answer distortion.

G. Differences in methodological rigor between studies may cause differences in results. To
to control for the effect of differences in the quality of studies we computed a measure for
‘study quality’. Four indicators of methodological rigor 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 students convenience sample (yes = 0, no = 1),
4. Where all coded variables retrievable from the publication (yes = 1, no = 0).
The sum of the scores on these four indicators was used as a measure for the quality of the
study.

COMPUTATION OF THE EFFECT SCORES

INDIVIDUAL VALIDATION STUDIES

The individual validation studies and the comparative studies are analyzed separately, because
their effect scores (coded outcomes) are not directly comparable.

In the individual validation studies, the deviation of the randomized response estimate
(\( \hat{\pi} \)) from 1 is used as effect score and is coded as the ‘percentage wrong answers’. All
estimates here are underestimations of the true population score of 100%, therefore the
individual validation studies deal only with false negatives. A smaller effect score is an
indication for a smaller deviation from the true score and thus for superior quality of the data collection method. An effect score of 0.85 can be read as a deviance from the true score of 85%, which can be interpreted as 5% better than an effect score of .80, compared to the true population score of 1. We included six studies in the first meta-analysis, the number of data collection methods within a study varied between 1 and 5, and a total of 34 sensitive questions could be coded. The mean effect score (percentage wrong answers) over 34 questions was .4122, with a standard deviation of .2620.

COMPARATIVE STUDIES
The interpretation of the results of the comparative validation studies is different from the individual validation studies, because different effect measures are used. Individual validation studies compared observed responses to a known true score, which results in a straightforward effect measure, the deviation of the true score. Comparative studies lack a known true score. Here, the effect of the randomized response condition is compared against the other data collection conditions, interpreting a higher incidence of sensitive characteristics as an indicator of obtaining more valid answers. The effect measure in the comparative studies is the standardized difference score for proportions. This is calculated by taking the cumulative standard normal value corresponding to the proportions $p_{rr}$ in the RRT group and $p_{control}$ in the control group, and calculating their difference:

$$d'_{probit} = Z_{rr} - Z_{control}$$

(3)

This standardized difference $d'_{probit}$ can be interpreted as the difference between the RRT proportion and the control data collection technique transformed to a probit scale. Rosenthal (1994) gives the corresponding sampling variance estimates as:

$$VAR(d'_{probit}) = \frac{2\pi p_{rr}(1 - p_{rr})e^{\sigma_{rr}^2}}{n_{rr}} + \frac{2\pi p_{control}(1 - p_{control})e^{\sigma_{control}^2}}{n_{control}}$$

(4)
Equation 4 does not take into account the extra variance added by the RRT condition, so it underestimates the sampling variance. Therefore we replace it by:

\[
VAR(d'_{probit}) = 2\pi(se_{n})^2 e^{\xi^2} + 2\pi(se_{control})^2 e^{\xi_{control}^2}
\]

(5)

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, Rhodes, Schwarz and Kilkenny 1979). If in a specific study an overestimation of the true population score is expected (boasting), a negative difference (estimate randomized response condition smaller than estimate control condition) was defined as a positive outcome\(^2\). As a consequence 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. We included 31 comparative studies in the second meta-analysis, the number of data collection methods included in a study varied between 1 and 4, and a total of 225 items could be coded. The mean effect score \((d'_{probit})\) over 225 sensitive items was .3775, with a standard deviation of .5033.

INTER-RATER RELIABILITY

Standard procedure in meta-analysis is to assess the reliability of the coding by calculating the intercoder reliability. Two raters coded a random sample of five publications. The inter-rater reliability for nominal variables was indicated using Cohen’s Kappa (Cohen 1960), and for scale variables as the coefficient alpha reliability (Nunnally 1967). In Table 2 the results of the inter-rater analysis are given. All Kappas and correlations were high.

Table 2 about here
DESIGN AND ANALYSIS METHOD

The analysis uses a multilevel approach to meta-analysis (Hox and de Leeuw 2002; Kalaian and Raudenbush 1996; Raudenbush 1994). We use a three-level weighted regression model, with studies at the highest level, data collection conditions within studies at the second level and effect scores per item within conditions at the lowest level. Three models are tested. The null or intercept only model has equation:

\[ Y_{ijk} = b_0 + v_{0k} + u_{0jk} + e_{ijk}, \]  

where \( Y_{ijk} \) is the effect score \( i \) of condition \( j \) in study \( k \). For individual validation studies \( Y_{ijk} \) is expressed as proportion wrong answer, and for the comparative studies \( Y_{ijk} \) is expressed as \( d'_{\text{probit}} \). The null-model estimates the effect score as the mean effect across all studies, conditions and items \( (b_{0jk}) \), plus residual error terms \( v \) on the study level \( (v_{\text{study}}) \), \( u \) on the ‘condition within studies’-level \( (u_{\text{condition}}) \), and \( e \) on the item level. The lowest level variance of \( e \), indicated by \( \sigma^2_e \), is the sampling error for each item known from the study publication and entered directly into the analysis as data input.

The variances of the residual errors across conditions within studies and across studies are given by \( \sigma^2_{\text{condition}} \) and \( \sigma^2_{\text{study}} \). These variances are tested for significance using a likelihood-based chi-square test. If \( \sigma^2_{\text{condition}} \) and/or \( \sigma^2_{\text{study}} \) is not equal to zero, the results are not homogeneous across conditions or studies. If this is the case, explanatory variables defined on the condition-level or the study level can be added into the model to explain the variance. This leads to the regression model

\[ Y_{ijk} = b_0 + b_1 X_{1jk} + \ldots + b_p X_{pkj} + b_{p+1} Z_{1jk} + \ldots + b_{p+q} Z_{qjk} + v_{0k} + u_{0jk} + e_{ijk}, \]  

where the \( X \)’s are condition-level explanatory variables (e.g., the type of RRT used), and the \( Z \) are study-level explanatory variables (e.g., the methodological quality of the study). The parameters are estimated by Restricted Maximum Likelihood (RIGLS) using MLwiN.
meta-analysis of randomized response research

(Goldstein, Rashbash, Plewis, Draper, Browne, Yang, Woodhouse and Healy 1998). The significance of the regression coefficients is determined by the Wald test at a significance level of 5%. For the variance components we report the standard errors, but their significance is assessed using a Likelihood-based chi-square test (Hox 2002; Raudenbush and Bryk 2002).

Homogeneity of the data was tested using M0, the null or intercept-only model. If the residual variances ($\sigma^2_{\text{condition}}$ and $\sigma^2_{\text{study}}$) are equal to zero, as determined by a non-significant Wald-test, all observed differences between effect scores are considered the result of sampling error. If a variance is not equal to zero, results are heterogeneous over conditions and/or studies. This means that we have systematic differences between conditions and/or studies. If the results are heterogeneous, the different data collection methods are added as explanatory variables at the condition level (M1), to obtain insight in the overall effects of the RRT versus non-RRT approaches. If after inclusion of the different data collection methods the residual error variances on the condition and/or study level are still significant, we add the social sensitivity of the research topic as explanatory variable at the condition level (M2). In applied research the social sensitivity of the topic is a given fact that contributes to the respondents’ willingness to answer a question. Differences in the social sensitivity of the topics can therefore be the cause of differential results across studies. A final model (M3) adds available covariates like cluster topic, data-quality, the chance one has to answer the sensitive question ($p$-true), and the randomization device (all as dummies) as explanatory variables at the study level. It turned out that none of these covariates contributes significantly to the model, therefore this model is not further elaborated.

RESULTS
GENERAL CHARACTERISTICS OF THE STUDIES

We start with an overview of the general characteristics of the studies. The 39 studies come from 20 different international journals, two working papers of the Research Triangle Institute (RTI, North Carolina) and 6 studies from unpublished literature databases. Two hundred sixty-four sensitive questions could be recorded. These questions are covering 10 socially sensitive topics: abortion (5.3%), sexual behavior (19.1%), drugs (9.5%), alcohol (5.3%), criminal offences (17.4%), ethical problems/attitudes (16.2%), charity (3.8%), cheating on exam (19.6%), environment (2.9%), and a diverse miscellaneous category (0.9%). Studies were conducted in the US (28), the Netherlands (6), Great Britain (2), Scandinavia (2), and Canada (1). The data collection methods used in these studies are Telephone interviewing (22), Self administered questionnaires (SAQ, 12), Computer assisted self administered interviews (CASI, 2), Scenario methods (1), the unmatched count technique (2), Face to face interviews (22), and the randomized response methods (all). The randomized response technique can be subdivided into the Forced Response-method (22), 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). As randomization device are used: Cards (3.4%), colored marbles (11.1%), dice (24.9%), coins (head or tails 36.4%), a banknote (2.7%), color lists (3.6%), a spinner (8.4%), a telephone book (8.9%), and social security numbers (0.6%).

Thirty-four percent of the research is based on population-based samples, 66% of the studies used convenience samples. The probability that one has to answer the sensitive question (p_true) in the randomized response conditions varied between 0.33 (Kerkvliet 1994) and 0.84 (Shotland, Lance and Yankovski 1982), with mean 0.67 and median .7.

RESULTS OF THE INDIVIDUAL VALIDATION STUDIES
For the individual validation studies, the effect score is the discrepancy between the known population probability of 1 and the observed value. All effect scores are underestimates of the true score. Small outcome values indicate a small discrepancy and hence a better quality of the data collection method. First, we test the null hypothesis that results are homogeneous over studies, i.e. $\sigma^2_{\text{study}}$ and $\sigma^2_{\text{condition}}$ are zero. 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\% between the known population value of 100\% and the observed proportions. The variances of the residual errors across conditions within studies and across studies are given by $\sigma^2_{\text{condition}}$ and $\sigma^2_{\text{study}}$. The chi-square test on the variances indicates that the effects are heterogeneous across both conditions ($\chi^2 = 246.69$, $p=.000$) and studies ($\chi^2 = 9.39$, $p=.001$). These results imply that differences in results across conditions and studies cannot be explained by sampling variation alone. There are systematic differences in outcomes at the study level as well as at the condition level, which, if possible, should be explained.

The first question in our meta-analysis is whether these differences can be explained by the different data collection methods, viz. RRT versus different non-RRT approaches. To test this hypothesis, a full set of five dummy variables is created for the variable ‘data collection method’. These dummies are added to the regression equation (Table 3, model M1). Since a full set of dummies is used, the intercept must be removed from this model. All explanatory variables have a significant effect, which means all produce a significant
underestimate of the known population value. Using randomized response methods produces the smallest standardized difference between the observed outcome and the known population score of 100%, with a mean underestimation of 38%, as can be derived from the regression coefficient of .38, across all randomized response conditions in all studies. The other data collection methods all deviate further from the true score, which indicates that randomized response techniques attain the most valid results. Using a self-administered questionnaire or a telephone interview to gather data produces approximately the same underestimation, 46-48% (b_{telephone} = .46 and b_{questionnaire} = .48). Face-to-face interviews show a mean underestimation 42% (b = .42) and Computer assisted self-interviewing (CASI) produced the largest discrepancy with an underestimation of 62% (b = .62). In M1, the study-level residual \( \sigma^2_{\text{study}} \) is significant (\( \chi^2 = 6.62, p=.01 \)), as well as the residual \( \sigma^2_{\text{condition}} \) (\( \chi^2 = 82.44, p=.000 \)). Adding data-collection methods explained 22% of the variance at the condition level ((.023-.018)/.023).

Model M2 adds ‘social sensitivity of the topic’ to the equation. This variable is significant (b = .12, Z=2.50, p=.006), and differences in data collection methods and social sensitivity together explained 33% of the variance of the ‘condition within studies’ level. The effect of ‘social sensitivity of the topic’ can be interpreted as follows: when the social sensitivity of a topic increases with one point, the effect score, the discrepancy between what respondents report and their true score, increases with 12%. The more sensitive a topic is, the larger the difference between the true score and the estimate. Furthermore, there is a large shift in the distribution of regression coefficients of the data collection methods. When we control for the social sensitivity of the research topic the effects of the data collection methods become smaller and for RRT and telephone interviews they become non-significant. Again the RRT condition resulted in the smallest difference between the observed outcome and the population value, with a mean difference of 4% across studies (b = .04). CASI still has the
largest difference between the true score and the estimate, 26% (b = .26). The underestimations of the other on-RRT approaches vary between 9% and 15%, when we control for social sensitivity.

Adding additional explanatory variables, like the indices for the quality of the study, the sensitive topic, or the characteristics of the randomized response technique (randomizer, p-true) employed, did not further improve the model significantly. For that reason this model (M3) is not included in Table 3.

RESULTS OF THE COMPARATIVE STUDIES

For the comparative studies, the effect variable is the $d'_{\text{probit}}$ for the comparison of the outcome of the RRT and one of the other data collection techniques. Positive outcomes indicate that in a specific comparison the RRT provided a more valid estimate, and negative outcomes indicate a less valid estimate for the RRT condition. From the fact that the residual error variances are significant at both levels (Table 4, M0: ($\chi^2 = 69.72$, $p < .001$) for $\sigma^2_{\text{study}}$ and ($\chi^2 = 54.12$, $p < .001$) for $\sigma^2_{\text{condition}}$) we conclude the effects are not homogeneous; differences in results across studies and data collection conditions cannot be attributed to sampling error. The significant intercept of .28 indicates an overall positive effect of randomized response methods as compared to non-RRT approaches ($b_{\text{intercept}} = .28$, $Z = 3.68$, $p < .001$).

--------------------------------
Table 4 about here
--------------------------------

To test for differences in the effects of RRT compared to different data collection
Methods a full set of five dummy variables was created. These 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 were added to the regression equation (Table 4, M1). Again, positive outcomes indicate that in a specific comparison the RRT provided a more valid estimate. Randomized response techniques produce significantly better population estimates than face-to-face interviewing ($b = .39, Z = 3.67, p<.001$) or self-administered questionnaires ($b = 0.24, Z = 1.76, p = 0.04$). Scenario-methods appear to produce better results than randomized response techniques, but not at a significant level ($b = -.14, Z = -.08, p = .53$). This is also the case for the unmatched count technique ($b = -.08, Z = 1.36, p = .68$). There is a significant residual variance on both levels (Table 4, M1: ($\chi^2 = 33.26, p<.001$) for $\sigma^2_{\text{study}}$ and ($\chi^2 = 77.73, p<.001$) for $\sigma^2_{\text{condition}}$).

In the third step, (Table 4, M2) ‘social sensitivity of the topic’ is added as an extra explanatory variable. The influence of social sensitivity on the effect scores is positive and significant ($b = .071, Z = 3.09, p = 0.002$); when the social sensitivity of a topic increases with one point, the effect score increases with 7%. This means that the advantage of randomized response methods becomes larger with increasing sensitivity of the topic. Table 5 demonstrates this result on a more basic level by presenting the relationship between social desirability and the mean $d_{\text{probit}}$ across studies and methods. As can be see there is a strong tendency of $d_{\text{probit}}$ to increase with increasing social desirability, which means that the difference between RRT results and the results from other conditions increases with increasing social sensitivity. The very small positive $d_{\text{probit}}$ for non-sensitive questions ($d_{\text{probit}} = .0062$, social sensitivity is 0) is the result of extrapolation, since all topics are sensitive.
Model M2 shows significant unexplained residual variances at the study level as well as at the data collection condition within studies level (Table 4, M2: \( \chi^2 = 71.00 \), \( p < .001 \)) for \( \sigma^2_{\text{study}} \) and (\( \chi^2 = 31.35 \), \( p < .001 \)) for \( \sigma^2_{\text{condition}} \). Adding data-quality, topic content, and characteristics of the randomized response procedures (randomizer, p-true) to the regression equation did not improve the model significantly. Therefore this model is not included in Table 4.

**CONCLUSIONS AND DISCUSSION**

The results of the individual validation studies show that all data collection methods present the researcher with some degree of discrepancy between the observed characteristic and the known population value, with RRT showing the smallest discrepancy. When a research topic becomes more sensitive, the results of the standard data collection methods become less valid, whereas the results of the RRT are more robust against fluctuations in social sensitivity of the topic. Even when a topic is very sensitive, RRTs still provide relatively valid estimators. This answers our first research question: RRT does provide more valid results than non-RRT approaches when sensitive topics are studied. When the social sensitivity of a research topic is very large, the advantage of using RRTs may outweigh the disadvantage of having to use a larger sample to obtain the same sampling variance. Adding characteristics like indices for data quality, topic content and specific RRT features did not significantly improve the model, which answers our second research question: there is no evidence that one specific RRT method is significantly superior or inferior to the other in its ability to obtain valid answers to
sensitive questions.

The results of the comparative studies corroborate the results of the individual validation studies. RRTs produce more valid population estimates than the question-answer methods face-to-face interviewing, self-administered questionnaires, and, although not significantly, telephone interviewing. When the social sensitivity of a topic increases the effect of using an RRT also increases, and compared to other data collection methods the results of randomized response studies become more valid.

This answers our third research question: the results of comparative validation studies are comparable to the results of the individual validation studies for the conditions face-to-face interviewing, telephone interviewing and self-administered questionnaires. Although individual validation studies are doubtlessly much stronger from a methodological point of view, it is obvious that they are more difficult to carry out, and using a comparative design appears to be a reasonable substitute.

Looking at the results of both meta-analyses, there are some important results that need clarification. In the individual validation studies, computer assisted methods do not work as well as they are often reputed to be (cf. Richman, Weisband, Kiesler and Drasgow 1999). In the comparative studies the ‘scenario method’-studies tend to provide the researcher with more valid estimates, although this result just fails to reach significance ($p=.08$). A scenario method involves the use of vignettes that describe a problem and then frame the questions to address the individual actions and the perception of other’s actions (Armacost, Hosseini, Morris, and Rehbein 1991, p. 1074). The problem with both results is that each originates from only a single study. This makes these results vulnerable to error, because there is no replication of these results. Both results call for further studies. Given our earlier conclusion that comparative studies yield results that corroborate the results of individual validation studies, such replications can be undertaken using a comparative design.
The unmatched count technique (Dalton, Wimbush and Daily 1994) provides better population estimates than RRTs in the comparative meta-analysis, although these results were also not significant. The unmatched count technique is often considered an extension of the classical randomized response procedures. It is also known as the item count paired list technique (Miller 1984), and the balanced incomplete block design (Raghavarao and Federer 1973). The method is compellingly simple. There are two lists with activities. One list contains only non significant activities, the other list contains one activity more, the sensitive activity of interest. Dalton Wimbush and Daily (1994) used this method to study the base rate for employer theft. They used the items (1) taking vitamin pills, (2) read book ‘The prince’, (3) have shotgun in my house, (4) have lived in two or more states, and (5) have lived outside the USA, in the short list. In the second condition the item (6) ‘I am involved in theft from my employer…’ was added. Respondents are randomly assigned to the short list or the longer list containing the sensitive activity. The respondents are asked to report the number of activities they have engaged in, but not which specific activities. The difference in mean between the two lists of activities is used as an estimate for the prevalence of employer theft. The advantage of the unmatched count technique is that it avoids the cognitive load and distrust associated with RRTs. The main disadvantage is that at present there are no techniques to relate covariates to the unobserved outcome of interest, unlike RRT, where techniques have been developed to relate the responses to explanatory variables (van der Heijden and van Gils 1996; Droitcour, Caspar, Hubbard and Ezzati 1991).

The results of this meta-analysis across 35 years of RRT validation studies show that using RRTs does produce better population estimates than using face-to-face interviews, self administered questionnaires, telephone interviews, and computer assisted methods. Since using RRTs increases the cost of data collection, the question is: Does the increase in data quality justify the extra costs? Our results show that with increasing social sensitivity of the
research topic the benefits of using RRTs increase (Table 5). We conclude that with increasing sensitivity of the topic the advantage of RRTs can counterbalance the costs.

A problematic factor in our results is the large amount of unexplained error variance between both conditions and studies. The meta-analysis literature points out the potential effect of data quality, but adding ‘data quality’ to the model did not significantly lower the residual variance at the study level. This can be the result of the small sample of studies included, especially in the individual validation studies. The small sample of studies results in a low power for explanatory variables at the study level, such as the overall methodological quality. Adding condition-level explanatory variables, such as features of the randomized response method, i.e. the probability that one has to answer the sensitive question (p-true), the content of the sensitive topic and the use of different randomizers, did not significantly lower the residual variance at the condition level. Part of the problem may be that we could not code for some known influences in our meta-analysis, due to insufficient detail in the original publications. For instance, respondents’ understanding of the randomized response procedure has been shown to enhance trust and co-operation (Landsheer et al. 1999), but it is seldom described in research reports. Given the modest impact of the variables we were able to code, we conclude that we can not explain all the study and condition variance. Translated to data collection on sensitive topics, we interpret this as an indication that the RRTs are not yet under adequate researcher control. It is known that responding to sensitive questions is sensitive to small variations in the actual data collection process (Jobe, Pratt, Tourangeau, Baldwin and Rasinski 1997). The interview situation in randomized response research is complex, and not much is known of the cognitions and motivations it triggers in respondents. This means that the researcher needs stronger control of the quality aspects of the actual data collection when using RRT. Much research in RRT has been aimed at improving statistical aspects of the method. The new challenge for researchers to take up is to improve the quality
of the RRT data collection process. We recommend for future research to establish a best practice procedure to study quality aspects of the data collection process in RRT. Our finding that comparative studies provide results that are basically similar to the methodologically stronger individual validation studies suggests that studies that aim to improve the RRT data collection process can use the comparative design, which is much easier to implement than an individual validation study.
NOTES

1. We will put our meta-analysis files on our homepage: www.fss.uu.nl/ms/gl. It will contain the variables: name author, year study, conditions within study, RRT, topic, items, study quality, randomizer, p-true, N per condition, population estimates from study, d-probit, and its standard error.

2. Although reversing a code is always arbitrary we felt that in this case it could be done, supported by the 100% concordance between our own ratings on boasting and the Himmelfarb and Lickteig (1982) data. The reversals are made for the topics Caritas (visit elderly people, doing voluntary work, blood donation and collect money for a good cause), environment (walking instead of taking car to save environment), and Opinions (attitude towards 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 two (Hox 2002).
References

References marked with an asterisk indicate studies included in the meta-analyses.


Table 1: Mean results of individual validation studies (1975-2000).

<table>
<thead>
<tr>
<th>Study</th>
<th>Method /Condition</th>
<th>Deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Lamb &amp; Stem 1978)</td>
<td>RRT</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Face to face</td>
<td>25</td>
</tr>
<tr>
<td>(Locander, Sudman and Bradburn 1976)</td>
<td>RRT</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Telephone</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Questionnaire</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>Face-to-face</td>
<td>32</td>
</tr>
<tr>
<td>(Van der Heijden, van Gils, Bouts and Hox 1998-2000)</td>
<td>RRT</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>CAI</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>Face-to-face</td>
<td>75</td>
</tr>
<tr>
<td>(Tracy and Fox 1980-1981)</td>
<td>RRT</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>Face-to-face</td>
<td>40</td>
</tr>
<tr>
<td>(Kulka, Weeks and Folsom 1981)</td>
<td>RRT</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>Face-to-face</td>
<td>16</td>
</tr>
</tbody>
</table>

Abbreviations: Deviance: Mean difference between the true population mean and the population estimate in percentages across different questions. CAI: Computer assisted interviewing.
Table 2: Inter-rater reliability

<table>
<thead>
<tr>
<th>Measure</th>
<th>Kappa (K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>1.00</td>
</tr>
<tr>
<td>Journal</td>
<td>1.00</td>
</tr>
<tr>
<td>Sample</td>
<td>.89</td>
</tr>
<tr>
<td>Format randomized response</td>
<td>1.00</td>
</tr>
<tr>
<td>Device</td>
<td>.92</td>
</tr>
<tr>
<td>Standard error RRT</td>
<td>1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scale measures</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of respondents in RRT-condition</td>
<td>1.00</td>
</tr>
<tr>
<td>Number of respondents in control-condition</td>
<td>1.00</td>
</tr>
<tr>
<td>Population estimate RRT-condition</td>
<td>.94</td>
</tr>
<tr>
<td>Population estimate control-condition</td>
<td>.95</td>
</tr>
</tbody>
</table>
Table 3: Results of the meta-analysis for individual validation studies

<table>
<thead>
<tr>
<th>Step</th>
<th>N</th>
<th>M0</th>
<th>M1</th>
<th>M2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(Intercept only)</td>
<td>(conditions added)</td>
<td>(social desirability added)</td>
</tr>
<tr>
<td>Intercept</td>
<td>.42</td>
<td>(.09)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RRT</td>
<td>6</td>
<td>.38 (.099)**</td>
<td>.04 (.130)</td>
<td></td>
</tr>
<tr>
<td>Telephone</td>
<td>2</td>
<td>.46 (.138)**</td>
<td>.13 (.151)</td>
<td></td>
</tr>
<tr>
<td>Questionnaire</td>
<td>1</td>
<td>.47 (.140)**</td>
<td>.15 (.150)**</td>
<td></td>
</tr>
<tr>
<td>CASI</td>
<td>1</td>
<td>.62 (.191)**</td>
<td>.26 (.141)*</td>
<td></td>
</tr>
<tr>
<td>Face-to-face</td>
<td>5</td>
<td>.42 (.099)**</td>
<td>.09 (.127)*</td>
<td></td>
</tr>
<tr>
<td>Social sensitivity</td>
<td></td>
<td></td>
<td>.12 (.036)**</td>
<td></td>
</tr>
</tbody>
</table>

| $\sigma^2_{\text{study}}$ | .042 (.028)* | .042 (.029) | .025 (.018) |
| $\sigma^2_{\text{condition}}$ | .023 (.010)** | .018 (.008)** | .013 (.005)** |

N = number of data collection conditions, standard error in parentheses, * p ≤ .05, ** p ≤ .01.
Table 4: Results of the multilevel analysis for comparative randomized response studies

<table>
<thead>
<tr>
<th>Step</th>
<th>N</th>
<th>M0</th>
<th>M1</th>
<th>M2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(Intercept only)</td>
<td>(conditions added)</td>
<td>(social desirability added)</td>
</tr>
<tr>
<td>Intercept</td>
<td></td>
<td>.28 (.077)**</td>
<td>.39 (.106)**</td>
<td>.21 (.119)**</td>
</tr>
<tr>
<td>D_{Face-to-face}</td>
<td>23</td>
<td>.39 (.106)**</td>
<td>.21 (.119)**</td>
<td></td>
</tr>
<tr>
<td>D_{Scenario}</td>
<td>1</td>
<td>-.13 (.224)</td>
<td>-.31 (.230)*</td>
<td></td>
</tr>
<tr>
<td>D_{Telephone interview}</td>
<td>3</td>
<td>.23 (.455)</td>
<td>.03 (.449)</td>
<td></td>
</tr>
<tr>
<td>D_{Questionnaire}</td>
<td>13</td>
<td>.24 (.136)*</td>
<td>.05 (.144)</td>
<td></td>
</tr>
<tr>
<td>D_{Unmatched count}</td>
<td>2</td>
<td>-.08 (.170)</td>
<td>-.24 (.177)</td>
<td></td>
</tr>
<tr>
<td>D_{Social sensitivity}</td>
<td></td>
<td>.07 (.023)**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>\sigma^2 study</td>
<td></td>
<td>.072 (.029)**</td>
<td>.17 (.05)**</td>
<td>.15 (.049)**</td>
</tr>
<tr>
<td>\sigma^2 condition in study</td>
<td></td>
<td>.031 (.009)**</td>
<td>.02 (.01)*</td>
<td>.03 (.008)*</td>
</tr>
</tbody>
</table>

N = number of data collection conditions, standard error in parentheses, * p ≤ 0.05, ** p ≤ 0.01, d'_{Face-to-face} is the d_{probit} for the difference between RRT and face-to-face interviewing, d'_{Scenario} is the d_{probit} for the difference between RRT and scenario methods, d'_{Telephone} is the d_{probit} for the difference between RRT and telephone interviewing, d'_{Questionnaire} is the d_{probit} for the difference between RRT and questionnaires, d'_{Unmatched count} is the d_{probit} for the difference between RRT and the unmatched count technique.
Table 5: The relation between the level of social sensitivity of a topic and the effectiveness of RRTs when compared to standard data collection methods (d-probit)

<table>
<thead>
<tr>
<th>Social sensitivity rating</th>
<th>Mean d-probit across topics and methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0062</td>
</tr>
<tr>
<td>1</td>
<td>0.1980</td>
</tr>
<tr>
<td>2</td>
<td>0.3254</td>
</tr>
<tr>
<td>3</td>
<td>0.3063</td>
</tr>
<tr>
<td>4</td>
<td>0.4037</td>
</tr>
</tbody>
</table>
APPENDIX: RANDOMIZED RESPONSE TECHNIQUES

One of the first improved randomized response designs following Warner (1965) was proposed by Simmons (Greenberg, Abul-Ela, Simmons and Horvitz 1969; Horvitz, Shah, and Simmons 1967), and is known as the ‘unrelated questions’-method. Respondents use a randomization device to decide which from two questions has to be answered. The main difference with Warner’s method is that the questions are not related to each other. For instance:

Question 1: Have you ever had an (illegal) abortion?

Question 2: Is /Was your mother’s birthday between January 1 and February 29?

When the occurrence of question 2 in the population is known, as in our example, only one sample is needed to compute the unbiased estimates for sensitive attribute A (for estimates and variance computation see equation 11 and 12). If the occurrence in the population of question 2 is not known (for instance: do you have a subscription on Time magazine?) than two samples are needed to compute estimate A, because in this design there are 2 parameters unknown, \( \pi_a \) (the prevalence of sensitive attribute A) and \( \pi_b \) (the prevalence of non-sensitive attribute B). The probabilities of Yes-answers are given by the equation:

\[
\lambda_a = p_1 \pi_a + (1-p_2) \pi_b
\]

were \( p_1 \) (probability of answering the sensitive question) and \( p_2 \) (the probability of answering the innocuous question) are not equal. \( a \) can be replaced by \( b \) to estimate the probability yes on question B. The estimate of \( \pi_a \) can be computed as:

\[
\hat{\pi}_a = \left[ \hat{\lambda}_1 (1-p_2) - \hat{\lambda}_2 (1-p_1) \right] / (p_1 - p_2)
\]

with variance:

\[
\text{var}(\hat{\pi}_a) = \left[ 1/(p_1 - p_2)^2 \right] \left[ \hat{\lambda}_1 (1-\hat{\lambda}_1) (1-p_2)^2 / n_1 + \hat{\lambda}_2 (1-\hat{\lambda}_2) (1-p_1)^2 / n_2 \right]
\]

A third statistical improvement is proposed by Boruch (1972), and is known as the
‘forced response’ or ‘contamination’-design. The respondents have to throw two dice. When they throw 5, 6, 7, 8, 9, or 10 (probability \( p \)) they have to answer the sensitive question truthfully. When they throw 2, 3, or 4 (probability \( \theta \)) they have to answer ‘yes’ and when they throw 11 or 12 they have to answer ‘no’ (probability \( 1 - p - \theta \)). The unbiased estimate for the population is given by:

\[
(\hat{\pi} = (\hat{\lambda} - \theta) / p)
\]

where \( \hat{\lambda} \) is the obtained proportion of ‘yes’-answers in the sample. The variance is given by:

\[
\text{var}(\hat{\pi}) = \frac{1}{p^2} \times \frac{\hat{\lambda}(1 - \hat{\lambda})}{N}
\]

All these methods require the respondent to answer ‘yes’ sometimes, and ‘yes’ can be incriminating. The absolute non-incriminating answer to give is always ‘no’. Kuk (1990) developed a randomized response method, which does not require direct answers. The respondents get two packs of cards and a sensitive question. The proportion of red cards is \( \theta_1 \) in the first pack of cards (yes) and \( \theta_2 \) in the second pack (no), where \( \theta_1 \neq \theta_2 \). The binary outcomes of a question are the colors of the card drawn. Since respondents do not have to answer the question with yes or no, they are presumably more willingly to co-operate in the procedure. Let \( r \) be the proportion of red cards reported by \( n \) respondents. Then

\[
r = (\theta_1 - \theta_2)\hat{\pi} + \theta_2
\]

with variance

\[
V_r = \frac{\phi(1 - \phi)}{n(\theta_1 - \theta_2)^2}
\]

However, since Kuk’s method is formally equal to Warner’s original proposal, it leads to large sampling errors, and is therefore one of the more inefficient RRTs.
Bibliography

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, where she teaches methodology and statistics for the social sciences. Her research focuses on research methods for socially sensitive topics, in surveys as well as experimental settings, and on doing meta-analysis.

Joop J. Hox is 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 include interviewer effects on nonresponse, meta-analysis, and a handbook on multilevel analysis.

Peter G. M. van der Heijden is a professor of social science statistics in the Faculty of Social Sciences at Utrecht University, the Netherlands. He is interested in estimation of population sizes in cases of sensitive topics and categorical data analysis. He also works with logistic regression to relate explanatory variables to RRT-outcomes.