MEASURING POLITICALLY SENSITIVE BEHAVIOR

Using Probability Theory in the Form of Randomized Response
to Estimate Prevalence and Incidence of Misbehavior
in the Public Sphere: A Test on Integrity Violations

Carel F.W. Peeters
Final thesis Political Science,
Vrije *Universiteit Amsterdam*
September 2005

Supervisor
Prof. Dr. Leo W.J.C. Huberts
Strategic Chair Integrity of Governance
Vrije Universiteit Amsterdam

Co-Supervisor
Dr. Andre Krouwel
Department of Political Science
Vrije Universiteit Amsterdam

Reading Committee

Drs. Karin Lasthuizen
Department of Public Administration and Organization Science
Vrije Universiteit Amsterdam

Prof. Dr. Raymond M. Lee
Social Research Methods, Department of Politics and International Relations
Royal Holloway, University of London

Dr. Gerty Lensvelt-Mulders
Department of Methodology and Statistics
Utrecht University

ISBN-10: 90-78223-01-4
ISBN-13: 978-90-78223-01-6

Published by: Dynamics of Governance /Department of Public Administration and Organization Science, Faculty of Social Sciences, VU Amsterdam
Cover designer: S. van der Ploeg, Room for ID's, Nieuwegein
Photography: iStockphoto™
Printed and bound by reprography VU, Amsterdam

MEASURING POLITICALLY SENSITIVE BEHAVIOR

Contents

List of Tables and Figures

Abbreviations

CARR	Computer Assisted Randomized Response
CASI	Computer Assisted Self-Interviewing
DQ	Direct Questioning
EM	Expectation Maximization (Algorithm)
FRR	Forced Randomized Response
LCM	Latent Class Model
ML	Maximum Likelihood
RR	Randomized Response
UQD	Unrelated Question Design

Preface and Acknowledgements

The social and political sciences in general, are concerned (whether qualitatively or quantitatively) with the determination of the relationships between and among various phenomena of interest. Any endeavor of this kind attends the issue of 'measurement', which in the social realm often crosses the problem of accurately representing the latent with the manifest. This points to problems of conceptualization and especially measurement as the most serious research issues in the social realm.

This text reviews the problem of measuring prevalence and incidence of sensitive behaviors. A specific conduct for which specific methodological effort is to be justified if the latent is to be accurately represented by a manifest numerical. The great problem in the measurement of sensitive behavior is that disclosure of information regarding them can be threatening to those involved in the sense of being severely intrusive, stigmatizing or incriminating, resulting in an incentive to retain anonymity. External measures regarding these behaviors, such as sexual or criminal behavior, then come to reflect more the contingencies in their production than the behaviors they purport to measure, while in self-reports, self-representational concerns peak. As information regarding sensitive issues is likely to remain limited to third persons, it is the terrain of enhancing the accurateness of self-reports where ground has to be gained.

This text proposes the *Randomized Response Technique*, to enhance accurateness regarding self-reports of sensitive behaviors. This, in empirically research endeavors, unjustifiably overlooked technique, uses the insertion of random error by an element of chance to inoculate individual responses. When the respondent understands and trusts that the disclosure of information is conditional only, and that identification can never occur on the basis of responses, it becomes at least theoretically possible to come to more accurate estimates of prevalence and incidence. These writings revolve around the translation of theoretical potentiality to practical feasibility in measuring behavior of a politically sensitive nature. In doing so, this marriage of mathematical/statistical technicality and the obvious social component, draws on certain behavioralist premises, in what is essentially a post-behavioral condition of the discipline of Political Science.

The 'preface and acknowledgements' section of an academic work is a place of relief and revelation. Not only because it is written lastly, at the end of sometimes grueling days, weeks and months and one has the opportunity to give some 'lasting' thanks to the many people who have contributed in some way to the work that lies before the reader. But also because it is the only place in a serious academic paper where, in doing so, one can use the word 'I'. Thus, 'I' would like to thank ...

My family. Firstly my parents, for their support and their production of the 1.42E -14 chance of my chromosome constellation ($.5^{23}$ x $.5^{23}$); for the simplicity of the argument even ignoring the absolute minute chances on the exact trajectory leading (in both my parents' lifespan as well as in the actual 'act') to the actual moment of conception of that exact constellation that culminated in 'me'. For that many thanks. And of course my 'little' brother, for simply being just that.

My friends, who will not be named separately because they are with too many and as 'friends', 'family' and 'colleagues' are not discrete categories in many instances. Thank you all for necessary distraction and for listening to undoubtedly boring stories regarding mathematical formulae and social measurement. But there are some people who must be named specifically. Many thanks to Jean Paul Duyx and Berry Remmers, for assisting, relieving and solving many computer-related problems. Special thanks to Jan T., Chris H., Paul D., Eelco H. and Jelle E., for their very valuable assistance with regard to our respondent group. I would also like to reserve the room to thank all officers and employees of the regional police force within which this research was conducted, for their acceptance and kind compliance. In this respect the chief constable of this regional police force may not go unmentioned, as it takes a lot of generosity and courage to allow for research such as this in an organization that is so often publicly battered. Bart Bannink and Rob de Vries of the computer-technical service of the Social Science Department Vrije Universiteit Amsterdam can also not go unmentioned, as they played an important part in the possibility of this research.

Gerty Lensvelt-Mulders and Rens Lensvelt of WetenschapWerkt. The former is deeply thanked for her support, ideas and comments on these texts as Randomized Response specialist, fellow researcher in this project and member of the 'reading committee'. The latter is thanked for the absolutely brilliant programming of the computerized questionnaires and randomization devices as developed in this text. The kind and very useful

extra readership by Raymond Lee of Royal Holloway, University of London, is also gratefully acknowledged here.

My colleagues at the research group Integrity of Governance of the Department of Public Administration and Organization Science, Vrije Universiteit Amsterdam. A group of which Gjalt de Graaf, Zeger van der Wal, Judith van der Veer and Ronald van Steden must be especially acknowledged for their useful comments, their interest and their lunch company. I am also deeply indebted to Judith van de Veer for her selfless and tireless work on the publication of this research. For that many, many thanks. Kim van Nieuwaal and Guda van Noort must be acknowledged for 'distraction' and 'good' conversation. And of course Franziska Cecon must not be forgotten, for a wonderful day in Bern in which time and space were explored, and for her discovery of the missing 'h'.

Leo Huberts, Karin Lasthuizen and Andre Krouwel must be given special acknowledgements. I want to express my deep gratitude towards Leo Huberts and Karin Lasthuizen for giving me the freedom and the means to express scientific ideas, and the acceptance and responsibilities granted to me as a researcher, even before any official graduation. Their trust and friendship is very valuable to me, not only as a researcher, but also as a person. Andre Krouwel is thanked analogously for his trust and the opportunity given to me to bore people with technical lectures.

The lovely Sabine, for keeping me – often unintentionally – within earth's gravitational influence and for giving me a feeling that, for a while there, I thought I had lost.

And Lieke Herpers. Who proved to be a crucial factor in my 'early years'. Life is not a box of chocolates; like Terence already noted, it's a game with dice. If the numbers that turn up are not the ones you wanted, hoped or wished for; strive to contrive to use it equally.

Amsterdam/Bern, september '05
- C.F.W.P.

Our Culture has one idiosyncratic feature that distinguishes it from most and perhaps all other cultures. It is a culture in which there is a general desire to make social life translucent, to remove opacity, to reveal the hidden, to unmask...A secret in our culture has become something to be told. And social science research cannot hope to avoid being in part an expression of this same tendency.

- Alisdair MacIntyre

Est in vita quasi cum ludas tesseris: si id quod jactu opus erat forte non cecidit, id quod cecidit arte corrigas

- Publius Terentius Afer [Terence]

1. INTRODUCTION: MEASUREMENT IN SOCIAL SCIENCE AND MEASURING SENSITIVE BEHAVIORS

Social Science in general is concerned with the determination of the empirical relationships between and among various social phenomena. Endeavors of this kind always attend the issue of 'measurement', which can be most commonly defined as the coherent assignment of numbers to empirical objects or events (Stevens 1951; as referred to by Carmines & Zeller 1979; and Luce & Narens 1987). But the fact that many constructs of interest to the Social Scientist are not directly observable and are thus latent (as for example attitudes and preferences) makes measurement in the social realm a problematic and complex undertaking. Social measurement cannot but involve greater consideration than the somewhat oversimplified notion of 'assigning numbers'.

As some constructs are latent, observable indicators have to be constructed which can accurately represent the unobservable latent concept one is interested in. Measurement in social science, when stated fully, is thus the process by which a concept is coherently linked to one or more latent variables, which in turn are connected to observable (numerical) measures (Carmines & Zeller 1979; Bollen 1989). Carmines & Zeller (1979) observe that the attention for the issue of measurement in the Social Sciences has been asymmetric. There seems to be a ritualistic concern with the issue, which approximates the simple notion of assigning numbers to objects or events, but systematic attention for the methodology and complexity of measurement is often lacking in its mainstream conduct. This is a great flaw in Social Science as its inherent nature of connecting concepts to observable indicators bears on what could be its greatest problem (Blalock 1979): the degree in which the

manifest is able to represent the latent. Given the status of measurement in Social Science, where we thus often have to rely on constructed observables such as an item designed on a questionnaire to indirectly represent those constructs in which we take an interest, makes it a crucial aspect of any endeavor involving measurement to question the accurateness of the former in representing the latter. A statement by which we arrive at the concept of 'measurement error'.

'Measurement error' refers to all factors that make an observable score or indicant deviate from its true value (in latent space). Measurement error is always present to some extent in any endeavor using measurement of any kind. Mostly they are induced by the method of measurement itself (Podsakoff *et al.* 2003), the make-up of which can lead to variable or systematic deviations of observed from true score. When systematic errors are present, the relationships between a concept and its indicator become weak or faulty, leading to incorrect inferences, misleading conclusions and a flawed understanding of the phenomenon under investigation (Carmines & Zeller 1979). This at least partly underlies the abundance of social theory when compared to the available data on social phenomena. Given the notions of the possibility of error in measurement and the potentially resulting limited notions of our theoretical indicants in relation to the concepts they purport to measure, it can be argued that the issue of measurement itself, is our most pressing one (Blalock 1979). A statement, which is especially apparent when researching sensitive topics.

LIMITED ACCURATENESS IN MEASURING SENSITIVE BEHAVIORS

A topic can be said to be 'sensitive' when the disclosure of information regarding this topic poses to be threatening for the respondent in the form of being potentially stigmatizing, incriminating or severely intrusive (Lee & Renzetti 1990; Lee 1993). Due to the potential threats, people have self-representational concerns and a resulting incentive to retain their anonymity with respect to the sensitive topic researched, leading to refusals to cooperate or evasive cooperation. Although behaviors of a sensitive nature are physically observable (in comparison with an attitude or preference), the mentioned self-representational concerns give these behaviors a certain latent status, as only part of its prevalence and incidence will move into observable space, giving its estimation the nature of estimating a hidden population. It is these behaviors and their

measurement in which we take an interest in this text as they can hamper on some of the most pressing problems of our time (as for example the spread of sexually transmittable diseases through sexual behavior and the loss of confidence in government due to misbehavior by officials) and as the research problems associated with these behaviors tend to inhibit their adequate measurement and subsequently our substantive and theoretical understanding of them (see for example Brewer 1990; Herzberger 1990; Lee & Renzetti 1990; Fenton *et al.* 2001).

Sampling theory generally assumes that the data collected on units in the sample are accurate representations of the values associated with the units sampled. This assumption forms the base of statistical inference directed to a certain population when only a part or sample of that same population can be studied. With most research this is an assumption that can be maintained, as most surveys do not deal with sensitive topics and as most respondents will not expect consequences due to participation. But in inquiries into sensitive behaviors, attitudes or preferences, respondents will be inclined not to respond or respond evasively, due to heightened concern over anonymity and/or possible consequences, leading a substantial part of observed values to deviate from true values. This constitutes major bias in research and theory. When assessing prevalence and incidence of certain sensitive behaviors one thus has to look beyond the reach of mere sampling as one runs into the problem of differential validity. This problem is centered around self-reported measures of individual sensitive behavior on the one hand where one assesses information directly obtained from the individual of interest, and external measures on the other where one does not so. External measures cannot validly pierce deep into latent space to come to a more accurate estimate of prevalence and incidence of certain behaviors, as they reflect the contingencies in their production more so than the behaviors they purport to measure. Most importantly due to the limited knowledge of certain behaviors available to third persons, as most want to keep information of a sensitive nature private. To be able to give more accurate estimates of prevalence and incidence of certain sensitive behaviors, one has to collect reliable and valid self-reports. But here one runs into the problem of observational statistics in the social and behavioral sciences (as opposed to experimental statistics): we have to rely on observations and/or reports given by respondents who can systematically distort results. In self-reports regarding sensitive behavior the self-representational concerns of the respondent peak, which potentially heightens the inclination not to respond or to respond evasively. To give an accurate estimate of prevalence and incidence of certain sensitive

behaviors, the potential distortion by the non-sampling biases of response refusal and evasiveness thus have to be reduced in inquiries of a sensitive nature which use self-reports in their design. This poses a substantive methodological challenge regarding the issue of measurement.

THE USE OF ELEMENTARY PROBABILITY AND GAINING ACCURATENESS

The methodological challenge posed by the desire to come to accurate representations of sensitive behaviors through a design making use of self-reports, could be (partly) answered by elementary probability theory. When probability theory is directed towards the conduct of asking questions itself, it is possible to render out individual meaning of answers in such a design.

A simple example is a situation in which respondents are instructed to answer a question regarding a certain sensitive behavior with the help of a die (Boruch 1971). Their answer should be dependent on the outcome of an independent individual throw with the die. Now consider a question to which only a 'yes' or 'no' response is possible and the specific instruction to each respondent is to answer 'yes' irrespective of their true status when the outcome is 1, to answer 'no' irrespective of their true status when the outcome is 6, and to respond truthfully with 'yes' or 'no' when the outcome is 2, 3, 4 or 5. When the researcher is unaware of the outcome of the throw with the die, it can never be known why a respondent answers with 'yes' or 'no' as it can be truthful or redirected by the outcome. This provides a respondent with full anonymity and should potentially relieve self-representational concerns, so that when directed to do so, respondents will disclose their true status on a probability basis (when the outcome is 2, 3, 4 or 5), which makes possible the accurate estimation of the prevalence of the sensitive behavior researched.

Assuming 600 respondent in our example and 200 'yes' responses, we can easily see that 100 persons answered 'yes' truthfully to the question regarding a certain sensitive behavior as the probability of a 'yes' redirected by the outcome '1' is 1/6, which means that 100 'yes' responses should be subtracted from the total number of affirmative responses. Exactly who the 100 persons are who answered truthfully can never be known. In this example, our desired estimate of the prevalence of the behavior for which we have asked would then culminate in 25 percent, derived by subtracting the probability of the outcome '1' from the total proportion of 'yes' responses, and subsequently dividing this

product by the total probability of the outcomes 2, 3, 4 and 5 (or – in this example – by dividing the truthful number of 'yes' responses by the total number of truthful responses).

The given example is a very simple representation of the *Randomized Response Technique*, which uses known probability distributions in the questioning procedure to purposively misclassify data. The misclassification rules out individual meaning of responses so that self-representational concerns are relieved. The rationale is that when respondents are assured that their privacy is guarded, they will be more inclined to cooperate and to respond truthfully. When respondents' privacy is fully protected, admitting to sensitive behavior becomes less threatening, so that more valid population estimates can be obtained as the potential of distortion due to response refusal and evasiveness reduces. The insights into the probabilities concerned make it then subsequently possible to accurately examine certain population parameters regarding the sensitive behaviors under review. This technique and its theoretical potential are directly linked to the ambitions and aims of this study.

AMBITIONS AND AIMS

Thus far, no study has been able to assess the accuracy of prevalence and incidences reported for in individual misbehavior regarding integrity (violations) in the public sphere, without the effects of social desirability and response refusal influencing the results (see Huberts, Lasthuizen & Peeters, forthcoming). Integrity violations in the public sphere are a form of sensitive behavior that is essentially political in nature, that is: behaviors hampering on the workings of the public arena, which are potentially incriminating and/or stigmatizing for individuals or groups who have vested political alignments/positions in that arena. The dissemination of information on behaviors of this kind could not only be incriminating or stigmatizing for the group or individual of interest, in case of officials it can also potentially undermine the public trust that is entrusted in them.

The direct motivation is to assess the viability of the – in empirical research endeavors unjustifiably and largely neglected – *Randomized Response Technique* to accurately gauge the level of certain types of wrongdoing in one organization in the public arena. The study however, will be no rigorous determination of the prevalence and incidence of wrongdoing in all similar organizations in the public arena as a whole.

The study as to be proposed seeks to hold as an 'experimental' test case. This motivation has multiple components and ambitions.

Firstly, and most abstractly, this study seeks to open up a relatively obscure technique for research on sensitive items in the social sciences, specifically to clarify the relevance of the Randomized Response (RR) technique for methodologically interested non-methodologists and social and political scientists. While this technique has been under construction for over 40 years since its first theoretical realization in a seminal article by Stanley Warner (1965), it has not found its way to mainstream social science toolkits. This can be a limitation as the RR method can improve observational statistics with regard to sensitive issues in terms of more accurateness in representing the latent with the manifest.

Secondly, a new unified RR model for eliciting sensitive dichotomous and multi-proportional data will be presented. Many of the most pressing questions regarding sensitive issues are those wanting to explore the relative number of people displaying the behavior in question (questions of prevalence) and those seeking to give a representation of the relative frequency of occurrence in a certain population (questions of incidence). As the dichotomous RR model explores questions of prevalence, the quantitative multi-proportional RR model makes possible the investigation of incidence. A unified framework for these models will be provided which also seeks to address the methodologically and statistically more advanced reader.

Thirdly, with the help of that unified framework it will be attempted to accurately gauge the level of certain types of wrongdoing in one organization in the public arena, in terms of prevalence and incidence. The viability of RR for providing a more accurate assessment of certain kinds of wrongdoing will be assessed, and the proposed study seeks to hold as an 'experimental' test case. To assess the relative merits of RR, the present study will take a comparative road, by assessing an RR design with the more traditional direct questioning technique in their viability to delve into latent space so as to gauge the level of certain sensitive traits, adding to the relatively scarce literature on field-tests of the RR method. A specific form of randomized response will thus for this purpose be incorporated with computer assisted self-interviewing, so that the randomization device which provides for the misclassification will have a virtual nature. This Computer Assisted Randomized Response Survey (CARR) will be set out in the finite population of a police force. The survey will contain questions regarding individual misbehavior in the integrity sphere. The potential of the randomized response technique to elicit sensitive information will be assessed by comparing the

experimental group with a control group within the same police force that will receive the same survey without the incorporation of the RR technique. This control group will thus be subjected to a Computer Assisted Self-Interview (CASI) in which the sensitive questions are posed directly.

The underlying ambition in the empirical part of this study is to modestly contribute to a more definitive protocol for and understanding of RR. Also it seeks to act as a very modest basis for much needed cumulative work on RR designs in actual research settings.

QUESTIONS AND STRUCTURE

The focus of present study as has become clear is the methodology of measurement in the political and social sciences and the reduction of systematic distortion in the measurement of behaviors of a (politically) sensitive nature through the use of elementary probability theory. This focus and the previously stated aims and ambitions find their expression in the *guiding research question:*

> *Can elementary probability theory in the form of Randomized Response Techniques be used to obtain more accurate population estimates regarding prevalence and incidence of sensitive behaviors of a political nature?*

While it may seem somewhat trivial at firsthand, it will be stated here that it is an appropriate question for a study, which eventually seeks to translate theoretical potentiality to practical feasibility. To enhance comprehensibility the main question can be broken down into the following *secondary-questions:*

a. What are sensitive behaviors and how does sensitivity relate to the accuracy of various estimates of prevalence and incidence?
b. What is Randomized Response and how does it use probability theory to give more accurate population estimates of sensitive behaviors?
c. In what framework will the Randomized Response technique be modeled to be theoretically able to efficiently give more accurate population estimates regarding prevalence and incidence?
d. In what empirical framework will the Randomized Response technique be field-tested to give more accurate population

estimates regarding prevalence and incidence of sensitive behaviors of a political nature and how will the accuracy of the estimates be evaluated?

e. How does the Randomized Response technique perform in obtaining population estimates of prevalence and incidence of sensitive behaviors of a political nature, in comparison with a more traditional questioning technique?
f. Which inroads do the provided Randomized Response framework and its subsequent field-testing and field-performance give, to further the translation of theoretical potentiality to practical feasibility?
g. Can the Randomized Response technique delve deeper into a certain latent status to give more accurate population estimates regarding prevalence and incidence of sensitive behaviors of a political nature?

The subsequent secondary questions lay bare the structure of the remainder of this text. Chapter 2 delves into sub-question *a*. Here a more elaborate discussion of the problems associated with the measurement of a certain latent status when the latency is induced by 'sensitivity' will be given, as touched upon at the beginning of this introduction. The argument is that the validity problems, which are intricately related to various external measures, leave self-reports often as the only road for the estimation of prevalence and incidence. To curb the problems associated with self-reports however, one must endeavor on specific methodological effort. Chapter 3 bears on the introduction of this methodological effort by reviewing the RR technique, which – as was also touched upon (and probably to the horror of the devotees of classical test theory) – uses the insertion of random error by an element of chance to provide a respondent with full anonymity when self-reporting on questions regarding sensitive issues. It can be expected that respondents' self-representational concerns are relieved when full privacy protection is guaranteed, so that with the use of this technique more accurate estimates of prevalence and incidence are at least theoretically within reach. Chapter 4 builds on the RR introduction by constructing a unified approach to eliciting sensitive dichotomous and quantitative data for procuring prevalence and incidence estimates. Chapter 5 subsequently builds on the RR introduction by building an empirical framework for RR with which its practical feasibility will be empirically tested with regard to integrity violations in a police force, which will hold as a test

case for sensitive behaviors of a political nature. As already stated, this framework entails comparing RR with the more traditional direct questioning technique by subjecting an experimental group to CARR and a control group to CASI. Chapters 4 and 5 together give a framework for theoretical potential and a possible translation to empirical practice. Chapter 6 then reports on the results of this empirical effort. In doing so this chapter already touches upon sub-question *f*, but a more elaborate discussion of the RR framework, it's field-testing and its field-performance will be provided for in Chapter 7. Chapter 8 forms the concluding chapter and the writings and findings of the preceding ones will culminate here, in a thorough reflection on the main research question.

THE RELEVANCE OF FOCUSING ON MEASUREMENT AND METHODOLOGY

The 'sensitivity' of a certain topic does not indefinitely amount to great social and theoretical significance (Lee 1993: 2). However, as Sieber & Stanley rightly point out, sensitive topics often address "some of society's most pressing social issues and policy questions" (1988: 55). While many studies approach the issue of sensitivity from an ethical or normative standpoint (see for example Barnes 1980; Melton & Gray 1988; Sieber & Stanley 1988) this text is not about the disclosure of ethical issues regarding the conduct of research on sensitive topics, but rather on the methodology of estimation regarding population parameters of behaviors which can be posed as sensitive for especially the respondent. As already argued, the real challenge with regard to these topics is to accurately represent the latent with the manifest. With research on sensitive topics this problem is especially apparent as the self-representational concerns of respondents can lead them to evade response or to respond evasively. While our theoretical understanding of and subsequently policy actions regarding sensitive topics are often built on our perceptions of their prevalence and occurrence, methodological effort to make up for leeway regarding their measurement, is deemed very necessary. This study finds its encompassing relevance connected to this remark.

The specific methodological relevance can be found in the method to be proposed. A latent quality due to sensitivity tends to inhibit adequate measurement and the RR method is potentially able to delve deeper into a certain latent status. Part of the potential gain here is the attempt to open up this technique to the practitioner in the field. Also, by presenting a

unified framework for eliciting sensitive dichotomous as well as multi-proportional data with RR, which also uses the incorporation of computer assistance, this study can modestly contribute to the more technical literature on the estimation of prevalence and incidence regarding latent traits.

The methodological relevance carries over into the results, by the empirical testing of the new RR framework. In doing so, it will be an empirical 'first' in many instances, as substantial tests on the RR technique are relatively rare, let alone the testing of a computer assisted RR model which also pursues to evaluate its relative merits in producing incidence estimates by incorporating a specific multi-proportional application. In this experimental translation of theoretical potentiality to practical feasibility, there is the potential to contribute to a better understanding of RR in field-research settings and to modestly spur cumulative work on an administrative protocol for RR which aids this necessity. Also, the framework for RR testing as proposed in this text could also be extended for research on other sensitive behaviors or even attitudes and preferences of an essentially political nature. One can think of the breaching of behavioral codes by high government officials or sympathy for political parties at the extremes of the political spectrum. Estimates of which are notoriously inaccurate.

Many researchers are shunned away from research on sensitive topics, due to the ethical problems involved and especially the methodological difficulties it poses. To shy away from these topics would be an avoidance of responsibility (Sieber & Stanley 1988: 55) as is ignoring their seemingly insurmountable methodological problems, as "this ignorance may potentially generate flawed conclusions on which both theory and policy subsequently may be build" (Lee & Renzetti 1990: 525). The Social Scientist thus must confront head on, the methodological problems these topics pose.

2. MEASURING SENSITIVE BEHAVIOR: DIFFERENTIAL VALIDITY AND THE NEED FOR ACCURATE SELF-REPORTS

Many issues that are of interest to social and behavioral scientists are of a sensitive nature. People's sexual behavior, voting preferences and propensity for involvement in criminal activities are just a few examples of sensitive behaviors and attitudes which hinge on some of the most pressing problems of our time (like the spread of sexually transmitted diseases and the rapid growth of violent crime). An accurate assessment of prevalence, incidence and intensity of these sensitive traits is crucial for our theoretical and empirical understanding of these pressing issues, which have the tendency to spill over into policy questions. This chapter delves into the relationship between sensitivity and accuracy by reviewing the question what sensitive behavior and research into sensitive behavior entails and how sensitivity relates to the accuracy of various estimates of prevalence (occurrence of certain behaviors: number or percentage of people involved in certain behaviors) and incidence (frequency of occurrence of certain behaviors). It will be argued that the threat of stigmatization and/or incrimination leads respondents to respond evasively or to not respond at all to questions of a sensitive nature. This leads to serious underestimation of sensitive behaviors in external measures as well as internal measures. However, the validity problems inherent in various external measures leave self-reports of certain behaviors as the only road for estimation of prevalence and incidence. This, as will become clear towards the end of this chapter, justifies specific methodological effort to enhance the validity of self-reports regarding sensitive behaviors.

SENSITIVITY AND ERRORS OF (NON)MEASUREMENT

Often, the term 'sensitivity' is used in the literature in a self-explanatory sense (Lee & Renzetti 1990: 510). When conducting research into socially sensitive questions, one has to have a notion however, of what constitutes a 'sensitive topic'. Sieber and Stanley proposed a definition of socially sensitive research based on consequences and implications:

> studies in which there are potential consequences or implications, either directly for the participants in the research or for the class of individuals represented by the research.
>
> (1988, p. 49)

The problem inherent in this definition is that it does not specify the nature of consequences and implications in relation to sensitivity. 'Sensitivity' in this definition thus encompasses all research, as any research can be said to have consequences or implications. Lee and Renzetti amend this definition by emphasizing the common thread in the literature on sensitive topics: the implicit assumption that topics of a sensitive nature involve an apparent or latent level of threat, which hampers on the process or conduct of research:

> a sensitive topic is one which potentially poses for those involved a substantial threat, the emergence of which renders problematic for the researcher and/or the researched the collection, holding, and/or dissemination of research data.
>
> (1990, p. 512)

The threat posed by research can potentially involve all those involved, and may in fact make itself felt in every stage of the research cycle. As will be argued, the threat posed by research on sensitive topics most obviously affects participants in the research (respondents) and the collection of data. These will be the main focus in the remainder of this text.[1]

The sensitive character of a certain research topic lies mainly in the relationship between the specified topic under review and the social context within which the research is conducted (Lee & Renzetti 1990: 512). 'Sensitivity' thus has an emergent character. Lee (1993: 4-9) builds on this idea by specifying three areas where research can be posed as 'threatening'. The first is where research is intrusive, piercing areas

which are "private, stressful or sacred" (Lee 1993: 4). The second area where research is seen as threatening is the area of deviance, misconduct and social control. Research into which is potentially incriminating. The third is where research impinges on political alignments, possibly trespassing on areas of social conflict. Note that the specific threat in these areas of sensitive topics and research lies in the possibility that the disclosure of certain information will be severely stressful, stigmatizing, or incriminating for the participants.

The concerns for the respondent in socially sensitive research shift to the confidentiality of information and one's own privacy needs, as most want to keep information of a sensitive nature private. The respondents' interest in control over the boundaries between him or herself and others heightens (Sieber & Stanley 1988: 51, 53; Fox & Tracy 1986: 9), specifically because certain information is possibly stigmatizing or incriminating. Where research is threatening, the relationship between the researcher and the researched thus loses openness, and "is likely to become hedged about with mistrust, concealment and dissimilation" (Lee 1993: 2). The heightened interest in controlling the boundaries between researched and researcher on part of the former affects the availability and also the quality of data, diminishing reliability and especially the validity of research findings.

Sensitivity thus becomes intricately related to the validity and reliability of estimates of certain social parameters. As 'validity' and 'reliability' are central concepts throughout all chapters, they will first be elaborated, before delving further into the specific relationship between sensitivity and diminishing validity of research findings.

Validity and reliability are two traits with which the quality of measurement is assessed: meaning the degree in which a certain empirical indicator represents the underlying concept it is intended to measure. Validity refers to the question if a certain indicator actually measures what it purports to measure (Carmines & Zeller 1979: 12; Bollen 1989: 184). An indicator or measure is thus said to be valid if it provides an accurate representation of the concept that is examined with the measure. Reliability refers to the consistency of measurement or the extent to which a certain measure yields the same results over repeated trails of the measurement procedure (Carmines & Zeller 1979: 11-12; Bollen 1989: 206-207). This concept implies the possibility of an infinite number of repeated measurement trials, where the degree of consistency in the repeated measurements denotes the degree of reliability (Bollen 1989: 207; Groves 1991: 2). 'Validity' then, refers to the *accuracy* of a

certain measure, whereas 'reliability' refers to its *precision* (Dutka & Frankel 1993: 475).

Reliability and validity are both central properties of empirical measurement, and their distinction is important, as one does not imply the other. It is possible to have a reliable measure that is not valid. Bollen gives an enlightening example of such a situation (1989: 207). Consider a scale, which always points to exactly the same measure of one's bodyweight but in reality, gives a systematic underestimation of one's bodyweight by 11 lbs. The scale thus gives a very reliable but totally invalid measure of one's bodyweight, pointing to the difference between the accuracy of correspondence between measure and concept, and the consistency of that measure. This example also points to the intricate relationship between validity, reliability and the concept of measurement error.

'Error' in social inquiries refers to "deviations of obtained (...) results from those that are true reflections of the population" (Groves 1991: 1). There are two encompassing forms of error, which affect empirical measurements of any kind to deviate from true population or respondent reflections: random error and nonrandom error. Random error designates all "chance factors that confound the measurement of any phenomenon" (Carmines & Zeller 1979: 13). It constitutes variable error. Because random error is induced by chance, it can be expected to cancel out over repeated measurements, thus not affecting the mean of an estimate by systematically distorting observed and true score. Random error however, reduces the reliability of measurement. Because of random faults, which can be an overestimation or underestimation of a true score, the dispersion of values from the expected value increases (again, without affecting its mean), diminishing the precision of an estimate. In probability and statistics, this is measured by the variance of a statistic. Nonrandom error points to the systematic factors confounding the measurement of a phenomenon (Carmines & Zeller 1979: 14; Groves 1991: 2).[2] Being a constant error, it does not cancel out over repeated measurements, thus systematically distorting observed and true score. A problem of measurement commonly referred to as 'bias'. Bias affects the mean of an estimate, reducing the validity of a certain measure. A statistic is then an unbiased estimate of a parameter if the expected value of the sampling distribution is equal to the parameter of which the statistic is an estimate.

Although bias can also affect the variance of a statistic, it can be said that "[R]eliability is that part of a measure that is free of purely random error" (Bollen 1989: 207), and that validity is that part of a measure that

is free of systematically distorting factors. Both constructs, it must be stressed, are a matter of degree. The former is mostly seen as an empirical issue, while validity is widely believed to be mostly a theoretical concern. In this text, as will become clear from this and following chapters – which will elaborate on a method for the accurate assessment of sensitive behaviors and the criteria for evaluating accuracy; both will have an empirical connotation.

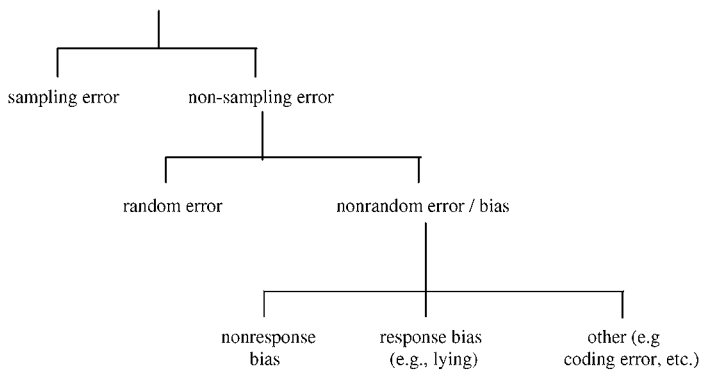
Survey estimates[3] are subject to random and nonrandom error, which can arise due to factors related to sampling or factors beyond the sampling procedure (see Figure 2.1). It is the latter, which in light of inquiries into sensitive topics has our concern.

Most research only harnesses itself against sampling error that arises from the variation that is present because a specific sample of the population is studied instead of a complete enumeration of the population. The aim of sampling is to select elements of study which adequately represent a certain population of interest. Probability theory in the form of probability sampling is directed towards this problem. In probability sampling each member of the sample is selected with a known probability, thus being more representative of the larger population (to a calculable degree). Samples, which are in compliance with the laws of elementary probability, are thus believed to be (more) free of (selection) biases. Sampling theory generally assumes that the data collected over the units in the sample, accurately represents the properties of the underlying population (see for example Sudman 1976; and Thompson 1997). But there is also another main source of error to which almost all inquiries are exposed, and for which no special measures are taken in most research: non-sampling error. Concerns over which are heightened when conducting inquiries of a sensitive nature.

Non-sampling error actually is more problematic. It can be of two main types. The first of which is the above discussed random error, which reduces the reliability of measurements but which can be expected to cancel itself out over many repeated measurements, thus not distorting the findings systematically (which means that precision reduces but that the accuracy is not affected). The second type of error has a non-random systematic nature, commonly referred to as 'bias'. In turn this bias can be broken down into two main sources: 'nonresponse bias' and 'response bias'. Nonresponse bias finds its source in respondents' refusal to respond or cooperate. Bias arises if there are systematic differences between the respondents who do and the respondents who do not respond to a survey or certain questions in it (Fox & Tracy 1986: 9; De Leeuw, Hox &

Huisman 2003: 155). The second source of bias stems from the deliberate falsification of information by respondents: response bias or evasive answer bias. These types of systematic distortion reduce the validity of measurements, influencing the estimate of a parameter in ways not to be eliminated by repeated trials. Throughout the years, many studies have indicated the distorting effects nonresponse and response bias can have in various areas of social inquiry (see for example Hansen *et al.* 1951; Sudman & Bradburn 1974; Filion 1975; Biemer *et al.* 1991; Niemi 1993; Smeets 1995; Voogt & Van Kempen 2002).

Figure 2.1: Types of Error in Surveys



Source: Fox & Tracy 1986, p. 9

Response bias is even likely to arise in surveys of relatively innocuous information (Philips 1971). When surveys ask for sensitive information the potential for response and nonresponse bias reaches a critical point due to the threat of stigmatization and/or incrimination, and the from these threats following concerns over anonymity and self-representation (Fox & Tracy 1981: 187-188). Self-representation is influenced by what psychologists call 'social desirability', which refers to the need for social acceptance and the tendency of individuals to present themselves in a favorable light so as to gain that acceptance (see Podsakoff 2003). Locander, Sudman and Bradburn (1976) have shown that response distortion sharply increases as threat increases. When asking respondents

to self-report certain behaviors we are actually dealing with a problem of 'informed consent' (Sudman & Bradburn 1982: 8-9; Sieber & Stanley 1988: 54). Potential respondents should receive enough information regarding what is being asked and how this information will be used, to judge for themselves if certain consequences follow as a result of disclosure. Most respondents in most survey research will have no problem regarding the issue of informed consent, as most surveys do not deal with sensitive topics and as most respondents thus will not expect consequences due to participation. But when respondents are asked to report sensitive behaviors (or attitudes and preferences), the respondent will be inclined to respond evasively or to not respond at all, due to fears over stigmatization and incrimination, and concerns over anonymity and self-representation. When assessing prevalence and incidence of certain sensitive behaviors, this leads to serious underestimation.[4]

The threat of reporting sensitive data for oneself is different and is often more prevalent then for example in proxy reporting. Self-representational concerns can be shown to be much higher when information is collected from the individual of interest (Eisenhower, Mathiowetz & Morganstein 1991: 130-131) affecting willingness to give and accuracy of certain answers. Here we come across a distinction in the obtainment of sensitive information: self-reported measures where information is collected from the individual of interest, and various external measures where information is not directly obtained from the individual of interest. For the latter one can think of a range of measures from official statistics to proxy reporting where an individual reports on the behaviors/attitudes of people in his or hers emotional, cultural or economical vicinity.

We will shift our attention to the differential validity in the estimation of prevalence and incidence of sensitive behavior, between self-reports and external measures. First by reviewing various external measures and indicating that while they may be valid in terms of representing organizational contingencies or the extent to which certain sensitive behaviors are visible, they cannot pierce deep into latent space. To be able to give more accurate representations of prevalence and incidence, one has to collect self-reports. But here, as will be shown next, self-representational concerns peak.

EXTERNAL MEASURES AND LIMITED ACCURATENESS IN ESTIMATING LATENT STATUS[5]

When assessing sensitive behaviors on prevalence or incidence, one is actually endeavoring on the estimation of a hidden population within the total sample or population. Because of the potential stigmatization and/or incrimination and the resulting self-representational concerns, individuals possessing certain sensitive traits will want to retain their anonymity. Sensitive behavior thus has a certain latent status, as only a part of it will move into observable space. Here we will analyze the accuracy of various external measures in the estimation of prevalence and incidence of latent status. For this purpose we will frame the elaboration in terms of various measures of integrity violations by officials (which will hold as a test case for the method to be proposed in this text; see following chapters).

Integrity can be defined as the quality of acting in accordance with the relevant moral values, norms and rules. Public integrity then, "denotes the quality of acting in accordance with the moral values, norms and rules accepted by the body politic and the public" (Fijnaut & Huberts 2002: 4; also see Huberts *et al.* 2004). Integrity is a quality of individuals (Klockars 1997; Solomon 1999) but, as is posed by Kaptein & Wempe (2002), it might be a quality of organizations as well. Additionally, ethics can be defined as the collection of values and norms, functioning as standards or yardsticks for assessing the integrity of one's conduct (Benjamin 1990). The moral nature of these values and norms refers to what is judged as right, just, or good conduct. An integrity violation thus breaches a certain law or code of acceptable behavior. Note, that in this respect, an integrity violation can be stigmatizing, incriminating and can even hamper on vested political alignments. Eight encompassing forms of integrity violations in the public sphere can be distinguished: corruption, fraud and theft, conflict of interest, improper use of authority, misuse and manipulation of information, discrimination and sexual harassment, waste and abuse of resources and private time misconduct (Huberts, Pijl & Steen 1999; Huberts, Lasthuizen & Peeters, forthcoming).[6]

To illustrate the discussion on the ability of various external measures to estimate prevalence and incidence of latent status, the ability of various measures to accurately gauge the level of corruption for a certain population is assessed. Corruption is "behavio[u]r on the part of officials in the public sector, whether politicians or civil servants, in which they improperly and unlawfully enrich themselves, or those associated with them, by the misuse of the public power entrusted to them" (Fijnaut & Huberts 2002: 4). Usually the level of corruption is estimated by using

proxy estimates or by research on official statistics, where various measures give various representations of the proverbial corruption 'iceberg'. The proverbial iceberg implies a certain sensitive problem, the lion's share of which finds itself in unobserved latent space, and of which only a part is observable. Various methods of measurement then give various levels of accurateness in representing true levels of corruption. Mostly, these endeavors take the form of reputation research, research on official statistics of criminal cases and internal investigations, or research on proxy estimates given by workers/officials on the extent of corrupt behavior in their work environment. Each will be discussed.

One of the most common and often discussed methods to measure corruption is reputation research. The common basis is that respondents considered being experts are asked to estimate the amount of corruption in their environment, usually their sector of society or their country. Most prominent is the last type, including Transparency International's Corruption Perception Index (CPI), which is an annual composite index of multiple surveys where the respondents are business people and country analysts, by which countries are ranked on the basis of their subjective notions regarding the level of corruption. By selecting specific respondents, these opinions and guesses are considered to be 'expert' opinions and 'educated' guesses.

The expert estimations have resulted in stable rankings of countries on the CPI.[7] Nevertheless, the stability of these expert estimations is also a pitfall in terms of validity. A reputation is something that is built up over the years, because exactly as the name 'reputation' says, it is about how a country or organization has come to be known. Therefore, reputations refer to general perceptions, not to specific (expert) experiences. In other words, the problem is that an order of reputations is based on already preconceived reputations. Known or felt reputations carry themselves into reputation research. This type of research is not about actual cases of corruption in a country, but about views on corruption from experts in the field. Views say something about reality but they must not be confounded with it. Corruption reputation research can provide interesting data regarding views on the extent of corruption and the mechanisms that underlie public wrongdoing. However, this type of research finds itself outside the proverbial 'iceberg' of corrupt administrative behavior, as it is concerned with opinions, instead of data on or evidence of actual cases of corruption, fraud or other integrity violations.

Official statistics on criminal cases form the top of our corruption iceberg, as in these cases the evidence of incidence is most fully

crystallized. There are several problems associated with these statistics however, when one wants to estimate prevalence and incidence of misbehavior. First of all, it could be that the criminal cases, which by law are directed towards individual suspects, are representative for even fewer corruption cases, as it is possible to have multiple suspects for one case of corruption. They provide data on offenders, not offences. There are no data available on this point. Most importantly, there is the issue of the 'dark numbers'. A part of what these statistics purport to measure goes unreported or unrecorded. With these data alone it is impossible to assess how criminal cases and possible convictions relate to the actual prevalence and extent of corruption (and other integrity violations) (Nelen & Nieuwendijk 2003). Data on criminal cases, as well as data on internal investigations, only provide circumstantial evidence regarding prevalence and incidence.

Research on internal investigations gives higher estimations of prevalence and incidence of corruption. Internal investigations might follow out of reports in the work environment, and can lead to the criminal cases we have just discussed. Research into internal investigations is valid for representing internal research activities, but says very little about the actual prevalence and extent of misconduct. There is a 'dark number' of misconduct in the organization, which can be expected to remain unknown to those conducting the investigation. An additional problem here is that the dark number varies for different types of integrity violations, as it is dependent on the chance of discovery and the willingness to conduct an internal investigation. This partly underlies the high incidence of investigations into, for example, the use of force by police officers (because any incidence of the use of force by an officer has to be reported and investigated) and the very low incidence of investigations into corruption.

Also, organizations that pay more attention to certain violations will be more prone to track and investigate these violations, so that they may appear more corrupt, while that might not be the case. In other words: a large number of investigations could say more about the level of attention and openness, than about the actual levels of corruption and other integrity violations (Huberts *et al* 2004). Thus research on internal investigations can raise an important interpretation problem: do more investigations into corruption and fraud indicate immaculate policy and attention or a highly corrupt institution?

Moreover, when the statistics on internal investigations are to be delivered by the organizations under investigation, there is always the

chance of nonresponse or the willing distortion of the figures to evade stigmatization or present the organization in a favorable manner.

Another type of research is the conduct of surveys with estimations of the extent of corrupt behavior in the work environment. This type of research, while still working with indirect estimates, focuses more directly on corruption itself by using perceptions of employees of certain behaviors in their working environment as a proxy estimate of the actual level of corruption in a certain organization. It often clarifies that there is much more corruption than the internal investigations detect (see Huberts, Kaptein & Lasthuizen 2005). The question is what the perceptions of employees tell us about the actual level of corruption and other breaches of integrity within organizations. This type of research too, has problems with regard to the validity of estimates regarding prevalence and incidence of misbehavior.

The first problem has to do with what respondents *can* observe. Employees can only have a limited view on what actually happens. This holds for some integrity violations more than for others. Unethical behavior that takes place outside the organization will be harder to observe than behavior that takes place inside the organization. This is particularly true for private time misconduct, but also for work that takes place outside the office (for instance police officers on patrol). The fact that offenders will try to hide their actions, (e.g. bribery) and that some violations have no victims or do not need a third party (e.g. fraud or theft), makes the validity problem of relying on perceptions as a true reflection of the actual level of unethical behavior even bigger.

The second problem has to do with what respondents *will* observe. When respondents are asked which types of integrity violations they perceive in their work environment, they will reflect in their answers their own specific experiences, but – without a doubt – more subjective impressions will play a role as well. Respondents need to label what they see. They need to be aware of the different aspects of the (asked) type of integrity violation and need to recognize behaviors as manifestations of the problem. Discrimination and sexual harassment are especially not always recognized as such. The attention management gives to unethical behavior (in policies, training, codes of conduct etc.) can increase awareness among employees and might be reflected in their observations as well.

Respondents' perceptions are not neutral. Besides awareness, the acceptability of behavior will play a role in estimating what happens. Although it cannot be known exactly how respondents' evaluations of the asked behaviors influence their estimations (they might lead to

underestimation as well as overestimation), it can be expected that it does make a difference if behavior is widely accepted, or not, within the organization.

The third problem has to do with what respondents *report.* Respondents might be reluctant to report what is really going on in their direct environment, because of loyalty to colleagues or loyalty to the organization, or fears over stigmatization or incrimination of oneself. The perceptions of employees are at best an indicator of the actual amount of integrity violations present in the organization.

The proverbial iceberg, depicting the actual structure of corruption and other integrity violations, is very complex: of all the actual cases of corruption, only a part makes itself felt or visible in the direct work environment; only a limited number of these cases are reported, leading to the occasional internal investigation; these investigations do not always lead to satisfactory evidence or a criminal investigation as a follow-up; and the criminal cases do not always lead to a conviction.

Although external measures can be valid indicators for the cases of corruption that eventually crystallize, the number of internal investigations that are conducted in a certain organization, or as a representation of the opinions on the extent of the corruption problem in a certain country or organization; their accurateness in revealing the true prevalence and incidence of corruption (and other sensitive behaviors) is likely to be (very) limited. These external measures may reflect the methodological and "organizational contingencies embedded in their production more clearly than they do the behavio[u]rs they are alleged to measure" (Lee 1993: 48).

Part of these contingencies can also be found in self-reports of victimization. The Netherlands Institute for the Study of Crime and Law Enforcement (NSCR) participated in a large-scale international comparative survey project called the International Crime Victims Survey (ICVS) that was conducted in 1989, 1992, 1996 and 2000. In more than 60 countries across the globe a representative sampling of 2,000 citizens each were questioned about their experiences with diverse forms of crime (Nieuwbeerta 2002). Since 1996 the respondents were also asked if they have experienced being a victim of corruption and by what type of government institution or civil servant (Nieuwbeerta, De Geest & Siegers 2002). Revealing that one has been a 'victim' of a corrupt practice is not as sensitive as revealing that one has been a perpetrator. Victimization research can catch data on the people who stumbled unwillingly upon various forms of misconduct but is not suited however, to uncover those

who were willingly involved in the realms of misconduct. There is ground to be gained on self-reports of deviant behavior.

THE PROBLEMS OF SELF-REPORTS OF DEVIANT BEHAVIOR

Outside documented behavior or reports, which do not directly assess one's own sensitive behavior, are less threatening to the reporting individual but they cannot give accurate estimates of prevalence and/or incidence of certain behaviors. Moreover, "[C]ertain kinds of information are not available through secondary data sources" (Hosseini & Armacost 1993: 464). For a more accurate exploration of latent space one would thus have to obtain unbiased self-reports of actual sensitive/deviant behavior.

Research on individual deviant behavior remains *terra incognita* in the science of misconduct, mainly because researchers have to overcome a problem that is felt in all research on latent variables, but which becomes pressing regarding individual behavior: non-response and evasive answer bias. To obtain accurate estimates of prevalence and incidence one needs valid self-reports, but self-representational concerns peak when one has to report one's own deviant behavior. The concerns regarding response and nonresponse bias become critical when asked for sensitive information, due to heightened concern over anonymity when using direct questioning of any sort (whether face-to-face, on a questionnaire or otherwise). The potential distortion due to this problem can easily be proven (as given by Lang, February 2004):

Consider a dichotomous population of which we draw a random sample of $i = 1,\ldots,i,\ldots n$ persons. Our goal is the estimation of a certain trait π_X ($0 < \pi_X < 1$) among the persons in the random sample. To this purpose each person in the sample is asked the question: "Do you possess trait X?"

Where

$$X_i \begin{cases} 1 & i\text{th person possessing trait X} \\ 0 & \text{otherwise} \end{cases}$$

and

$$Z_i \begin{cases} 1 & i\text{th person answering question with 'yes'} \\ 0 & i\text{th person answering question with 'no'} \end{cases}$$

When the question asks for (relatively) innocuous information (as for example the possession of a yellow colored coat)[8], we can expect correct answers to our question. This means that the chances of combinations of certain answers and possible possession of trait X (denoted as P), can be stated as

$$\begin{aligned}
&\text{P(yes answer | possession of trait X)} &&= P(Z_i = 1 \mid X_i = 1) = 1\\
&\text{P(yes answer | no possession of trait X)} &&= P(Z_i = 1 \mid X_i = 0) = 0\\
&\text{P(no answer | possession of trait X)} &&= P(Z_i = 0 \mid X_i = 1) = 0\\
&\text{P(no answer | no possession of trait X)} &&= P(Z_i = 0 \mid X_i = 0) = 1
\end{aligned}$$

so that

$$P(Z_i = 1) = P(Z_i = 1 \mid X_i = 1) \cdot P(X_i = 1) = \pi_X \tag{2.1}$$

which means that

$$\pi_X' = \frac{1}{n} \sum_{i=1}^{n} Z_i \tag{2.2}$$

is an unbiased estimator of the prevalence of π_X , where

π_X' = the estimation of π_X

$\sum_{i=1}^{n} Z_i$ = the sum of yes responses (note that a yes response is scored with 1)

n = the number of respondents in the sample

But what happens when we ask for sensitive information? Consider the situation that the sample we have just drawn consists of public civil servants, to which the following question is asked:

> "Have you ever improperly or unlawfully enriched yourself, or those associated with you, by the misuse of the public power that is entrusted to you?" (Our corruption example from above)

In this situation we can be sure that those not possessing trait π_X (corruption) will answer truthfully, but because a positive response is incriminating and/or stigmatizing we must consider that persons who do possess trait π_X, are not always going to give a truthful 'yes' response, so that:

$$\begin{aligned} &P(\text{yes answer} \mid \text{possession of trait X}) &&= P(Z_i = 1 \mid X_i = 1) = q_x < 1 \\ &P(\text{yes answer} \mid \text{no possession of trait X}) &&= P(Z_i = 1 \mid X_i = 0) = 0 \end{aligned}$$

where q_x denotes the underestimation of the possession of trait X, so that

$$P(Z_i = 1) = P(Z_i = 1 \mid X_i = 1) \cdot P(X_i = 1) = q_x \cdot \pi_x \tag{2.3}$$

which means that the estimator for the prevalence of π_x

$$\pi_x' = \frac{1}{n} \sum_{i=1}^{n} Z_i$$

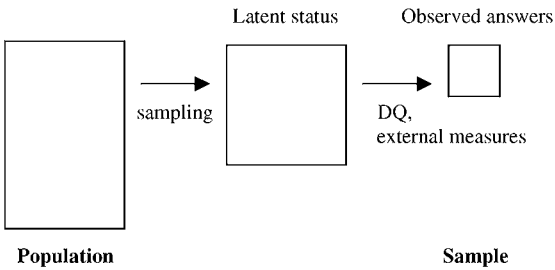
is biased with

$$E(\pi_x') = q_x \cdot \pi_x < \pi_x \tag{2.4}$$

The prevalence of trait X will thus be systematically underestimated. The smaller q_x (thus the smaller the probability that a person who possesses trait X, will answer 'yes' and disclose his or hers true status), the greater the underestimation and the greater the bias. This example deals with response bias only, but it is easy to see that nonresponse will further the distortion of results. The example makes clear that it is impossible to achieve accuracy consistent with the theory of random sampling when asking for self-reports of sensitive behaviors in a direct manner, because this exclusive use of probability theory cannot capture distortion due to nonresponse and response bias. It also makes clear that a careless design of sensitive surveys which hope on making use of self-reports, will be disastrous in terms of the accurateness of the information it will elicit (Hosseini & Armacost 1993). It is exactly the use of any form of direct questioning (DQ) in inquiries of a sensitive nature making use of self-reports, which makes respondents question the anonymity and feel the potential threat of stigmatization and/or incrimination. A great deal of the error found in surveys of a sensitive nature, which lean on the assumption that potentially stigmatizing or incriminating information can be elicited directly from the individual under investigation, is attributable to common method bias: variance in measures that is attributable more to the contingencies in the method used, rather than the constructs these measures purport to represent (Podsakoff *et al.* 2003); a problem also prevalent in the external measures. Theoretically, when taking into account the argument in this and the previous section, the accuracy between various external measures or self-reports which make use of

various forms of direct questioning, with regard to the estimation of a latent status can be represented as in Figure 2.2.

Figure 2.2: The Accuracy of External Measures or Direct Questioning in Estimating Latent Status of a Sensitive Nature



Adapted from: Van Den Hout 2004, p. 1

THE NEED FOR ACCURATE SELF-REPORTS

When one wants to obtain accurate estimates of prevalence and/or incidence of specific sensitive behaviors in certain populations one has to ask for self-reports instead of relying on external measures, because the latter are more accurate reflections of the organizational contingencies that lead to their production then of the extent of the behaviors they ought to measure. However, self-representational concerns in research regarding self-reports of sensitive topics become critical when a more traditional design is handled which makes use of direct questioning of any sort under the assumption that threatening information can be obtained directly from the respondent. Concerns over systematic distortion due to systematic non-sampling errors then become pressing. What is needed is specific methodological effort to curb these effects. More specifically, in order to avoid or minimize the potential dangers of non-sampling errors so as to come to more accurate population estimates of prevalence and incidence, a data collection approach is needed that, when asking for self-reports of sensitive behavior: (a) has the potential to draw hesitating and suspicious respondents across the imaginary line

when conducting research regarding sensitive topics; and (b) protects the respondents' information in such a way that they have enough trust in the method to cooperate without willingly distorting their response.

As stated before, the use of elementary probability theory with regard to most research has almost exclusively been directed towards sampling procedures so as to curb sampling bias. However, it is impossible to achieve accuracy consistent with the theory of random sampling, because this exclusive use of probability theory cannot capture distortion due to nonresponse and response bias. Therefore, it is proposed to direct the attention of elementary probability theory also to the conduct of asking sensitive questions itself. Using a scheme of misclassification via probability makes possible the dissemination of sensitive information from individuals under the protection of full anonymity, potentially reducing response and nonresponse bias. The technique which makes possible the investigation of private activities which have a latent status due to sensitivity, through an examination of individual disclosure on one's own behavior under the protection of full anonymity from the use of probability, is known as 'Randomized Response': to which we now turn.[9]

NOTES

[1] Clearly, the threat sensitive research poses can be felt in all aspects of scientific activity, ranging from the formulation of research questions to the dissemination and publication of research findings, and can hamper on both participant and researcher as well as on society or certain groups at large. This asks for specific ethical dimensions in the conduct of sensitive research. These ethical dimensions will not be specifically dealt with here. For a review and taxonomy of ethical dimensions of socially sensitive research we confine by referring to Boruch & Cecil (1979) and Sieber & Stanley (1988).

[2] Measurement error is discussed here in terms of variable and systematic error. This is a statistical interpretation of measurement theory, as this text will try to pose statistical solutions to certain measurement problems. It is acknowledged that from a psychometric perspective however, which draws on alternative notions of test theory, only variable errors exist. For a small discussion, see Groves (1991: 6-9).

[3] In much of the European literature the term 'survey' is used as a synonym for 'questionnaire'. The Anglo-Saxon literature has a broader view of surveys, meaning all systematic inquiries into human attitudes, preferences of behaviors. It is stressed here, that the meaning of 'survey' as used in this text, is in no way limited to questionnaires, method of conduct or finite populations. The term does however have a quantitative connotation in the remainder of this chapter.

[4] It must be noted that in effect socially desirable behavior can be overestimated due to the described factors. Sensitive information as respondent income can also be, due to the urge to

present oneself in a favorable way, overestimated. While this text is preoccupied with behavior of which can be expected that they will be underestimated, the basic premises of the argument in this chapter and the technique described in following chapters can also be applied in these situations.

[5] Parts of this section will appear elsewhere in a chapter co-authored with L.W.J.C. Huberts and K. Lasthuizen for an edited volume by Charles Sampford, Arthur Shacklock and Carmel Connors (accepted, planned publication 2005).

[6] There is no intention to start a discussion on integrity and related violations, or to review the debate regarding these issues. As this text is focused on accurate measurement, such would be out of scope for this text, and – as will become clear in Chapter 5 – different behaviors will be measured which are widely accepted as clear breaches of integrity in the group in which these will be assessed. A review of integrity as such is thus not deemed necessary.

[7] See http://www.transparency.org for publications of Transparency International and the CPI country rankings.

[8] Although it can be argued that every question is sensitive to some, it will be clear that the possession of a yellow colored coat holds, for most, no potential threat of stigmatization or incrimination.

[9] The Randomized Response technique can thus be seen as a de-jeopardizing technique. While there are more such techniques, ranging from administrative strategies as archival procedures to other statistical setups such as 'multiplicity techniques', some cannot be employed in a one-to-one setting while others do not provide full anonymity. It is the great advantage of the Randomized Response technique that it makes possible the investigation of essentiality private activities where the amount of knowledge available to others or in other sources is severely limited. As this text is concerned with the use of probability theory so as to inoculate individual responses, a full review of other de-jeopardizing techniques is not deemed necessary. We confine by referring to Lee (1993) for an overview of de-jeopardizing techniques.

3. RANDOMIZED RESPONSE: USING PROBABILITY THEORY TO INOCULATE INDIVIDUAL RESPONSES

When one wants to assess prevalence and incidence of certain sensitive behaviors one runs into the problem of differential validity, which centers around self-reported measures of individual sensitive/deviant behavior on the one hand and external measures on the other. As external measures are severely limited in their accurateness in representing prevalence and incidence of sensitive behaviors, and as certain kinds of information are not available through secondary data sources, one has to turn to the collection of self-reports.

With self-reports one (again) runs into the problem of observational statistics in the social and behavioral sciences (as opposed to experimental statistics): the reliance on observations and/or reports given by respondents who can systematically distort results. In self-reports regarding sensitive behavior the respondent has self-representational concerns, potentially leading to response (evasive answers) and nonresponse bias (item-specific or full refusal to answer questions). To give an accurate estimate of prevalence and incidence of certain sensitive behaviors, the potential distortion by the non-sampling biases of nonresponse and response effects thus have to be reduced in inquiries of a sensitive nature which ask for self-reports.

'Randomized Response' (RR) refers to techniques, which as a core characteristic use the insertion of random error by an element of chance to provide a respondent optimal privacy protection when answering questions of a sensitive nature. Through an artificially generated variability in responses, individual answers are inoculated as the probability of misclassification cancels out individual meaning of

answers. When fully protected from stigmatization or incrimination, it can be expected that respondents cooperate and do so without willingly distorting their responses, making possible the investigation of private activities through an examination of individual disclosure, in order to come to more accurate estimates of prevalence and incidence. This chapter delves into RR by describing the technique and its use of probability theory as to more accurately estimate latent status. To enhance comprehensibility first an intuitive explanation of RR will be given, as well as the basic outline of two 'classic' RR models, preceding a more abstract theoretical explanation of the technique and its underlying assumptions. After this more in-depth explanation of RR, further methodological advances and RR performance in field-tests as well as gaps in RR literature and research will subsequently be dealt with, serving as a stepping-stone for the methodological framework provided in Chapters 4 and 5.

INTUITIVE EXPLANATION OF THE RANDOMIZED RESPONSE TECHNIQUE

In a firing squad, not all the marksmen have live bullets. Upon capital punishment, one of the sharpshooters at random receives a rifle that is loaded with a blank. The purpose is that each shooter can maintain to have fired the blank bullet, when his conscience would force him to do so. The element of chance (the probability that one has fired the blank bullet) makes it possible to rationalize that the individual act of using the rifle, was not necessarily fatal for the unlucky one (example taken from Fox & Tracy 1986: 17). The same use of probability theory can be applied to surveys of sensitive information. To illustrate this claim, consider the following hypothetical situation.

One finds oneself in a room full of academic colleagues. A speaker asks the attendants to raise their hands if they have ever twisted data in ways so that the results became statistically significant or conform a hypothesis (example taken from Lensvelt-Mulders & De Leeuw 2002a). Probably nobody will raise his or her hand. But then the situation is changed. The speaker asks every attendant to flip a coin and to look if the throw tells heads or tails, without revealing the outcome to others. Subsequently everyone who has tails and/or ever has reworked data in an 'un-academic' way, is asked to raise his or her hand. Statistically, half of all attendants will raise their hand due to a 'tails' score on their coin flip. When we assume 100 attendants that would be 50 persons raising their

hand. When 56 persons raise their hand, we can very easily compute that 12 persons in the auditorium (12%) ever have unscientifically revised their data. The surplus of 6 represents those who obtained heads on the coin but did rework their data in academically unaccepted ways. We can assume 50 academics with heads, meaning a 12% prevalence (6 out of 50) of academic lying with statistics. The twelve percent prevalence exists for those obtaining heads; as well as for those obtaining tails on their coin flip. Out of the total group of 100 attendants, we can then estimate 12 persons to have ever unscientifically revised their data. But exactly who those persons are however, will never be revealed, as basically all the persons with tails on their coin flip protect the privacy of the frauds.

While this is a strongly simplified example of the RR technique, the basic argument is clear. The method of probability guarantees that individual respondents cannot be identified on the basis of their response to a certain sensitive question. The anonymity of responses is thus maintained. The rationale is that when respondents are assured that their privacy is guarded, they will be more inclined to cooperate and to respond truthfully. Under the assumption of non-evasive cooperation and as the statistical relations between the posed question and the observed answers are partly known, it is possible to obtain certain accurate estimates of aggregate parameters.

The above-described procedure can also be applied in a one-to-one setting. This was the initial idea of Stanley Warner (1965): using the element of chance stemming from a randomization device to inoculate individual responses to sensitive inquiries and thus reduce non-sampling bias.

THE WARNER MODEL

Consider the problem given in Chapter 2, where we want to estimate the prevalence of a certain sensitive trait π_x. Regarding trait X the population is dichotomous, either possessing X or not possessing X (-X). When drawing a simple random sample, one cannot achieve accuracy consistent with random sampling, as it cannot capture distortion due to nonresponse and response bias (as given in equations 2.3 and 2.4).

Warner (1965) proposed in a seminal article a specific use of probability theory, directed at the conduct of asking sensitive questions itself, so as to overcome the problems of distortion of results due to nonsampling biases. His suggestion was that the connection between

question and response could be stigmatizing or incriminating in nature, more so than the 'yes' or 'no' response itself. Probability could thus be used to curb self-representational concerns arising from the connection between a certain question and response. For this purpose, Warner proposed the RR technique.

In the RR technique each respondent in a sample receives a randomization device that follows a Bernoulli distribution[1], with known probability p. On the basis of this randomization device (which can be a die, a deck of cards or otherwise) the respondent answers one of two questions: (1) "Do you possess X?" or its converse (2) "Do you not possess X?" (-X). As for example:

> Question 1 (X): "Have you ever improperly or unlawfully enriched yourself, or those associated with you, by the misuse of the public power that is entrusted to you?"
> Question 2 (-X): "Have you never improperly or unlawfully enriched yourself, or those associated with you, by the misuse of the public power that is entrusted to you?"

Question 1 and 2 have respective selection probabilities p and $1 - p$ ($0 < p < 1$). The researcher, as part of the design, chooses the parameter p.

Each respondent may thus for example be supplied with a deck of cards in which each card is imprinted with the question regarding X or -X in proportions p and $1 - p$. The respondent is asked to draw one card and to reply with 'yes' or 'no' to the respective question imprinted on the card. This specific question (the outcome of the randomization) should be known only to the respondent. As only the respondent is aware which of the two logical opposite questions is answered, the response is believed to be no longer revealing. When a response is no longer revealing, truthful cooperation can be assumed.

When assuming truthful responses, the total proportion of affirmative responses can be expressed in terms of possession of the sensitive attribute in question (as given by Fox & Tracy 1986: 19):

$$P(\text{Yes}) = P(\text{Question1})P(\text{Yes} \mid \textbf{Question1}) + P(\text{Question2})P(\text{Yes} \mid \textbf{Question2})$$

or, more formally

$$\lambda = p\pi_x + (1 - p)(1 - \pi_x) \tag{3.1}$$

where

λ = probability of 'yes' response = P(Z_i = 1)
π_x = the probability of possessing the sensitive attribute in question
p = probability of selection of (sensitive) question

When we denote the total number of affirmative responses in our sample by $\sum Z_i$, an unbiased estimator of λ can be given by (similar to equation 2.2)

$$\lambda' = \frac{1}{n} \sum_{i=1}^{n} Z_i$$

denoting the observed sample proportion answering 'yes'.[2] In terms of known p and λ (e.g. λ'), an unbiased estimate of π_x (the true probability of X in the population) can be given by solving equation 3.1 for π_x:

$$\pi_x' = \frac{(\lambda' + p - 1)}{(2p - 1)} \qquad (p \neq .5)^3 \tag{3.2}$$

with sampling variance

$$\text{var}(\pi_x') = \frac{\pi_x(1 - \pi_x)}{n} + \frac{p(1 - p)}{n(2p - 1)^2} \tag{3.3}$$

The expression for var (π_x') involves the unknown parameter π_x. The variance of π_x' can be unbiasedly estimated by substituting π_x for its unbiased estimate π_x':

$$\text{var}'(\pi_x') = \frac{\pi_x'(1 - \pi_x')}{n - 1} + \frac{p(1 - p)}{(n - 1)(2p - 1)^2}. \tag{3.4}$$

The element of chance thus inoculates individual responses, but at the same time the technique provides the researcher with sufficient data for analysis. The statistical properties regarding the parameters p and λ', lead the estimate regarding parameter π_x theoretically to become an unbiased estimate of the true population prevalence regarding sensitive trait X. Warner thus posed a true innovation in survey research, proposing an alternative procedure rather than an amendment to conventional techniques. But despite the innovation, Warner's model is somewhat crude.

The first crudeness in the Warner model is the inflation of variance. The term $\pi_x(1 - \pi_x)/n$ denotes normal sampling variance in a DQ setting, where $[p(1 - p) / n(2p - 1)^2]$ is additional variance due to the insertion of

random error via the randomization procedure. This means that the dispersion of values from the expected value increases, resulting in a less precise and thus less efficient estimation. As the variance is a function of parameter p, the efficiency of an estimate is dependent (in a certain model) on its choice. When $p = 1$ or 0, the variance will equal normal sampling variance, and the estimate will thus be 'efficient'. But in this setting, as Chaudhuri and Mukerjee (1988: 5) point out, a 'yes' or 'no' response by the respondent will be maximally revealing as a specific response implies a belonging to a certain group with a very high probability (X or -X).[4] The privacy of the respondent is then not protected and one can assume, as this situation equals a DQ setting, reporting on a sensitive question not to be truthful. As p will be closer to .5, a respondent will increasingly see his privacy protected[5], but the variance will reach maximum inflation. 'Efficiency' and level of 'respondent protection' are thus inversely related in a RR setting (a point to which will be returned more deeply in Chapter 4). But, due to the construction of the Warner model, the variance of the estimator is still considerably inflated when p approaches 1 (Fox & Tracy 1986: 20; Lensvelt-Mulders & De Leeuw 2002b).

The second crudeness in Warner's model is that it still carries the burden of potential distortion because both questions (X and -X) are sensitive in nature (sensitive complements). In this setup "the threatening nature of the topic itself may encourage a refusal to participate" (Fox & Tracy 1986: 20) or encourage evasiveness. It is mostly this issue which has led to a first methodological innovation that culminated in the second 'classic' RR model: the Unrelated Question Model.

THE UNRELATED QUESTION MODEL

Simmons (see Horvitz, Shah & Simmons 1967; and Greenberg *et al.* 1969) proposed to curb the crudeness of posing two sensitive inversely related questions in the Warner model by pairing a sensitive question (X) to a question completely unrelated to the potentially stigmatizing trait (Y). As for example:

> Question 1 (X): "Have you ever improperly or unlawfully enriched yourself, or those associated with you, by the misuse of the public power that is entrusted to you?"
> Question 2 (Y): "Were you born in the Netherlands?"

The latter statement is unrelated to the former and an affirmative or negative response can be said not to be stigmatizing or incriminating. Now, not only the proportion of the sensitive attribute π_x has to be estimated, but also the prevalence of the non-sensitive attribute π_y (proportion of 'yes' responses to innocuous question). Thus two independent random samples of size $n1$ and $n2$ are required to estimate π_x and π_y. As in the Warner model, each respondent (in each sample) receives a randomization device with known probability p. On the basis of this randomization device, the respondent answers 'yes' or 'no' to the question regarding X with probability pi or to the question regarding Y with probability $1 - pi$, where $i = 1,2$ (sample 1 or sample 2) and where the probability of selecting the question regarding X in sample 1 is not equal to the analogous probability in sample 2 ($p1 \neq p2$). Again, which question is answered is only known to the respondent.

The probability of a 'yes' response in the ith sample is then

$$\lambda i = pi\,\pi x + (1 - pi)\pi y \qquad (i = 1,2) \tag{3.5}$$

The overall probability of a 'yes' response in the ith sample can unbiasedly be estimated by λi', the observed sample proportion answering 'yes' in the ith sample (again as in equation 2.2). The prevalence of sensitive attribute X can then be unbiasedly estimated by solving 3.5 for π_x with λ'

$$\pi x' = \frac{\lambda 1'(1 - p2) - \lambda 2'(1 - p1)}{p1 - p2} \tag{3.6}$$

with an unbiased estimate of variance

$$\text{var}'(\pi x') = \frac{\left[\dfrac{\lambda 1'(1 - p2)^2\,(1 - \lambda 1')}{n1 - 1} + \dfrac{\lambda 2'(1 - p1)^2\,(1 - \lambda 2')}{n2 - 1} \right]}{(p1 - p2)^2} \tag{3.7}$$

The great advantage of this specific method is that it may enhance truthful reporting more so than the Warner model, due to the greater comprehensibility of using an innocuous question as opposed to using sensitive complements (Horvitz, Shah & Simmons 1967; Greenberg *et al.* 1969). The disadvantage is its inefficiency, which exceeds the inefficiency of the Warner model under most design parameters, making very large samples necessary to obtain acceptable confidence intervals (Capaciteitsgroep Methodenleer en Statistiek, Universiteit Utrecht 2003: 18).

The efficiency problem can be curbed when the innocuous question regarding Y involves a characteristic whose true proportion π_y is known (Greenberg *et al.* 1969; also see Horvitz, Greenberg & Abernathy 1976: 183-185). An example may be to ask for a person's month of birth. Birth registration files contain the proportion of the population born in each month of the year (clearly, this proportion will approximate 1/12). When π_y is known, just one sample is required to estimate π_x. Thus one random sample from the population is drawn and each respondent reports a 'yes' or 'no' to either the question regarding X or Y on the basis of the outcome of a randomization device with respective probabilities p and $1 - p$. The probability of a 'yes' response can then be stated by

$$\lambda = p\pi_x + (1 - p)\pi_y \tag{3.8}$$

and equations 3.6 and 3.7 simplify to

$$\pi_x' = \frac{\lambda' - (1 - p)\pi_y}{p} \tag{3.9}$$

and

$$\text{var}'(\pi_x') = \frac{\lambda'(1 - \lambda')}{(n - 1)p^2} \tag{3.10}$$

The design where π_y is known is very efficient and potentially better understandable for the respondent in comparison with the Warner model. The disadvantage is that is has proven to be very difficult to find enough innocuous questions where the parameter π_y is known to accompany sensitive questions in surveys which develop inquiries into a relatively substantial number of sensitive traits (Capaciteitsgroep Methodenleer en Statistiek, Universiteit Utrecht 2003: 18). In Chapter 4 we will build a model of RR on an extension of the Unrelated Question Design (UQD), which subsequently curbs the above stated problem by letting the randomization device provide for π_y.

THEORETICAL EXPLANATION OF THE RANDOMIZED RESPONSE TECHNIQUE

The RR method has in fact two basic premises (Van Den Hout 2004: 3), which are accompanied by underlying assumptions regarding a respondent and his or her disclosure of true status. The first premise is

that the insertion of random error via a randomization design protects the privacy of the respondents. The probability of a required true statement (regarding the sensitive question) is less than 1, which is less revealing than the requirement to answer truthful under the condition of $p = 1$. As part of the data is misclassified, individual meaning of a certain response is canceled out, meaning that the individual cannot be identified on the basis of his or her answer. The assumption is that if the respondent understands the privacy given by the method or trusts the method to provide full anonymity in the disclosure of information (whether fully understanding it or not), he or she is relieved from self-representational concerns and will be more inclined to cooperate and will do so in a non-evasive manner (Greenberg *et al.* 1969; Deffaa 1982; Chaudhuri & Mukerjee 1988; Landsheer, Van Der Heijden & Van Gils 1999; Lensvelt-Mulders 2003). When a respondent is directed to answer a sensitive question in a RR design, and he possesses sensitive trait X for which an estimate is desired, he is believed to answer 'yes' as he understands that information is furnished on a probability basis only.

Now we can restate the problem of estimating π_X as given in Chapter 2, so that upon asking a sensitive question now:

$$X_i \begin{cases} 1 & i\text{th person possessing trait X} \\ 0 & \text{otherwise} \end{cases}$$

and

$$W_i \begin{cases} 1 & i\text{th respondent who randomly chooses sensitive question}^6 \\ 0 & \text{otherwise} \end{cases}$$

and

$$Z_i \begin{cases} 1 & i\text{th person answering question with 'yes'} \\ 0 & i\text{th person answering question with 'no'} \end{cases}$$

Given the assumptions described above regarding the inoculation through probability theory which relieves self-representational concerns, the probability of a 'yes' answer when possessing X and when asking for a self-report regarding this sensitive trait can be represented as

$$P(Z_i = 1 \mid W_i = 1, X_i = 1) = 1$$

so that π_X can be unbiasedly estimated through λ'.

Thus in a RR design "uncontrolled uncertainty caused by self-conscious subjects is exchanged for controlled uncertainty produced by a simple chance event" (Kundert 1989: 409). This leads us to the second premise of RR designs: as the conditional probability of misclassification is known through the specific model used and the choice of p (and possibly π_y in the UQD), the data can be analyzed in order to arrive at unbiased population estimates. To make this clearer, we will look into the properties of random error and its relation to statistical parameters.

Let X_o be the observed value of our sensitive trait, X_t being its true value and ε the random error we induce with our method of randomization. Under the assumption that respondents will state their true status when directed to answer the sensitive question on a probability basis, the proportion of true values (X_t) equals p with $p < 1$. The remaining proportion is random error. While the remaining proportion may be seen as consisting of true statements on a non-sensitive or complementary question, these statements are randomly induced error surrounding the parameter we want to estimate. As random error is independent from all other variables[7] it does not affect the mean, so that the expected (E) observed value equals the expected true value (also see Carmines & Zeller 1979):

$$\begin{aligned} E(X_o) &= E(X_t + \varepsilon) \\ &= E(X_t) + E(\varepsilon) \\ &= E(X_t) \end{aligned} \tag{3.11}$$

stated in RR terms

$$E(X_i) = \pi_x$$

We are thus theoretically able in a RR design to give a fully unbiased estimate of sensitive trait X, as the expected value of the sampling distribution of the statistic is equal to the parameter of which it is an estimate. The insertion of random error does however affect the variance of the estimate. Let V denote the variance, then

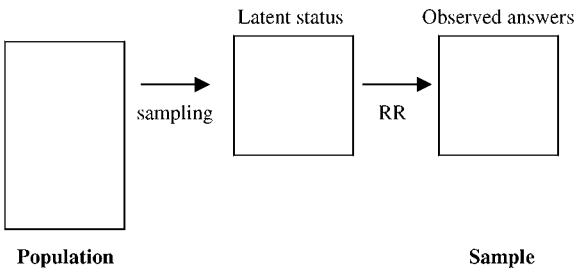
$$\begin{aligned} V(X_o) &= V(X_t + \varepsilon) \\ &= V(X_t) + V(\varepsilon) \end{aligned} \tag{3.12}$$

Equation 3.12 makes clear that variance is inflated due to the insertion of random error, which we already saw in the exploration of the Warner and Unrelated Question RR models. This means that larger samples have to be drawn in comparison with DQ (when assuming completely honest

answers in a DQ setting) to obtain similar levels of precision in terms of confidence intervals. The level of variance in a RR setting is thus dependent on the specific model and randomizer used, the choice of p and the scope of the sample.

Taking into account the argument and assumptions in the preceding text, the theoretical accuracy of RR in estimating a certain latent status of a sensitive nature can be presented as in Figure 3.1.

Figure 3.1: The Accuracy of Randomized Response in Estimating Latent Status of a Sensitive Nature



Source: Van Den Hout 2004, p. 1

It must again be stated here that validity/accuracy is a matter of degree. The assumptions that everyone understands and/or trusts the RR procedure and that everyone will disclose his or her true status on a probability basis are very strong. They will not have an encompassing status in actual research. However – as we will also see in the following section – it can be expected that more people will give an answer that corresponds to their true status in comparison with a DQ setting, due to the level of protection that is offered with RR. In effect, RR purports to keep the error term q_x (denoting the probability that someone will state his or her true status when possessing X; see equation 2.3) as big as possible. Meaning that the probability of a 'yes' answer when possessing trait X will approximate 1 as closely as possible and that the underestimation of X will be kept as small as possible. When the random error inserted by the randomization device is subsequently filtered out,

RR designs are able to come to more accurate estimates of prevalence (and as we will see; also incidence) of certain sensitive behaviors as opposed to inquiries using DQ settings or external measures.

The two basic premises of RR point us to two important issues (Van Den Hout 2004: 3). The first is the design of a RR model and the accompanying randomizer. This issue also includes the balance between efficiency, respondent protection and the choice of parameters. The second issue revolves around the mode of analysis and interpretation of RR data. Both issues will be thoroughly dealt with in Chapter 4 and 5 respectively.

FURTHER METHODOLOGICAL ADVANCE AND FIELD-TESTS OF THE RANDOMIZED RESPONSE METHOD

The lion's share of RR literature in the 40 odd years of the technique's existence has been devoted to mathematical and statistical technicalities regarding improvement and extensions of the method. Many publications have sought to refine the 'classic' Warner model and UQD mostly by proposing modifications so as to come to a more efficient desired estimate of π_x (see for example Moors 1971; Folsom *et al.* 1973; Lanke 1975; Raghavarao 1978; Kim & Elam 2005). Also, many have proposed alternative procedures to the Warner and Unrelated Question designs for indirect questioning making use of probability theory (cf. Takahasi & Sakasegawa 1977; Kuk 1990; Mangat & Singh 1990; Mangat 1994; Chua & Tsui 2000; Padmawar & Vijayan 2000; Singh 2002; Christofides 2003; Chaudhuri 2004; Christofides 2005; Kim, Tebbs & An, article in press).

The Warner model and the UQD as presented in the previous pages, as well as most of the literature referred to above, deal with a single dichotomous attribute. The RR technique can be extended to deal with sensitive characteristics having more than two categories. Several sensitive characteristics of interest are still categorical but have an essentially polychotomous nature, as for example sexual preference (which can be categorized in heterosexual, homosexual and bisexual). Abul-Ela *et al.* (1967) were the first to demonstrate a RR model based on an extension of the Warner technique, able to handle polychotomous populations. Additionally, the RR technique can be extended to deal with multiple sensitive attributes simultaneously, as in some investigations several incriminating or stigmatizing traits have to be considered in relation (see Tamhane 1981).

An even greater improvement of the RR technique came with its extension to quantitative measures. The 'classic' models as well as the extension of RR techniques to polychotomous and multiattribute situations are examples of qualitative RR techniques, where the characteristics investigated are inherently nominally categorical and do not involve numerical values as such. They are preoccupied with the investigation of prevalence. It is the establishment of quantitative RR techniques, which makes possible the "investigation of incidence rather than just prevalence" (Fox & Tracy 1986: 44) as they attempt to measure frequency (or intensity) of discrete or continuous quantitative traits. To stay with our corruption example as used several times in this and the previous chapter: not only do we want to know if someone has ever improperly or unlawfully enriched him or herself, or those associated with them, by the misuse of the public power that is entrusted to them, but also how many times one has done so. Greenberg *et al.* (1971) were the first to suggest a quantitative RR model on an extension of the UQD. For further reading on quantitative RR models, we confine by referring to Poole (1974), Sen (1974), Liu, Chow & Mosley (1975), Pollock & Bek (1976), Eichhorn & Hayre (1983) and Bar-Lev, Bobovitch & Boukai (2004).

Much of the above given literature has stimulated innovation in the administration of the RR method. While early usage of the RR always involved a researcher providing a randomization device to the respondent, Orwin & Boruch (1982) and Stem & Steinhorst (1984) have explored alternative administration techniques where the researcher would not have to be physically present (application via mail or telephone), where the respondent could supply for the randomizing device (like the coin from the example given at the beginning of this chapter) and where the randomizer itself would not have to have a physical nature (using proxy randomizers as for example the second digit of a respondents telephone number).

In its comparatively short history of existence, the RR method has also spurred extensive publication on various miscellaneous topics, as for example the use of Bayesian estimation in RR modeling (Winkler & Franklin 1979; Spurrier & Padgett 1980; O'Hagan 1987); the establishment of a unified framework for RR, based on an extension of general linear regression models (Warner 1971; Bellhouse 1980); the combination of RR with forms of DQ (Gupta, Gupta & Singh 2002; Arnab 2004; Chaudhuri & Saha, article in press; Padmawar, article in press); and the use of RR in data-mining and cryptography (Du & Gopalakrishna 2001; Du & Zhan 2003; Ambainis, Jakobsson & Lipmaa

2004). For an in-depth review of most technicalities touched upon in this section we refer to Deffaa (1982), Fox & Tracy (1986), Chaudhuri & Mukerjee (1988) and Tracy & Mangat (1996).

Methodological innovation carries the danger of absorption with the technical fix (Lee 1993: 207), as is also the case with the RR literature. Notwithstanding, there have been numerous field-tests of the RR method. Especially in early applications of the technique, the mode of administration was more exploratory, preoccupied with the applicability of the technique in the field and the obtainment of population estimates of certain topics previously overlooked by social science. Most of these substantive applications revolved around the topic of induced abortion (see for example Horvitz, Shah & Simmons 1967; Abernathy, Greenberg & Horvitz 1970; Liu & Chow 1976a; Shimizu & Bonham 1978).

A most interesting question is of course if the theoretical advantage of RR – more accurate population estimates in comparison with more traditional questioning techniques – can also be proven in field research efforts. Two sorts of studies have been directed towards this issue: comparison studies and validation studies (Fox & Tracy 1986; Lenvelt-Mulders *et al.* 2005). Comparative studies compare RR with more traditional questioning techniques on their relative success in obtaining accurate estimates of sensitive behaviors. Mostly, these studies view higher estimates as the more accurate ones, as no criteria measures for which true values were known, could be used. Validation studies have the same mode as the comparison studies, but here the results can be compared to a known criterion as in validation studies the true status of each individual in the sample is known.

Comparison studies have been conducted to establish the relative merit of RR in obtaining estimates of wide-ranging sensitive issues such as induced abortion (I-Cheng, Chow & Rider 1972; Lara *et al.* 2004); sexual history (Krotki & Fox 1974); substance use and abuse (Goodstadt & Gruson 1975; Zdep *et al.* 1979; Duffy & Waterton 1984; Danermark & Swensson 1987); child abuse (Zdep & Rhodes 1976); racial and political opinion (Wiseman, Moriarty & Schafer 1975); tax ethics and fraud (Larkins, Hume & Garcha 1997; Houston & Tran 2001; Gils *et al.* 2003); rule transgressing for instrumental laws (Elffers, Van Der Heijden & Hezemans 2003); academic cheating and whistle blowing (Burton & Near 1995); and stealing and other criminal behaviors (Beldt, Daniel & Garcha 1982). To date, there have also been about half a dozen validation studies. Again, these studies examined wide ranging phenomena, such as drunken driving, involvement in bankruptcy and voting registration (Locander,

Sudman & Bradburn 1976); course failure in college (Lamb & Stem 1978); arrest history (Fox & Tracy 1980; Tracy & Fox 1981); and social security fraud (Van Der Heijden *et al.* 1998, 2000). Lensvelt-Mulders *et al.* also point to a validation study conducted by Kulka, Weeks and Folsom (1981).

Although there is ample evidence that RR leads to less biased estimates of sensitive behaviors (see for example Krotki & Fox 1974; Tracy & Fox 1981; Van Der Heijden *et al.* 1998, 2000), the evidence is not conclusive as some comparison as well as validation research efforts could not report on the superiority of the RR method (see for example Larkins, Hume & Garcha). Part of the variability in research outcomes is due to the lack of consensus regarding best method and protocol. This variability in outcomes formed the motivation for two formal meta-analyses on the relative merit of RR in comparison and validation settings performed by Lensvelt-Mulders *et al.* (2005).

A meta-analysis statistically combines results from different studies so as to assess (the performance of) a dependent variable or condition. The main goal of the meta-analyses by Lenvelt-Mulders *et al.* (2005) was to assess if RR produces more valid population estimates of sensitive topics than conventional designs making use of DQ (as for example the questionnaire, telephone interviewing and face-to-face interviews). The dependent variable of the meta-analysis on individual validation studies was the percentage of incorrect answers, meaning: the difference between the population probability of 1 and the point estimate (for the various research conditions). In comparison studies higher estimates are seen as more valid estimates. The effect measure of the meta-analysis on these studies is based on the difference between the standard normal value (Z) of π_x in the RR and control conditions (also referred to as standardized difference score for proportions). For the validation studies a small discrepancy (between population probability and point estimate) indicates a high validity in estimating population parameters. The setup of the effect score for the comparison studies means that a significant positive score points to RR as providing more valid populations estimates in comparison with non-RR conditions. The research by Lenvelt-Mulders *et al.* contained multiple studies, research conditions (RR and various conventional DQ designs) and sensitive items, so that a sequenced multilevel meta-analysis was performed on both validation and comparison setups. The results are given in Table 3.1.

The column denoted with M0 gives the first test-sequence with an intercept only model. The residual variances (σ^2_{study} and $\sigma^2_{condition}$) in the meta-analyses for both comparison and validation studies are

significantly greater than zero, which implies that the differences in results across research conditions (RR and other) and studies are systematic, as they cannot be explained by sampling variance alone.

Table 3.1: Lensvelt-Mulders et al. Multilevel Meta-Analysis for Individual Validation and Comparative Randomized Response Studies

Meta-Analysis for Individual Randomized Response Validation Studies				
Step	n	M0 (Intercept only)	M1 (Conditions Added)	M2 (Sensitivity Added)
Intercept		.42 (.09)*		
RR	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^2study		.042 (.028)*	.042 (.029)	.025 (.018)
σ^2condition		.023 (.010)*	.018 (.008)**	.013 (.005)**
Meta-Analysis for Comparative Randomized Response Studies				
Step	n	M0 (Intercept only)	M1 (Conditions Added)	M2 (Sensitivity Added)
Intercept		.28 (.077)**		
Telephone	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)
Sensitivity				.07 (.023)**
σ^2study		.072 (.029)**	.17 (.05)**	.15 (.049)**
σ^2condition		.031 (.009)**	.02 (.01)*	.03 (.008)*

NOTES: n = number of data collection conditions; beta's given in columns with standard errors in parentheses. RR = Randomized Response; CASI = Computer Assisted Self-Interview; Scenario = scenario design; Unmatched count = unmatched count technique[8]. In the meta-analysis regarding the comparative studies the effect variable is represented by dummies, which represent the differences between RR and the more conventional data collection techniques; 'telephone' then is the dummy for the difference between RR and telephone interview results regarding their estimate of a sensitive trait. σ^2study = variance of residual errors across studies; σ^2condition = variance of residual errors across data collection conditions within studies. * $p \leq .05$; ** $p \leq .01$.
Source: Lensvelt-Mulders et al. 2005, p. 336, 338.

The residual variances justify the inclusion of a set of dummy variables representing the various (more traditional) data collection methods at the condition level giving the second sequence (M1), so as to delve into their explanation. The meta-analysis for the validation studies then shows that all data-collection methods underestimate sensitive traits in a certain population, but that the use of RR leads to the smallest underestimation (38 percent across all studies). The analogous analysis for the comparison studies indicates similar results by pointing to significantly more valid population estimates obtained by RR in comparison with face-to-face interviewing and self-administered questionnaires. The significant intercept for the meta-analysis on comparison studies then indicates the overall positive effect of RR in furnishing information of a sensitive nature as compared to all non-RR approaches included in the analysis. These results reveal that RR is able to obtain more valid population estimates than the traditional techniques. A conclusion maintained in the third sequence.

As there still is unexplained residual variance in M1 the researchers subsequently control for item sensitivity (a measure of the degree of social sensitivity of the topics researched in the studies included in the meta-analyses). As can be seen in Table 3.1 column M2, sensitivity sorts a positive and significant influence on the difference between the population probability and the point estimates in the validation study analysis as well as on the standardized difference score in the comparison study analysis. Thus: the greater the degree of sensitivity of a certain topic, the more valid the results that are obtained with the use of RR. The relative merit of RR increases as the sensitivity of the topic increases. The study by Lensvelt-Mulders *et al.* showed overall that, although both meta-analyses have unexplained variances, compared to other methods usage of RR results in more valid data.

GAPS IN RANDOMIZED RESPONSE LITERATURE AND RESEARCH

Although the RR technique has been tested in field research settings on numerous occasions, the lion's share of the literature has been preoccupied with the technical fix and the theoretical exploration of the assumptions underlying the method. This absorption with technicalities however, has not brought consensus on best practice and protocol of administration. This lack of consensus and the use of a wide variability of specific models and protocols at least partly underlie the variability in

field performance of individual studies as described in the preceding section. The conducted field-tests themselves also point out certain gaps in RR literature and research. Firstly, RR has mostly been field-tested with regard to criminal behavior, alcohol and drug use and abuse, demographic variables such as induced abortion, sexual behaviors and rule transgressing with regard to fraud. The technique remains to be tested on sensitive behaviors of an essentially political nature. Secondly, most studies do not delve into the relative merits of RR in obtaining quantitative characters. The research by Fox and Tracy on arrest history (1980, 1981), is one of the very few field-tests of a quantitative RR model. Additionally, when using RR, one has to deliberate on the extra costs such an endeavor entails. We have seen that the insertion of random error inflates variance, making the RR technique less efficient than traditional methods making use of DQ (when assuming fully honest responses in the latter, a point to which we will return in Chapter 7). Larger samples are thus necessary in a RR design to obtain an acceptable level of precision (in terms of confidence intervals). Also, RR designs are relatively more complex in comparison to traditional methods, laying a burden on researcher as well as researched. On the former to develop good instructions so that the respondent will fully understand and/or trust the method. On the latter as RR increases the cognitive load for respondents (Lensvelt-Mulders *et al.* 2005). The increased complexity an RR design brings, allows for new sources of error, which have to be dealt with. These remarks on the gaps in RR literature and research, and the extra costs an RR design can bring, come together in the goals and ambitions of present study.

To assess the relative merits of RR, the present study will take a comparative road, by assessing an RR design with the traditional direct questioning technique in their viability to delve into latent space so as to gauge the level of certain sensitive traits, adding to the relatively scarce literature on field-tests. This comparative study will be conducted on sensitive behaviors of a political nature, meaning: behaviors hampering on the workings of the public arena, which are potentially incriminating and/or stigmatizing for individuals or groups who have vested political alignments/positions in that arena. The dissemination of information on behaviors of this kind could not only be incriminating or stigmatizing for the group or individual of interest, in case of officials it can also potentially undermine the public trust entrusted in them. Integrity violations in the public sphere are a perfect test case for behaviors of this kind. We will try to accurately gauge the level of certain types of wrongdoing in one organization in the public arena. In doing so we will

not only delve into questions of prevalence, but also into questions of incidence. The study will be no rigorous determination of the prevalence and incidence of wrongdoing in all similar organizations in the public arena as a whole. It will try to assess the viability of RR for providing a more accurate assessment of certain kinds of wrongdoing, and the proposed study will hold as an 'experimental' test case. For this purpose also a quantitative RR model will be proposed. The ambition here is to present a unified RR model for the dissemination of sensitive dichotomous and quantitative information. In doing so, this study will hope to modestly contribute to a more definitive protocol for and understanding of RR and seeks to act as a very modest basis for much needed cumulative work on RR designs in actual research settings. It is these issues for which a framework will be provided in Chapters 4 and 5.

NOTES

[1] Bernoulli's theorem states the probability of r wins under the assumption that the probability of that win (p) remains constant over repeated trails. This holds when in a sample the items are replaced and is approximately true if the sample is large. Thus a binomial distribution can be fitted to a distribution obtained by sampling, where

p = probability of a win in a single trail
q = 1 – p
n = number of independent trails
P(r) = probability of r wins, with r being bigger or equal to zero and smaller or equal to n

so that

$$P(r) = \binom{n}{r} p^{r} q^{n-p}$$

For a more in depth review of the binomial distribution and Bernoulli's theorem, see Lambde (1967: 26-29).

[2] As stated before, the choice of p is at the disposal of the researcher. When $p = 1$, this equals a DQ survey. In case of a non-sensitive question in a DQ setting, equation 2.2 is equivalent to an unbiased estimator of πx.

[3] p should not equal .5 as in this situation no information regarding πx would be furnished because when $p = .5$, then $\lambda = .5$. This means that the probability of a 'yes' answer no longer involves the possession of trait X (see Warner 1965; Verdooren 1976: 9; Chaudhuri & Mukerjee 1988: 5).

[4] For example: when $p = 0$, a 'no' response implies in the certainty of probabilistic terms that the respondent possesses trait X.

[5] Note, as Warner points out: "[I]t is a feature of the dichotomous nature of a population that telling the truth .2 of the time is equivalent to telling the truth .8 of the time" (1965: n2).

[6] Or the positive version as in Warner's model: "Do you possess X?"
[7] The mean of a random error term is thus 0.
[8] Although some see the unmatched count technique as a specific alternative within the RR family (Fox & Tracy 1986), Lenvelt-Mulders *et al.* (2005) do so otherwise. In this technique there are two groups. Each group receives a list with the same innocuous behaviors. One of the groups however, also has a sensitive behavior on their list. The respondents simply reply to the researcher how many behaviors in total he or she has participated in. The difference between the mean counts in the two samples is then an indication of the prevalence of the sensitive behavior (Capaciteitsgroep Methodenleer & Statistiek, Universiteit Utrecht: 27-28). Although its simplicity may seem to have a definite appeal, there are many problems to be considered, for an overview of which we confine by referring to Fox & Tracy (1986: 41-42).

4. MEASURING PREVALENCE AND INCIDENCE: A UNIFIED APPROACH TO ELICITING SENSITIVE DICHOTOMOUS AND MULTI-PROPORTIONAL DATA

Usage of a questioning design, which misclassifies data with known probability distributions, can improve the accuracy of population estimates regarding sensitive behaviors, as the inoculation through chance factors renders out the need for self-representational concerns. While there is empirical evidence for the superiority of RR compared to more traditional methods making use of DQ, there is still ground to be gained in the translation of RR theory to practice. A framework for the dissemination of sensitive dichotomous and quantitative multi-proportional data is provided for in this and the following chapter. This framework seeks to address the ambition of modestly acting as a basis for cumulative work on RR in research settings. This chapter deals with sub-question *c*, laying bare the construction of RR models, which at least theoretically can efficiently handle eliciting sensitive information of a dichotomous and quantitative nature, so that prevalence as well as incidence estimates are possible. Chapter 5 embarks on the placement of the presented models in a comparative design, enabling to contribute to the relatively rare empirical RR endeavors.

Here, a RR design is proposed in which the distribution of π_Y is provided for by the randomization device itself. It is on this design a RR model for the dissemination of dichotomous data, and a new model for the dissemination of multi-proportional data, both with newly developed randomizers, will be built. In doing so, we will also discuss more

thoroughly the issues of 'efficiency' and 'respondent protection', before moving to Chapter 5 in which our framework is amended with an experimental design for empirical testing.

FORCED RESPONSES FOR MISCLASSIFICATION OF SENSITIVE DATA

In Chapter 3 the UQD was discussed, which, when π_y is known, has considerable advantages over the Warner model in terms of efficiency as well as comprehensibility. It is very difficult to find enough innocuous questions where this parameter is known to accompany sensitive questions in surveys, which develop inquiries into a relatively substantial number of sensitive traits. A solution for this problem is to let the randomization device provide for π_y. Richard Morton (see Greenberg *et al.* 1969) and Robert Boruch (1971, 1972) almost simultaneously proposed an equivalent procedure that has this specific design characteristic, now commonly referred to as 'forced randomized response' (FRR).

We have already touched upon the FRR in Chapter 1. The RR example given there describes a procedure in which, according to certain outcomes, the respondent is directed to answer 'yes' or 'no' irrespective of his or her true status. Here, that example will be extended according to the procedure proposed by Boruch (1971, 1972).

Again, we are interested in estimating the prevalence of a certain sensitive dichotomous trait X. Each respondent is provided with a pair of dice. The respondent throws the dice and keeps the outcome hidden for the researcher. The instructions are to answer truthfully if the outcome is 5, 6, 7, 8, 9 or 10; to always answer with 'yes' if the outcome is 2, 3 or 4; and to always answer with 'no' if the outcome is 11 or 12. The respective probabilities are p_1, p_2 and p_3, where $p_1 + p_2 + p_3 = 1$. The unrelated character π_y is thus automatically induced by

$$\pi_y = \frac{p_2}{p_2 + p_3} \tag{4.1}$$

Table 4.1 gives the possible outcomes of the dice throw and their permutations, along with the three selection probabilities.

Table 4.1: Outcomes and Permutations in Forced Randomized Response With Pair of Dice as a Randomization Device

Outcome	2	3	4	5	6	7	8	9	10	11	12
Permutations	1	2	3	4	5	6	5	4	3	2	1
	forced 'yes'			*truthful response*						*forced 'no'*	
	*p*2			*p*1						*p*3	

From Table 4.1 we see that under the given instructions the probability of a forced 'yes' is .167 (6/36), the probability of a truthful response is .75 (27/36) and the probability of a forced 'no' is .08 (3/36). If we denote our observed answers with i and a true status with j, and assume 1 = yes and 0 = no, then the RR matrix with conditional misclassification probabilities is given by

$$P_{ij} = \begin{pmatrix} p11 & p10 \\ p01 & p00 \end{pmatrix} = \begin{pmatrix} 33/36 & 6/36 \\ 3/36 & 30/36 \end{pmatrix} \tag{4.2}$$

where p_{10} and p_{01} represent the probability that a respondent is misclassified while he or she does not have trait X, and the analogous probability while he or she does have X, respectively. Due to these probabilities it can never be known if a certain response corresponds to the latent status of X, so that the privacy of each respondent is fully guarded (Van Den Hout & Van Der Heijden 2004: 386).

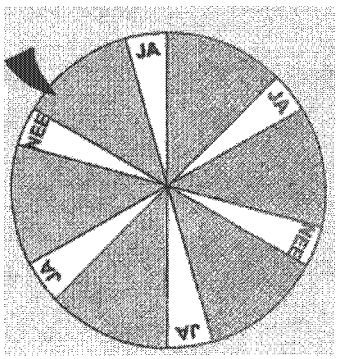
Due to its efficiency, its simplicity when it comes to inoculation and its practicability when developing inquiries into large numbers of sensitive traits, the FRR is seen as a viable model. It is on extensions of this design on which a new randomizer for a dichotomous RR model and a new quantitative RR model are built.

Dichotomous Forced Randomized Response Technique with New Randomizer for the Estimation of Prevalence

Consider a sensitive trait X for which the population is dichotomous. A random sample of $i = 1,\ldots,i,\ldots n$ persons is drawn. The goal is the estimation of the population prevalence of X (π_X) from the persons in our

random sample. To this purpose each person in the sample is asked the question: "Do you possess trait X?" Each respondent receives the randomization device given in Figure 4.1 along with the following instructions: turn the spinner; if it stops on an empty area, respond with 'yes' or 'no' truthfully; if it stops on an area imprinted with 'yes', always answer with 'yes'; if it stops on an area imprinted with 'no', always answer 'no'.

Figure 4.1: Spinner for Dichotomous Forced Randomized Response Inquiries



In this setup we have

$$X_i \begin{cases} 1 & i\text{th respondent possessing trait X} \\ 0 & \text{otherwise} \end{cases}$$

$$W_i \begin{cases} 1 & i\text{th respondent randomly requested to respond truthfully} \\ 2 & i\text{th respondent randomly directed to 'yes' response} \\ 3 & i\text{th respondent randomly directed to 'no' response} \end{cases}$$

$$Z_i \begin{cases} 1 & i\text{th person answering question with 'yes'} \\ 0 & i\text{th person answering question with 'no'} \end{cases}$$

$p1$ = probability of a request on a truthful answer
$p2$ = probability of a forced 'yes'
$p3$ = probability of a forced 'no'

with

$$p1 + p2 + p3 = 1$$

Taking into account the assumptions regarding RR, events can be expressed in probabilities, so that

$$\begin{aligned} P(Z_i = 1) = \lambda &= P(Z_i = 1 \mid W_i = 1, X_i = 1) \cdot P(W_i = 1) \cdot P(X_i = 1) + \\ &\quad P(Z_i = 1 \mid W_i = 2) \cdot P(W_i = 2) \\ &= p1\ \pi x + p2 \end{aligned} \tag{4.3}$$

As the probability of a 'yes' response can be estimated by the sample proportion answering 'yes' ($\sum Z_i/n$, as in equation 2.2) and as $p1$ and $p2$ are fixed by design, an unbiased estimate of the population prevalence of X can be obtained by

$$\pi x' = \frac{\lambda' - p2}{p1} \tag{4.4}$$

with sampling variance

$$\text{var}(\pi x') = \frac{\lambda(1-\lambda)}{np1^2} \tag{4.5}$$

and its unbiased estimate

$$\text{var}'(\pi x') = \frac{\lambda'(1-\lambda')}{(n-1)p1^2} \tag{4.6}$$

Note that via this technique 'yes' and 'no' responses aren't definitive in meaning. Individually it cannot be traced if one answers with 'yes' or 'no' because one is forced to do so, or because it is a truthful response. Admitting to certain behaviors will thus become less threatening. It will then become at least theoretically possible to give more accurate population estimates of prevalence of sensitive behaviors.

The FRR design has a setup that is analogous to the UQD where π_y is known. As π_y is induced in the FRR by the randomization itself, the efficiency is greatly enhanced, and the cognitive load for the respondent potentially reduces (just one question posed). A possible drawback can also be found in its essentially theoretical nature (Chaudhuri & Mukerjee

1988: 17; Capaciteitsgroep Methodenleer en Statistiek, Universiteit Utrecht 2003): it can prove to be counterintuitive to give an answer that is the opposite of one's true status (an issue to which we will return in this and the next chapters).

New Discrete Quantitative Forced Randomized Response Technique with New Randomizer for the Estimation of Incidence

In inquiries into sensitive behaviors not only questions of prevalence are of interest, but also questions of incidence. For example: we are not only interested whether one has indulged in corruption, but also how many times one has done so. To evaluate frequency of occurrence, RR models must be used that deal with multi-proportional or quantitative data. Here, a discrete quantitative FRR model is developed for which previous work by Greenberg *et al.* (1971), Eriksson (1973), Liu & Chow (1976b) and Stem & Steinhorst (1984) is acknowledged.

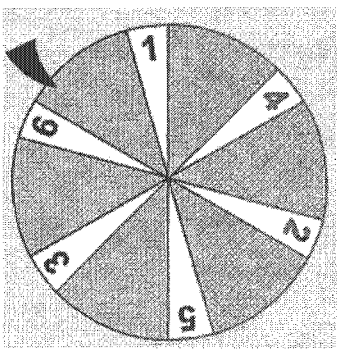
Consider a sensitive trait X, for which the population is supposed to be continuous or multi-proportional. A random sample of $i = 1,\ldots,i,\ldots n$ persons is drawn. Our desire is to estimate μ_X, the unknown population mean of X. To this purpose each person in the sample is asked the question: "How many times have you indulged in X?" Consider the randomization device as given in Figure 4.2. Each respondent receives this device with the following instructions: turn the spinner; if it stops on an empty area, respond with '1', '2', '3', '4', '5' or '6' truthfully; if it stops on an area imprinted with '1', '2', '3', '4', '5' or '6', respond accordingly. The selection probabilities for the truthful answer and the forced response are p and p_j respectively, with $j = 1,2,3,4,5,6$ and $\sum p_j = 1 - p$.

The use of this randomization device redirects X to be discrete, assuming values $x_1,\ldots, x_6$ with respective unknown true proportions $\pi_1,\ldots, \pi_6$. Each of those values can be assigned a category of numbers, for example: the responses '1', '2', '3', '4', '5' and '6' could correspond with the respective categories 0, 1 time, 2 to 3 times, 4 to 5 times, 5 to 10 times and more than 10 times.

The randomization device provides for a one-step simulation of 2 throws with 3 dice, where the selection probabilities of p_2 and p_3 (as in the previous section) require for the third die to be thrown. The number that turns up in this throw gives the forced response. It provides a setup where any piece of quantitative information can be misclassified in 6

discrete categories. As the possible options for forced answers have the same magnitude as the possible options for the request to answer truthfully (sensitive question), and as the probability distribution of the forced answers can be fixed by the researcher, respondents' privacy is fully guarded with respect to quantitative responses and our desired population mean can be estimated using the known parameters.

Figure 4.2: Spinner for Discrete Quantitative Forced Randomized Response Inquiries



In this setup we thus have

X_i { frequency of participation in X by ith respondent

Y_i { frequency reported by ith respondent according to inoculating forced response

W_i { 1 ith respondent randomly requested to respond truthfully
0 ith respondent randomly directed to forced quantitative answer

Z_i { number reported (1 to 6) by ith respondent

where

p = probability of a request for a truthful answer
θ = mean of the forced responses
Z^* = total mean response in the sample

so that, according to RR assumptions,

$$\begin{aligned} E(Z_i) &= E(Z_i \mid W_i = 1) \cdot P(W_i = 1) + E(Z_i \mid W_i = 0) \cdot P(W_i = 0) \\ &= E(X_i) \cdot p + E(Y_i) \cdot (1 - p) \\ &= \mu_X p + \theta(1 - p) \end{aligned} \tag{4.7}$$

When we denote the sum of numerical responses in our sample by $\sum Z_i$, the conditional expectation of Z_i, Z^*, can again be obtained as in equation 2.2. As p and θ are fixed by design, an unbiased estimate of the population mean of X can then be obtained by

$$\mu_X' = \frac{Z^* - (1 - p)\theta}{p} \tag{4.8}$$

with sampling variance

$$\text{var}(\mu_X') = \frac{\sigma_Z^2}{np^2} \tag{4.9}$$

where σ_Z^2 denotes the variance of the quantities reported for the sensitive question as well as the forced responses. It involves the unknown parameter μ_X. The variance of μ_X' can then be unbiasedly estimated by using the sample variance of all responses as an estimate of σ_Z^2, so that

$$\text{var}'(\mu_X') = \frac{\Sigma(Z_i - Z^*)^2}{n(n-1)p^2} \tag{4.10}$$

As already stated, this model redirects X to be discrete, assuming values $x_1, \ldots, x_6$ with respective unknown true proportions $\pi_1, \ldots, \pi_6$. Instead of focusing on the mean of our ordinal categorization, this setup also makes possible independent estimates of the population proportion in each of our discrete categories separately. We are then interested in estimating π_i, where $i = 1,2,3,4,5,6$. Again, let p denote the selection probability for the truthful answer and $\sum p_j$ give the selection probability of a forced quantitative response with $j = 1,2,3,4,5,6$ and $\sum p_j = 1 - p$. If we let γ_i denote the probability of a quantitative response, again with i denoting the response pointing to one of our 6 discrete quantitative categories, it can then be shown with proof analogous to (4.3) that

$$\gamma_i = p\,\pi_i + p_j \qquad (i,j = 1,2,3,4,5,6) \tag{4.11}$$

As the probability of a certain quantitative response within our 6 category parameter space can be estimated by the sample proportion giving that

specific quantitative answer, and as p and p_j are fixed by design, each of our 6 population proportions can be unbiasedly estimated by solving (4.11) for π_i, so that

$$\pi_i' = \frac{\gamma_i' - p_j}{p} \tag{4.12}$$

with unbiased estimate of variance

$$\text{var}'(\pi_i') = \frac{\gamma_i'(1-\gamma_i')}{(n-1)p^2} \tag{4.13}$$

Note that in this form, a quantitative answer again is not definitive in meaning. It cannot be traced individually if one answers with for example '3' because one is forced to do so, or because it is a truthful response regarding the frequency of individual behavior. Admitting ones frequency of certain behaviors becomes less threatening, and at least theoretically more accurate estimates of incidence can be given as a result.

The great advantage of this design is its efficiency (see next section). Also, with this model, sensitive quantitative information can be misclassified in 6 discrete categories, where the scope of the frequencies denoted with each category can be adopted to the researchers need. In previous designs the desired populations means regarding certain sensitive behaviors had to be estimated before any field research efforts, so as to adapt the setup of the randomizer to the expected range of frequencies in the population, as the range of responses to the innocuous question must be similar to the range of possible responses to the sensitive question so as to be able to unbiasedly estimate incidence while still providing sufficient respondent protection (see for example Liu & Chow 1976b; Stem & Steinhorst 1984). In this model, the categories can be very easily adopted to expected frequencies while retaining setup and efficiency, as the forced responses have a range that is equal to the range of possible responses on the sensitive question, due to the multi-proportional construction. Full respondent protection is thus guaranteed. Again, a possible drawback is that it can prove to be counterintuitive to give an answer that is the opposite of one's true status.

ON THE CHOICE OF p (AND π_Y)

When endeavoring on the use of RR one must deliberate on the design parameters, as these are closely connected to 'respondent jeopardy', 'risk

of suspicion' and the sampling precision of the model (efficiency). Respondent jeopardy and risk of suspicion refer to 'respondent hazards', where the former denotes the degree in which a 'yes' or affirmative response implies possession of the sensitive trait, and where the latter gives the degree in which a 'no' or negative response still leaves the possibility of possession of trait X (Fox & Tracy 1986: 32). Especially the choice of p and π_y force the researcher to simultaneously consider these three critical issues. Clearly, when p approaches 1, efficiency increases[1], but respondent jeopardy also increases as the conditional probability of possession of X increases with an affirmative response under these conditions. When p increases risk of suspicion decreases as a 'no' response then has a more definitive meaning (Fox & Tracy 1986: 32). When π_y decreases, clearly, respondent jeopardy and risk of suspicion both increase.

Each of the three concepts can be quantified. The relative efficiency can be obtained by dividing the variance of conventional DQ estimates by the variance of the RR model used[2] and the respondent hazards can be quantified by using conditional probabilities based on Bayes's rule[3] (see Anderson 1976; Lanke 1976; Leysieffer & Warner 1976; Warner 1976a; Fox & Tracy 1986: 32-37; Ljunqvist 1993). However, the boundaries of the quantified levels of efficiency, respondent jeopardy and risk of suspicion have to be subjectively determined by the researcher. Moreover, these quantifications are dependent on π_x, which cannot be objectively known in advance. The relevance of these measures is essentially theoretical, as a respondents' reaction cannot be caught with reference to these indices alone. For example, Fox & Tracy (1986: 32-37) give the quantifications of respondent hazards for the UQD with known π_y, a model statistically similar to the dichotomous FRR model. Their indexes give that a π_y of 1 would be allowable. In the FRR this would mean that $p_3 = 0$, and that the forced response would thus always be 'yes'. While it can still be argued that the FRR offers protection from full disclosure in this setup, in the sense that a respondent could always rationalize that a certain 'yes' response was forced, the respondent who possesses trait X will actually feel very little protection as he will always be required to answer affirmatively. This example makes clear that indexing quantifications of the possible respondent hazards cannot capture a respondents' reaction. In fact, only efficiency is a function of actual p, while the respondent hazards are a function of p as perceived by the respondent (Moors 1976; Scott 1976; Fox & Tracy 1986: 38). It will be attempted to appreciate these comments in terms of the FRR models

presented in the previous subsections.

Moriarty & Wiseman (1976) give evidence that respondents, due to the involved permutations leading to a certain outcome, misperceive selection probabilities when dice are used as a randomization device. If the misperceiving of probabilities can be incorporated in the randomization device, it would be possible to provide for sufficient protection, while pertaining relatively high efficiency. In this study the forced dichotomous randomization spinner was constructed so that p = 3/4. The respective probabilities for p_2 and p_3 were 1/6 and 1/12, making π_y 2/3. These selection probabilities were translated into the randomization device by converting probabilities into degrees. To enhance the misperceiving of selection probabilities the empty and 'yes' and 'no' imprinted areas were evenly divided over the spinner, which is divided in 24 sub-areas. This gives every empty area a selection probability of 3/24 ($6 \cdot 3/24 = 3/4$) and every area imprinted with 'yes' or 'no' a selection probability of 1/24, which makes for a 1/6 selection probability of a forced 'yes' ($4 \cdot 1/24$) and a 1/12 selection probability of a forced 'no' ($2 \cdot 1/24$). Figure 4.1 corresponds to this setup. The objective was to retain a relatively high efficiency with a relatively large p and simultaneously distract respondents from the actual selection probabilities so as to let them perceive sufficient protection. The forced multi-proportional quantitative randomization spinner was constructed analogously, with a p of 3/4 and a $\sum p_j$ of 1/4. Each empty area having a selection probability of 3/24 and the analogous selection probability of each of the 6 forced numerical responses being 1/24. Figure 4.2 corresponds to this setup. Table 4.2 gives an overview of the parameters as used in the empirical testing of the models.

Table 4.2: Design Parameters Dichotomous and Discrete Quantitative Forced Randomized Response Models

Design parameter	Dichotomous FRR Model	Discrete Quantitative FRR Model
$p1$	3/4	3/4
$p2$	1/6	-
$p3$	1/12	-
$\sum p_j$	-	1/4
π_y	2/3	-
$\boldsymbol{\theta}$	-	3.5

From Table 4.2 it is easy to see that the conditional misclassification matrix for the dichotomous FRR model equals the matrix in equation 4.2. Again, if observed status is denoted with i, and true status with j, with possible answering categories 1, 2, 3, 4, 5 and 6, then the conditional misclassification matrix for the discrete quantitative FRR model becomes

$$P_{ij} = \begin{pmatrix} p11 & p12 & p13 & p14 & p15 & p16 \\ p21 & p22 & p23 & p24 & p25 & p26 \\ p31 & p32 & p33 & p34 & p35 & p36 \\ p41 & p42 & p43 & p44 & p45 & p46 \\ p51 & p52 & p53 & p54 & p55 & p56 \\ p61 & p62 & p63 & p64 & p65 & p66 \end{pmatrix}$$

$$= \begin{pmatrix} 19/24 & 1/24 & 1/24 & 1/24 & 1/24 & 1/24 \\ 1/24 & 19/24 & 1/24 & 1/24 & 1/24 & 1/24 \\ 1/24 & 1/24 & 19/24 & 1/24 & 1/24 & 1/24 \\ 1/24 & 1/24 & 1/24 & 19/24 & 1/24 & 1/24 \\ 1/24 & 1/24 & 1/24 & 1/24 & 19/24 & 1/24 \\ 1/24 & 1/24 & 1/24 & 1/24 & 1/24 & 19/24 \end{pmatrix} \quad (4.14)$$

Given the misclassification matrices, the models as developed here can be generally captured in the function (analogous to Bourke & Moran 1988 and Van Den Hout & Van Der Heijden 2004)

$$\eta^* = P\eta \quad (4.15)$$

so that

$$\eta' = P^{-1}\eta^{*'} \quad (4.16)$$

where η^* is the vector denoting the probabilities of the observed responses with $1,\ldots,K$ categories, η is the vector denoting the probabilities of true answers/status with $1,\ldots,K$ categories, and P is the K x K matrix of conditional misclassification probabilities given in (4.2) and (4.14).

Now that we have presented the construction of a unified approach to the obtainment of sensitive dichotomous and quantitative data, and have deliberated on the design parameters for empirical testing of the models, we will turn to Chapter 5, which gives the experimental design for our empirical endeavor.

NOTES

[1] As already stated, when $p = 1$, we actually have a DQ setting. One can make the remark regarding greater efficiency, when one would assume honest responses in a DQ setting. This assumption is the standard for efficiency comparison between RR and DQ conditions. As seen in Chapter 3, untruthful responses add to the error variance, leading to variance inflation. When conducting inquiries into sensitive behaviors, the assumption of truthful reporting in a DQ setting poses hard to maintain. It does not take a very high proportion of untruthful responses in a DQ setting to let its mean square error (variance plus square of the bias induced by deviations from true population score) exceed the mean square error of RR, when assuming that more truthful answers will be obtained with the latter (Warner 1965). We will return to this issue in Chapter 7.

[2] An analysis based on this endeavor will show that the FRR design is most efficient compared to other RR designs, irrespective of the assumed πx in the population (see Capaciteitsgroep Methodenleer en Statistiek, Universiteit Utrecht 2003; Lensvelt-Mulders, Hox & Van Der Heijden 2005).

[3] Bayes theorem is a likelihood approach that differs from the frequentist approach in its use of prior distributions. The basic theorem states that

$$p(\mathrm{B}|\mathrm{A}) = \frac{p(\mathrm{A}|\mathrm{B})\, p(\mathrm{B})}{p(\mathrm{A})}$$

where B implies certain parameters and A certain data. A posterior distribution in this form of likelihood is then a function of the prior distribution $p(\mathrm{B})$ and the likelihood function $p(\mathrm{A}|\mathrm{B})$ (see Lynch & Western 2004). This type of conditional probability has proven to be functional for quantifying respondent hazards in RR models.

5. FORCED RANDOMIZED RESPONSE MODELS IN A COMPARATIVE DESIGN: EMPIRICAL FRAMEWORK FOR TESTING

A unified approach to eliciting sensitive dichotomous and quantitative data was presented in the preceding chapter. As there is still ground to be gained in the translation of RR theory to practice, the second part of our framework for the dissemination of sensitive dichotomous and quantitative multi-proportional data is provided for here, which deals with sub-question *d*, laying bare the placement of the dichotomous and quantitative FRR models in a comparative design, so as to contribute to the relatively rare empirical RR endeavors.

Firstly, the research trajectory for the empirical testing of the RR models is discussed, subsequently dealing with the test-group and their behaviors for which prevalence and incidence estimates are desired, the comparative design in which RR performance will be compared to a DQ setting and official statistics, and the specific field-trajectory of this endeavor. Secondly, we will delve into the mode of analysis of our data, the hypotheses that guide the analysis and the evaluation of the estimates that will be obtained.

RESEARCH TRAJECTORY

Here, we will lay bare the structure of the research trajectory in which the FRR models were empirically tested. First, the test-group will be reviewed along with our reasons for testing the viability of RR with

regard to integrity violations within that group. Next, we will delve more specifically into the research design comparing the RR results. A 2-factorial design with 2 interview conditions was used (namely RR and DQ) which incorporated a computer-assisted environment. As will become clear, the results of these questioning conditions were also compared to official statistics. Lastly, this section will embark on the specific field trajectory of the empirical endeavor testing the FRR models.

Estimating Prevalence and Incidence of Integrity Violations in a Police Force

Many field studies have demonstrated that the police in general, is a difficult group to study (see for example Klockars 1985; Van Maanen 1988; Punch 1989; Brewer 1990; Huberts 1998). The difficulties associated with researching police practice become especially pressing when embarking on an assessment of police deviance. The police organization is a very secluded part of the criminal justice system and has many defenses that protect the concealed reality of its daily practice (Skolnick, as referred to by Punch 1989), while there are many opportunities for misconduct (Klockars *et al.* 2000). As Klockars *et al.* state (2000: 1): "[p]olicing is a highly discretionary, coercive activity, that routinely takes place in private settings, out of the sight of supervisors, and in the presence of witnesses who are often regarded as unreliable". Police practice gives rife opportunities for breaches of a certain law or code of acceptable behavior, while knowledge to third persons on these breaches is very likely to be limited. Regarding police deviance, we encounter the problem of representing the latent with the manifest in research on sensitive topics.

Disclosure of information on integrity violations by police officers can be potentially stigmatizing, intrusive or even incriminating. Also, as it concerns government officials, disclosure can also undermine their position in the public arena and the public trust entrusted in their practice. Integrity violations by members of the police are thus an appropriate test case for the viability of RR in representing the latent with the manifest regarding behaviors of a political nature, as police deviance hampers on the workings of the public arena, and possible disclosure of information regarding integrity violations can pose to be intrusive, stigmatizing, incriminating or politically undermining for this politically vested group and its individuals.

Our target sample included every member of one of the 25 regional police forces in the Netherlands. The specific force will remain anonymous. Each potential respondent in the sample was randomly assigned to one of two groups: a group that was subjected to the RR questioning condition, or the second group subjected to a more traditional DQ condition. Every respondent across the questioning conditions received exactly the same questions regarding their personal behavior regarding certain integrity violations. Self-reports were thus desired across both questioning conditions.

The integrity violations for which self-reports were desired were based on the typology of categories of integrity violations as developed by Huberts, Pijl & Steen (1999). This typology was the outcome of an analysis of the literature on police integrity and was assessed against the results of empirical research on internal investigations in the Dutch police force. The typology can be found in Appendix A. The specific questions regarding the integrity violations were developed with the cooperation of police officials. In doing so, all specific formulations referred to behaviors, which are commonly and clearly acknowledged, by officers themselves, as breaches of the behavioral norms and laws fitting individuals in the police organization. Also, this endeavor made sure that the terminology used, corresponded to daily police practice. Importantly, basing the questions on the typology makes sure a variability in violations in terms of their expected prevalence and occurrence, which can act as an intuitive check on the compliance with the RR and DQ formats (we will return to this point in Chapter 6). To avoid misunderstanding of the questions and to enhance compliance, all questions were posed as neutral as possible and referred to specific situations and behaviors. All questions are presented, along with the desired estimates of prevalence and incidence, in Chapter 6, and can also be found in Appendix B.

Computer Assisted Surveying with Randomized Response and Direct Questioning Conditions

A computer-assisted environment aided the mode of administration of the RR and DQ conditions. The use of various forms of Computer Assisted Self-Interviewing (CASI) has previously proven to be very promising with regard to eliciting sensitive information (Tourangeau & Smith 1996; Lensvelt-Mulders & De Leeuw 2002b) with audio-CASI being one of the latest developments (Couper, Singer & Tourangeau 2003). In the latter, the respondent listens to digitized voice-recordings of question and

answer formats over a headphone, and keys his or her answer into a computer.

It is clear that self-reporting overall will lead to more honest reporting than interviewer administration. But the use of computers, when addressing sensitive issues, has several additional advantages, one of the most important being that it can hypothetically further enhance truthful reporting due to lacking social context cues. Also, as many survey tasks are automated with the use of computers, time needed to complete a survey can be reduced considerably for the respondent. Computers make possible the design of complex questionnaires, while maintaining standardized administration (Kinsey *et al.* 1995: 1023). A possible drawback, which is especially apparent in audio-CASI, is potentially high costs. Also, in a comparative study by Van Der Heijden *et al.* (1998, 2000), the use of CASI resulted in least valid outcomes with regard to the sensitive context of welfare and unemployment benefit, in comparison to RR and face-to-face DQ. Lara *et al.* (2004) found similar performance by CASI in their comparison of RR, self-administered questionnaires, audio-CASI and face-to-face interviews, in eliciting information regarding induced abortion in Mexico.

The literature gives several possible reasons for the lack of a consistent positive effect of CASI. Firstly, the effect of CASI on eliciting sensitive information could decrease over time, as the usage of computers loses its magic due to its penetration of mainstream culture (Lenvelt-Mulders & Boeije 2002; Lenvelt-Mulders 2003: 68; Lenvelt-Mulders, Van Den Hout & Van Der Heijden, forthcoming). As 'computer-literacy' increases, respondents' awareness of the negative possibilities of computers regarding his or her privacy issues also increases (such as the collection of data in online databases), which can contribute to a time-dependent decrease of trust in computer aided interviewing. Secondly, and very importantly, as Moon (1998) points out, the social desirability benefits associated with mode of administration are more convincing than the analogous benefits associated with the use of computers alone.

The refutations of CASI, give a clear justification for the effort to incorporate the standing benefits of CASI – lack of social context, complex designs with preservation of standardization and increased response convenience – with the inoculating benefits of RR designs. This incorporation of CASI and RR in Computer Assisted Randomized Response Interviewing (CARR), has been proposed and successfully tested by Musch, Bröder & Klauer (2001), Lenvelt-Mulders & Boeije (2002) and Lenvelt-Mulders, Van Den Hout & Van Der Heijden (forthcoming). CARR could prove to be very important for future

inquiries into socially sensitive issues, as other computer based techniques – as argued – lose their 'magic', and as the use of internet makes possible the dissemination of data on a large scale (Musch, Bröder & Klauer 2001), which can also possibly relieve the efficiency problems inherent in RR designs. As the studies of Lenvelt-Mulders and associates were conducted in cognitive laboratories, we will try to provide further preliminary testing of CARR in actual research settings.

As every Dutch police officer is expected to have a minimum level of computer literacy, the aid of a computer-assisted environment in present study was deemed useful. The study thus made use of a 2-factorial design, having 2 interview conditions: CARR and CASI. Both respondent groups were granted anonymity. In the CASI condition this was achieved by informing the respondents of the researchers' grant on confidentiality of data. The respondents in the CARR condition were also assured of their anonymity by the incorporation of RR. Both respondent groups received the same questionnaire. Respondents were provided with instructions on-screen, in a computer environment, which was fully controllable with the mouse pad, and thus asked for minimal computer literacy.

After an introduction explaining the nature and aim of the survey, the respondents were asked 18 general questions regarding their work and their work environment, immediately followed by 10 prevalence questions regarding individual misbehavior in police function. Subsequently, 10 incidence questions were posed regarding individual misbehavior by police officials, in which the numbers 1 to 6 referred to the numerical categories as given in the preceding section on the quantitative FRR: 0 times, 1 time, 2 to 3 times, 4 to 5 times, 5 to 10 times, and more than 10 times. The survey was concluded with several questions regarding the respondents' evaluation of perceived anonymity and user-friendliness of our CARR and CASI endeavors.

The difference between the CARR and CASI conditions was that in the former, an introduction and explanation of the RR technique were provided for along with instructions as given in earlier sections of this chapter, and – preceding the questions regarding prevalence and incidence – few trail questions to let the respondent get acquainted with the dichotomous and quantitative FRR method. Also, the evaluation questions with which the survey was concluded included additional questions regarding the RR method in the CARR condition. No questions were asked that could reveal a persons' identity by reviewing individual or chained responses.

As stated before in the subsections on the FRR models, it can pose to be counterintuitive for respondents to give an answer that does not correspond to one's true status, especially when one has to say 'yes' when the actual answer is 'no'. This can spur 'cheating' in the RR procedure, meaning not following forced directions, as respondents can have the feeling that a forced answer is not honest or still incriminating. To decrease the chance of cheating on our FRR procedure we followed the recommendations by Lenvelt-Mulders & Boeije (2002) to 1) acknowledge that a forced answer that does not correspond to one's true status can sometimes be counterintuitive and difficult, and 2) to redefine the construct of 'being honest' in terms of following the FRR instructions (see also Stem & Steinhorst 1984: 558). These instructions significantly decreased cheating in their cognitive laboratory study of CARR. Appendix B contains the full texts of the CARR and CASI questionnaire, along with an impression of the virtual make-up of the computer assisted questionnaires as given to the respondent.

Due to the computer-assisted environment the RR procedure and the accompanying randomizer had a virtual make-up. Lenvelt-Mulders & Boeije (2002) showed that respondents preferred virtual randomizers when responding to a computer-assisted questionnaire. The virtual make-up of the randomizers as given in Figure 4.1 and 4.2 corresponded to the chance distributions as given in the section on the choice of p and Table 4.2. Two important problems are additionally solved concerning the randomizers, when a virtual make-up is used. The first of which is randomness. Randomness has proven to be hard to achieve with physical randomizers. For example, Stem & Steinhorst (1984) used a physical spinner in a mail questionnaire. This spinner could however get bend in the mail, and subsequently affect the randomness of the device. Appendix C contains several test-runs proving the randomness of the randomizers as used in this study. The second problem alleviated with a virtual make-up of the randomizing device is the problem of inconclusive outcomes. In the same study by Stem & Steinhorst, an answer could land on a line on their physical spinner, proving the need for additional instructions, which increases the cognitive load for the respondent. Sufficient programming can curb this problem when using virtual randomizers.

The randomizers could be activated by clicking with the mouse pad on 'start spinner' next to the randomizer. A red arrow in the upper-left corner pointed out the area where the spinner stopped, and subsequently directed the respondent to a certain answer according to instructions ('yes' or 'no', or '1' to '6'). The spinners' programming was such that

when started, its speed decreased steadily before stopping on a certain area. This construction was intended to increase perceiving of randomness. Also, the spinner could be started multiple times. This could enhance cheating among respondents when a certain outcome was perceived as counterintuitive. But this drawback was felt to be outweighed by the advantage given by this setup, being that respondents could experience that her or "his selection of questions is determined by chance" (Brown 1975: 27, as quoted by Fox & Tracy 1986: 61). If only one throw preceding each sensitive question were possible, the respondent could very easily develop the notion of the randomizer not being random but acting via a pre-fixed distribution. When suspicious, a respondent must be able to see that his selections are random. Also, when a respondent fully understands the instructions of FRR and the protection it provides, multiple throws looses its necessity. It must be stated however, that neither RR nor the specific programming of the randomizers can curb all distrust regarding sensitive inquiries and the use of computers.

The setup of RR and DQ conditions in a computer-assisted environment makes possible a comparative design for the evaluation of the relative merit of RR in eliciting sensitive information of a dichotomous and quantitative nature. 'Comparative', as the true status on integrity violations of each individual in the sample can never be known, adding to the necessity of testing methods as RR on their relative merits in eliciting sensitive information. In doing so, the estimates obtained with the CARR and CASI formats will also be compared to an external measure, namely: official statistics on the internal investigations conducted by the police force into certain integrity violations by its members. These official statistics provide an external measure regarding prevalence only.

Field Trajectory

Our target sample consisted of 828 persons, all being salaried police employees.[1] These potential respondents had to be divided over the CARR and CASI interviewing conditions. Due to the inefficiency of the RR design, induced by the insertion of random error, which lets its variance inflate, the RR condition requires a sample size larger that the CASI condition to obtain similar levels of precision. If we assume a constant n, it can easily been seen from equations 4.5, 4.9 and 4.13 that the efficiency of the models depends on p (p_1). With our p (p_1) of .75, the

models require a sample size 1.78 times larger in comparison with DQ conditions to achieve the desired precision. This study maintained a sampling size 3 times larger in comparison with the DQ condition. Firstly, to heighten the chances of usable response to this pilot study. Secondly, and connected to the former, to anticipate on possible lower response rates due to the newness of the method to the sample group and to simultaneously meet other factors hampering on possible response (see next section). The CARR and CASI groups were constructed by classifying each fourth name on the full alphabetical employee address list of the police force, in the latter.

Next to the expectation that every officer had to have at least minimal levels of computer literacy, we also found that Internet ownership at the home address was very high among the respondents in our target sample, and that the computer-assisted questionnaires could not be operated from the working address due to specifics in programming which could not be handled by the computer-technical service of the police force. The force banned the use of Flash computer programming as it was seen as unsafe for their protected local intranet. Flash programming however, comes almost standard to every home computer with Internet. Taking into account these factors, and to encourage further cooperation due to convenience and further improvement of anonymity, the respondents were encouraged to cooperate from their home address. To this purpose the CARR and CASI questionnaires were made accessible on two separate Internet sub-domains protected by username and password, which were maintained by the university.

Following an announcement on the police force local intranet introducing the setup and aim of this study, two separate letters were mailed to the respondents' home address, corresponding to their status with regard to the questioning condition to which they were randomly assigned. The letter for the CARR group contained a more elaborate introduction into the nature and aim of our study, along with assurances over their anonymity and instructions regarding the questionnaire. Importantly, the letter announced the Internet address of the sub-domain containing the questionnaire along with username and password for accession. Concluding the letter were two autographs, from the research coordinator and the 'privacy and information confidentiality functionary' of the police force, to enhance trust in the aims and assurances of the research endeavor. The CASI letter was constructed analogously. Both the CARR and CASI respondent letters can be found in Appendix D.

Over a course of four weeks the respondents received several reminders via postings on the police force intranet or via their personal

police email-accounts. Subsequently, a follow-up study was conducted which consisted of announced visits to three larger district bureaus and a two day visit to police headquarters. These visits had three main goals. Firstly, to make possible cooperation to those who had not yet been able to complete the questionnaires, due to undeliverable mail, unavailability of Internet at the home address, or other factors. Secondly, to boost response on our CARR condition, as it was slightly lacking. These two aims were made possible by usage of a laptop, which was turned into a stand-alone server, which contained the CARR and CASI questionnaires. And lastly, to experience first-hand, how the questionnaires and the various questioning conditions were received and perceived by various members of the police force. This experience was gained by conducting short open interviews with police employees on reasons for response or response refusal, on their thoughts and confidence with regard to the questioning methods and on their experiences with the mode of administration.

ANALYSIS AND COMPARISON OF DATA

Taking into account the assumptions regarding RR, previous comparative and validation research on RR and other techniques, and the framework for empirical testing as elaborated in preceding sections, the basis of analysis is provided for here in our research hypotheses. Subsequently, the mode of analysis will be reviewed, dealing with method of estimation and the basis for comparison of obtained estimates.

Research Hypotheses

Although the general assumption states that the use of RR would also reduce nonresponse, as the assurance of anonymity would reduce the burden of cooperation with regard to a survey on a sensitive topic, this will not be an assumption maintained within this specific empirical endeavor regarding RR. As one, especially with the absence of a researcher explaining the technique beforehand, has to actually endeavor on responding to a RR survey before fully grasping its relative merit and assurances with regard to anonymity and data confidentiality, it can be expected that RR will actually have a stronger stance in reducing evasiveness than in reducing response refusal. This expectation, which is also reflected in our sampling procedure, results in the first hypothesis:

H.1: Nonresponse in the randomized response condition (CARR) will not be significantly reduced in comparison with the direct questioning condition (CASI).

Taking into account previous empirical work on RR, along with the assumptions regarding RR in general and the dichotomous FRR model as developed here, the expectation is expressed that RR will lead to more accurate estimates of prevalence. Respondents completing the randomized response questionnaire (CARR) will be expected to give more accurate answers regarding prevalence of their own misconduct than respondents completing the direct questionnaire (CASI), resulting in the second hypothesis:

H.2: The randomized response condition (CARR) will lead to higher estimated proportions of prevalence than the direct questioning condition (CASI) when assessing self-reports of individual misconduct.

The CARR and CASI results on prevalence of certain integrity violations will be compared to official police statistics on the internal investigations conducted by our regional police force into misconduct by her own employees. All Dutch police forces have a structural provision for conducting investigations into police misconduct whenever "there is a suspicion of an infraction of disciplinary or criminal law" (Lamboo *et al.* 2002: 9). From this external measure prevalence estimates can be distilled. An endeavor connected to the third hypothesis. As argued in Chapter 2, there will be a 'dark number' of misconduct in the police organization, which can be expected to remain unknown to those conducting an investigation. Also, one can imagine, as these investigations often follow from reports from the work environment, the reluctance of officers to report misconduct by their colleagues, which is also known as the 'blue curtain' (Klockars *et al.* 2000). As both the CARR and CASI conditions have the advantage of lacking social context cues, one can then expect the inferiority of official statistics in representing prevalence of misconduct, resulting in the third hypothesis:

H.3: The prevalence estimates distilled from official police statistics on internal investigations into misconduct, will be lowest in comparison with CARR and CASI condition estimates.

Taking into account previous empirical work on RR, along with the assumptions regarding RR in general and the discrete quantitative FRR

model as developed here, superiority of the RR condition is expected in eliciting sensitive information of a quantitative nature. It is to be believed that respondents completing the randomized response questionnaire (CARR) will give more accurate answers regarding incidence of their own misconduct than respondents completing the direct questionnaire (CASI), giving the thrust of the fourth hypothesis:

> *H.4*: The randomized response condition (CARR) will lead to higher estimated population proportions for the discrete quantitative categories denoting more frequent incidence than the direct questioning condition (CASI) when assessing self-reports of individual misconduct.

We also delve into respondents' confidence in the various questioning conditions. As respondents in the RR condition receive the assurance of confidentiality of responses by this specific method, and as the assumption is handled that most respondents will understand the protection given with a RR condition, it is expected that there is more trust in the RR condition. Our last hypothesis states:

> *H.5*: Respondents completing the randomized response questionnaire (CARR) will have greater confidence in the anonymity of their answers than respondents completing the direct questionnaire (CASI).

Comparison and Estimation

As this study comprises RR field-testing in a comparative design, where the true status of each individual in the sample regarding the behaviors asked for is not known as there is no criterion measure to which our estimates can be compared, a different method of comparison is needed. As is clear from the hypotheses as formulated above, higher estimates are seen as more accurate estimates. This method of comparing estimates from RR and other condition is used in all RR studies with a comparative design.

Umesh & Peterson (1991) criticize comparative RR designs. Their first critique is that comparative studies are often posed as validating RR when higher estimates are obtained. It is made very clear here, that we acknowledge that a validation effort is not conducted or possible. The comparative design is elaborated in preceding sections. Their second critique is that higher reported prevalence or incidence is not necessarily more accurate as without criterion measures, RR estimates can, as they argue, be overestimated as well as underestimated. This critique is

rejected on two grounds. Firstly, methods like RR are intended to provide accurateness of representation where other measures are unavailable, biased or incomplete. To demand criterion measures would not only deny the necessity of methodological advance in enhancing the accurateness of representing the latent with the manifest. It also renders out as useful and necessary the collection on, from a scientific and/or policy view interesting behaviors on which no criterion measures are available or possible (such as for example prevalence and incidence of unprotected sexual intercourse). Moreover, the argument by Umesh & Peterson seems to mix arguments of social desirability and undesirability. As we are assessing behaviors, which are generally perceived as undesirable (by law or code) they are likely to be underestimated. Overestimation can only occur if a (substantive) part of the respondents state that they indulge in certain undesirable behaviors, to an extent that exceeds their true status. Dealing with socially undesirable traits, and having clear RR instructions, it is very safe to assume that displaying such exhibitionistic behavior – if any – is very unlikely to occur. A remark empirically backed by the validation study of Locander, Sudman & Bradburn (1976). Given the status of sensitive behaviors and the arguments given in preceding chapters and sections, it can be expected that the behaviors of which estimates are desired will still be underestimated, as it is safe to assume that not everyone possessing a certain sensitive trait will disclose information even on a probability basis, but that the underestimation will be smaller in comparison with other, more traditional measures. The assumption of higher estimates as more accurate estimates is thus perceived to be very safe.

The estimation of prevalence in our DQ setting is straightforward. They can be obtained by using equation 2.2 where Z_i can then denote a dichotomous or quantitative observed numerical. In RR settings estimation is somewhat harder as the given equations reveal relatively more complexity. Also, all given RR estimates are moment-estimates. These are capable of producing estimates outside the probability range [0,1] for the dichotomous RR models and outside our categorical quantitative range [1,...,6]. Maximum likelihood (ML) estimates have to be produced to keep the estimates within the specified parameter space (Bourke & Moran 1984, 1988).[2] As immediate solutions for finding ML estimates in complex RR models are not available, iterative optimization methods are needed.

RR data can be seen as purposively misclassified data (Chen 1979, Van Den Hout & Van Der Heijden 2002, 2004; Van Den Hout 2004).

Van Den Hout & Van Der Heijden (2004) have shown that the general RR design given in equations 4.15 and 4.16 can be described by a latent class model (LCM), as the misclassification in a RR design can be seen as analogous to a log-linear[3] model with one or more categorical latent variables. A program, which has proved to be very useful for fitting log-linear models when one or more categorical variables are measured in a RR condition, is the ℒEM program by Vermunt (1997).[4] This general program for categorical data has proven to be able to handle the log-linear parameterization of an LCM, which was subject to a RR condition (see Van Den Hout & Van Der Heijden 2004, for this specific parameterization). It uses the expectation maximization (EM) and Newton-Rhapson algorithms for computing ML estimators.[5]

In many cases the ML estimates obtained with the LCM analogy in the ℒEM program will be the same as those obtained with the moment estimates. The use of ML estimates as in ℒEM restricts the estimates however, in numerically perverse situations, to their natural boundaries. Appendix E gives the basic ℒEM input and explanation for obtaining ML estimates for our dichotomous and quantitative FRR models. We now turn to Chapter 6 to review the results of our empirical comparative RR endeavor.

NOTES

[1] It must be stated here, that our target sample is actually a finite population. Most studies assume a simple random sample scheme with replacement. When drawing a sample from a population from which not every member can be identified, this is an assumption that can be maintained. When sampling in a finite population of N persons, where we have N identifiable units, the variance of the given estimators need a correction for sampling without replacement. This is the so-called 'finite population correction', which consists of

$$\left[1 - \frac{n-1}{N-1} \right]$$

with N denoting the population, and n giving the sample size. This correction, when taken into account in a RR setting, is multiplied with the normal sampling variance and added to the variance induced by the specific RR model (in this case dichotomous and quantitative FRR models). Its multiplication with normal sampling variance alone gives the finite population sampling corrected variance of DQ settings.

This finite population correction has a negligible effect on the variance and thus the standard error (square root of variance divided by square root of n), unless n is (very) large relative to N. As this is not the case in this study, while mostly ignored in most studies, and

expressing the necessity to not further complicate matters in this pilot study, we will confine by referring to this specificity with this note. For more on finite population corrections and RR sampling see Kim & Flueck (1978).

[2] ML parameter estimation determines the parameters that maximize the probability of sample data.

[3] Log-linear analysis is a multivariate extension of the Chi-square test of independence. For more on ML estimation we confine by referring to Eliason (1993), for more on log-linear analysis we confine by referring to Hagenaars (1993).

[4] The £EM program and manual can be downloaded for free from URL: http:// www. uvt.nl /faculteiten /fsw/organisatie/departementen/mto/software2.html

[5] The program starts with the EM algorithm, which views the data as mixture or incomplete data, which iteratively constructs optimized lower bounds for prior distributions. When close to the maximum £EM uses the root finding Newton-Rhapson algorithm. For more on the EM and Newton-Rhapson algorithms we confine by referring to Louis (1982), McArthur (1993) and Delleart (2002).

APPENDICES

6. TESTING THE FORCED RANDOMIZED RESPONSE MODELS ON INTEGRITY VIOLATIONS: RESULTS FROM THE FIELD EXPERIMENT

A unified approach to eliciting sensitive dichotomous and quantitative data was presented in Chapter 4. New randomizers for dichotomous and quantitative FRR endeavors were developed along with a new multi-proportional RR approach within the framework of FRR modeling to sensitive quantitative characteristics. A framework was presented in which the practical feasibility of RR as translated to the given models was empirically tested. This framework entailed comparing RR with the more traditional DQ technique by subjecting an experimental group to CARR, and a control group to CASI. The translation of theoretical potentiality to practical feasibility was empirically tested with regard to integrity violations in a police force, which will hold as a test case for sensitive behaviors of a political nature. This chapter reports on the results of this empirical effort, by reviewing how the RR technique performs in obtaining estimates of prevalence and incidence of sensitive behaviors of a political nature, in comparison with a DQ condition.

The results will be presented analogous to the structure of the hypotheses given in Chapter 5. Firstly, inspired by hypothesis 1, we will delve into questions of response and nonresponse, and the factors that influenced both in our fieldwork efforts. Secondly, we will delve into the height of the prevalence and incidence estimates obtained with the CARR and CASI conditions, and the prevalence estimates obtained with official police statistics. In doing so, we will review the second to the fourth hypothesis. Lastly, we will evaluate reported trust and confidence by

respondents in the two questioning conditions of our experiment, embarking on our last hypothesis. The results and their discussion as presented here, serve as a stepping-stone for the discussion of RR and the framework of our empirical endeavor regarding RR, presented in Chapter 7.

RESPONSE AND NONRESPONSE

The general assumption in most empirical work on RR is that the technique, next to evasiveness, is also able to reduce response refusals, as the assurance of anonymity would reduce the burden of cooperation with regard to sensitive topics. We hypothesized that nonresponse would not be significantly reduced in our RR setting in comparison with the DQ setting, as we expected that RR will have a stronger stance in reducing evasiveness than in reducing response refusal. This could especially be the case when the RR technique is incorporated into a computer-assisted environment, with no researcher present to express the relevance of the method. Table 6.1 gives the sample characteristics of our study with which we can assess the first hypothesis.

Table 6.1: Sampling Characteristics

Characteristic	CARR	CASI	Total
Target sample	621	207	828
Number of questionnaires completed via Internet	78	45	123
Number of questionnaires completed during follow-up visits	45	-	45
Total number of completed questionnaires	123	45	168
Effective response rate in %	19.8%	21.7%	20.3%

As stated in the preceding chapter, our target sample consisted of all salaried employees of one regional police force in the Netherlands. Our experimental group subjected to the CARR questionnaire was randomly constructed to be three times larger than the control group subjected to

the CASI questionnaire, giving a potential respondent sample of 621 and 207 persons for the CARR and CASI conditions respectively. The CARR and CASI questionnaires were made accessible on two separate Internet sub-domains protected by username and password.

Over a course of four weeks 78 and 45 questionnaires were retrieved for CARR and CASI respectively via the Internet option. Subsequently, a follow-up study was conducted, consisting of announced one-day visits to three larger district bureaus and a two-day visit to police headquarters. This follow-up study sought to enable cooperation to those who were willing but not previously able, and to experience first-hand how the questionnaires and the various questioning conditions were received and perceived by various members of the police force. Also, as can be seen in Table 6.1, the follow-up study sought to boost response to the CARR condition, so as to obtain a response with acceptable statistical power. While from a methodological view, this matter of conduct cannot be considered optimal as additional factors are brought into the research (such as the presence of a researcher), it must be considered important as it is (as will become clear in this and the subsequent chapter) able to enhance our understanding of the practical properties of RR in field-research settings. As this study can be considered a pilot, these endeavors are justified on the basis of additional learning.

Through the follow-up, an additional 45 respondents were included in the CARR condition, giving an effective response rate of 19.8% and 21.7% for the CARR and CASI conditions respectively, and an overall effective response rate of 20.3%. As, as well as on the Internet sub-domains as on the laptop used for the follow-up study, respondents' answers were only stored in a sealed database when all questions where answered, the total number of completed questionnaires give the effective response rate. Evaluating the response rates in Table 6.1, where we find the effective response rate of RR – even after follow-up study – to be below the effective response rate of the DQ setting, we can confirm H1: the randomized response condition did not reduce nonresponse in comparison with the direct questioning condition.

Some researchers, who were also unable to establish superiority of RR in reducing nonresponse, have argued that the additional time associated with completing a RR survey (due to additional instructions and operating the randomization procedure) could account for their findings (Armacost *et al.* 1991; Houston & Tran 2001). This is not a viable explanation in this study, as the virtual setup of the questionnaires provided for a RR setting, which was not significantly more time-consuming than the DQ setting. Both questionnaires, as the personal encounters with the respondent

group proved, took no longer than 15 minutes to complete. It can be argued that the RR technique has a stronger stance in reducing evasiveness than in reducing response refusal. One has to actually endeavor on responding to questions in a RR condition before fully grasping its relative merit and assurances with regard to anonymity and data confidentiality. This could especially be the case in the research design as developed in the previous chapters, which made use of a computer-assisted environment, so that initially no researchers were present who could explain the technique beforehand so as to overcome initial burdens with regard to response.

Overall, the response rate was rather low. Several factors could account for this. Firstly, from the open interviews conducted during our follow-up study it became clear that the police employees suffered from 'questionnaire fatigue'. Police employees receive several questionnaires each year, from within the police force, but also from public and commercial research institutes. This has spurred some disinclination or aversion against research. Moreover, some weeks prior to the survey reported on here, our regional police force received a questionnaire also handling integrity related subjects. This scenario-based questionnaire, as was learned, took more than 50 minutes to complete. This research regarding the same subject so shortly preceding our own has – as several police employees have stated – furthered reluctance towards our endeavor. There were also several officers who mistook our comparative RR research for the preceding scenario-based research. Another factor, which has hampered response, was negative media reports. There were several nationwide media messages in popular dailies and news programs during the field-process of this research, which stated that the integrity problem in the Dutch police force could be greater than previously assumed. Negative news exposure can be said to enhance feelings of 'group jeopardy' and to negatively affect disclosure of information. We will return to this topic in subsequent sections.

More specifically, Table 6.1 makes clear that the response rate to the CARR questionnaire was significantly smaller initially. The RR procedure could have a counterproductive effect, namely that of inducing suspicion. We have elaborately assured respondents of their privacy, with an assurance of privacy by the RR procedure for the CARR respondents. However, there is evidence that emphasizing assurances of confidentiality can spur perceiving of sensitivity (Reamer 1979; Singer, Von Thurn & Miller 1995). This effect, coupled with the relative obscurity of the RR technique, could offset suspicion in the police as a respondent group, which by its very nature operates on doubt and questioning of assurances.

The follow-up study confirmed these emotions among respondents in the CARR condition, pointing to a portion of officers who seriously doubted the relative merit of RR or its potential for estimating certain behaviors. The factor of suspicion has, aided by other specifics traits of our respondents group, hampered response to the CARR questionnaire, and – as we will see in the remainder of this chapter – left its mark on the desired estimates.

ESTIMATING PREVALENCE WITH THE DICHOTOMOUS FORCED RANDOMIZED RESPONSE MODEL

Taking into account previous empirical work on RR, along with the assumptions regarding RR in general and the dichotomous FRR model as developed in Chapter 4 specifically, the expectation was expressed that the RR condition would lead to more accurate estimates of prevalence. The second hypothesis gave the expression of this expectation, stating that with the CARR questionnaire prevalence estimates would be obtained that are superior to the CASI condition, when assessing self-reports of individual misconduct.

The data obtained could be specified on the basis of the distinction 'executives/non-executives'. One question in the survey was asked to divide between executive and non-executive police employees, as with some prevalence and incidence questions, this distinction could prove to be important. The data however, showed that executives as well as non-executives display all behaviors. With the benefit of hindsight it must also be stated that the question may not have delineated the distinction between executive and non-executive in a manner that is unambiguous. Nonetheless, tests were conducted to establish differences between prevalence estimates for executive and non-executive police employees within the questioning conditions. For the RR data this test entailed a LCM model in the ℒEM program, which compared prevalence for the categorical subgroups 'executive' and 'non-executive'. By fitting a saturated LCM model which takes into account main and interaction effects between the specified latent (the socially sensitive behavior) and manifest variables (response to question regarding socially sensitive behavior, and our categorical variable giving 'executives' and 'non-executives'), it was possible to obtain precise prevalence estimates for both subgroups. For specifications of these models see Appendix E. No significant differences were found.

As our CARR response was attributable to response obtained via the Internet sub-domain and response obtained during our follow-up study, there was a possibility of differences between the heights of the prevalence estimates distilled from these sub-conditions. This specification in our response obtainment was treated as a categorical variable so as to make possible the comparison of prevalence estimates obtained via the Internet and those obtained via the follow-up study. Again, LCM analogies in the ℒEM program were constructed to test the difference and precise estimates of our latent (the socially sensitive behavior) and manifest variables (response to question regarding socially sensitive behavior, and our categorical variable giving response via Internet and response during follow-up). Again, for specifications of these models see Appendix E. No significant differences were found between the prevalence estimates that were obtained with the CARR questionnaires that were completed on the Internet and those obtained with the CARR questionnaires that were completed on the laptop during the follow-up study. The direction of the differences did favor the laptop in almost all cases. However, in comparison with the prevalence estimates obtained with the CASI condition, the image of comparison would not be fundamentally altered when using the slightly higher CARR estimates obtained with response during follow-up study. We therefore compare prevalence estimates for the CARR and CASI conditions obtained over the total sub-samples.

Table 6.2 (page 84) gives an overview of the desired prevalence estimates regarding ten sensitive behaviors for both the CARR and CASI conditions. The affirmation rates yielded with the RR method are, contrary to our expectation, not appreciably higher than the affirmation rates obtained with the DQ setting. Only regarding 'personally taken possessions of the police force' does the CARR condition lead to an appreciably higher prevalence estimate. The prevalence estimates in all other cases are highly similar between the questioning conditions, with CASI slightly outperforming CARR in yielding truthful affirmative responses.

Differences in prevalence estimates between CARR and CASI conditions were tested through a z-score normal curve deviate of the difference between two proportions. This difference of proportions test in our setup is denoted by

$$z = \frac{\pi x'carr - \pi x'casi}{\sqrt{\frac{\lambda'(1-\lambda')}{(n-1)p1^2} + \frac{\pi x'(1-\pi x')}{n-1}}} \tag{5.1}$$

where $\pi x'carr$ denotes $\pi x'$ for the CARR condition and $\pi x'casi$ denotes $\pi x'$ for the CASI condition. The difference between proportions is divided by the square root of the sum of the variance of the dichotomous FRR model and the normal sampling variance of a DQ setting. The critical value for a one-sided test of size $\alpha = .05$ equals ± 1.645.

We have already seen from Table 6.2 that the CARR condition slightly outperformed the RR setting when neglecting statistical significance. When applying the difference of proportions test, none of the differences tested significant however. Our second hypothesis is thus rejected and on the basis of our findings here, the null hypothesis seems tenable, as the prevalence estimates obtained with the CARR condition do not differ statistically from the prevalence estimates obtained with the DQ method in the CASI condition.

The lack of difference in performance could indicate that we have approached the upper bound of misconduct with the two techniques, as we have no criterion measures to which our estimates can be compared. From the reasoning in preceding chapters it is stated that such is highly unlikely. Another possible explanation for the similar performance of the questioning conditions is that the questions are not sensitive enough to the specific respondent group so that the relative merits of RR are offset by a lack of representational incentives. A RR study by Berman, McCombs & Boruch (1977) partially suffered from such a situation. However, as stated in Chapter 5, the questions were devised in cooperation with police officers to represent integrity violations with a broad variability in sensitivity and expected occurrence. For the above given reasoning to be empirically backed we would then expect increasing performance of RR when subjective sensitivity rises. No such trend can be discovered in Table 6.2 however. A more plausible explanation could be that the counter-intuitiveness of a forced answer and the suspicion induced by the RR method itself, in combination with certain specific traits of the police as a respondent group, have nullified the theoretical potential of our dichotomous FRR model in eliciting sensitive information on individual behavior. We will delve thoroughly into this issue after reviewing the incidence estimates obtained with our quantitative FRR model.

Table 6.2: Prevalence Estimates Obtained with CARR, CASI and Official Statistics

Have you ever in the past twelve months	π_x' CARR	π_x' CASI	z-value of difference	prevalence estimates distilled from official statistics
1. accepted gifts with a value below 25 euros from externals (as citizens and shop holders)?	.092 (.051)	.111 (.047)	-.274	-
2. unjustifiably declared working hours?	.00 (.042)	.00 (.00)	-	-
3. been careless with confidential police information?	.016 (.046)	.022 (.022)	-.110	.006
4. twisted the true cause of a case?	.016 (.046)	.089 (.043)	-1.153	.001
5. threatened with the use of violence during an interrogation?	.00 (.033)	.044 (.031)	-.972	-
6. in exchange for compensation (in money or kind) given confidential police information to third persons?	.00 (.036)	.00 (.00)	-	-
7. incorrectly taken care of arrestees?	.00 (.038)	.022 (.022)	-.501	-
8. personally taken possessions of the force?	.157 (.054)	.111 (.047)	.645	.006
9. participated in the harassing and bullying of a colleague?	.049 (.049)	.044 (.031)	.083	-
10. in your free time associated with persons, which are known to have criminal antecedents?	.049 (.049)	.089 (.043)	-.619	.001

NOTES: Given proportions are estimates of prevalence for the CARR and CASI conditions. Standard errors in parentheses. Difference in proportions tested with z-score normal curve deviate of the difference between two proportions. CARR = Computer Assisted Randomized Response; CASI = Computer Assisted Self-Interviewing.

All Dutch police forces have a structural provision for conducting investigations into police misconduct whenever a certain infraction of a behavioral code justifies such an endeavor. From these official police statistics, prevalence estimates can be distilled. The fourth column of Table 6.2 gives the estimates distilled from official statistics on the

regional police force in which we have found our respondent group. Several remarks must however be made.

Firstly, our research covers misconduct over the twelve months ranging from mid-2004 to mid-2005. As official statistics are not yet available for the year 2005, these statistics cover the twelve months of 2004. As no trends were reported in these official statistics by the structural police provision conducting the internal investigations, the data are assumed to be representative. Secondly, the forms of misconduct investigated can slightly differ from our exact formulations in the questionnaires. Therefore, assignment of investigations to the categories of misconduct denoted with our specific questions cannot but be somewhat arbitrary. Thirdly, as stated before, these data regard investigations, and not all infractions were confirmed.

In total 8 investigations were conducted that were analogous to (a part of) the misconduct sought for with the questionings. They took place over a span of twelve months and 12 police employees were involved in the investigations as subjects. Prevalence estimates were established by simply dividing the number of police employees involved in a certain category of misconduct by the total number of employees in the force. When no estimate is given, no research into a certain type of misbehavior was conducted.

The official data make clear that there is a 'dark number' of misconduct, which remains unknown to those conducting investigations. All estimates are substantially smaller than the prevalence estimates yielded with the CARR and CASI questionnaires. Also, into many forms of misconduct no investigations were enacted, possibly pointing to contingencies inherent to the statistics and the investigations. The third hypothesis is posed to be confirmed. As both the CARR and CASI conditions have the advantage of less social context cues, the official statistics in comparison give a representation of prevalence that is inferior.

ESTIMATING INCIDENCE WITH THE MULTI-PROPORTIONAL DISCRETE QUANTITATIVE FORCED RANDOMIZED RESPONSE MODEL

Taking into account previous work on RR along with the assumptions regarding RR in general and the multi-proportional discrete quantitative FRR model as developed in Chapter 4 specifically, the expectation was expressed that the RR condition would give superior performance in

eliciting sensitive information of a quantitative nature. The fourth hypothesis expressed this expectation by stating that CARR in comparison with CASI would lead to higher estimated population proportions for the discrete quantitative categories denoting more frequent incidence, when assessing self-reports on individual misconduct.

Again the data could be specified on the basis of the 'executive/non-executive' and 'CARR questionnaire completed via Internet/CARR questionnaire completed during follow-up study' distinctions. The estimated proportions $\pi_1,\ldots, \pi_6$ were compared for the indicated categorical subgroups (see Appendix E for specifications). No significant differences between incidence estimates were found for eight of the questions regarding frequency of misconduct when sub-categorizing data along the lines of questionnaire completion. There were significant differences favoring laptop for 'applying inappropriate and disproportional violence', and 'utilization of unlawful investigative methods'. In almost all other non-significant comparisons the direction of the estimates also favored the estimates derived from the laptop used during the follow-up study. However, in comparison with the incidence estimates obtained with CASI, the image of comparison would not be fundamentally altered when using the slightly higher CARR incidence estimates obtained during follow-up study. Also, no unambiguous differences between incidence estimates were found when sub-categorizing data along the lines of the executive/non-executive division. The incidence estimates are therefore compared over the total CARR and CASI sub-samples. Table 6.3 gives the overview of our desired incidence estimates.

Differences in estimated proportions for $\pi_1,\ldots, \pi_6$ between CARR and CASI conditions were again tested through a z-score normal curve deviate. In the multi-proportional setup, this difference of proportions test is given by

$$z = \frac{\pi i\text{'carr} - \pi i\text{'casi}}{\sqrt{\frac{\gamma i'(1-\gamma i')}{(n-1)p^2} + \frac{\pi i'(1-\pi i')}{n-1}}} \qquad (5.2)$$

where πi'carr denotes πi' for the CARR condition and πi'casi denotes πi' for the CASI condition, with $i = 1,2,3,4,5,6$. The difference between the two proportions is divided by the square root of the sum of the variance of the multi-proportional quantitative FRR model and the normal

sampling variance of a DQ setting. The critical value for a one-sided test of size $\alpha = .05$ equals ± 1.645.

If we translate the expectations expressed in the fourth hypothesis to our data, we would then ideally expect to find a significant higher estimated population proportion π_1 (frequency of 0) for the CASI condition and significantly higher estimated population proportions π_2, π_3, π_4, π_5 and π_6 for the CARR condition. From Table 6.3 we see that with regard to 'wrongly reporting sick', and 'acceptance of money for emphasizing or neglecting different tasks', the CARR and CASI conditions give exactly similar performances with the population proportion in π_1 being 1 for both.

Table 6.3: Incidence Estimates Obtained with CARR and CASI

How many times in the past twelve months have you	π_1' 0 times	π_2' 1 time	π_3' 2 to 3 times	π_4' 4 to 5 times	π_5' 5 to 10 times	π_6' more than 10 times
1. used organizational resources for private purposes?						
π_i CARR	.747(.059)	.107(.040)	.064(.034)	.053(.033)	.010(.026)	.020(.028)
π_i CASI	.622(.073)	.089(.043)	.20 (.060)	.089(.043)	.00 (.00)	.00 (.00)
z-value of difference	1.332	.306	-1.972*	-.664	.385	.714
2. consulted police records for family and friends?						
π_i CARR	.903(.052)	.090(.038)	.00 (.024)	.01 (.026)	.00 (.015)	.00 (.021)
π_i CASI	.933(.038)	.044(.031)	.022(.022)	.00 (.00)	.00 (.00)	.00 (.00)
z-value of difference	-.466	.938	-.676	.385	-	-
3. wrongly reported sick?						
π_i CARR	1 (.044)	.00 (.011)	.00 (.021)	.00 (.021)	.00 (.024)	.00 (.024)
π_i CASI	1 (.00)	.00 (.00)	.00 (.00)	.00 (.00)	.00 (.00)	.00 (.00)
z-value of difference	-	-	-	-	-	-
4. applied inappropriate and disproportional violence towards citizens, suspects or arrestees?						
π_i CARR	.991(.043)	.004(.026)	.00 (.019)	.004(.026)	.00 (.015)	.00 (.011)
π_i CASI	.978(.022)	.022(.022)	.00 (.00)	.00 (.00)	.00 (.00)	.00 (.00)
z-value of difference	.269	-.528	-	0.154	-	-
5. accepted money or gifts to emphasize or neglect certain tasks in your function as police officer?						
π_i CARR	1 (.041)	.00 (.021)	.00 (.021)	.00 (.021)	.00 (.011)	.00 (.021)
π_i CASI	1 (.00)	.00 (.00)	.00 (.00)	.00 (.00)	.00 (.00)	.00 (.00)
z-value of difference	-	-	-	-	-	-
6. (sexually) intimidated a colleague?						
π_i CARR	.983(.047)	.017(.028)	.00 (.024)	.00 (.021)	.00 (.024)	.00 (.015)
π_i CASI	1 (.00)	.00 (.00)	00 (.00)	00 (.00)	00 (.00)	00 (.00)
z-value of difference	-.362	.607	-	-	-	-
7. reported untruthful information in charges or reports?						
π_i CARR	.994(.045)	.00 (.019)	.00 (.024)	.006(.026)	.00 (.015)	.00 (.021)
π_i CASI	.978(.022)	.00 (.00)	00 (.00)	00 (.00)	00 (.00)	.022(.022)
z-value of difference	.319	-	-	.231	-	-.723

Table 6.3 continued

8. taken found or confiscated goods or items?						
πi CARR	.958(.049)	.018(.028)	.00 (.015)	.007(.026)	.018(.028)	.00 (.021)
πi CASI	1 (.00)	.00 (.00)	.00 (.00)	.00 (.00)	.00 (.00)	.00 (.00)
z-value of difference	-.857	.643	-	.629	.643	-
9. utilized unlawful investigative methods?						
πi CARR	.954(.050)	.018(.028)	.00 (.024)	.029(.030)	.00 (.021)	.00 (.018)
πi CASI	1 (.00)	.00 (.00)	.00 (.00)	.00 (.00)	.00 (.00)	.00 (.00)
z-value of difference	-.920	.643	-	.967	-	-
10. used drugs in your free time?						
πi CARR	1 (.038)	.00 (.018)	.00 (.021)	.00 (.015)	.00 (.015)	.00 (.018)
πi CASI	.978(.022)	.00 (.00)	.00 (.00)	.00 (.00)	.022(.022)	.00 (.00)
z-value of difference	.501	-	-	-	-.826	-

NOTES: Given proportions are the estimated populations proportions πi for CARR and CASI conditions. Standard errors in parentheses. Difference in proportions tested with z-score normal curve deviate of the difference between two proportions. CARR = Computer Assisted Randomized Response; CASI = Computer Assisted Self-Interviewing. * $p \leq .05$

With regard to 'use of organizational resources for private purposes', 'reporting untruthful information in charges and reports', and 'usage of drugs in free time', CASI slightly outperforms CARR with the former having a smaller estimated population proportion in π_1 and slightly higher estimated population proportions in one or more of the subsequent categories. Only with regard to 'consultancy of police records for family or friends', 'taking of found or confiscated goods or items', and 'utilization of unlawful investigative methods', do we find a performance of CARR that is in the direction of our expectation. None but one of the differences proved to be statistically significant in favor of CASI (π_2' for usage of organizational resources for private purposes). Our fourth hypothesis is thus rejected on the basis of our findings. The null hypothesis seems tenable on the basis of the given data: the incidence estimates obtained with the RR method in the CASI condition do not differ appreciably from the incidence estimates obtained with the DQ method in the CASI condition.

Again, the lack of statistical significance in the difference in performance could indicate approaching the upper bound of possible estimates or a lack of sensitivity in our questions. These explanations are rejected on the same grounds as given in the preceding section. As stated, the counter-intuitiveness of a forced answer and the suspicion induced by

the RR method itself, in combination with certain specific traits of police employees, could have nullified the theoretical potential of the FRR models as developed in Chapter 4. It is to this issue we now turn.

TRUST, UNDERSTANDING AND THE COUNTER-INTUITIVENESS OF A FORCED ANSWER

In the preceding section the superiority of RR in eliciting information of a sensitive nature, could not be proven statistically. With no statistically significant differences established between CARR and CASI our null hypotheses regarding performances between the two questioning conditions in eliciting information of a sensitive dichotomous and quantitative nature seem tenable. When neglecting the properties of statistical significance however, the CASI condition even performs better in several instances. With our research design the possible explanations of reaching upper bounds of possible estimates and lack of sensitivity in questions, are rejected. Our explanation lies in the psychological properties of 'trust', 'understanding' and the related issues of 'counter-intuitiveness' and 'suspicion'.

It is safe to assume that in our RR condition the respondents who were directed to give a truthful answer were not less inclined to do so, than the respondents in our DQ setting. Under this assumption, the lower performance of the CARR condition can only be explained by response falsification on the part of those who were redirected to give an answer that was opposed to their true status. This 'cheating' on the FRR procedure can be expected to be especially apparent for those redirected to give an answer that exceeded their true status, or, in other words, for the 'innocent' who are encouraged to falsify their answers (Miller 1981). Thus, while the RR technique is directed to relieve self-representational concerns, in the setup as presented in this text, it can also actually spur concerns regarding the representation of self for those who are directed to say 'yes' when their true status is 'no', or those directed to a quantitative answer that severely exceeds their own. This effect is able to nullify the relative merit of RR for the total sample researched. We have several empirical indicants for this explanation.

Firstly, during our open interviews with police employees who had already cooperated or agreed to cooperate, that were conducted while endeavoring on the follow-up study, a majority declared that they were unwilling to give a forced answer that exceeded their true status when directed to do so. They found such a redirected answer to be

counterintuitive. This counter-intuitivism is, secondly, also represented by the data on the propositions for the evaluation of our questioning methods, with which the CARR and CASI questionnaires were concluded. These data can be found in Table 6.4.

Table 6.4: Propositions on the Evaluation of the CARR and CASI Questioning Conditions

Proposition	(strongly) disagree	do not disagree, but also do not agree	(strongly) agree
1. I found it easy to answer questions with the help of a computer			
CARR	4%	5.7%	90.3%
CASI	0%	2.2%	97.7%
2. I am confident that all answers are fully anonymous on the individual level			
CARR	4.9%	14.6%	80.4%
CASI	8.9%	11.1%	80%
3. I am confident that nobody can know if I responded truthfully or according to the spinner directions.			
CARR	8.1%	29.3%	62.6%
CASI	-	-	-
4. With the use of the computer-assisted questionnaire it became easier for me to respond truthfully to sensitive questions on which one has to answer with 'yes' or 'no'.			
CARR	..	..	..
CASI	35.6%	33.3%	31.1%
5. It was clear to me when to answer with 'yes' and when to answer with 'no'.			
CARR	1.6%	4.9%	93.5%
CASI	-	-	-
6. With the use of the spinner it became easier for me to respond truthfully to sensitive questions on which one has to answer with 'yes' or 'no'.			
CARR	52.8%	29.3%	17.8%
CASI	..	..	..

Table* 6.4 *continued

7. With the use of the computer-assisted questionnaire it became easier for me to respond truthfully to sensitive questions on which one has to answer with a certain number.			
CARR	-	-	-
CASI	35.6%	37.8%	26.6%
8. It was clear to me when to answer with a certain number.			
CARR	5.7%	13%	81.3%
CASI	-	-	-
9. With the use of the spinner it became easier for me to respond truthfully to sensitive questions on which one has to answer with a certain number.			
CARR	52.9%	29.3%	17.9%
CASI	-	-	-

NOTES: Propositions 1 and 2 were posed to both CARR and CASI respondents. Propositions 4 and 7 were posed to CASI respondents only. Propositions 3, 5, 6, 8 and 9 were posed to CARR respondents only. CARR = Computer Assisted Randomized Response; CASI = Computer Assisted Self-Interviewing.

The first proposition proves the levels of computer literacy needed to adequately cooperate with computer-assisted research to be present in both the CARR and CASI respondent groups. The second proposition shows that with similarly high levels of confidentiality in the anonymity of elicited data on the individual level (80.4% and 80% respectively for the CARR and CASI conditions), confidence is not the crux of our story. This also leads to a rejection of the fifth hypothesis, which assumed superior performance to be related to superior level of confidence in the anonymity of data. Other constructs are of importance here. Again, we can trace the issue of self-representational concerns due to the counter-intuitiveness of the forced answer. More than 90 percent of respondents in the CARR condition declare to understand when to answer with 'yes' or 'no' and more than 80 percent analogously declares to understand

when to answer with a certain number. Simultaneously, more than 50 percent of CARR respondents state that the use of the 'spinner method' (as the FRR procedure was called in the questionnaires) does not make it easier to respond truthfully to dichotomous or quantitative questions of a sensitive nature (propositions 6 and 9). The data on the analogous propositions posed to the CASI respondents regarding their propensity for responding truthfully to sensitive questions (propositions 4 and 7) reveal more appreciable percentages.

Our third empirical indicator for the nullification of the relative merit of RR in eliciting sensitive information by the self-representational concerns spurred by forced answers is given by raw cell counts. When we would assume that none of the behaviors sought for are present in our population, then the proportion of affirmative responses to our dichotomous question would be equal to the probability of a forced 'yes', namely 1/6. The analogous population proportions π_2, π_3, π_4, π_5 and π_6 for our quantitative questions would then amount to 1/24 each. Raw cell counts for both types of questions differ more often from these probabilities than found by chance, pointing to cheating on the RR procedure.

There are also other factors, which can interact with the cheating found on the FRR and the subsequent nullification of the relative merit of RR. Firstly, it is possible that, when not being able to grasp that unbiased estimates are mathematically discernable from RR data, respondents are encouraged to not respond truthfully as they feel that there is random error in the data anyway. Several respondents' reactions pointed to this factor. Although it is clear to these respondents when to answer with 'yes', 'no' or a certain number, it is not fully understood that the RR procedure protects anonymity and still makes possible the estimation of group parameters. No systematic research was developed into this explanation, but it seems a plausible additional factor as the average police officer is cognitively less sophisticated. Secondly, an elaborate emphasizing of confidentiality can raise suspicion by itself, enhancing perceived sensitivity (Berman, McCombs & Boruch 1977; Reamer 1979; Singer, Von Thurn & Miller 1995). The presence of a researcher could then account for the slightly better performance of CARR during the follow-up study, as a researcher can have a motivational and moderating effect and is able to provide for extra explanation not possible with fully computerized research settings. No systematic research into this effect was however conducted.

Moreover, the specific traits of the police officer could have furthered the incentives for non-compliance with the FRR method. Police officers

are procured for their expected honesty and their natural inclination towards suspicion. This could make them less inclined towards cooperation with RR than other groups. More importantly, previous RR studies were performed on general populations, where the threat of disclosure of sensitive information can be said to be mainly individual. Our empirical endeavor tested the RR in an organizational population, which could be – especially when dealing with police – sensitive to issues of 'collective jeopardy'. Here, the disclosure of information does not hold a threat for the individual only, but also for the police as a group, which – when recalling the negative media-coverage of police during our fieldwork – can have furthered incentives for non-compliance with FRR rules.

Reflecting on the empirical findings the limitations of present study are clear. It is the translation of theoretical potentiality to practical feasibility where RR in general and FRR specifically find its sore spots. The underlying ambition in the empirical part of this study was to modestly contribute to a more definitive protocol for and understanding of RR and to modestly spur cumulative work on RR designs in actual research settings. Given our findings, it is the cognitive processes underlying response to different RR settings in relation to subtleties in technicalities where RR needs further attention. The findings here, give ample opportunity to recommence and revitalize research on RR regarding translation of theory to practice. It is to these issues we direct our attention in a discussion of the statistical and psychological properties of RR, with an emphasis on the empirical trajectory as developed in this text, and the subsequent possibilities for further research.

7. DISCUSSION: STATISTICAL AND PSYCHOLOGICAL PROPERTIES OF RANDOMIZED RESPONSE DESIGNS

A critique on many applications of the RR technique is that they are primarily demonstrative (Umesh & Peterson 1991), aimed at statistical illustrations of certain highly technical derivations. While certain parts of the preceding text were explicitly aimed at statistical and technical improvements in estimating sensitive characteristics with RR, the empirical endeavor regarding the testing of the developed FRR models also sought to bridge the gap between technicalities and application for practitioners in the field. In doing so, a more substantive application was provided for, by comparing the practical feasibility of RR as translated to the given models with the more traditional DQ technique, within a computer-assisted environment outside the walls of the cognitive laboratory, on eliciting information on sensitive behaviors of a more political nature.

Superiority of RR in eliciting sensitive information of a dichotomous and quantitative nature could not be established within our empirical framework. Performances between the CARR and CASI conditions were statistically similar. When statistical significance was ignored, CASI even outperformed CARR in several instances. Fox & Tracy (1986: 64) have questioned the comparison of RR with other anonymous methods[1], and have stated that a more appropriate standard could be found in the comparison of RR with non-anonymous survey techniques. For the RR method to justify its relatively higher costs however, it must be able to further confidence in the assurance of anonymity and subsequently reduce evasiveness in response to a level unachievable by other, more traditional methods for assuring anonymity. Given these remarks, the

findings of the empirical endeavor regarding the developed framework for RR, can be used to explore inroads for recommencing and revitalizing research on the technique regarding the translation of its theory to substantive practice. This chapter embarks on this endeavor.

A discussion of RR and RR designs must cover characteristics of the two domains that actually form the pillars of the technique when substantively used: the statistical and the psychological domain. Properties of both will be assessed with emphasis on the models and empirical trajectory as developed earlier and possibilities for further research. Firstly, several comments will be made on the statistical properties of 'efficiency' and 'analytic capabilities', often negatively associated with the RR technique. For RR in general as well as for the specific RR models as developed in Chapter 4, it will be argued that they are theoretically fully sound. However, a purely statistical rationale is difficult to translate to practical research settings. Then, secondly, we move into the psychological realm arguing that it is here where RR finds its main limitations. The translation of RR theory to practice reverberates on certain psychological properties, which the strictly statistical rationale of RR does not fully take into account in its calculations. It is exactly here where further ground has to be cleared in assessing the viability of the RR technique.

COMMENTS ON STATISTICAL PROPERTIES

The RR technique is mainly statistically criticized on the two properties of 'efficiency' and 'analytic capabilities'. The first of these critiques of a statistical nature asserts that due to the random error induced in the data by the randomizing device, variance is inflated significantly. The insertion of random error increases the dispersion of values from the expected value, resulting in less precise and thus less efficient estimations. Subsequently, costs rise, as larger samples are needed to achieve a statistical power comparable to that of a DQ setting.

But efficiency comparisons between RR and DQ techniques are based on the standard assumption of honest responses in both settings (Greenberg *et al.* 1969). Untruthful responses add to the error variance, which also leads to variance inflation due to deviations from true population score. Here again we come across the two traits with which quality of measurement is assessed, being validity expressed in level of bias due to deviations from true score, and reliability expressed in variance. When conducting inquiries into sensitive behaviors, the

assumption of truthful reporting poses hard to maintain in the DQ setting. RR justifies itself on the assumption that due to inoculation of responses through insertion of random error, self-representational concerns are relieved, leading to more truthful responses being expected when a respondent is directed to do so. Theoretically, it does not take a very high proportion of untruthful responses in a DQ setting to let its mean square error – which consists of variance plus square of the bias induced by deviations from true population score – exceed the mean square error of RR, when assuming more truthful responses in the latter (Warner 1965; Verdooren 1976; Tracy & Fox 1981). In purely theoretical language, bias reduction through RR would nullify concerns over efficiency (Tracy & Fox 1981).

Moreover, variance is a function of p, and the efficiency of a certain estimate in a certain RR model, is then directly dependent on its choice. When $p = 1$, we have a setting that equals the DQ method and for reasons given above, only under the assumption of fully truthful responses would the RR then reach maximum efficiency, as total measurement error can be expected to increase due to responses being maximally revealing. When respondent protection drops, evasiveness can be expected to increase. Theoretically, it is possible to give a construction of RR, which pertains a relatively high efficiency, and still provide for sufficient respondent protection, if the misperceiving of selection probabilities can be incorporated into the randomization device. The dichotomous and multi-proportional quantitative FRR models as developed in Chapter 4 had spinners for randomizing devices, which provided for π_y, and were the selection probabilities were converted into degrees and were evenly divided over the spinner surface area. This setup theoretically enhanced misperceiving and made possible high statistical efficiency by a large p while simultaneously potentially reducing the cognitive load for the respondent (as just one question needed to be posed because π_y was provided for by the randomizing device).

Furthermore, specific improvements were proposed for eliciting quantitative information through RR. In Chapter 4 a multi-proportional appreciation of quantitative RR data was developed. In the given model, sensitive quantitative information can be misclassified in K discrete categories (our endeavor entailed 6 categories), in which each category can denote a certain scope of frequencies. In previous quantitative RR designs the desired estimate consisted of a mean[2], and the desired population means had to be estimated before any field research efforts so as to adapt the setup of the randomizer to the expected frequencies in the population, as the range of responses to the innocuous question must be

similar to the range of possible responses to the sensitive question to be able to unbiasedly estimate incidence while still providing sufficient respondent protection. In the multi-proportional discrete quantitative FRR model, the scope of the frequencies denoted with each category can be easily adapted to the researchers need (expected frequencies) while setup and efficiency remain unaltered, as the forced responses have a range equal to the range of possible responses to the sensitive question posed. Statistically, this provides for an optimal design which does not involve excessive levels of respondent hazards, retains relative simplicity, and where no information is lost during estimation as – in addition to population mean – also the population proportion in each of the quantitative categories can be estimated separately. Also, due to the multi-proportional setup, both the dichotomous and quantitative FRR models could be presented in a unified model, which makes possible ML estimates through a log-linear LCM analogy. This remark takes us to the second statistical critique on RR, that of 'analytic capabilities'.

By its very nature, the RR technique does not give disclosure on individual data. A common critique is that due to this restrictive nature the analytic capabilities of RR would be restricted to the univariate summary measures of proportions and means (Fox & Tracy 1984; Dutka & Frankel 1993: 479). A critique that is often even forwarded by those who have used RR themselves (see for example Sudman & Bradburn 1982: 81; Burton & Near 1995: 21; Inspectie voor de Rechtshandhaving 1998: 60). However, the analytical capabilities of RR are mainly restricted by the statistical ingenuity of the researchers.

This paper has limited analyses, due to its empirical setup wishing to test the relative merits of RR, to estimates of proportions and the comparison of proportions across subgroups. Due to the setup of the dichotomous and multi-proportional discrete quantitative FRR models in a unified framework which views RR data as misclassified data, it was additionally shown however, that a log-linear latent class analysis is possible that restricts both dichotomous as well as quantitative RR estimates to their natural boundaries. This unified setup also makes possible to delve into the degree of association between two different sorts of sensitive behaviors measured under a RR condition, for both the dichotomous (Capaciteitsgroep Methodenleer en Statistiek, Universiteit Utrecht 2003) and quantitative FRR models.

It is exactly the view of RR data as data perturbed by misclassification due to the insertion of random error, which makes possible bivariate or multivariate approaches to RR. The RR data are then seen as individual level data on which bivariate and multivariate measures are applicable if

they are corrected for this contamination by probabilistic elements (Capaciteitsgroep Methodenleer en Statistiek, Universiteit Utrecht 2003; Van Den Hout 2004; Van Den Hout & Van Der Heijden 2004). Thus when the right corrections are established it becomes possible to calculate correlations between pairs of variables where one or both are measured under a RR condition (Fox & Tracy 1984; Han & Warde 1994). Also, by adapting the logistic regression technique to deal with dependent RR variables, the relation between prevalence of certain behavior and one or more continuous explanatory variables can be assessed (cf. Maddala 1983: 54-56; Scheers & Dayton 1988; Van Der Heijden *et al.* 2000; Van Den Hout & Van Der Heijden 2004).[3] These bivariate and multivariate approaches have focused mainly on dichotomous RR data (with the 1984 Fox & Tracy article as a notable exception). Due to the multinomial setup of the quantitative FRR model as proposed in Chapter 4 however, the given analytic techniques can be easily adapted to deal with quantitative data submissive to probability elements as given by (4.12).

The RR technique additionally has certain advantages with regard to the ethics of research (Boruch 1971, 1972). It is possible that research data on sensitive behaviors (such as criminal behavior) "might be appropriated and used in civil or criminal action" (Boruch 1972: 412). As researchers have no legal mandate regarding the promise of data confidentiality, obtained data may be subpoenaed if a certain infraction of civil or criminal law by (a part of) the respondents in the research calls for such an act. When the assurance of confidentiality is then subsequently breached, the research becomes ethically and professionally compromised (*Ibidem*). When the data are gathered with RR however they are conditional only as responses have no individual meaning due to insertion of random error. As the implications of certain answers are probabilistic only, both respondent and researcher are likely to be free from legal prosecution and judicial interrogation (Boruch 1971: 310).[4] These properties, which carry themselves into the ethics of research, could be furthered by the empirical setup as given in previous chapters. By using a computer-assisted environment which can be visited from the home address and where data is stored in a protected database which does not track IP addresses, not only are data conditional, but also it cannot be known exactly who cooperates with the research or in which specific order. A number of ethical concerns regarding sensitive research or the assurance of confidentiality can thus be said to be nullified.

The additional advantage of ethical nullification through RR can also be applied to data that has already been collected. When microfile data[5] are released that are collected with traditional methods and where the

privacy of the respondents should be respected, identifying variables (such as for example gender and age) can be misclassified with RR. Misclassification of identifying variables makes the re-identification of individuals statistically improbable and individual privacy is protected, while analysis can still be conducted as the misclassification probabilities can be taken into account. The use of RR in problems of statistical disclosure control is generally referred to as the 'post randomization method' and was introduced by Kooiman, Willenborg & Gouweleeuw (1997), with further elaboration by De Wolf *et al.* (1997), De Wolf & Van Gelder (2004) and Van Den Hout (2004).

The remarks made above give that statistically, the RR technique is very well developed. Variance inflation is nullified by reduction in bias, there is ample opportunity for bivariate and multivariate analyses with RR data as well as the use of sophisticated ML estimations, and, due to its statistical properties, certain ethical concerns regarding sensitive research are (partly) nullified. However, the practical feasibility of RR has been established on its statistical reasoning only. When endeavoring on gathering data with RR it is believed, due to the protection offered by statistical randomization, that "the primary reason for either a refusal or an evasive answer does not exist" (Sen 1976: 223). The rationale of RR is thus purely based on statistics. However, the general population (at least) is not statistically inclined or statistically literate. Cognitively and intuitively there is much demanded from the respondent when using RR. Substantively, where problems arise, is specifically in the translation of theoretical potentiality to practical feasibility, where we arrive at the (socio-) psychological properties related to RR. Its limitations are not so much statistical, as they are (socio-)psychological.

COMMENTS ON PSYCHOLOGICAL PROPERTIES

When answering questions of sensitive nature, respondents may have concerns beyond those covered by confidentiality assurances (Rasinski *et al.* 1999). Mostly, these take the form of concerns over revealing information to the interviewer and his or her subsequent response, and the fear that others may still be aware of the responses that are given (*Ibidem*). The use of RR and the empirical framework in which the technique was tested should have nullified these concerns, as responses were stochastically misclassified for even the researcher, and as compliance with the research took place without the researcher present

and with the possibility to utilize the privacy of one's own home. However, in absolute terms, the RR technique could not outperform a DQ setting in which confidentiality assurances can be said to be of a significantly lesser degree. There thus seems to be a difficult interaction between the statistical and psychological properties of RR designs.

As we have seen from the results of the empirical experiment in Chapter 6, the psychological characteristics 'trust' and 'understanding' may, as previously stated by Landsheer, Van Der Heijden & Van Gils (1999: 11), "provide an explanation for the mixed results that have been reported in the past". 'Trust' is the confidence of the respondent that his or her response is protected in its anonymity by the use of RR (Landsheer, Van Der Heijden & Van Gils 1999). 'Understanding' refers to the insight of the respondent in the RR instructions and the protection RR subsequently offers (*Ibidem*). Also, the results give further evidence for two basic forms of non-compliance with RR rules: 'lying' and 'cheating'. The former refers to answering contrary to fact, while the latter refers to non-compliance with the instructions of RR (Edgell, Himmelfarb & Duchan 1982; Boeije & Lensvelt-Mulders 2002). Note that lying is always non-compliance with the rules but that a cheater does not necessarily have to be a liar. The former is thus especially important with regard to inducing bias when respondents do not give their true status with regard to a certain sensitive trait when directed to answer truthfully, where the latter is especially important with regard to inducing bias in the RR method when respondents do not give an answer that is redirected (forced) by the RR instructions. 'Trust', 'understanding', 'lying' and 'cheating' have different interactions with each other and it will be argued here, that these interactions may be of greater complexity and greater subtlety, than previously assumed.

In the connection between trust, understanding and non-compliance with RR further subtleties can be defined. The silent assumption in research dealing specifically with psychological aspect of RR is that trust and subsequent compliance are intricately developed when understanding is present (cf. Landsheer, Van Der Heijden & Van Gils 1999; Boeije & Lensvelt-Mulders 2002). However, while there is likely to be a plausible connection between the development of understanding and trust as a precursor for RR tolerance, compliance with RR rules may require a leap of faith beyond pure understanding or trust. It is possible that one understands the RR instructions, but not the protection the technique offers. It is possible that one understands both instructions and subsequent protection, but does not develop trust in the method. It could also be possible that one understands both instructions and protection,

subsequently trusts the method, but is still not inclined to comply with the rules of RR. Although the opposite can also be said to be possible, that one trusts without full understanding, psychologically such situations are less likely (especially when dealing with sensitive issues). In all situations, lying and cheating can be expected to be present in a degree disproportionate to that expected on the basis of the strictly statistical rationale. It is still largely unknown to what degree the use of RR demands from cognitive abilities. But the research findings as presented in the previous chapter point to a complexity in the understanding, trust and subsequent compliance with RR, that exceeds the complexity previously assumed to be connected to these characteristics.

Firstly, as previously stated in the chapter dealing with the results, the RR method may invoke suspicion by itself. As the technique is not well known (especially in the respondent groups generally surveyed) and requires some cognitive leaps by the respondent, it is possible that in a part of the sample it may provoke considerable suspicion or confusion (Boruch 1971; Berman, McCombs & Boruch 1977). The relative curiosity of the technique has – as our results point out – respondents wandering contrary to its purpose, the trustworthiness of the technique and/or its related instructions. Coupled with the increasing level of perceived sensitivity due to elaborate confidentiality assurances (Reamer 1979; Singer, Von Thurn & Miller 1995), which characterize the RR method, the induction of suspicion can alter responses towards lying or cheating. Suspicion could be an explanation for those respondents who come to understand RR instructions and protection, but still do not develop trust in the method.

Secondly, full honesty by all respondents can never be achieved by any method (Li 1976). Those who have the sensitive trait asked for, but are fully unwilling to state their true status in any circumstance, even under protection from additional probability, will give a negative response. As we are dealing with socially undesirable behaviors, lying of this kind can be expected to be in the same direction, giving a further factor for bias in RR estimates.

Moreover, if a respondent is anxious to hide any hint of indulgence in a certain sensitive behavior, the respondent would have an incentive to answer simply 'no' or '1' regardless of the outcome of the randomizing device (Campbell 1987). We have given an argument of this form for lying above. But there is also a specific distinction for the case of cheating. As becomes clear from our data, there can be considerable response falsification on the part of those who where directed by the outcome of the randomizing device, to give an answer opposed to their

true status. A behavior that is especially apparent for those who are redirected to give an answer that exceeds their true status. This can even be present when there is understanding and subsequent trust, as the open interviews have indicated. The rationale is that one wants to be sure, even when the protection offered by RR is understood and trusted, to hide any hint of indulgence, so that negating responses ('no' or '1') are the safe option. A forced answer that exceeds true status was perceived to be counterintuitive, which led to (especially in the 'innocent group') new representational concerns when using RR.

The specific setup of the RR models could have spurred incentives for lying and especially cheating behavior of the kind as described above. FRR designs confront respondents directly with the sensitive question (Hosseini & Armacost 1993). The FRR design does not have symmetry of response. A RR design "is said to have symmetry of response with respect to a characteristic if each possible response, of itself, conveys no information on the state of the respondent with respect to that characteristic" (Bourke & Dalenius 1976: 219; also see Bourke 1984). For example, the Warner design is symmetric as for both 'yes' and 'no' responses no information is conveyed with regard to the status of a certain individual respondent. A 'yes' response may be affirmative to either possession of X or no possession of X, and a 'no' response may be negative to either possession of X or no possession of X (i.e. there are no safe responses). The UQD, which is statistically related to FRR designs, lacks symmetry of response as a 'yes' means that the respondent either possesses X or Y, where a 'no' conveys the information that the respondent either does not possess X or Y (Greenberg *et al.* 1977; Bourke 1984). In both the UQD and the FRR design lying and cheating may thus be spurred as a negative response (or 1 in the quantitative FRR model) is always the safe option as there is a transfer of risk from the respondents possessing X to the respondents not possessing X, due to lack of symmetry. A negative response is never associated with the sensitive trait. Although there is no symmetry of response, the UQD still has the psychological advantage that there are two meaningful interpretations for a certain response (in terms of the sensitive or the innocuous question). An additional drawback of the FRR design may thus be that a certain response has only a meaningful interpretation in terms of the sensitive question (Miller 1981). Where the design of the UQD and the FRR can both encourage non-compliance with RR rules, it is the latter where such incentives may peak.

Additionally, certain subject attributes can also hamper adequate or truthful responses to RR surveys. The research as presented here indicates

that especially cognitive sophistication and the level of possible threat (individual or collective) could be of importance. It is still unknown in which way the understanding of RR instructions and protection, and the psychological ease with RR are related to cognitive abilities on part of the respondents. Although it is highly probable that use of a RR procedure puts higher demands on the cognitive abilities of the respondents, cognitive laboratory experiments by Boeije & Lensvelt-Mulders (2002) led to the (contradicting) conclusion that instructions or reasoning accompanying RR were not too difficult for the general respondent to understand. However, a cognitive laboratory setting is divorced from an actual research setting were threats can be more direct, or where researchers do not have to be present (as in the setup given in previous chapters). It seems highly probable that the leap from the statistical rationale to psychological appreciation and subsequent compliance with RR is mediated by cognitive sophistication. This expectation, previously worded by Warner (1976b) could be backed by the evidence that probabilities are difficult to appreciate and comprehend, even by well-educated persons (Kahneman & Tversky 1973, 1974).[6] The results in Chapter 6 give at least preliminary evidence – as the average police officer is cognitively less sophisticated (e.g. has not received extensive formal education) – that cognitive abilities are an additional factor in raising suspicion towards the RR method and can simultaneously hamper full understanding of RR instructions, protection and/or purpose. As we will reason below, it could be initial instructions that aid this effect.

Also, previous empirical endeavors regarding RR were performed on general populations, where the threat stemming from the possible disclosure of information can be said to be mainly individual. The research reported on here, evaluated the viability of RR in an organizational population. It is very well possible that individual respondents perceive collective jeopardy as well. In an organizational setting, population incentives for non-compliance may not be individual only, but may also be linked to representational concerns regarding the group of which one is a member. Such incentives are likely in groups where the individual has strong identifications with his or her group. The disclosure of information that is potentially stigmatizing or incriminating then becomes threatening for both individual and group. Such strong group identifications are likely among police officers, which derive part of their identity from being an officer. Part of this identity, such as a natural inclination towards suspicion and expected honesty, which are seen as attributes of the police officer, could also have furthered incentives towards non-compliance.

Lastly, it must be stated that even when a RR condition would nullify all incentives for lying and cheating, not all sources of error in measurement (of sensitive behaviors) would be captured. Some response errors may have their roots in ignorance and suppression, and thus have sources deeper than that of mere self-representational concerns (Kish 1976).

In this section we have reviewed the limitations of RR that lie mainly within the mostly overlooked psychological domain of RR designs. Substantively, where problems arise, is when the practical feasibility or RR is established on its statistical reasoning only. We have seen that, while statistically assumed, the nullification of certain concerns may not be psychologically feasible. It can be argued that ground in the realm of measurement is gained regarding the estimation of prevalence and incidence when there are subjects who understand the RR method, develop trust in its protection and subsequently decide to cooperate in a non-evasive manner with socially sensitive research. However, when the incentives from within the psychological domain for suspicion, lying and cheating dwarf the main effects of RR and if subsequently the gain in accurateness is nullified by new sources of error, then the RR method cannot be justified. If we are to anticipate the success of RR relative to other methods in eliciting sensitive information, then we must conduct further research in its psychological properties in connection to technical subtleties.

RESEARCH RECOMMENDATIONS AND AGENDA FOR ADVANCE IN TRANSLATING THEORETICAL POTENTIALITY TO PRACTICAL FEASIBILITY

Research should be directed towards if and how the theoretical potential of RR as given in its statistical assumptions can be translated to (socio-)psychological feasibility. Further research must make up for leeway in the excessive attention for technicalities regarding the RR technique. The finding as presented previously give opportunity to verbalize some research recommendations in connection to the knowledge gaps which exist with regard to the cognitive dynamics underlying RR.

Firstly, and most basically, it must be assessed which kind of assurance enhances feelings of confidentiality of responses. As stated several times, elaborate assurances of anonymity can even enhance suspicion and perceived levels of sensitivity, whatever the content of

research. It must be systematically assessed if an elaborate assurance really produces psychological security, and if the psychological security provided with RR is appreciably greater than with other confidentiality assurances (such as the simple researchers' word).

Secondly, it should be assessed which RR design is psychologically strongest. The Warner design has the advantage that is symmetric, the UQD has the advantage that the alternative explanation for a certain answer can be worded in terms of a non-sensitive unrelated question, and the FRR design has – at least in theory – the advantage that its setup has the least initial information to be processed (just one question posed). An evaluation should be directed towards which setup spurs the least psychological hazards and the most trust. In this respect these designs should also be tested against optional RR models to assess which design has the least non-compliance.

The issues of assurance and differing designs have a close connection to initial RR instructions provided to the respondents. Although the issue of 'being honest' was redefined in terms of following RR rules, cheating still appeared to be a very significant factor in responses. It would be useful to evaluate if different wordings of the instructions can convince respondents of the protection offered by RR, while simultaneously give a convincing appreciation of the necessity of a forced answer (when one uses FRR). Within experimentation with different wordings of instructions a simple example to introduce RR may be useful. The demanding part of the comprehensibility of RR could well lie in the fact that an understanding of RR calls for a cognitive appreciation of Bayesian probability. To understand the method respondents (must) develop for themselves a notion of 'jeopardy' and 'risk of suspicion'.[7] Comprehension of the degree in which a 'yes' or 'no' gives the probability of possession of X, calls for an understanding of conditional probability which may well be too demanding for most respondents, and which can hamper compliance. Cosmides & Tooby (1996) however, give preliminary evidence that by expressing problems in frequentist terms[8] correct Bayesian reasoning can be elicited in many subjects. We also have preliminary evidence for this effect as during the follow-up study certain respondents asked for additional explanation of RR. After giving a simple numerical explanation of RR in terms of frequencies of events, the mentioned respondents stated enhanced comprehension of the subtleties of the RR method. A promising inroad for further research would then be to adapt instructions of RR with different wordings and with simple examples of RR where the conditional probabilistic information associated with RR (the probability of the event 'outcome of randomizing

device' and subsequent 'respondent jeopardy' and 'risk of suspicion'), is explained with frequencies. The question then becomes if suspicion can be reduced and if comprehension and compliance can be enhanced. Additionally, encouraging respondents to articulate their concerns can also improve insight in the occurrence or risk of certain events (Rasinski *et al.* 1999).[9]

The ideas regarding research on RR instructions give further inroads for experiments regarding respondents with varying degrees of cognitive abilities. Although the RR technique is assumed not to be too difficult to understand for the general respondent, no systematic research has thus far been developed into the connections between understanding of RR and cognitive abilities. It seems highly probable, that greater cognitive abilities will lead to greater comprehension of RR instructions and protection. A question that needs systematic attention is then if the increased comprehension through greater cognitive sophistication is negatively related to suspicion and positively related to trust and compliance with RR. If more knowledge is accumulated on this point, greater deliberation can be given to the selection of respondent groups.

Understanding of RR through the setup of subsequent instructions and degree of cognitive abilities is also mediated by the possible presence of a researcher. Due to less social context cues the absence of a researcher is assumed to lead to greater feelings of confidentiality and thus to more truthful self-reporting. However, as was learned from the follow-up study, the presence of a researcher can also have a motivational effect. Additionally, in case of partial or no understanding of the RR technique, a researcher can provide for extra explanation, which can also potentially heighten trust and/or compliance. It must be assessed if more social context cues by the presence of a researcher are counterbalanced by the positive effects this presence can have on understanding of and compliance with RR. In doing so, one can also look for optimal combinations of computerization and researcher presence.

More broadly, an assessment needs to be made if incentives for non-compliance in situations of collective jeopardy, transcend incentives for non-compliance in situations where mainly individual jeopardy can be argued to exist. This calls for a systematic evaluation of (non-)compliance between populations with differing levels of possibly perceived threat. In such an assessment great use could be made of the FRR design. By fixing the outcome of the randomizing device for several questions in a sequence of questions across populations with differing levels of threat, prevalence of non-compliance can be evaluated.[10] Also, Clark & Desharnais (1998) have devised a RR model statistically similar

to the FRR design where a certain sample is split into two subgroups. Each group then is subjected to the same RR condition, but are assigned differing randomizing probabilities. In this setup they proved that it is possible to detect significant cheating and its subsequent extent, while still being able to protect the status of every individual. This model (which can be extended to deal with quantitative characters of a multi-proportional nature) could be used similarly to assess incentives for non-compliance in populations with differing levels of perceived threat.

Also, it is generally assumed that the relative merits of RR increase, when the sensitivity of the topics increase. However, the data as presented in Chapter 6 did not point to such an effect. Additionally, it would be psychologically feasible that the incentives for non-compliance when the behavior asked for is of greater subjective sensitivity. Incentives to hide any hint of suspicion, whether by lying and/or cheating can be psychologically expected to be greater if a certain behavior is of a very sensitive nature to the respondent. Edgell, Himmelfarb & Duchan (1982) found that non-compliance with directed response was more sizeable when the questions were of greater sensitivity. It is still unclear if these increasing incentives for non-compliance with increasing sensitivity are less pressing in a RR condition than the analogous incentives when using other questioning techniques. This calls for more systematic attention for the relationship between subjective relative perceived sensitivity of a question and incentives for non-compliance. A possible inroad for such an endeavor would be an experiment coupling a range of sensitive behaviors with differing scores on socio-psychological scales of attribute sensitivity with the cheating models discussed above.

Reflecting on the gaps in knowledge regarding the psychological dynamics underlying RR, what is needed is a series of experiments, which systematically delve into one or more of the psychological aspects discussed above. The use of criterion measures is superior. However, these are not always available. Umesh & Peterson (1991) have argued for the use of physiological tests or biochemical measures to obtain criterion measures to which RR responses can be compared. However, such measures can offset by themselves new factors that hamper response to RR and can give new ethical concerns. Best is to use criterion measures where possible and otherwise design comparative frameworks where it is safe to assume that higher estimates are more accurate. Broadly speaking, if we want to anticipate on the success of RR we need systematic attention for its psychological dynamics with a range of experiments which test these dynamics with differing designs and within differing

populations to be able to state with more confidence if, and for which research situations RR's theoretical potential has practical feasibility.

NOTES

[1] The Respondents in the CASI condition were also granted anonymity. In the CASI condition this was achieved by informing the respondents of the researchers' grant on confidentiality of data.

[2] In the quantitative FRR model as developed in Chapter 4, estimation of a population mean is also still possible.

[3] The adapted logistic regression analysis that connects dependent RR data to one or more continuous explanatory variables can be conducted with SPSS. A specific macro for such an endeavor can be found in Capaciteitsgroep Methodenleer en Statistiek, Universiteit Utrecht (2003).

[4] For a more in-depth analysis of RR and the possible subpoena of data, see Boruch (1972).

[5] Microfile data are data where each row corresponds to an individual respondent and where the columns represent the variables.

[6] It is generally assumed in the psychological literature that the human mind does not encompass a calculus of probability.

[7] See Chapter 4.

[8] The frequentist approach to probability derives the concept of probability from the concept of relative frequency. Where Bayesian statistics has a conditional appreciation of probability and states that a single event can be captured in a probability, frequentist approaches state that probability is derived from the relative frequency of the occurrence of a certain event over an infinite or very large number of iterations. When assuming capability in calculating probabilities, the human mind is likely to represent probabilities in frequentist terms (Gigerenzer 1991, as cited in Cosmides & Tooby 1996). For an introduction to the frequentist approach to probability, see Rudas (2004).

[9] This technique was also proposed by Raymond M. Lee in a private communication.

[10] Edgell, Himmelfarb & Duchan (1982) have used such a setup to test the viability of forced responses in a general population.

8. SUMMARY AND CONCLUSION: ON THE VIABILITY OF RANDOMIZED RESPONSE

Having reviewed the results from the experiment and having discussed the specific endeavor regarding the testing of the viability of randomized response (RR) in a broader perspective, time has come to resonate the findings in a reflection on the main research question. Here, in a concise and comprehensive manner, an answer will be formulated on the ability of the RR technique to give more accurate population estimates regarding prevalence and incidence of sensitive behaviors of a political nature, while simultaneously elaborating on a perspective for RR in general. In doing so, we will first summarize the purpose and main findings of the previous chapters, before culminating these findings in a unified assessment of the viability of RR.

SUMMARY OF PREVIOUS CHAPTERS

Representing the latent with the manifest was posed to be one of the greatest problems in social science. The reliance on constructed observables to indirectly represent the construct in which one takes an interest, makes it a crucial aspect of any endeavor involving measurement to question the accurateness of the former in representing the latter as there are many factors that can make an observable score deviate from its true value (in latent space). When such factors are systematic, the relationship between concept and indicator become weak or faulty, and subsequently our understanding will be hampered. This issue of measurement was argued to be especially apparent with regard to

research on sensitive behaviors, where an accurate assessment of prevalence and incidence of certain behaviors, are crucial for our theoretical and empirical understanding of many pressing issues.

Chapter 2 delved into the meaning of 'sensitivity' in relation to behaviors and explored the relationship of 'sensitivity' to the accuracy of various estimates of prevalence and incidence. A topic was said to be sensitive if the disclosure of information regarding an individuals' status on this topic, poses to be threatening for the respondent in the form of being potentially stigmatizing, incriminating or severely intrusive. Due to these potential threats, people have self-representational concerns and a resulting incentive to retain their anonymity with regard to the sensitive behaviors researched, leading to refusals to cooperate (nonresponse bias) or evasive cooperation (evasive answer bias). As a substantial part of observed values can deviate from true scores due to these biases, sensitive behaviors of a generally undesirable nature are thus underestimated, and the researcher has to subsequently look beyond the reach of mere sampling issues as the problem of differential validity is encountered, which centers around external and internal measures of behavior. Theoretically, external measures can be argued not to be able to validly pierce deep into latent space as they may reflect the contingencies in their production more so than the behaviors they purport to measure, and as certain kinds of information are unavailable from secondary data sources. The validity problems inherent in various external measures often leave self-reports of certain behaviors as the only road for estimation of prevalence and incidence. But here another problem of observational statistics is encountered, as the self-representational concerns of the respondent peak and (part of) responses can be expected to be systematically distorted. Specific methodological effort to enhance the validity of prevalence and incidence measures through self-reports of sensitive behaviors is thus justified.

A technique that makes possible the investigation of sensitive private activities through an examination of individual disclosure under the protection of full anonymity is the RR technique. Chapter 3 elaborated on this technique and its use of probability theory to give more accurate population estimates of sensitive behaviors. The core characteristic of the technique is the insertion of random error by an element of chance to provide respondents optimal privacy protection when answering questions of a sensitive nature. Due to the probability of misclassification individual meaning of answers is rendered out so that respondents cannot be known with regard to their status on a certain sensitive trait. The rationale is that when fully protected from stigmatization or

incrimination, respondents are more willing to cooperate and will do so in a non-evasive order. At least theoretically, more accurate population estimates regarding prevalence and incidence are then possible. Certain gaps in the method and its research literature where however acknowledged. Firstly, the absorption with the technical fix and theoretical exploration of the assumptions underlying the method has left no cumulative work on field-practice, which partly underlies the wide variability in previous (relatively rare) field-performances. Secondly, most studies did not delve into the relative merits of RR in obtaining quantitative characters. Thirdly, the RR design allows for new sources of error. And lastly, sensitive behaviors of a political nature were not previously assessed with RR. These gaps, which all revolve around the substantive use of RR in actual research settings made clear that there is still ground to be gained in the translation of RR theory to practice, and provided the background for the methodological framework of this study.

Chapter 4 provided the first part of the methodological framework, giving the specific technical constellation in which the RR technique was modeled to be theoretically able to efficiently give more accurate population estimates regarding prevalence and incidence. A new randomizer was developed for a dichotomous forced randomized response (FRR) model for the estimation of prevalence. Additionally a new quantitative FRR model was developed which gave a multi-proportional appreciation of sensitive quantitative data, making possible independent separate estimates of the population proportion in K discrete quantitative categories. This gave an efficient modeling, where the range of possible inoculating responses is similar to the range of possible responses to the sensitive question, so that categories could be easily adapted to expected population frequencies while setup and efficiency were retained. As both the dichotomous and quantitative FRR models were based on proportion estimations from misclassified data, they could be captured in a general function allowing for a latent class model (LCM) analogy on which maximum likelihood (ML) estimates could be based.

Chapter 5 provided the concluding part of the methodological framework by elaborating on an empirical framework in which the RR models as developed in Chapter 4 were field-tested to give more accurate population estimates regarding prevalence and incidence of sensitive behaviors of a political nature. This empirical endeavor entailed the assessment of the viability of RR in a comparative framework. A computer-assisted randomized response survey (CARR) was set out in the finite population of a regional Dutch police force. The virtual questionnaire contained questions regarding individual misbehavior in the

integrity sphere. This experimental group was compared to a control group within the same police force that received the same questionnaire but without the incorporation of the RR technique. This control group was subjected to a computer assisted self-interview (CASI), which was actually a direct questioning (DQ) condition. The estimates obtained from the questions regarding prevalence, were also compared to official police statistics. This setup was proposed as an 'experimental test case' for the evaluation of RR in eliciting information of a politically sensitive nature. By taking into account the assumptions and previous empirical work regarding RR; expectations were expressed regarding the superiority of RR estimates.

The performance of RR in obtaining population estimates regarding sensitive behaviors in comparison withy the more traditional DQ technique was reported on in Chapter 6. Both the RR and DQ conditions outperformed official statistics in obtaining prevalence estimates, pointing to the contingencies in the latter and the advantage of a lack of social context cues. However, the superiority of RR in eliciting sensitive information of a dichotomous and quantitative nature could not be established in the given framework. Performances between CARR and CASI conditions were similar, and – when ignoring statistical significance as an issue – CASI even outperformed CARR in several instances. The performance contrary to expectation was not due to an approaching of upper bounds of possible estimates or lack of sensitivity in the questions, but was attributable to psychological factors underlying (response to) RR. Especially the counter-intuitiveness of a forced answer that exceeds true status and subsequent non-compliance with RR rules provided an explanation for the nullification of the relative merits of RR.

Chapter 7 placed the findings of the empirical endeavor in a broader perspective. It was argued that statistically the RR technique is very well developed. Substantively where problems arise – as became clear from the empirical testing – is in the translation of RR theory to practical feasibility, as this translation is generally performed on the strictly statistical rationale of RR. Its limitations are not so much statistical as they are (socio-)psychological, as the statistical assumptions may not prove to be psychologically feasible. The interactions between 'trust' and 'understanding' in relation to RR and forms of non-compliance are of greater complexity and greater subtlety than previously assumed. Thus substantive attention needs to be directed towards the psychological properties of RR, in connection to related technical subtleties.

MEASURING (POLITICALLY) SENSITIVE BEHAVIOR THROUGH RANDOMIZED RESPONSE

The RR approach remains somewhat overlooked in field-applications wishing to measure the extent of certain deviant behaviors. Although the technique was proposed by some as a valuable educational tool in introductory statistics and probability courses (Campbell & Joiner 1973), such endeavors did not come about and the technique has claimed a domain that lies outside the toolkit of field-practitioners. A possible reason may be that the technique proposes a methodological advance, which is "an alternative to, rather than an extension of, traditional methods"(Fox & Tracy 1981). New methods are always confronted with considerable resistance or a certain cultural lag. An additional reason for this lag may well lay in the specifics of the domain that RR has claimed for itself. This domain employs a mainly technical vision on RR. The absorption with the technical fix and the lack of cumulative work in actual field-research settings has left the cognitive dynamics underlying RR severely underexposed. As RR has proven not to be under sufficient researchers' control when used substantively, non-methodologists may be scared away by seemingly insurmountable technicalities and the anticipation of mixed results.

A consequence of the mainly technical view on RR has been that many previous empirical endeavors have often set out to champion the method. The encompassing ambition of this study was to give a more substantive application, with a special view towards actual working and practicality of the technique in field-research settings. It is here that the cognitive dynamics underlying RR were found to be crucial in connection to the main research question and in connection to the use of RR in general.

The main research focus questioned if elementary probability theory in the form of RR techniques could be used to obtain more accurate population estimates regarding prevalence and incidence of sensitive behaviors of a political nature. While stated very basically, when a study seeks to translate theoretical potential to practical feasibility on partly unexplored grounds, basic formulations are appropriate. Nonetheless, such a question also involves a deliberation regarding 'extent' as degrees of accurateness can only be established in comparison. It is on such we will elaborate here.

The viability of RR in giving accurate representations of prevalence and incidence of behaviors of a politically sensitive nature was assessed against a more traditional DQ condition where anonymity was granted

verbally only, and against official police statistics. Both the RR and DQ condition proved the prevalence estimates distilled from the police statistics to be inferior, pointing to contingencies in external measures and the advantage of lacking social context cues. But if the RR method is to justify its existence and its relatively high costs, it must be able to further confidence in the assurance of anonymity and subsequently reduce evasiveness in response to an extent unachievable by other, more traditional existing methods for assuring anonymity. However, superiority of RR could not be established against a DQ condition where respondents received no more than the researchers' grant to ensure anonymity. Preliminary, reflecting on the extent in which RR was able to pierce latent space to come to accurate representations of prevalence and incidence in comparison with the DQ condition in which confidentiality assurances can be said to be of a significantly lesser degree, the answer to the main research question must thus be negative. The reason for a poorer performance than expected was argued to lie in the psychological domain. Elaborate assurances of confidentiality, as is characteristic of RR, can invoke suspicion by itself and can enhance the perceiving of sensitivity. Additionally, the RR technique can involve new sources of error stemming from non-compliance with the RR procedure; the most prominent of which was found to be cheating on part of those who were directed to give an answer that exceeded their true status. Especially this counter-intuitiveness of the forced answer can be mediated by cognitive abilities and other respondent attributes, pointing to the possibility that RR can actually spur new self-representational concerns. These findings give important opportunities to revitalize research on RR regarding the translation of its theory to practice for the political sciences as well as the behavioral sciences in general.

With the benefit of hindsight it can be stated that the choice of police officers to assess the viability of RR with regard to measuring the extent of integrity violations was not the best option for testing RR's viability with regard to politically sensitive behavior. The specific attributes of police officers such as expected honesty, on average less cognitive abilities and perceiving of group jeopardy due to strong group-identifications, could have furthered incentives or motives for non-compliance with RR rules leading to the subsequent nullification of its relative merits. As the specific cognitive dynamics of collective jeopardy and intellectual sophistication remain (relatively) unexplored in relation to the RR method, exploring behaviors for groups in the political realm with vested political alignments and organizational like features with RR may be too soon. Especially if collective jeopardy takes prevalence over

individual jeopardy, RR may be of diminished use in these settings. However, particularly in the political realm it would be interesting to explore RR's merits in eliciting information on preferences which, psychologically speaking, can be said to be mainly individual. Not only would one then explore new latent domains with the RR technique as one moves beyond the focus on behaviors of a sensitive nature towards which people have self-representational concerns, but one would then be able to assess the viability of RR with regard to preferences on which other estimates are notoriously inaccurate (such as preferences for political parties at the extremes of the political spectrum, or intolerant ideas regarding foreigners). The quantitative (F)RR modeling could then easily be adopted to let a certain numerical state intensity of preferences.

The empirical findings also have important implications for RR in relation to the behavioral sciences in general. Theoretically the RR technique is fully sound. Simultaneously, the anticipation of its success is established on its statistical reasoning only. The actual research setting in which the viability of the RR technique was tested in this study points to the important conclusion that the psychological dynamics underlying RR are neglected. The translation of RR theory to practice reverberates on certain psychological properties, giving an explanation for the mixed results regarding RR performance in the past. The statistical assumptions prove that they are not psychologically feasible in all situations, pointing to the fact that the RR technique is not under full researchers' control. Further research needs to be directed towards the psychological dynamics underlying (response to) RR. Although the main limitations lie in the translation of theoretical potential to practical feasibility, it is exactly due to its potential that RR deserves further substantive attention of the kind elaborated in Chapter 7. The biases inherent in external measures or traditional questioning techniques call for a creativeness such as displayed by the RR technique. If we want to anticipate on RR's success in giving accurate representations of sensitive traits and if we want to make up for leeway regarding research, time has come to establish its structuring psychological features.

APPENDICES

Appendix A: Typology of Integrity Violations

This appendix contains the typology of categories of integrity violations as developed by Huberts, Pijl & Steen (1999). This typology was the outcome of an analysis of the literature on police integrity and was assessed against the results of empirical research on internal investigations in the Dutch police force. It was used to construct the self-report questions regarding individual misconduct by police officials.

1 corruption: bribing
misuse of public power for private gain; asking, offering, accepting bribes;

2 corruption: nepotism, cronyism, patronage
misuse of public authority to favour friends, family, party

3 fraud and theft
improper private gain acquired from the organization (with no involvement of external actors)

4 conflict of (private and public) interest
personal interest (through assets, jobs, gifts etc.) interferes (or might interfere) with public interest

5 improper use of authority (for noble causes)
to use illegal/improper methods to achieve organizational goals (within the police for example illegal methods of investigation and disproportionate violence)

6 misuse and manipulation of information
lying, cheating, manipulating information, breaching confidentiality of information

7 discrimination and sexual harassment
misbehaviour towards colleagues or citizens and customers

8 waste and abuse of resources
failure to comply with organizational standards, improper performance, incorrect or dysfunctional internal behavior

9 private time misconduct
conduct in one's private time which harms the publics trust in administration/government.

Appendix B: CARR and CASI Questionnaires

This appendix contains the translated CARR and CASI questionnaires. It reveals the specific instructions and questions posed to the respondents in both questioning conditions. The appendix also contains an impression of the virtual make-up of the questionnaires.

After receiving a general introduction in the CARR condition the respondents received several general questions regarding their work and work environment. These were questions regarding topics with a possible relationship with prevalence and incidence of certain integrity violations. These possible relations, who can be assessed with an adapted logistic regression procedure, will be reported elsewhere. Subsequently, the respondents were provided with 10 prevalence and 10 incidence questions and the accompanying introduction to the dichotomous and quantitative FRR procedures. The questionnaire was concluded with a few questions evaluating user-friendliness and trust in the method. The CASI questionnaire was similar to the CARR questionnaire with exactly the same questions asked. As the CASI condition had no RR incorporation, the difference lay in the absence of extra text blocks introducing RR and subsequent trial questions in the CASI condition. Also, the CASI condition had a differing setup with regard to the evaluation questions.

Both questionnaires were accessible on separate Internet sub-domains with provided username and password. Both questionnaires were fully operable with the mouse pad. Each text block and each question was given on a separate page. It took respondents no more than 15 minutes to complete the questionnaire in the CARR condition, and no more than 10 minutes to answer all questions in the CASI condition.

Note that due to translation nuances can be lost. For preservation of the anonymity of the police force and its employees, names are feigned with 'X'.

CARR QUESTIONNAIRE

Upon accession of the questionnaire on the Internet sub-domain, the respondents were firstly thanked for their willingness. By clicking on 'to the questionnaire' the question process was started with a general introduction.

B.1 Introduction

We want to ask you a few questions regarding integrity problems, which can occur in your work or work environment. Many people have mixed feelings regarding these sorts of questions. One the one hand one does not want to be a squeaker, on the other hand integrity dilemmas can hamper on you and your work. These dilemmas can also encompass one's own behavior. An integrity problem that you have encountered in your work can be difficult to openly discuss with colleagues or superiors, as it cannot be known how your environment will react. But when there is no overview of the integrity problems that exist in the work environment, it will be very hard to appropriately react upon them.

With this questionnaire we seek to address two aims. Firstly, we want to give everyone the chance to report integrity problems. Next, we want to acquire an overview of the integrity violations that occur within the police force. On the basis of this data, it can be assessed how frequently occurring problems can be influenced.

We do **not** want to embarrass anyone. This questionnaire is not about which police employee ever crosses the line, but about insights in the occurrence of certain unacceptable behaviors. That is why the researchers of the Vrije Universiteit van Amsterdam along with the researchers of the Universiteit van Utrecht have developed a method with which such sensitive questions can be asked in a way that fully protects your privacy. XX (the privacy and information confidentiality functionary of the XX police force) has approved the method, as it fully protects your anonymity.

Before we explain this method to you, we will firstly ask a few general questions regarding your work and work environment.

The questions regarding work and work environment were divided in 'perceiving of work environment' and 'perceiving of rules'. Both received a short introduction along with instructions. Each question was posed on a separate page with a text block in the upper corner in which the instructions were shortly repeated. Each possible answering category was given directly under the question (numbers accompanied by meaning).

B.2 General Questions Regarding Your Work and Work Environment

B.2.1 Perceiving of work environment

We want to ask you a few questions regarding your work environment. These revolve around topics, which may have a possible relation with the integrity of the organization. We are interested in *your own opinion.*

The answering possibilities are given on a so-called 5-point scale. Pick the answer that is closest to your opinion and click on this alternative with the mouse pad. Try to answer all questions and remember that there are no right or wrong answers; we are interested in your personal opinion. It is therefore important that you respond to the questionnaire individually.

The numbers have the following meaning:

1 = strongly disagree
2 = disagree
3 = do not disagree, but also do not agree
4 = agree
5 = strongly agree

Next the questions were posed as positions. Each question was posed on a separate page.

(1/50) The employees in our force are expected to strictly abide the law.
(2/50) Within this force everyone is expected to abide internal rules and procedures.

(3/50) The force management gives good example with regard to integrity.
(4/50) My direct superior gives good example with regard to integrity.
(5/50) My direct superior gives me clear directions regarding the norms and values to be followed
(6/50) My direct superior has sufficient attention for possible unwanted behavior within the organization.
(7/50) If certain incidents occur within my unit, then it is good to find out who has been negligent.
(8/50) In order to be successful within this organization, I sometimes have to sacrifice my own values and norms.
(9/50) I am satisfied with my salary and my career possibilities.
(10/50) I enjoy going to work.

B.2.2 Perceiving of rules

Also, we would like to ask you a few questions regarding *your perceiving of rules within your* work environment. Again these questions revolve around topics, which may have a possible relation with the integrity of the organization, and again we are interested in your *personal* opinion.

Again, the answering categories are given on a 5-point scale, the meaning of which can be directly read from the screen.

Next the questions were posed. Each question was posed on a separate page.

(11/50) How familiar are you with the behavioral rules as they apply within your force?

With answering categories:
1 = not at all familiar
2 = not familiar
3 = fairly familiar
4 = familiar
5 = very familiar

(12/50) How clear do you perceive the behavioral rules as they apply to your force?

With answering categories:
1 = very unclear
2 = unclear
3 = fairly clear
4 = clear
5 = very clear

(13/50) Do you always know the boundaries of good police conduct?

With answering categories:
1 = I rarely know
2 = I do not know for all situations
3 = I know for a fair amount of situations
4 = I know for most situations
5 = I always know

(14/50) How important is it to you that the norms for good police conduct are not violated within your force?

With answering categories:
1 = unimportant
2 = fairly unimportant
3 = important
4 = very important
5 = of absolute importance

(15/50) How big do you perceive the chance that someone will get caught when he or she violates the behavioral rules?

(16/50) How big do you perceive the chance that sanctions will follow if someone is caught violating the behavioral rules?

Question 15 and 16 both with answering categories:
1 = very small
2 = small
3 = not so big
4 = big
5 = very big

(17/50) How harsh are violators of behavioral rules punished according to you?

With answering categories:
1 = very mildly
2 = mildly
3 = not so harsh
4 = harshly
5 = very harshly

One questions was asked to divide between executive and non-executive police employees, as with some prevalence and incidence questions, this distinction can prove to be important. Clearly, this was a dichotomous question and was posed on a separate page.

B.2.3 Executive and Non-Executive Function

(18/50) Can you point out if you have an executive or a non-executive function?

Next, the dichotomous FRR procedure was introduced. The introduction provided an explanation of the method along with instructions. After this introduction 3 trial questions were asked to let the respondent get acquainted with the dichotomous FRR procedure. Following these trial questions, 10 prevalence questions were asked.

All questions were posed on a separate page. The page make-up was such that a text block in the upper left corner repeated the dichotomous FRR instructions and the upper right corner contained the spinner for dichotomous forced randomized response inquiries as given in Figure 4.1. The spinner could be started by clicking on 'start spinner'. The bottom half of the screen page contained the question, along with the answering categories and their meaning. Clearly, all questions were dichotomous 'yes' or 'no' questions. This setup can be found in the impression at the end of this appendix.

B.3 Introduction of the Dichotomous Spinner Method

Now we arrive at the questions that you can answer with the help of the newly developed method by the Vrije Universiteit van Amsterdam and

the Universiteit van Utrecht. This method of answering questions works with the help of a spinner, which you can spin yourself. Your answer is dependent on the area on which the spinner stops, so that your answer is completely anonymous.
How Does this Work?
On screen there will appear a circle, which is divided into several areas. By clicking with your mouse pad on 'start spinner', you will let the spinner spin. The red arrow marks the area where the spinner eventually stops. Your answer is dependent on the area where the spinner stops.

- If the spinner stops on an area imprinted with 'yes', then always answer 'yes' by clicking on 'yes'
- If the spinner stops on an area imprinted with 'no', then always answer 'no' by clicking on 'no'
- If the spinner stops on an *empty* area, then answer 'yes' or 'no' truthfully

As nobody knows on which area the spinner stopped, nobody can know why you have answered with 'yes' or 'no'. Your answer is thus a secret, as even the researchers cannot know why you have given a 'yes' or 'no' answer. Nevertheless, it is useful to answer according the given rules, as the researchers of the Vrije Universiteit and the Universiteit Utrecht can estimate the proportion of people answering 'yes' because of spinner directions and the proportion of people answering 'yes' truthfully.

First, three trial questions will be given.

Each trial question was posed on a separate page according to setup described above.

(19/50) Have you read a newspaper today?
(20/50) Have you run through a red traffic light past week?
(21/50) Is/was your mother's birthday between January 1 and March 31?

After these trial questions further comments were delivered to acknowledge that a forced answer that does not correspond to one's true status can sometimes be counterintuitive and to redefine the construct of 'being honest' in terms of following the FRR instructions.

You have just made all trial questions. It could be that you were directed to answer 'yes', as the spinner stopped on an area imprinted with 'yes', while your true answer would be 'no'. Or, that the spinner stopped on an area imprinted with 'no', so that you were directed to answer accordingly, while your true answer would be 'yes'. From previous research we know that people can find these instructions strange or even dishonest. You do not have to worry about these issues. You can view the spinner method as a game, and you play honestly when playing by the rules.

Now, our questions will follow.

Each of the 10 prevalence questions were posed on a separate page according to setup described above

(22/50) Have you ever in the past twelve months accepted gifts with a value below 25 euros from externals (as citizens and shop holders)?
(23/50) Have you ever in the past twelve months unjustifiably declared working hours?
(24/50) Have you ever in the past twelve months been careless with confidential police information?
(25/50) Have you ever in the past twelve months twisted the true cause of a case?
(26/50) Have you ever in the past twelve months threatened with the use of violence during an interrogation?
(27/50) Have you ever in the past twelve months in exchange for compensation (in money or kind) given confidential police information to third persons?
(28/50) Have you ever in the past twelve months incorrectly taken care of arrestees?
(29/50) Have you ever in the past twelve months personally taken possessions of the force?
(30/50) Have you ever in the past twelve months participated in the harassing and bullying of a colleague?
(31/50) Have you ever in the past twelve months in your free time associated with persons, which are known to have criminal antecedents?

Subsequently the quantitative FRR procedure was introduced. The introduction provided an explanation of quantitative method along with instructions. After this introduction 2 trial questions were asked to let the respondent get acquainted with the quantitative FRR procedure. Following these trial questions, 10 prevalence questions were asked.

All questions were posed on a separate page. The page make-up was analogous as in the dichotomous FRR. The spinner corresponded to Figure 4.2. All questions were multi-proportional with possible answer possibilities 1, 2, 3, 4, 5 and 6. These numbers corresponded with the respective categories 0 times, 1 time, 2 to 3 times, 4 to 5 times, 5 to 10 times and more than 10 times. This setup can also be found in the impression at the end of this appendix.

B.4 Introduction of the Quantitative Spinner Method

You have just answered some questions with the spinner method. Up till now, all spinner questions were 'yes or no' questions. Now we want to ask some questions to which you must respond with a certain number. These questions will also be asked with the help of the spinner method.

How must you answer these questions with the spinner method?
On your screen the spinner will appear again, which can be operated as previously. The areas imprinted with 'yes' or 'no' are now replaced with areas imprinted with numbers. Again, your answer is dependent on the area where the spinner stops.

- If the spinner stops on a numbered area, the number points to your according answer. For example, if the spinner stops on an area imprinted with '1', then please click on answering category '1'. If the spinner stops on an area imprinted with '2', answer accordingly, etc.
- If the spinner stops on an *empty* area, we then again ask you to respond truthfully.

By answering according to the rules of this method, your privacy is again fully protected. For example, if you respond with '3', it cannot be known if that is your truthful answer, or an answer redirected by the spinner as it stopped on an area imprinted with '3'. Again, your answers are secret.

Nevertheless, it is useful to answer according the given rules, as the researchers of the Vrije Universiteit and the Universiteit Utrecht can estimate the mean of all responses.

First, two trial questions will be given.

Each trial question was posed on a separate page according to setup described above.

(32/50) How many times in the past twelve months have you dodged fares when using public transport?
(33/50) On which day of the month is or was your father's birthday?

The second trial question was the only one with deviating answering categories:
1 = 1 – 5
2 = 6 – 10
3 = 11 – 15
4 = 16 – 20
5 = 21 – 25
6 = 26 – 31

After these trial questions further comments were delivered to acknowledge that a forced answer that does not correspond to one's true status can sometimes be counterintuitive and to redefine the construct of 'being honest' in terms of following the FRR instructions.

You have just made all trial questions. Again, it could be that you were asked to give a certain answer that did not correspond with your true answer (when the spinner stops on a numbered area). We emphasize again that this method can be seen as a game, and you play honestly when playing by the rules.

Now, our questions will follow.

(34/50) How many times in the past twelve months have you used organizational resources for private purposes?
(35/50) How many times in the past twelve months have you consulted police records for family and friends?
(36/50) How many times in the past twelve months have you wrongly reported sick?

(37/50) How many times in the past twelve months have you applied inappropriate and disproportional violence towards citizens, suspects or arrestees?
(38/50) How many times in the past twelve months have you accepted money or gifts to emphasize or neglect certain tasks in your function as police officer?
(39/50) How many times in the past twelve months have you (sexually) intimidated a colleague?
(40/50) How many times in the past twelve months have you reported untruthful information in charges or reports?
(41/50) How many times in the past twelve months have you taken found or confiscated goods or items?
(42/50) How many times in the past twelve months have you utilized unlawful investigative methods?
(43/50) How many times in the past twelve months have you used drugs (with the exception of alcohol, tobacco or medicine) in your free time?

Lastly, the evaluation questions were posed in the form of propositions. They encompassed a general question regarding the use of computers, 3 questions regarding the dichotomous FRR and 3 questions regarding the quantitative FRR.

B.5 Use of the Spinner Method

To conclude we would like to know your opinion on this questionnaire and its method. A few propositions will follow and we are interested to what degree you agree.

Each proposition was posed on a separate page.

(44/50) I found it easy to answer questions with the help of a computer.

With regard to the dichotomous questions
(45/50) It was clear to me when to answer with 'yes' and when to answer with 'no'.

(46/50) I am confident that all answers are fully anonymous on the individual level.
(47/50) With the use of the spinner it became easier for me to respond truthfully to sensitive questions on which one has to answer with 'yes' or 'no'.

With regard to the quantitative questions
(48/50) It was clear to me when to answer with a certain number.
(49/50) I am confident that nobody can know if I responded truthfully or according to the spinner directions.
(50/50) With the use of the spinner it became easier for me to respond truthfully to sensitive questions.

All with answering categories:
1 = strongly disagree
2 = disagree
3 = do not disagree, but also do not agree
4 = agree
5 = strongly agree

By clicking on 'end' all data were simultaneously stored in a php database. The respondents were then directed to the final screen, which thanked them for their cooperation and gave the email account and phone number of the project coordinator for questions or remarks.

CASI QUESTIONNAIRE

The CASI questionnaire had essentially the same setup and make-up as the CARR questionnaire. The same questions were asked as in the CARR questionnaire. Also, the operation of the questionnaire was the same as in the CARR condition. Due to being just a computer-assisted questionnaire, which had no RR incorporation, all questions were posed directly. The CASI questionnaire thus had a slightly different general introduction, no RR introductions and trial questions, and had a different setup for the evaluation questions. We will look into the text block formulations of B.1, B.3, B.4 and B.5 for CASI here.

B.1 CASI: Introduction

The first two paragraphs of the introduction were the same as in the CARR questionnaire. The third and fourth paragraph had a slightly different formulation, which, naturally, did not glimpse on the RR method.

We do **not** want to embarrass anyone. This questionnaire is not about which police employee ever crosses the line, but about insights in the occurrence of certain unacceptable behaviors. The researchers of the Vrije Universiteit van Amsterdam along with the researchers of the Universiteit van Utrecht guarantee that everyone's privacy will be respected. XX (the privacy and information confidentiality functionary of the XX police force) has approved the questionnaire, as our method of data collection fully protects your anonymity. The police force will neither receive the individual answers nor the data file containing all answers. These will remain with the researchers. Also, no personal traits will have to be revealed in this questionnaire.

Before ask our questions regarding integrity problems which can occur in your work, we will firstly ask a few general questions regarding your work and work environment.

B.2 was identical in CARR and CASI conditions

B.3 CASI: Prevalence Questions Regarding Integrity Problems in Work and Work Environment

The prevalence questions received the following introduction.
Now, we arrive at the questions regarding integrity problems, which can occur in your work and work environment. These concern questions regarding *your own* conduct and which can be answered with 'yes' or 'no'. We emphasize again that we do not want to embarrass anyone. This questionnaire is not about which police employee ever crosses the line, but about insights in the occurrence of certain unacceptable behaviors.

You can answer by clicking on 'yes' or 'no' with your mouse pad.

No trial questions were given. All prevalence questions in the CASI condition were identical to the non-trial dichotomous questions in the CARR condition.

B.4 CASI: Incidence Questions Regarding Integrity Problems in Work and Work Environment

The incidence questions received the following introduction.

You have just answered some questions regarding integrity problems. Until now, those were all questions, which could be answered with 'yes', or 'no'. Now, we want to ask some questions regarding *your own* conduct, which revolve around *frequency* of occurrence.

You can answer by clicking on the right answering category with your mouse pad.

No trial questions were given. All incidence questions in the CASI condition were identical to the non-trial quantitative questions in the CARR condition.

B.5 CASI: Trust in the Anonymity of the Questionnaire

To conclude we would like to know your opinion on the method of questioning in this survey. A few propositions will follow and we are interested to what degree you agree.

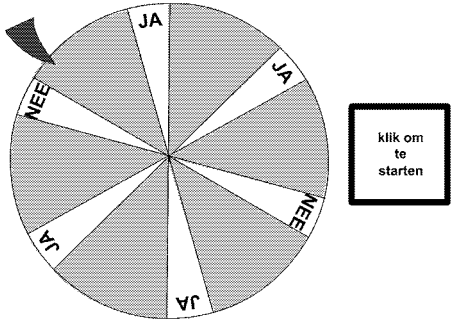
The first two propositions posed were identical to question 44/50 and 44/60 in the CARR questionnaire. Then the following propositions were posed (note that the CASI questionnaire had less questions due to the absence of trial questions.

(41/42) With the use of the computer-assisted questionnaire it became easier for me to respond truthfully to sensitive questions on which one has to answer with 'yes' or 'no'.

(42/42) With the use of the computer-assisted questionnaire it became easier for me to respond truthfully to sensitive questions on which one has to answer with a certain number.

Uitleg

- Als de spinner stopt op een vlak met het opschrift 'ja', antwoordt u dan altijd met 'ja' door dit antwoord aan te klikken;
- Als de spinner stopt op een vlak met het opschrift 'nee', antwoordt u dan altijd met 'nee' door dit antwoord aan te klikken;
- Als de spinner stopt op een vlak zonder opschrift, beantwoordt u dan de volgende vraag naar waarheid:



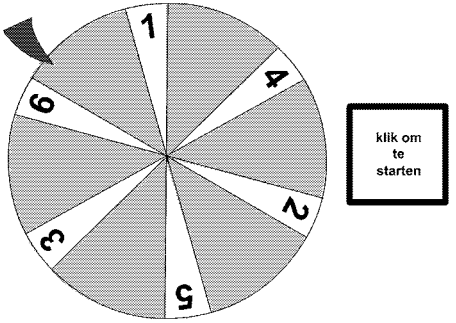
(22/50) Heeft u in de afgelopen 12 maanden wel eens geschenken met een waarde van onder de 25 euro van externen (zoals burgers, winkeliers en leveranciers) aangenomen ?

Ja Nee

volgende vraag

Uitleg

- Als de spinner stopt op een vlak met het opschrift '1', '2', '3', '4', '5' of '6' dan klikt u dat antwoord rechtstreeks aan bij de keuzevakken
- Als de spinner stopt op een vlak ***zonder*** opschrift, beantwoordt u dan de volgende vraag naar waarheid:



(34/50) Hoe vaak heeft u in de afgelopen twaalf maanden organisatiemiddelen voor privé-doeleinden gebruikt ?

a
a

1	2	3	4	5	6
0 keer	1 keer	2-3 keer	4-5 keer	5-10 keer	10+ keer
◉	○	○	○	○	○

volgende vraag

Appendix C: Test-Run Results Randomizer

This appendix contains the test-run results of the new randomizer for the dichotomous forced RR model. Three consecutive test-runs with differing iterations are presented along with χ^2. Note that the randomizer for the multi-proportional forced RR model has a make-up that is analogous with the dichotomous randomizer. Test results for this specific randomizer were thus similar.

The numbers 1 to 24 represent each of the 24 virtual areas of the total randomizer, which means that every area with 'yes' or 'no' or a number (1 to 6) has a 1/24 chance, making the total chance for a forced response 1/4 (6 · 1/24). The chance on an empty area and thus a truthful response is 3/24 each, making a total chance of 3/4 (6 · 3/24). The second column represents the absolute frequency of occurrence of each area out of the total number of iterations. The third column presents the χ^2 (Chi Square) test.

RUN 1

1	2129	1,013983
2	2023	1,730643
3	2085	0,001841
4	2131	1,104155
5	2093	0,047607
6	2068	0,108616
7	2136	1,346389
8	2053	0,433261
9	1994	3,806174
10	2076	0,023804
11	2026	1,56202
12	1964	6,802993
13	2135	1,296022
14	2081	0,002001
15	2149	2,088533
16	2078	0,012203
17	2081	0,002001
18	2077	0,017523
19	2063	0,192828
20	2127	0,927651
21	2062	0,212551
22	2086	0,004201
23	2153	2,34953
24	2123	0,766508
Total	49993	25,85304
Expected	2083,04167	

RUN 2

1	9554	0,039371305
2	9608	0,564667265
3	9442	0,899814164
4	9382	2,443136529
5	9638	1,120798209
6	9469	0,451684322
7	9415	1,500860351
8	9588	0,298794197
9	9529	0,003318497
10	9569	0,123931526
11	9534	4,09691E-05
12	9686	2,403281789
13	9463	0,538053738
14	9555	0,04354032
15	9341	3,932051929
16	9714	3,374583754
17	9652	1,444932614
18	9543	0,007356412
19	9423	1,306830696
20	9672	1,979300772
21	9421	1,354079539
22	9546	0,013570605
23	9487	0,237884618
24	9600	0,448249472
Total	228831	24,53013359
Expected	9534,625	

RUN 3

1	1039	0,016312
2	1053	0,093484
3	999	1,866522
4	1020	0,512657
5	1083	1,524281
6	1058	0,212118
7	1033	0,098277
8	1022	0,427816
9	1085	1,681022
10	1010	1,051902
11	1061	0,306306
12	1039	0,016312
13	1016	0,705348
14	1018	0,605168
15	1041	0,004329
16	1087	1,845431
17	1092	2,290009
18	1055	0,135186
19	1030	0,165144
20	1077	1,100075
21	1064	0,41775
22	1033	0,098277
23	987	3,019787
24	1033	0,098277
Total	25035	18,29179

Expected 1043,125

TOTAL RUNS

1	12722	0,29591
2	12684	0,042543
3	12526	1,435044
4	12533	1,289865
5	12814	1,853975
6	12595	0,341886
7	12584	0,465766
8	12663	0,000385
9	12608	0,220125
10	12655	0,002649
11	12621	0,125061
12	12689	0,062848
13	12614	0,172932
14	12654	0,003643
15	12531	1,330555
16	12879	3,760814
17	12825	2,129754
18	12675	0,015945
19	12516	1,65587
20	12876	3,658115
21	12547	1,022728
22	12665	0,001399
23	12627	0,09019
24	12756	0,715961
Total	303859	20,69396
Expected	12660,79167	

Appendix D: CARR and CASI Respondent Letters

This appendix contains the original respondent letters for the respective CARR and CASI respondent groups. They were sent immediately following an announcement introducing this study on the regional police force local intranet. The letters provided further introduction into the setup and aims of the study, along with assurances regarding confidentiality and specifications regarding the sub-domain Internet address containing the specific questionnaire as well as username and password for accession.

For preservation of the anonymity of the police force and its employees, names are feigned with 'X'.

Faculteit der Sociale Wetenschappen

Datum	Ons kenmerk	Bijlage(n)
26-05-2005		-
Uw brief van	Uw kenmerk	

Telefoon	Fax	E-mail
(020) 598 6805	(020) 598 6820	CFW.Peeters@fsw.vu.nl

Postadres: Afdeling Bestuur en Organisatie, De Boelelaan 1081c, 1081 HV Amsterdam

Aan: de leden van politiekorps XX

vrije Universiteit *amsterdam*

Beste korpsmedewerker,

Graag willen we u uitnodigen mee te werken aan een onderzoek naar de persoonlijke beleving van integriteitsvoorvallen in het politiewerk. Dit onderzoek wordt uitgevoerd door de onderzoeksgroep Politiestudies en Integriteit van Bestuur van de Vrije Universiteit Amsterdam en de capaciteitsgroep Methodenleer en Statistiek van de Universiteit Utrecht. Het betreft een nieuw soort onderzoek dat nog nooit eerder binnen de politie is uitgevoerd, waarbij gebruik wordt gemaakt van een nieuwe vragenlijstmethode, die uw privacy volledig beschermt. Een integriteitsprobleem dat u bent tegengekomen in uw werk kan moeilijk zijn om te bespreken met collega's en/of leidinggevenden. Met het onderzoek willen we twee doelen bereiken. Het gebruik van deze vragenlijstmethode geeft politiemedewerkers de kans volledig anoniem integriteitsvoorvallen in het werk te melden. Daarnaast willen wij een goed inzicht verwerven in de soorten integriteitsproblemen die voorkomen bij de politie. Het gaat ons niet om de vraag welke medewerker wel eens onaanvaardbaar gedrag vertoont, maar om inzicht in welke onaanvaardbare gedragingen relatief vaak voorkomen. De korpsleiding heeft ingestemd om in XX de nieuwe methode uit te proberen, omdat zij het belang inziet van een methode die de gelegenheid biedt om zonder verdere gevolgen voor u als politiemedewerker, gevoelige informatie over eigen handelen te geven.

Controlegroep
Om deze nieuwe methode goed te kunnen testen hebben we ook een controlegroep met respondenten nodig. Uw gegevens zijn via het personeelsbestand willekeurig geselecteerd om deel uit te maken van deze controlegroep. We willen u vragen de vragenlijst in te vullen, waarbij geen gebruik is gemaakt van de nieuwe vragenlijstmethode. Ook deze vragenlijst is anoniem. De korpsleiding heeft niet de beschikking over de individuele antwoorden en ook niet over het databestand met de antwoorden van alle respondenten. Deze blijven bij de onderzoekers van de universiteiten. Daarnaast hoeft u geen persoonlijke kenmerken in te vullen in deze vragenlijst. Het spreekt voor zich dat we uw gegevens volledig vertrouwelijk behandelen.

Waar vindt u de vragenlijst en hoe moet u deze invullen?
Het onderzoek maakt gebruik van een vragenlijst op internet, welke zich bevindt op een internet subdomein waarvan alleen u het adres weet. Om uw privacy optimaal te waarborgen, willen we u de gelegenheid geven thuis via uw internetverbinding in te loggen. U vindt hieronder het internetadres waarop de vragenlijst zich bevindt en het wachtwoord waarmee u kunt inloggen.

Bezoekadres: Buitenveldertselaan 3

Internet: http://www.fsw.vu.nl

Het kan natuurlijk zijn dat u op uw privé-adres geen internetverbinding heeft. Over enige tijd zullen de onderzoekers met een laptop de wijkbureaus bezoeken, om u daar alsnog in de gelegenheid te stellen uw medewerking te verlenen aan het onderzoek. U ontvangt ten aanzien van deze mogelijkheid bericht per e-mail op uw werk. Als u thuis wel de beschikking heeft over internet, willen wij u vragen de vragenlijst vóór 17 juni aanstaande in te vullen.

Het invullen van de vragenlijst is simpel. Zodra u het onderstaande internetadres heeft ingetypt boven in de internet adresbalk, wordt u gevraagd een gebruikersnaam en een wachtwoord te geven. Deze gegevens staan ook hieronder aangegeven. Hierna komt u vanzelf in de vragenlijst die u met eenvoudige instructies uitlegt hoe u de vragen invult. De vragenlijst is kort gehouden. Naast enkele algemene vragen over uw werkbeleving en het interne beleid van het korps volgen twee maal tien vragen over uw persoonlijke beleving ten aanzien van integriteit. We willen ook graag weten wat u van de manier van vragen stellen vindt. Daarom stellen wij aan het einde van de vragenlijst hierover enkele vragen. Probeert u de hele vragenlijst in te vullen, er zijn geen goede of foute antwoorden, het gaat om *uw* beleving en om wat *u* aan ons wilt laten weten. U sluit de vragenlijst af door na de laatste vraag 'einde' met de muis aan te klikken.

Internetadres waarop de vragenlijst zich bevindt:
casi.fsw.vu.nl [Let op!: geen www.]

Uw inloggegevens:
- gebruikersnaam: **casi**
- wachtwoord: **politie2** [politie twee]

Contact
Mocht u voor of na het invullen van de vragenlijst vragen hebben met betrekking tot de privacy van de door u ingevulde gegevens, dan kunt u contact opnemen met XX, privacy en informatiebeveiligingsfunctionaris van korps XX. U kunt ten aanzien van alle vragen die u over het project mocht hebben ook contact opnemen met Carel Peeters, coördinator van dit onderzoek en verbonden aan de Vrije Universiteit Amsterdam. Contactgegevens van beide personen bevinden zich onder.

Wij hopen ten zeerste op uw medewerking. Niet alleen vanwege het inzicht dat zo wordt verkregen in de soorten integriteitsproblemen bij de politie. Maar ook omdat het de onderzoekers inzicht geeft in de vraag of het via de gebruikte methode gemakkelijker wordt persoonlijke integriteitsvoorvallen aan te kaarten.

Alvast hartelijk dank voor uw medewerking, namens alle betrokken onderzoekers van de Vrije Universiteit en de Universiteit Utrecht,

Carel Peeters
Coördinator integriteitsonderzoek
Vrije Universiteit Amsterdam

CFW.Peeters@fsw.vu.nl
06-42126287
www.fsw.vu.nl/integriteit

XX
Privacy- en Informatiebeveiligingsfunctionaris
Korps XX

XX@XX.politie.nl
XX

Faculteit der Sociale Wetenschappen

Datum	Ons kenmerk	Bijlage(n)
26-05-2005		-
Uw brief van	Uw kenmerk	

Telefoon	Fax	E-mail
(020) 598 6805	(020) 598 6820	CFW.Peeters@fsw.vu.nl

Postadres: Afdeling Bestuur en Organisatie, De Boelelaan 1081c, 1081 HV Amsterdam

Aan: de leden van politiekorps XX

vrije Universiteit *amsterdam*

Beste korpsmedewerker,

Graag willen we u uitnodigen mee te werken aan een onderzoek naar de persoonlijke beleving van integriteitsvoorvallen in het politiewerk. Dit onderzoek wordt uitgevoerd door de onderzoeksgroep Politiestudies en Integriteit van Bestuur van de Vrije Universiteit Amsterdam en de capaciteitsgroep Methodenleer en Statistiek van de Universiteit Utrecht. Het betreft een nieuw soort onderzoek dat nog nooit eerder binnen de politie is uitgevoerd, waarbij gebruik wordt gemaakt van een nieuwe vragenlijstmethode, die uw privacy volledig beschermt. Een integriteitsprobleem dat u bent tegengekomen in uw werk kan moeilijk zijn om te bespreken met collega's en/of leidinggevenden. Met het onderzoek willen we twee doelen bereiken. Het gebruik van deze vragenlijstmethode geeft politiemedewerkers de kans volledig anoniem integriteitsvoorvallen in het werk te melden. Daarnaast willen wij een goed inzicht verwerven in de soorten integriteitsproblemen die voorkomen bij de politie. Het gaat ons niet om de vraag welke medewerker wel eens onaanvaardbaar gedrag vertoont, maar om inzicht in welke onaanvaardbare gedragingen relatief vaak voorkomen. De korpsleiding heeft ingestemd om in XX de nieuwe methode uit te proberen, omdat zij het belang inziet van een methode die de gelegenheid biedt om zonder verdere gevolgen voor u als politiemedewerker, gevoelige informatie over eigen handelen te geven.

Waar vindt u de vragenlijst en hoe moet u deze invullen?
Het onderzoek maakt gebruik van een vragenlijst op internet, welke zich bevindt op een internet subdomein waarvan alleen u het adres weet. Om uw privacy optimaal te waarborgen, willen we u de gelegenheid geven thuis via uw internetverbinding in te loggen. U vindt hieronder het internetadres waarop de vragenlijst zich bevindt en het wachtwoord waarmee u kunt inloggen. Het kan natuurlijk zijn dat u op uw privé-adres geen internetverbinding heeft. Over enige tijd zullen de onderzoekers met een laptop de wijkbureaus bezoeken, om u daar alsnog in de gelegenheid te stellen uw medewerking te verlenen aan het onderzoek. U ontvangt ten aanzien van deze mogelijkheid bericht per e-mail op uw werk. Als u thuis wel de beschikking heeft over internet, willen wij u vragen de vragenlijst vóór 17 juni aanstaande in te vullen.

De vragen in deze vragenlijst worden gesteld met behulp van een kansmechanisme, een soort spel dat ervoor zorgt dat uw antwoorden geen enkele individuele betekenis meer hebben en waardoor het ook voor ons als onderzoekers niet meer mogelijk is terug te rekenen waarom u met 'ja' of 'nee' heeft geantwoord. De korpsleiding heeft niet de beschikking over de individuele

Bezoekadres: Buitenveldertselaan 3

Internet: http://www.fsw.vu.nl

antwoorden en ook niet over het databestand met de antwoorden van alle respondenten. Deze blijven bij de onderzoekers van de universiteiten. Daarnaast hoeft u geen persoonlijke kenmerken in te vullen in deze vragenlijst.

Het invullen van de vragenlijst is simpel. Zodra u het onderstaande internetadres heeft ingetypt boven in de internet adresbalk, wordt u gevraagd een gebruikersnaam en een wachtwoord te geven. Deze gegevens staan ook hieronder aangegeven. Hierna komt u vanzelf in de vragenlijst die u met eenvoudige instructies uitlegt hoe u de vragen invult en hoe het kansmechanisme precies werkt. De vragenlijst is kort gehouden. Naast enkele algemene vragen over uw werk en het interne integriteitsbeleid van het korps, volgen twee maal tien vragen over uw persoonlijke beleving ten aanzien van integriteit, welke gesteld zullen worden met behulp van het kansmechanisme. We willen ook graag weten wat u van de nieuwe methode vindt. Daarom stellen wij aan het einde van de vragenlijst hierover enkele vragen. Probeert u de hele vragenlijst in te vullen, er zijn geen goede of foute antwoorden, het gaat om *uw* beleving. U sluit de vragenlijst af door na de laatste vraag 'einde' met de muis aan te klikken.

Internetadres waarop de vragenlijst zich bevindt:
rrt.fsw.vu.nl [Let op!: geen www.]

Uw inloggegevens:
- gebruikersnaam: **rrt**
- wachtwoord: **politie1** [politie een]

Contact
Mocht u voor of na het invullen van de vragenlijst, vragen hebben met betrekking tot de privacy van de door u ingevulde gegevens, dan kunt u contact opnemen met XX, privacy- en informatiebeveiligingsfunctionaris van korps XX. U kunt ten aanzien van alle vragen die u over het project mocht hebben, ook contact opnemen met Carel Peeters, coördinator van dit onderzoek en verbonden aan de Vrije Universiteit Amsterdam. Contactgegevens van beide personen bevinden zich onderaan deze brief.

Wij hopen ten zeerste op uw medewerking. Niet alleen vanwege het inzicht dat zo wordt verkregen in de soorten integriteitsproblemen bij de politie. Maar ook omdat het de onderzoekers inzicht geeft in de vraag of het via de gebruikte methode voor u gemakkelijker wordt uw persoonlijke integriteitsvoorvallen aan te kaarten.

Alvast hartelijk dank voor uw medewerking, namens alle betrokken onderzoekers van de Vrije Universiteit en de Universiteit Utrecht,

Carel Peeters
Coördinator integriteitsonderzoek
Vrije Universiteit Amsterdam

CFW.Peeters@fsw.vu.nl
06-42126287
www.fsw.vu.nl/integriteit

XX
Privacy- en Informatiebeveiligingsfunctionaris
Korps XX

XX@XX.politie.nl
XX

Appendix E: Basic LEM Commands

This appendix contains the ℒEM commands that give the ML estimates of the proportions in our dichotomous and quantitative FRR models. The ℒEM interface has three windows: an input window, an output window and a log window. The program is operated from the input window. Here, the basic input files are given for the obtainment of ML estimates of the models for prevalence and incidence estimates as given in (4.4) and (4.11). The basic model for ML estimates in both our dichotomous and quantitative FRR models can be stated as in (4.15) and (4.16) due to the (multi-)proportional setup. Also, input files are given for subgroup prevalence and incidence estimates. Each input file is accompanied by an output example.

For more on the ℒEM program in general see Vermunt (1997). For more on the program in analyzing RR data see Capaciteitsgroep Methodenleer en Statistiek, Universiteit Utrecht (2003) and Van Den Hout (2004). The ℒEM program and manual can be downloaded for free from URL: http://www.uvt.nl/faculteiten/fsw/organisatie/departementen/mto/software2.html.

ℒEM COMMANDS FOR PREVALENCE OR INCIDENCE ESTIMATES

First the basic input file for the obtainment of ML prevalence estimates in the dichotomous FRR model is given, followed by its output. Subsequently the basic input file for the obtainment of ML incidence estimates in the multi-proportional quantitative FRR is given also followed by its output. Input and output receive basic description.

E.1 Input File for Prevalence Estimates with Dichotomous Forced Randomized Response Model

```
lat 1
man 1
dim 2 2
lab P A

mod P {P}
A|P  {wei(PA)}
sta wei(PA)[.9167 .0833
            .1667 .8333]

dat [29 94]

nec
nfr
nR2
npa
```

The input file is a general input file for a ML prevalence estimate with ℒEM, when our variable is measured under a RR condition. The commands 'lat' and 'man' give the number of latent and manifest variables in the analysis. We are interested in a prevalence estimate and both are '1', as the latent construct we are interested in is abstracted from the answers on our sensitive question. The command 'dim' gives the number of categories associated with our latent and manifest variables. Both can be labeled with the 'lab' command.

The 'mod' label gives which variable we wish analyzed. Next we find a specification of the relationship between our manifest and latent constructs. Subsequently, we find our conditional misclassification matrix

as in (4.2) with rows denoting '1'. The 'mod' command through the specification of the matrix with the conditional misclassification probabilities gives the model specification.

Data in LEM are the observed frequencies of our manifest variable ('yes' and 'no') and can be specified with the 'dat' command. The commands 'nec', 'nfr', 'nR2' and 'npa' are not necessary, but suppress certain output so that interpretation of output is enhanced.

The input file gives the following output:

```
*** STATISTICS ***

  Number of iterations = 24
  Converge criterion   = 0.0000006102
  Seed random values   = 778

  X-squared            = 0.0000 (0.0000)
  L-squared            = 0.0000 (0.0000)
  Cressie-Read         = 0.0000 (0.0000)
  Dissimilarity index  = 0.0000
  Degrees of freedom   = 0
  Log-likelihood       = -67.17739
  Number of parameters = 1 (+1)
  Sample size          = 123.0
  BIC(L-squared)       = 0.0000
  AIC(L-squared)       = 0.0000
  BIC(log-likelihood)  = 139.1670
  AIC(log-likelihood)  = 136.3548

WARNING: no information is provided on identification of
parameters


*** (CONDITIONAL) PROBABILITIES ***

* P(P) *

  1              0.0922
  2              0.9078

* P(A|P) *

  1 | 1          0.9167
  2 | 1          0.0833
  1 | 2          0.1667
  2 | 2          0.8333
```

```
*** LATENT CLASS OUTPUT ***

          P  1    P  2
        0.0922  0.9078
  A  1  0.9167  0.1667
  A  2  0.0833  0.8333

E = 0.0922, lambda = 0.0000
```

The output file gives the subsequent output of the above given input file with data from the manifest variable 'Have you ever in the past twelve months accepted gifts with a value below 25 euros from externals (as citizens and shop holders)?' as an example. First the statistics regarding the fit of the model is given, of which especially X-squared (Chi Square) is of importance (measure of discrepancy between model and data). The warning is given due to the suppression of data and can be ignored. Next we find our conditional probabilities. The section 'latent class output' gives the data we are looking for. We see our conditional misclassification matrix, and directly under 'P 1' and 'P 2' the estimated prevalence of acceptance and non-acceptance of gifts by officers respectively. Estimated population prevalence of police officer acceptance of gifts by externals is thus 9.22%.

E.2 Input File for Incidence Estimates with Multi-Proportional Discrete Quantitative Forced Randomized Response Model

```
lat 1
man 1
dim 6 6
lab P A

mod P {P}
A|P  {wei(PA)}
sta wei(PA)[.7917 .0417 .0417 .0417 .0417 .0417
            .0417 .7917 .0417 .0417 .0417 .0417
            .0417 .0417 .7917 .0417 .0417 .0417
            .0417 .0417 .0417 .7917 .0417 .0417
            .0417 .0417 .0417 .0417 .7917 .0417
            .0417 .0417 .0417 .0417 .0417 .7917]

dat [74 15 11 10 6 7]

nec
nfr
nR2
npa
```

The input file is an input file for ML incidence estimates with ℒEM, when our variable is measured under a RR condition that gives a multi-proportional appreciation of quantitative data. The interpretation of the input file is analogous to the input file for prevalence estimates. Due to our multi-proportional setup of the quantitative FRR model, which redirects data into 6 discrete categories, we have 6 dimensions for our latent and manifest constructs and 6 separate frequencies behind our 'dat' command. In our model specification we find our matrix of conditional misclassification probabilities as given in (4.14).

The input file gives the following output:

```
*** STATISTICS ***

  Number of iterations = 41
  Converge criterion   = 0.0000009362
  Seed random values   = 223

  X-squared            = 0.0000 (0.0000)
  L-squared            = 0.0000 (0.0000)
  Cressie-Read         = 0.0000 (0.0000)
  Dissimilarity index  = 0.0001
  Degrees of freedom   = 0
  Log-likelihood       = -159.00248
  Number of parameters = 5 (+1)
  Sample size          = 123.0
  BIC(L-squared)       = 0.0000
  AIC(L-squared)       = 0.0000
  BIC(log-likelihood)  = 342.0659
  AIC(log-likelihood)  = 328.0050

WARNING: no information is provided on identification of
parameters


*** (CONDITIONAL) PROBABILITIES ***

* P(P) *

  1              0.7467
  2              0.1070
  3              0.0637
  4              0.0528
  5              0.0095
  6              0.0203
```

```
* P(A|P) *

  1 | 1          0.7915
  2 | 1          0.0417
  3 | 1          0.0417
  4 | 1          0.0417
  5 | 1          0.0417
  6 | 1          0.0417
  1 | 2          0.0417
  2 | 2          0.7915
  3 | 2          0.0417
  4 | 2          0.0417
  5 | 2          0.0417
  6 | 2          0.0417
  1 | 3          0.0417
  2 | 3          0.0417
  3 | 3          0.7915
  4 | 3          0.0417
  5 | 3          0.0417
  6 | 3          0.0417
  1 | 4          0.0417
  2 | 4          0.0417
  3 | 4          0.0417
  4 | 4          0.7915
  5 | 4          0.0417
  6 | 4          0.0417
  1 | 5          0.0417
  2 | 5          0.0417
  3 | 5          0.0417
  4 | 5          0.0417
  5 | 5          0.7915
  6 | 5          0.0417
  1 | 6          0.0417
  2 | 6          0.0417
  3 | 6          0.0417
  4 | 6          0.0417
  5 | 6          0.0417
  6 | 6          0.7915


*** LATENT CLASS OUTPUT ***

           P  1    P  2    P  3    P  4    P  5    P  6
         0.7467  0.1070  0.0637  0.0528  0.0095  0.0203
  A  1   0.7915  0.0417  0.0417  0.0417  0.0417  0.0417
  A  2   0.0417  0.7915  0.0417  0.0417  0.0417  0.0417
  A  3   0.0417  0.0417  0.7915  0.0417  0.0417  0.0417
  A  4   0.0417  0.0417  0.0417  0.7915  0.0417  0.0417
  A  5   0.0417  0.0417  0.0417  0.0417  0.7915  0.0417
  A  6   0.0417  0.0417  0.0417  0.0417  0.0417  0.7915

E = 0.1698, lambda = 0.3297
```

The output file gives the subsequent output of the above given input file with data from the manifest variable 'How many times in the past twelve months have you used organizational resources for private purposes?' as an example. The interpretation of the output file is analogous to the output file for prevalence estimates. Under 'P 1' to 'P 6' we find the estimated population proportions in each of our 6 discrete quantitative categories.

ℒEM COMMANDS FOR SUBGROUP PREVALENCE AND INCIDENCE ESTIMATES

It can prove to be necessary to divide the obtained sample in subgroups and compare prevalence or incidence estimates between categorical groupings. Here basic ℒEM commands are given for subgroup prevalence and incidence estimates. First the basic input file for the obtainment of ML subgroup prevalence estimates in the dichotomous FRR model is given, followed by its output. Subsequently the basic input file for the obtainment of ML subgroup incidence estimates in the multi-proportional quantitative FRR is given also followed by its output. Input and output receive basic description. The examples given here have CARR questionnaire completion via Internet or during follow-up study as the subgroup specification.

E.3 Input File for Subgroup Prevalence Estimates with Dichotomous Forced Randomized Response Model

```
lat 1
man 2
dim 2 2 2
lab P A V

mod PV {P,V}
A|P  {wei(PA)}
sta wei(PA)[.9167 .0833
            .1667 .8333]
```

```
dat [10 19 35 59]

nec
nfr
nR2
npa
nco
nlo
```

The input file is a general input file for testing the significance of the relation between a certain categorical variable and a dichotomous latent construct measured under a RR condition. The input file has a setup that is analogous to the input file that gives prevalence estimates. But here we have an additional manifest variable, namely our categorical variable for which we want to explore if prevalence estimates differ over its categories. Here, that variable has received the label 'V'. The notation of the model specification {P,V}, gives a model which tests its main effects only. Also, as we have two subgroups, the data have to be specified over the categories of those subgroups. The data are ordered to subsequently give the frequency of 'yes' in subgroup 1, the frequency of 'yes' in subgroup 2, the frequency of 'no' in subgroup 1, and frequency of 'no' in subgroup 2.

The input file gives the following output:

```
*** STATISTICS ***

  Number of iterations = 24
  Converge criterion   = 0.0000005809
  Seed random values   = 4369

  X-squared            = 0.0723 (0.7880)
  L-squared            = 0.0727 (0.7874)
  Cressie-Read         = 0.0724 (0.7878)
  Dissimilarity index  = 0.0099
  Degrees of freedom   = 1
  Log-likelihood       = -147.95296
  Number of parameters = 2 (+1)
  Sample size          = 123.0
  BIC(L-squared)       = -4.7394
  AIC(L-squared)       = -1.9273
  BIC(log-likelihood)  = 305.5303
  AIC(log-likelihood)  = 299.9059

WARNING: no information is provided on identification of
parameters
```

The output file gives the subsequent output of the above given input file with data from the manifest variable 'Have you ever in the past twelve months accepted gifts with a value below 25 euros from externals (as citizens and shop holders)?' as an example, and CARR questionnaire completion via internet or during follow-up study as the subgroup specification. The non-significant p-value .7880 of the Chi Square gives that there is no significant difference between the prevalence estimates for acceptance of gifts between the subgroup that completed the CARR questionnaire via the Internet and the subgroup that completed the CARR questionnaire during the follow-up study.

It is also possible to obtain the specific prevalence estimates for the various subgroups. One would than have to fit a saturated model, which, next to the main effects, also takes into account the interaction effects between the variables 'P' and 'V'. An input file, which specifies a saturated model, is given by:

```
lat 1
man 2
dim 2 2 2
lab P A V

mod PV {PV}
A|P  {wei(PA)}
sta wei(PA)[.9167 .0833
            .1667 .8333]

dat [10 19 35 59]

nec
nfr
nR2
npa
```

The input file for the saturated model is analogous to the input file given above. The model notation {PV} specifies a saturated model. This input file gives the following output:

```
*** STATISTICS ***

  Number of iterations = 25
  Converge criterion   = 0.0000009238
  Seed random values   = 4541

  X-squared            = 0.0000 (0.0000)
  L-squared            = 0.0000 (0.0000)
  Cressie-Read         = 0.0000 (0.0000)
  Dissimilarity index  = 0.0000
  Degrees of freedom   = 0
  Log-likelihood       = -147.91659
  Number of parameters = 3 (+1)
  Sample size          = 123.0
  BIC(L-squared)       = 0.0000
  AIC(L-squared)       = 0.0000
  BIC(log-likelihood)  = 310.2697
  AIC(log-likelihood)  = 301.8332

WARNING: no information is provided on identification of
parameters


*** (CONDITIONAL) PROBABILITIES ***

* P(PV) *

  1 1            0.0271
  1 2            0.0650
  2 1            0.3387
  2 2            0.5691

* P(A|P) *

  1 | 1          0.9167
  2 | 1          0.0833
  1 | 2          0.1667
  2 | 2          0.8333


*** LATENT CLASS OUTPUT ***

          P  1    P  2
        0.0922  0.9078
  A  1  0.9167  0.1667
  A  2  0.0833  0.8333
  V  1  0.2944  0.3731
  V  2  0.7056  0.6269

E = 0.0922, lambda = -0.0000
```

Again, the input file gives the subsequent output of the above given input file with data from the manifest variable 'Have you ever in the past twelve months accepted gifts with a value below 25 euros from externals (as citizens and shop holders)?' as an example, and CARR questionnaire completion via internet or during follow-up study as the subgroup specification. The interpretation of the output is analogous to the regular output file for prevalence estimates. Under 'P 1' we find our prevalence estimate as found earlier. With the conditional probabilities given under 'P (PV)' we can calculate the prevalence estimates for our specific subgroups. For example, the prevalence estimate of accepting gifts with a value below 25 euros from externals within the subgroup who completed the CARR questionnaire during the follow-up study equals .0271/(.0271 + .3387) = .074. The analogous prevalence estimate in the subgroup that completed the CARR questionnaire via the Internet equals .065/(.065 + .5691) = .1025.

E.4 Input File for Subgroup Incidence Estimates with Multi-Proportional Discrete Quantitative Forced Randomized Response Model

```
lat 1
man 2
dim 6 6 2
lab P A V

mod PV {P,V}
A|P  {wei(PA)}
sta wei(PA)[.7917 .0417 .0417 .0417 .0417 .0417
            .0417 .7917 .0417 .0417 .0417 .0417
            .0417 .0417 .7917 .0417 .0417 .0417
            .0417 .0417 .0417 .7917 .0417 .0417
            .0417 .0417 .0417 .0417 .7917 .0417
            .0417 .0417 .0417 .0417 .0417 .7917]

dat [28 46 4 11 5 6 2 8 3 3 3 4]

nec
nfr
nR2
npa
nco
nlo
```

The input file is a general input file for testing the significance of the relation between a certain categorical variable and a multi-proportional latent construct measured under a RR condition. The input file has a setup that is analogous to the input file that gives incidence estimates. The notation of the model specification {P,V}, gives a model which tests its main effects only. Again, as we have two subgroups, the data have to be specified over the categories of those subgroups. The data are ordered to subsequently give the frequency of respondents affirming π_1 in subgroup 1, the frequency respondents affirming π_1 in subgroup 2, the frequency of respondents affirming π_2 in subgroup 1, the frequency of respondents affirming π_2 in subgroup 2, etc.

The input file gives the following output:

```
*** STATISTICS ***

  Number of iterations = 41
  Converge criterion   = 0.0000009513
  Seed random values   = 1886

  X-squared            = 2.8284 (0.7264)
  L-squared            = 2.9458 (0.7083)
  Cressie-Read         = 2.8607 (0.7215)
  Dissimilarity index  = 0.0512
  Degrees of freedom   = 5
  Log-likelihood       = -239.77806
  Number of parameters = 6 (+1)
  Sample size          = 123.0
  BIC(L-squared)       = -21.1151
  AIC(L-squared)       = -7.0542
  BIC(log-likelihood)  = 508.4292
  AIC(log-likelihood)  = 491.5561

WARNING: no information is provided on identification of
parameters
```

The output file gives the subsequent output of the above given input file with data from the manifest variable 'How many times in the past twelve months have you used organizational resources for private purposes?' as an example, and CARR questionnaire completion via internet or during follow-up study as the subgroup specification. The non-significant p-value .7264 of the Chi Square gives that there is no significant difference between the incidence estimates for use of organizational resources for private purposes between the subgroup that completed the CARR

questionnaire via the Internet and the subgroup that completed the CARR questionnaire during the follow-up study.

It is also possible to obtain the specific prevalence estimates for the various subgroups. One would than have to fit a saturated model, which, next to the main effects, also takes into account the interaction effects between the variables 'P' and 'V'. An input file, which specifies a saturated model in our multi-proportional quantitative setup, is given by:

```
lat 1
man 2
dim 6 6 2
lab P A V

mod PV {PV}
A|P  {wei(PA)}
sta wei(PA)[.7917 .0417 .0417 .0417 .0417 .0417
            .0417 .7917 .0417 .0417 .0417 .0417
            .0417 .0417 .7917 .0417 .0417 .0417
            .0417 .0417 .0417 .7917 .0417 .0417
            .0417 .0417 .0417 .0417 .7917 .0417
            .0417 .0417 .0417 .0417 .0417 .7917]

dat [28 46 4 11 5 6 2 8 3 3 3 4]

nec
nfr
nR2
npa
```

The input file for the saturated model is analogous to the input file given above. The model notation {PV} specifies a saturated model. This input file gives the following output:

```
*** STATISTICS ***

  Number of iterations = 116
  Converge criterion   = 0.0000009674
  Seed random values   = 2480

  X-squared            = 0.0204 (0.0000)
  L-squared            = 0.0209 (0.0000)
  Cressie-Read         = 0.0206 (0.0000)
  Dissimilarity index  = 0.0021
  Degrees of freedom   = 0
```

```
  Log-likelihood       = -238.31562
  Number of parameters = 11 (+1)
  Sample size          = 123.0
  BIC(L-squared)       = 0.0209
  AIC(L-squared)       = 0.0209
  BIC(log-likelihood)  = 529.5653
  AIC(log-likelihood)  = 498.6312

WARNING: no information is provided on identification of
parameters


*** (CONDITIONAL) PROBABILITIES ***

* P(PV) *

  1 1            0.2832
  1 2            0.4618
  2 1            0.0230
  2 2            0.0836
  3 1            0.0339
  3 2            0.0296
  4 1            0.0013
  4 2            0.0512
  5 1            0.0122
  5 2            0.0000
  6 1            0.0122
  6 2            0.0080

* P(A|P) *

  1 | 1          0.7915
  2 | 1          0.0417
  3 | 1          0.0417
  4 | 1          0.0417
  5 | 1          0.0417
  6 | 1          0.0417
  1 | 2          0.0417
  2 | 2          0.7915
  3 | 2          0.0417
  4 | 2          0.0417
  5 | 2          0.0417
  6 | 2          0.0417
  1 | 3          0.0417
  2 | 3          0.0417
  3 | 3          0.7915
  4 | 3          0.0417
  5 | 3          0.0417
  6 | 3          0.0417
  1 | 4          0.0417
  2 | 4          0.0417
  3 | 4          0.0417
  4 | 4          0.7915
```

```
  5 | 4          0.0417
  6 | 4          0.0417
  1 | 5          0.0417
  2 | 5          0.0417
  3 | 5          0.0417
  4 | 5          0.0417
  5 | 5          0.7915
  6 | 5          0.0417
  1 | 6          0.0417
  2 | 6          0.0417
  3 | 6          0.0417
  4 | 6          0.0417
  5 | 6          0.0417
  6 | 6          0.7915


*** LATENT CLASS OUTPUT ***

          P  1    P  2    P  3    P  4    P  5    P  6
         0.7450  0.1066  0.0634  0.0525  0.0122  0.0201
  A  1   0.7915  0.0417  0.0417  0.0417  0.0417  0.0417
  A  2   0.0417  0.7915  0.0417  0.0417  0.0417  0.0417
  A  3   0.0417  0.0417  0.7915  0.0417  0.0417  0.0417
  A  4   0.0417  0.0417  0.0417  0.7915  0.0417  0.0417
  A  5   0.0417  0.0417  0.0417  0.0417  0.7915  0.0417
  A  6   0.0417  0.0417  0.0417  0.0417  0.0417  0.7915
  V  1   0.3802  0.2159  0.5338  0.0256  0.9998  0.6047
  V  2   0.6198  0.7841  0.4662  0.9744  0.0002  0.3953

E = 0.1612, lambda = 0.3678
```

Again, the input file gives the subsequent output of the above given input file with data from the manifest variable 'How many times in the past twelve months have you used organizational resources for private purposes?' as an example, and CARR questionnaire completion via internet or during follow-up study as the subgroup specification. The interpretation of the output is analogous to the regular output file for incidence estimates. With the conditional probabilities given under 'P (PV)' we can calculate the proportions $\pi_1,\ldots, \pi_6$ for our specific subgroups. For example, the population proportion who has never used organizational resources for private purposes in the past twelve months (π_1) within the subgroup who completed the CARR questionnaire during the follow-up study equals .2832/(.2832 + .0230 + .0339 + .0013 + .0122 + .0122) = .7742. The analogous population proportion in the subgroup that completed the CARR questionnaire via the Internet equals .4618/(.4618 + .0836 + .0296 + .0512 + .008) = .7282, etc.

References

Abernathy, J.R., B.G. Greenberg & D.G. Horvitz (1970) 'Estimates of Induced Abortion in Urban North Carolina', *Demography* 7: 19-29.

Abul-Ela, A-L.A., B.G. Greenberg & D.G. Horvitz (1967) 'A Multi-Proportions Randomized Response Model', *Journal of the American Statistical Association* 62: 990-1008.

Ambainis, A., M. Jakobsson & H. Lipmaa (2004) 'Cryptographic Randomized Response Techniques', *Public Key Cryptography, Proceedings of PKC '04*, pp. 425-438

Anderson, H. (1976) 'Estimation of a Proportion through Randomized Response', *International Statistical Review* 44: 213-217.

Armacost, R.L., J.C. Hosseini, S.A. Morris & K.A. Rehbein (1991) 'An Empirical Comparison of Direct Questioning, Scenario, and Randomized Response Methods for Obtaining Sensitive Business Information', *Decision Sciences* 22: 1073-1090.

Arnab, R. (2004) 'Optional Randomized Response Techniques for Complex Survey Designs', *Biometrical Journal* 46: 114-124.

Bar-Lev, S.K., E. Bobovitch & B. Boukai (2004) 'A Note on Randomized Response Models for Quantitative Data', *Metrika* 60: 255-260.

Barnes, J.A. (1980) *Who Should Know What? Social Science, Privacy and Ethics*. Cambridge [etc.]: Cambridge University Press.

Beldt, S.F., W.W. Daniel & B.S. Garcia (1982) 'The Takahasi-Sakasegawa Randomized Response Technique: A Field Test', *Sociological Methods and Research* 11: 101-111.

Bellhouse, D.R. (1980) 'Linear Models for Randomized Response Designs', *Journal of the American Statistical Association* 75: 1001-1004.

Benjamin, M. (1990) *Splitting the Difference: Compromising and Integrity in Ethics and Politics*. Kansas: University Press of Kansas.

Berman, J., H. McCombs & R. Boruch (1977) 'Notes on the Contamination Method: Two Small Experiments in Assuring Confidentiality of Responses', *Sociological Methods and Research* 6: 45-62.

Biemer, P.P., R.M. Groves, L.E. Lyberg, N.A. Mathiowetz & S. Sudman (Eds.) (1991) *Measurement Errors in Surveys*. New York [etc.]:

John Wiley & Sons (Wiley Series in Probability and Mathematical Statistics).

Blalock, H.M. (1979) 'The Presidential Address: Measurement and Conceptualization Problems: The Major Obstacle to Integrating Theory and Research', *American Sociological Review* 44: 881-894.

Boeije, H. & G.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.

Bollen, K. (1989) *Structural Equations with Latent Variables.* New York [etc.]: John Wiley & Sons (Wiley Series in Probability and Mathematical Statistics).

Boruch, R.F. (1971) 'Assuring Confidentiality of Responses in Social Research: A Note on Strategies', *American Sociologist* 6: 308-311.

Boruch, R.F. (1972) 'Relations Among Statistical Methods for Assuring Confidentiality of Social Research Data', *Social Science Research* 1: 403-414.

Boruch, R.F. & J.S. Cecil (1979) *Assuring the Confidentiality of Social Research Data.* Philadelphia: University of Philadelphia Press.

Bourke, P.D. (1984) 'Estimation of Proportions Using Symmetric Randomized Response Designs', *Psychological Bulletin* 96: 166-172.

Bourke, P.D. & T. Dalenius (1976) 'Some New Ideas in the Realm of Randomized Response Inquiries', *International Statistical Review* 44: 219-221.

Bourke, P.D. & M.A. Moran (1984) 'Application of the EM Algorithm to Randomized Response Data', *Proceedings of the American Statistical Association, Survey Research Methods Section,* pp. 788-739.

Bourke, P.D. & M.A. Moran (1988) 'Estimating Proportions From Randomized Response Data Using the EM Algorithm', *Journal of the American Statistical Association* 83: 964-968.

Brewer, J.D. (1990) 'Sensitivity as a Problem in Field Research: A Study of Routine Policing in Northern Ireland', *American Behavioral Scientist* 33: 578-593.

Brown, G.H. (1975) *Randomized Inquiry vs. Conventional Questionnaire Methods in Estimating Drug Usage Rates Trough Mail Surveys.* Alexandria, VA: Human Recourses Research Organization, HUMRO Technical Report 75-14.

Burton, B.K. & J.P. Near (1995) 'Estimating the Incidence of Wrongdoing and Whistle-blowing: Results of a Study Using

Randomized Response Technique', *Journal of Business Ethics* 14: 17-30.

Campbell, A.A. (1987) 'Randomized Response Technique', *Science* 236: 1049.

Campbell, C. & B.L. Joiner (1973) 'How to Get the Answer without Being Sure You've Asked the Question', *The American Statistician* 27: 229-231.

Capaciteitsgroep Methodenleer en Statistiek, Universiteit Utrecht (2003) *Handleiding dataverzameling met behulp van Randomized Response.* URL: http://www.randomized response.nl [Handbook data gathering with Randomized Response].

Capaciteitsgroep Methodenleer en Statistiek, Universiteit Utrecht (2003) *Handleiding data-analyse bij het gebruik van Randomized Response.* URL: http://www.randomized response.nl [Handbook data-analysis when using Randomized Response].

Carmines, E.G. & R.A. Zeller (1979) *Reliability and Validity Assessment.* Beverly Hills [etc]: Sage Publications (Sage University Paper Series on Quantitative Applications in the Social Sciences, no. 17).

Chaudhuri, A. (2004) 'Christofides' Randomized Response Technique in Complex Sample Surveys', *Metrika* 60: 223-228.

Chaudhuri, A. & R. Mukerjee (1988) *Randomized Response: Theory and Techniques.* New York: Marcel Dekker (Statistics Series, vol. 85).

Chaudhuri, A. & A. Saha (article in press) 'Optional versus Compulsory Randomized Response Techniques in Complex Surveys', *Journal of Statistical Planning and Inference.*

Chen, T.T. (1979) 'Analysis of Randomized Response as Purposively Misclassified Data', *Proceedings of the American Statistical Association, Survey Research Methods Section,* pp. 158-163.

Christofides, T.C. (2003) 'A generalized Randomized Response Technique', *Metrika* 57: 195-200.

Christofides, T.C. (2005) 'Randomized Response in Stratified Sampling', *Journal of Statistical Planning and Inference* 128: 303-310.

Chua, T.C. & A.K. Tsui (2000) 'Procuring Honest Responses Indirectly', *Journal of Statistical Planning and Inference* 90: 107-116.

Clark, S.J. & R.A. Desharnais (1998) 'Honest Answers to Embarrasing Questions: Detecting Cheating in the Randomized Response Model', *Psychological Methods* 3: 160-168.

Cosmides, L. & J. Tooby (1996) 'Are Humans Good Intuitive Statisticians after all? Rethinking some Conclusions from the Literature on Judgment under Uncertainty', *Cognition* 58: 1-73.

Couper, M.P., E. Singer & R. Tourangeau (2003) 'Understanding the Effects of Audio-CASI on Self-Reports of Sensitive Behavior', *The Public Opinion Quarterly* 67: 385-395.

Danermark, B. & B. Swensson (1987) 'Measuring Drug Use Among Swedish Adolescents. Randomized Response Versus Anonymous Questionnaires', *Journal of Official Statistics* 3: 439-448.

Deffaa, W. (1982) *Anonymisierte Befragungen mit Zufallsverschlüsselten Antworten. Die Randomized Response Technik (RRT): Methodische Grundlagen, Modelle und Anwendungen.* Frankfurt am Main: Verlag Peter Lang [Anonymous Questioning with Misclassified Responses. The Randomized Response Technique (RRT): Methodological Assumptions, Models and Uses].

Delleart, F. (2002) *The Expectation Maximization Algorithm.* College of Computing, Georgia Institute of Technology. Technical Report GIT-GVU-02-20.

Du, W. & R. Gopolakrishna (2001) *Application of Randomized Response Strategy in Privacy-Preserving Survey.* CERIAS Technical Report 2001-14.

Du, W. & Z. Zhan (2003) 'Using Randomized Response Techniques for Privacy-Preserving Data Mining', *Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data-mining*, pp. 1-6.

Duffy, J.C. & J.J. Waterton (1984) 'Randomized Response Models for Estimating the Distribution Function of a Quantitative Character', *International Statistical Review* 52: 165-171.

Dutka, S. & L.R. Frankel (1993) 'Measurement Errors in Organizational Surveys', *American Behavioral Scientist* 36: 472-484.

Edgell, S.E., S. Himmelfarb & K.L. Duchan (1982) 'Validity of Forced Responses in a Randomized Response Model', *Sociological Methods & Research* 11: 89-100.

Eichhorn, B.H. & L.S. Hayre (1983) 'Scrambled Randomized Response Methods for Obtaining Sensitive Quantitative Data', *Journal of Statistical Planning and Inference* 7: 307-316.

Eisenhower, D., N.A. Mathiowetz & D. Morganstein (1991) 'Recall Error: Sources and Bias Reduction Techniques', in: P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz & S. Sudman (Eds.) *Measurement Errors in Surveys.* New York [etc.]: John Wiley & Sons, pp. 127-144 (Wiley Series in Probability and Mathematical Statistics).

Elffers, H., P. van der Heijden & M. Hezemans (2003) 'Explaining Regulatory Non-Compliance: A Survey Study of Rule Transgression

for Two Dutch Instrumental Laws, Applying the Randomized Response Method', *Journal of Quantitative Criminology* 19: 409-439.

Eliason, S.R. (1993) *Maximum Likelihood Estimation: Logic and Practice.* Beverly Hills [etc]: Sage Publications (Sage University Paper Series on Quantitative Applications in the Social Sciences, no. 96).

Eriksson, S.A. (1973) 'A New Model for Randomized Response', *International Statistical Review* 41: 101-113.

Fenton, K.A., A.M. Johnson, S. McManus & B. Erens (2001) 'Measuring Sexual Behavior: Methodological Challenges in Survey Research', *Sexually Transmitted Infections* 77: 84-92.

Fijnaut, C. & L.W.J.C. Huberts (2002) 'Corruption, Integrity and Law Enforcement', in: C. Fijnaut & L.W.J.C. Huberts (Eds.) *Corruption, Integrity and Law Enforcement.* The Hague [etc.]: Kluwer Law International, pp. 3-34.

Filion, F.L. (1975) 'Estimating Bias due to Nonresponse in Mail Surveys', *Public Opinion Quarterly* 39: 482-492.

Folsom, R.E., B.G. Greenberg, D.G. Horvitz & J.R. Abernathy (1973) 'The Two Alternate Questions Randomized Response Model for Human Surveys', *Journal of the American Statistical Association* 68: 525-530.

Fox, J.A. & P.E. Tracy (1980) 'A Field-Validation of a Quantitative Randomized Response Model' *Proceedings of the American Statistical Association, Survey Research Section*, pp. 299-304.

Fox, J.A. & P.E. Tracy (1981) 'Reaffirming the Viability of the Randomized Response Approach', *American Sociological Review* 46: 930-933.

Fox, J.A. & P.E. Tracy (1984) 'Measuring Associations with Randomized Response', *Social Science Research* 13: 188-197.

Fox, J.A. & P.E. Tracy (1986) *Randomized Response. A Method for Sensitive Surveys.* Beverly Hills [etc]: Sage Publications (Sage University Paper Series on Quantitative Applications in the Social Sciences, no. 58).

Gigerenzer, G. (1991) 'How to make Cognitive Illusions Disappear: Beyond Heuristics and Biases', *European Review of Social Psychology* 2: 83-115.

Gils, G. van, P. van der Heijden, O. Laudy & R. Ross (2003) *Regelovertreding in de sociale zekerheid: de tweede meting van het periodiek onderzoek naar regelovertreding in de sociale zekerheidsregelingen WAO, WW en Abw.* 's-Gravenhage : Ministerie

van Sociale Zaken en Werkgelegenheid [Rule transgressing in social security].
Goodstadt, M.S. & V. Gruson (1975) 'The Randomized Response Technique: A Test on Drug Use', *Journal of the American Statistical Association* 70: 814-818.
Greenberg, B.G., A-L.A. Abul-Ela, W.R. Simmons & D.G. Horvitz (1969) 'The Unrelated Question Randomized Response Model: Theoretical Framework', *Journal of the American Statistical Association* 64: 520-539.
Greenberg, B.G., R.R. Kuebler, J.R. Abernathy & D.G. Horvitz (1971) 'Application of the Randomized Response Technique in Obtaining Quantitative Data', *Journal of the American Statistical Association* 66: 243-250.
Greenberg, B.G., R.R. Kuebler, J.R. Abernathy & D.G. Horvitz (1977) 'Respondent Hazards in the Unrelated Question Randomized Response Model', *Journal of Statistical Planning and Inference* 1: 53-60.
Groves, R.M. (1991) 'Measurement Error Across the Disciplines', in: P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz & S. Sudman (Eds.) *Measurement Errors in Surveys.* New York [etc.]: John Wiley & Sons, pp. 1-25 (Wiley Series in Probability and Mathematical Statistics).
Gupta, S., B. Gupta & S. Singh (2002) 'Estimation of Sensitivity Level of Personal Interview Survey Questions', *Journal of Statistical Planning and Inference* 100: 239-247.
Hagenaars, J.A. (1993) *Loglinear Models with Latent Variables.* Beverly Hills [etc]: Sage Publications (Sage University Paper Series on Quantitative Applications in the Social Sciences, no. 94).
Han, G-S. & W.D. Warde (1994) 'Correlations in Randomized Response Surveys', *Proceedings of the American Statistical Association, Survey Research Methods Section*, pp. 491-495.
Hansen, M.H., W.N. Hurwitz, E.S. Marks & W.P. Maudlin (1951) 'Response Error in Surveys', *Journal of the American Statistical Association* 46: 147-190.
Heijden, P.G.M. van der, G. van Gils, J. Bouts & J.J. Hox (1998) 'A Comparison of Randomized Response, CASAQ, and Direct Questioning', *Kwantitatieve Methoden* 19: 15-34.
Heijden, P.G.M. van der, G. van Gils, J. Bouts & J.J. Hox (2000) 'A Comparison of Randomized Response, Computer-Assisted Self-Interview, and Face-to-Face Direct Questioning: Eliciting Sensitive

Information in the Context of Welfare and Unemployment Benefit', *Sociological Methods and Research* 28: 505-537.

Herzberger, S.D. (1990) 'The Cyclical Pattern of Child Abuse: A Study of Research Methodology', *American Behavioral Scientist* 33: 529-545.

Horvitz, D.G., B.G. Greenberg & J.R. Abernathy (1976) 'Randomized Response: a Data-Gathering Device for Sensitive Questions', *International Statistical Review* 44: 181-196.

Horvitz, D.G., B.V. Shah & W.R. Simmons (1967) 'The Unrelated Question Randomized Response Model', *Proceedings of the American Statistical Association, Social Statistics Section*, pp. 65-72.

Hosseini, J.C. & R.L. Armacost (1993) 'Gathering Sensitive Data in Organizations', *American Behavioral Scientist* 36: 443-471.

Houston, J. & A. Tran (2001) 'A Survey of Tax Evasion Using the Randomized Response Technique', *Advances in Taxation* 13: 69-94.

Hout, A.D.L. van den (2004) *Analyzing Misclassified Data: Randomized Response and Post Randomization*. [S.l. : s.n.].

Hout, A. van den, P.G.M. van der Heijden (2002) 'Randomized Response, Statistical Disclosure Control, and Misclassification: A Review', *International Statistical Review* 70: 269-288.

Hout, A. van den, P.G.M. van der Heijden (2004) 'The Analysis of Multivariate Misclassified Data with Special Attention to Randomized Response Data', *Sociological Methods and Research* 32: 384-410.

Huberts, L.W.J.C. (1998) *Blinde vlekken in de politiepraktijk en de politiewetenschap*. Gouda Quint: Deventer [Blind spots in police practice and police science].

Huberts, L.W.J.C., H. Hulschebosch, K. Lasthuizen & C.F.W. Peeters (2004) *Nederland fraude- en corruptieland? De omvang, achtergronden en afwikkeling van corruptie- en fraudeonderzoeken in Nederlandse gemeenten in 1991 en 2003*. Amsterdam: Vrije Universiteit [The Netherlands, a fraudulent and corrupt country? The magnitude, backgrounds and measures regarding corruption and fraud investigations in Dutch municipalities in 1991 and 2003].

Huberts, L.W.J.C., M. Kaptein & K. Lasthuizen (2005) *Leadership and Integrity: A Preliminary Study of the Impact of Management Behavior on Employee Conduct*. Concept manuscript.

Huberts, L.W.J.C., K. Lasthuizen & C.F.W. Peeters (forthcoming) 'Measuring Corruption: Exploring the Iceberg', in: C. Sampford, A. Shacklock & C. Connors (Eds.) *Measuring Corruption*. Hampshire, London [etc.]: Ashgate Publishing.

Huberts, L.W.J.C., D. Pijl & A. Steen (1999) 'Integrity and corruption', in: C. Fijnaut, E. Muller & U. Rosenthal (Eds.) *Police: Studies on the Organization and its Functioning*. Alphen aan den Rijn: Samsom, pp. 57-79.

I-Cheng, C, L.P. Chow & R.V. Rider (1972) 'The Randomized Response Technique as Used in the Taiwan Outcome of Pregnancy Study', *Studies in Family Planning* 3: 265-269.

Inspectie voor de Rechtshandhaving (1998) *Monitoring van beleidsinstrumentele wetgeving*. Den Haag: Ministerie van Justitie, Inspectie voor de Rechtshandhaving [Monitoring policy-instrumental legislation].

Kahneman, D. & A. Tversky (1973) 'On the Psychology of Prediction', *Psychological Review* 80: 237-251.

Kaptein, M. & J. Wempe (2002) *The Balanced Company: A Corporate Integrity Approach.* Oxford: Oxford University Press.

Kim, J-I. & J.A. Flueck (1978) 'Modification of the Randomized Response Technique for Sampling Without Replacement', *Proceedings of the American Statistical Association, Social Statistics Section*, pp. 346-350.

Kim, J-M. & M.E. Elam (2005) 'A Two-Stage Stratified Warner's Randomized Response Model Using Optimal Allocation', *Metrika* 61: 1-7.

Kim, J-M., J.M. Tebbs & S-W. An (article in press) 'Extensions of Mangat's Randomized-Response Model', *Journal of Statistical Planning and Inference.*

Kinsey, S.H., J.S. Thornberry, C.P. Carson & A.P. Duffer (1995) 'Respondent Preferences Toward Audio-CASI and How that Affects Data Quality', *Proceedings of the American Statistical Association, Social Research Methods Section*, pp 1023-1028.

Kish, L. (1976) 'Discussion on the Invited and Contributed Papers', *International Statistical Review* 44: 227.

Klockars, C.B. (1985) *The Idea of Police*. Beverly Hills [etc.]: Sage Publications.

Klockars, C.B. (1997) *Conceptual and methodological issues in the study of police integrity.* Paper presented to the Advisory Panel Meeting of the Project on Police Integrity. December 16, Washington DC.

Klockars, C.B., S. Kutnjak Ivkovich, W.E. Harver & M.R. Haberfeld (2000) *The measurement of police integrity*. National Institute of Justice, U.S. Department of Justice.

Kooiman, P., L.C.R.J. Willenborg & J.M. Gouweleeuw (1997) *PRAM: A Method for Disclosure Limitation of Microdata.* Voorburg/Heerlen: Statistics Netherlands, Research paper no. 9705.

Krotki, K. & B. Fox (1974) 'The Randomized Response Technique, the Interview and the Self-Administered Questionnaire: An Empirical Comparison of Fertility Reports', *Proceedings of the American Statistical Association, Social Statistics Section*, pp. 367-371.

Kuk, A.Y.C. (1990) 'Asking Sensitive Questions Indirectly', *Biometrika* 77: 436-438.

Kulka, R.A., M.F. Weeks & R.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.

Kundert, K.R. (1989) 'Sensitive Questions and Randomized Response Techniques', *The College Mathematics Journal* 20: 409-411.

Lamb, C.W. & D.E. Stem (1978) 'An Empirical Validation of the Randomized Response Technique', *Journal of Marketing research* 15: 616-621.

Lambe, C.G. (1967) *Statistical Methods and Formulae.* London: The English University Press (Applied Mathematics Series).

Lamboo, M.E.D., L.W.J.C. Huberts, M. van der Steeg & A. Nieuwendijk (2002) *The Monitor Internal Investigations Police: Dimensions of Police Misconduct.* Paper presented at the 2002 American Society of Criminology Conference. Chicago, November 12-16, 2002.

Landsheer, J.A., P. van der Heijden & G. van Gils (1999) 'Trust and Understanding, Two Psychological Aspects of Randomized Response: A Study of a Method for Improving the Estimate of Social Security Fraud', *Quality and Quantity* 33:1-12.

Lang, S. (February 2004) *Randomized Response: Befragungstechniken zur Vermeidung von Verzerrungen bei sensitiven Fragen.* Habilitations-Probevorlesung Universität München, Unstitut für Statistik [Randomized Response: Questioning Techniques for Curbing Bias when asking Sensitive Questions].

Lanke, J. (1975) 'On the Choice of the Unrelated Question in Simmons' Version of Randomized Response', *Journal of the American Statistical Association* 70: 80-83.

Lanke, J. (1976) 'On the Degree of Protection in Randomized Interviews', *International Statistical Review* 44: 197-203.

Lara, D., J. Strickler, C. Díaz Olavarrieta & C. Ellertson (2004) 'Measuring Induced Abortion in Mexico: A Comparison of Four Methodologies', *Sociological Methods and Research* 32: 529-558.

Larkins, E.R., E.C. Hume & B.S. Garcha (1997) 'The Validity of the Randomized Response Method in Tax Ethics Research', *Journal of Applied Business Research* 13: 25-32.

Lee, R.M. (1993) *Doing Research on Sensitive Topics*. London [etc.]: Sage Publications.

Lee, R.M. & C.M. Renzetti (1990) 'The Problems of Researching Sensitive Topics: An Overview and Introduction', *American Behavioral Scientist* 33: 510-528.

Leeuw, E.D., J. Hox & M. Huisman (2003) 'Prevention and Treatment of Item Nonresponse', *Journal of Official Statistics* 19: 153-176.

Lenvelt-Mulders, G.J.L.M. (2003) 'Randomized Response Technieken voor het Onderzoek van Sociaal Gevoelige Onderwerpen', in: MarktOnderzoeksAssociatie (Ed.) *Ontwikkelingen in het Markt-onderzoek Jaarboek 2003*. Haarlem: Uitgeverij de Vrieseborch [Randomized Response Techniques for Research into Socially Sensitive Topics].

Lensvelt-Mulders, G.J.L.M., A. van den Hout & P.G.M. van der Heijden (forthcoming) 'An Empirical Test of a Computer-Assisted Randomized Response Survey to Study Sensitive Behavior', *Journal of the Royal Statistical Society, series A*.

Lensvelt-Mulders, G.J.L.M., J. Hox & P.G.M. van der Heijden (2005) 'Meta-Analysis of Randomized Response Research: 35 Years of Validation', *Sociological Methods and Research* 33: 319-348.

Lensvelt-Mulders, G.J.L.M., J. Hox & P.G.M. van der Heijden (2005) 'How to Improve the Efficiency of Randomized Response Designs', *Quality and Quantity* 39: 253-265.

Lenvelt-Mulders, G.J.L.M. & E. de Leeuw (2002a) 'Vragen naar gevoelige informatie', *Facta* 10: 34-35 [Asking for sensitive behavior].

Lenvelt-Mulders, G.J.L.M. & E. de Leeuw (2002b) 'Beschermd door een dobbelsteen, *Facta* 10: 28-30 [Protected by dice].

Leysieffer, F.W. & S.L. Warner (1976) 'Respondent Jeopardy and Optimal Designs in Randomized Response Models', *Journal of the American Statistical Association* 71: 649-656.

Li, R.K.C. (1976) 'Discussion on the Invited and Contributed Papers', *International Statistical Review* 44: 227.

Liu, P.T. & L.P. Chow (1976a) 'The Efficiency of the Multiple Trial Randomized Response Technique', *Biometrics* 32: 607-618.

Liu, P.T. & L.P. Chow (1976b) 'A New Discrete Quantitative Randomized Response Model', *Journal of the American Statistical Association* 71: 72-73.

Liu, P.T., L.P. Chow & W.H. Mosley (1975) 'Use of the Randomized Response Technique with a New Randomizing Device', *Journal of the American Statistical Association* 70: 329-332.

Ljungqvist, L. (1993) 'A Unified Approach to Measures of Privacy in Randomized Response Models: A Utilitarian Perspective', *Journal of the American Statistical Association* 88: 97-103.

Locander, W., S. Sudman & N. Bradburn (1976) 'An Investigation of Interview Method, Threat and Response Distortion', *Journal of the American Statistical Association* 71: 269-275.

Louis, T-A. (1982) 'Finding the Observed Information Matrix when Using the EM Algorithm', *Journal of the Royal Statistical Society, Series B* 44: 226-233.

Luce, R.D. & L. Narens (1987) 'Measurement Scales on the Continuum', *Science* 236: 1527-1532.

Lynch, S.M. & B. Western (2004) 'Bayesian Posterior Predictive Checks for Complex Models', *Sociological Methods and Research* 32: 301-335.

Maanen, J. van (1988) *Tales of the Field.* Chicago: University of Chicago Press.

Maddala, G.S. (1983) *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge [etc.]: Cambridge University Press.

Mangat, N.S. (1994) 'An Improved Randomized Response Strategy', *Journal of the Royal Statistical Society, Series B* 56: 93-95.

Mangat, N.S. & R. Singh (1990) 'An Alternative Randomized Response Procedure', *Biometrika* 77: 439-442.

McArthur, D. (1993) 'An Introduction to Log-Linear Analysis and Implementing the Newton-Rhapson Algorithm in APL2' *Proceedings of the International Conference on APL*, pp. 159-163.

Melton, G.B. & J.N. Gray (1988) 'Ethical Dilemmas in AIDS Research. Individual Privacy and Public Health', *American Psychologist* 43: 60-64.

Miller, J.D. (1981) 'Complexities of the Randomized Response Solution', *American Sociological Review* 46: 928-930.

Moon, Y. (1998) 'Impression Management in Computer-Based Interviews: The Effects of Input Modality, Output Modality, and Distance', *The Public Opinion Quarterly* 62: 610-622.

Moors, J.J.A. (1971) 'Optimization of the Unrelated Question Randomized Response Model', *Journal of the American Statistical Association* 66: 627-629.

Moors, J.J.A. (1976) 'Discussion on the Invited and Contributed Papers', *International Statistical Review* 44: 229.

Moriarty, M. & F. Wiseman (1976) 'On the Choice of a Randomization Technique with the Randomized Response Mode', *Proceedings of the American Statistical Association, Social Statistics Section*, pp. 624-626.

Musch, J., A. Broder & K.C. Klauer (2001) 'Improving Survey Research on the World-Wide Web using the Randomized Response Technique', in: U.D. Reips & M. Bosnjak (Eds.) *Dimensions of Internet Science*. Lengerich, Germany: Pabst Science Publishers, pp. 179-192.

Nelen, H. & A. Nieuwendijk (2003) *Geen ABC. Analyse van Rijksrecherche-onderzoeken naar ambtelijke en bestuurlijke corruptie*. Den Haag: Boom Juridische Uitgevers [No ABC. Analysis of federal police investigations of official and political corruption].

Niemi, I. (1993) 'Systematic Error in Behavioural Measurement: Comparing Results From Interview and Time Budget Studies', *Social Indicators Research* 30: 229-244.

Nieuwbeerta, P. (Ed.) (2002) *Crime Victimization in comparative perspective. Results from the International Crime Victims Survey, 1989-2000.* Den Haag: Boom Juridische uitgevers.

Nieuwbeerta, P., G. de Geest & J. Siegers (2002) 'Corruption in Industrialised and Developing countries', in: P. Nieuwbeerta (Ed.) *Crime victimization in comparative perspective. Results from the International Crime Victims Survey, 1989-2000.* Den Haag: Boom Juridische Uitgevers, pp. 163-182.

O'Hagan, A. (1987) 'Bayes Linear Estimators for Randomized Response Models', *Journal of the American Statistical Association* 82: 580-585.

Orwin, R.G. & R.F. Boruch (1982) 'RRT Meets RDD: Statistical Strategies for Assuring Response Privacy in Telephone Surveys.

Padmawar, V.R. (article in press) 'A Note on Combining Randomized Response and Direct Response', *Journal of Statistical Planning and Inference.*

Padmawar, V.R. & K. Vijayan (2000) 'Randomized Response Revisited', *Journal of Statistical Planning and Inference* 90: 293-304.

Phillips, D.L. (1971) *Knowledge From What? Theories and Methods in Social Research.* Chicago: Rand-McNally.

Podsakoff, P.M., S.B. MacKenzie, J-Y. Lee & N.P. Podsakoff (2003) 'Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies', *Journal of Applied Psychology* 88: 879-903.

Pollock, K.H. & Y. Bek (1976) 'A Comparison of Three Randomized Response Models for Quantitative Data', *Journal of the American Statistical Association* 71: 884-886.

Poole, W.K. (1974) 'Estimation of the Distribution Function of a Continuous Type Random Variable Through Randomized Response', *Journal of the American Statistical Association* 69: 1002-1005.

Punch, M. (1989) 'Researching Police Deviance: A Personal Encounter with the Limitations and Liabilities of Field-Work', *British Journal of Sociology* 40: 177-204.

Raghavarao, D. (1978) 'On an Estimation Problem in Warner's Randomized Response Technique', *Biometrics* 34: 87-90.

Rasinski, K.A., G.B. Willis, A.K. Baldwin, W. Yeh & L. Lee (1999) 'Methods of Data Collection, Perception of Risks and Losses, and Motivation to Give Truthful Answers to Sensitive Survey Questions', *Applied Cognitive Psychology* 13: 465-484.

Reamer, F.G. (1979) 'Protecting Research Subjects and Unintended Consequences: The Effect of Guarantees of Confidentiality', *The Public Opinion Quarterly* 43: 497-506.

Rudas, T. (2004) *Probability Theory: A Primer*. Thousand Oaks [etc]: Sage Publications (Sage University Paper Series on Quantitative Applications in the Social Sciences, no. 142).

Scheers, N.J. & M.C. Dayton (1988) 'Covariate Randomized Response Models', *Journal of the American Statistical Association* 83: 969-974.

Scolnick, J.H. (1975) *Justice Without Trial*. New York [etc.]: John Wiley & Sons.

Scott, C. (1976) 'Discussion on the Invited and Contributed Papers', *International Statistical Review* 44: 229.

Sen, P.K. (1974) 'On Unbiased Estimation for Randomized Response Models', *Journal of the American Statistical Association* 69: 997-1001.

Sen. P.K. (1976) 'Asymptotically Optimal Estimators of General Parameters in Randomized Response Models', *International Statistical Review* 44: 223-224.

Shimizu, I.M. & G.S. Bonham (1978) 'Randomized Response Technique in a National Survey', *Journal of the American Statistical Association* 73: 35-39.

Sieber, J.E. & B. Stanley (1988) 'Ethical and Professional Dimensions of Socially Sensitive Research', *American Psychologist* 43: 49-55.

Singer, E., D.R. von Thurn & E.R. Miller (1995) 'Confidentiality Assurances and Response: A Quantitative Review of the Experimental Literature', *The Public Opinion Quarterly* 59: 66-77.

Singh, S. (2002) 'A New Stochastic Randomized Response Model', *Metrika* 56: 131-142.

Smeets, I. (1995) 'Facing Another Gap: An Exploration of the Discrepancies Between Voting Turnout in Survey Research and Official Statistics', *Acta Politica* 30: 307-334.

Solomon, B. (1999) *A Better Way to Think About Business: How Personal Integrity Leads to Corporate Success.* New York: Oxford University Press.

Spurrier, J.D. & W.J. Padgett (1980) 'The Application of Bayesian Techniques in Randomized Response', *Sociological Methodology* 11: 533-544.

Stem, D.E. & R.K. Steinhorst (1984) 'Telephone Interview and Mail Questionnaire Applications of the Randomized Response Model', *Journal of the American Statistical Association* 79: 555-564.

Stevens, S.S. (1951) 'Mathematics, Measurement and Psychophysics', in: S.S. Stevens (Ed.) *Handbook of Experimental Psychology.* New York [etc.]: John Wiley & Sons, pp. 1-49.

Sudman, S. (1976) *Applied Sampling.* New York: Academic Press (Quantitative Studies in Social Relations series).

Sudman, S. & N.M. Bradburn (1974) *Response Effects in Surveys: A Review and Synthesis.* Chicago: Aldine.

Sudman, S. & N.M. Bradburn (1982) *Asking Questions: A Practical Guide to Questionnaire Design.* San Francisco: Jossey-Bass.

Takahasi, K. & H. Sakasegawa (1977) 'A Randomized Response Technique Without use of any Randomizing Device', *Annals of the Institute of Statistical Mathematics* 29: 1-8.

Tamhane, A.C. (1981) 'Randomized Response Techniques for Multiple Sensitive Attributes', *Journal of the American Statistical Association* 76: 916-923.

Thompson, M. (1997) *Theory of Sample Surveys.* London: Chapman & Hall (Monographs on Statistics and Applied Probability, vol. 74).

Tourangeau, R. & T.W. Smith (1996) 'Asking Sensitive Questions: The Impact of Data Collection Mode, Question Format, and Question Context', *The Public Opinion Quarterly* 60: 275-304.

Tracy, P.E. & J.A. Fox (1981) 'The Validity of Randomized Response for Sensitive Measurements', *American Sociological Review*, 46: 187-200.

Tracy, D. & N. Mangat (1996) 'Some Developments in Randomized Response Sampling During the Last Decade – A Followup of Review by Chaudhuri and Mukerjee', *Journal of Applied Statistical Science* 4: 533-544.

Tversky, A. & D. Kahneman (1974) 'Judgement under Uncertainty: Heuristics and Biases', *Science* 185: 1124-1131.

Umesh, U.N. & R.A. Peterson (1991) 'A Critical Evaluation of the Randomized Response Method', *Sociological Methods & Research* 20: 104-138.

Verdooren, L.R. (1976) 'Loten bij delicate vragen; een overzicht van "randomized response"-technieken', *Statistica Neerlandica* 30: 7-24 [Drawing by lot for sensitive questions; an overview of "randomized response" techniques].

Vermunt, J.K. (1997) *LEM: A General Program for the Analysis of Categorical Data.* User's Manual Tilburg: Tilburg University.

Voogt, R.J.J. & H. van Kempen (2002) 'Nonresponse Bias and Stimulus Effects in the Dutch National Election Study', *Quality and Quantity* 36: 325-345.

Warner, S.L. (1965) 'Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias', *Journal of the American Statistical Association* 60: 63-69.

Warner, S.L. (1971) 'The Linear Randomized Response Model', *Journal of the American Statistical Association* 66: 884-888.

Warner, S.L. (1976a) 'Optimal Randomized Response Models', *International Statistical Review* 44: 205-212.

Warner, S.L. (1976b) 'Discussion on the Invited and Contributed Papers', *International Statistical Review* 44: 230.

Winkler, R.L. & L.A. Franklin (1979) 'Warner's Randomized Response Model: A Bayesian Approach', *Journal of the American Statistical Association* 74: 207-214.

Wiseman, F., M. Moriarty & M. Schafer (1975) 'Estimating Public Opinion With the Randomized Response Model', *The Public Opinion Quarterly* 39: 507-513.

Wolf, P-P. de & I. van Gelder (2004) *An Empirical Evaluation of PRAM.* Voorburg/Heerlen: Statistics Netherlands, Discussion paper no. 04012.

Wolf, P-P. de, J.M. Gouweleeuw, P. Kooiman & L.C.R.J. Willenborg (1997) *Reflections on PRAM.* Voorburg/Heerlen: Statistics Netherlands, Research paper no. 9742.

Zdep, S.M., I.N. Rhodes (1976) 'Making the Randomized Response Technique Work', *The Public Opinion Quarterly* 40: 531-537.

Zdep, S.M., I.N. Rhodes, R.M. Schwarz & M.J. Kilkenny (1979) 'The Validity of the Randomized Response Technique', *The Public Opinion Quarterly* 43: 544-549.

Dedicated to the curse and blessing of 'what if' until time and expectation run out