Bayesian Randomized Item Response Modeling for Sensitive Measurements

M. Avetisyan

Bayesian Randomized Item Response Modeling for Sensitive Measurements

M. Avetisyan

December 6, 2012

Graduation Committee

Chair	Prof. Dr. K. I. van Oudenhoven-van der Zee
Promotor	Prof. Dr. C. A. W. Glas
Assistant promotor	Dr. Ir. G. J. A. Fox
Members	Prof. Dr. W. Albers
	Prof. Dr. P. G. M. van der Heijden
	Prof. Dr. M. J. IJzerman
	Prof. Dr. J. A. M. van der Palen
	Prof. Dr. J. K. Vermunt

Avetisyan, Marianna
Bayesian Randomized Item Response Modeling for Sensitive Measurements
PhD Thesis University of Twente, Enschede. - Met samenvatting in het Nederlands.
ISBN: 978-90-365-3480-2
doi: 10.3990/1.9789036534802
printed by: Ipskamp Drukkers B.V., Enschede

BAYESIAN RANDOMIZED ITEM RESPONSE MODELING FOR SENSITIVE MEASUREMENTS

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof. dr. H. Brinksma,
on account of the decision of the graduation committee,
to be publicly defended
on Thursday, December 6, 2012 at 12.45

by

Marianna Avetisyan

born November 28, 1975
in Yerevan, Armenia

This dissertation is approved by the following promoters:

Promotor: Prof. Dr. C. A. W. Glas
Assistant promotor: Dr. Ir. G. J. A. Fox

Acknowledgements

Present work is the result of my PhD project at the Research Methodology, Measurement and Data Analysis (OMD) group of the University of Twente. It was a turbulent journey full of new experiences and events.

First of all I would like to thank Jean-Paul for guiding me through this process. I highly value his suggestions and feedback that were invaluable for successful completion of this thesis. I would also like to express my gratitude to Cees for providing support and advice throughout and especially at the last stages of working on this project.

This thesis would not be complete without cooperation with Job who saw interesting opportunities in applied research at Medical Spectrum Twente (MST) in Enschede. Special thanks go to the Department of Pulmonology at MST, in particular to the team of pulmonologists, to the Longfunctieafdeling: Poli 12, and Stoppen met roken: Poli 10, for providing me with the opportunity to collect data within a framework of sometimes a bit strange randomized response PimPamPet study.

I would like to thank everyone at OMD for a pleasant working environment; in particular, Josine, Iris and Erika for being wonderful office mates and Bernard and Stphanie for the interest they have shown in my project.

Last but not least, I would like to thank my family. Mom and dad, you always believe in me and support my every endeavor. Ira and Ralf, it would be impossible to complete this thesis without your help and support at the critical moments finishing writing. I dedicate this thesis to my son Lex!

Enschede, November 2012
Marianna Avetisyan

Contents

1 Introduction **1**
- 1.1 Self-reports and Response Bias in Sensitive Research 1
- 1.2 Methods for Neutralizing Response Bias 3
- 1.3 Randomized Item Response Theory Models 5
- 1.4 Bayesian Approach to IRT Modeling 7
- 1.5 Outline 8

2 The Dirichlet-Multinomial Model for Multivariate Randomized Response Data and Small Samples **11**
- 2.1 Introduction 12
- 2.2 Multivariate Randomized Response Techniques 13
- 2.3 The Beta-Binomial Model for Multivariate Binary RR data 14
- 2.4 The Dirichlet-Multinomial Model for Multivariate Categorical RR Data 16
- 2.5 Empirical Bayes and Full Bayes Estimation 18
  - 2.5.1 Empirical Bayes Estimation 18
  - 2.5.2 Full Bayes Estimation 20
- 2.6 Restricted Dirichlet-Multinomial Modeling 22
- 2.7 Application of the Dirichlet-Multinomial Model 22
  - 2.7.1 Simulation Study 22
  - 2.7.2 Response Rates of Alcohol-Related Negative Consequences 26
- 2.8 Discussion 30

3 Mixture Randomized Item Response Modeling: A Smoking Behavior Validation Study **33**
- 3.1 Introduction 34
- 3.2 Method 36
  - 3.2.1 Mixture Randomized Item Response Model 37
  - 3.2.2 Bayesian Latent Variable Methods for Diagnostic Accuracy 39
- 3.3 Results 41
  - 3.3.1 RRT Validation 43
  - 3.3.2 Bayesian Diagnostic Evaluation of Randomized Response Testing 47
- 3.4 Discussion and Conclusions 49

4 A Multidimensional Randomized Item Response Model 51
4.1 Introduction . . . 52
4.2 Modeling Individual Response Probabilities . . . 54
4.3 The Model . . . 55
4.3.1 Probit Response Functions . . . 55
4.3.2 Forced Randomized Response Design . . . 55
4.3.3 Structural Multivariate Latent Model . . . 56
4.3.4 Identification Issues . . . 57
4.4 Bayesian Inference . . . 58
4.4.1 Implementation Issues . . . 60
4.5 Simulation Study . . . 61
4.6 Measuring Drinking Problems and Alcohol-Related Expectancies among College Students . . . 63
4.6.1 Multi-Dimensional Scale Analysis . . . 64
4.6.2 Structural Model Analysis . . . 66
4.7 Discussion . . . 69

5 Randomized Response Techniques 71
5.1 Introduction . . . 71
5.2 Randomizing Device . . . 74
5.3 The Type of Data . . . 74
5.4 Single-Item Randomized Response Techniques . . . 75
5.4.1 Opposite-Question Method (Warner) . . . 75
5.4.2 Unrelated-Question Method (UQM) . . . 77
5.4.3 Forced Response Method (FRR) . . . 80
5.4.4 Smoke Study: "Do you smoke?" . . . 82
5.4.5 Smoke Study: "How many cigarettes are you smoking per day?" . . . 84
5.5 Nonrandomized Response Techniques . . . 85
5.5.1 Takahasi's RR Technique . . . 85
5.5.2 The Triangular and Crosswise RR Methods . . . 87
5.5.3 The Hidden Sensitivity RR Method . . . 88
5.6 Randomized Response Methods and Multi-Item Measurements . . . 90
5.6.1 Multi-Item Randomized Response Models . . . 90
5.6.2 The Beta-Binomial and Dirichlet-Multinomial Modeling Approach . . . 91
5.6.3 The Randomized Item Response Theory Modeling Approach 92
5.6.4 Mixture Modeling for Compliance and Non-Compliance . . . 94
5.7 Discussion . . . 95

A Derivation of the Marginal Log-Likelihood Function 97

B WinBUGS Code: Multinomial-Dirichlet Model Specification 99

C CAPS-AEQ Questionnaire 101

D Smoking Scale Questionnaire 103

E WinBUGS Code: Mixture Randomized Item Response Model 105

F WinBUGS Code: Dichotomous FRR with Gender Effect 107

G WinBUGS Code: Polytomous FRR 109

References 111

Samenvatting 123

Chapter 1

Introduction

In behavioral, health, and social sciences, any endeavor involving measurement is directed at accurate representation of the latent concept with the manifest observation. However, when sensitive topics, such as substance abuse, tax evasion, or felony, are inquired, substantial distortion of reported behaviors, attitudes and opinions might occur due to the self-representational issues. One major concern is the impact of the response distortion on the survey or test results.

Reporting about socially undesirable or disapproved behaviors often involves systematic misreporting. For example, being strongly advised by a pulmonologist to cease smoking, a lung patient that is failing to quit will feel strong incentive to lie about his smoking behavior. Without validation measures, it is not possible to assess the amount of misreporting, and the resulting data can be exceedingly misleading when drawing inferences. In anticipation of response distortion, an alternative method of data collection, assuring confidentiality of individual responses, can lead to more accurate observations. When dealing with sensitive topics so-called randomized response techniques for data collection can provide the necessary degree of response protection.

The models presented in this thesis are meant for multivariate randomized response data analysis. The models are useful for sensitive topic research, where a randomized response data collection method is used to neutralize systematic response bias. A distinction is made between models suited to small and large-scale surveys. First, for small data samples, Bayesian estimation procedures are developed for ordinal count data. Second, for mixed large-scale survey data, Bayesian randomized item response theory models are developed for measuring single and multiple latent respondent characteristics.

1.1 Self-reports and Response Bias in Sensitive Research

Studies of individual behaviors and attitudes involve constructs, which cannot be observed directly. Observable indicators, such as items, have to be constructed, which can accurately represent the unobservable latent concept one is interested

in. Measurement is defined as the process of linking a concept, that is connected to one or more latent variables, with observable measures. A constructed observable, or simply an item, has to be questioned on its accurateness of representation of the concept of interest. Departures from its true value are defined as measurement error.

The process of (self-)reporting on an item comprises cognitive steps ranging from question comprehension to judgement formation and response formulation (Cannell, Miller, & Oksenberg, 1981; Tourangeau, Rips, & Rasinski, 2000). Response variability is always present due to the context dependent variability in judgement. After a response is formulated it is revealed to a researcher.

Self-report data collection using the conventional direct-questioning mode is the most common survey method. Under direct questioning, information is elicited in a direct manner. A respondent is asked to respond to one or more items and the response is recorded. The direct-questioning mode is believed to provide the necessary level of reliability when measuring opinions, attitudes and behaviors. Figure 1.1 presents the life cycle of the direct-questioning process.

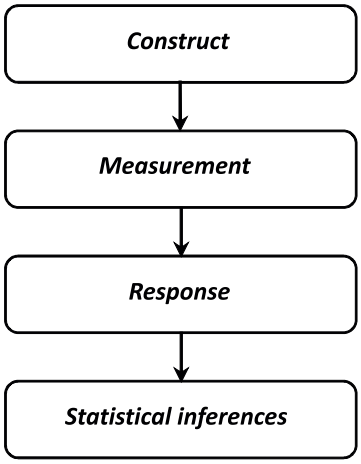

Figure 1.1: Direct questioning survey life cycle.

Obtaining valid and reliable information is a prerequisite for obtaining meaningful results. Standard procedures for making statistical inferences operate on empirical observations, which are assumed to represent accurate outcome values. However, self-reports do largely depend on the level of cooperation and truthful responding of participants, and self-representational concerns may induce response distortion.

People often try to guess the research objective and report accordingly in a socially desirable manner. The self-representational concerns peak, when the survey inquires on a sensitive behavior. This can relate to normatively charged values or embarrassing behaviors. There is a wide variety of potentially sensitive topics such as substance abuse (e.g., drugs, alcohol, and tobacco (e.g., Avetisyan & Fox, 2012; Fox & Wyrick, 2008)), sexual activity (e.g., De Jong, Pieters, & Fox, 2010),

welfare fraud (e.g., van der Heijden, van Gils, Bouts, & Hox, 2000) and tax evasion (e.g., Corstange, 2009; Elffers, Robben, & Hessing, 1992).

When a topic of inquiry relates to a socially sensitive behavior, truthful self-reporting is not likely to be the general norm. Survey researchers have to rely on what individual respondents are willing to disclose. When asking sensitive questions, this can lead to a unit nonresponse, an item nonresponse, or falsification of results. Two kinds of the latter are recognized, namely under- or overreporting. Underreporting often occurs due to the stigmatization of behavior in question, such as drug usage, illegal practices, and alcohol consumption. Overreporting is often the result of an attempt to improve self-representation when for example questioned on usage of seat-belt, voting behavior, or charitable giving (Bradburn, Sudman, & Wansink, 2004).

Undesirable attitudes and stigmatized behaviors are often not only misreported but also misreported in a systematic and unmeasurable way. Obviously, any intentional misrepresentation of the actual behaviors will result in systematically incorrect inferences.

1.2 Methods for Neutralizing Response Bias

As mentioned in the former section, respondents are inclined to supply answers in the direction of the perceived goal of the research. However, the data collection method can influence the cognitive process of (self-)reporting, addressed in Section 1.1. Furthermore, different stages of the data collection process can be adapted to improve the self-reported data.

The question comprehension can be positively affected by a familiar wording of questions. For example, Bradburn and Sudman (1979) discussed significant effects of word choice in research on drinking and sexual experiences. Colloquial expressions in questions on both sensitive topics resulted in increased truthful responding.

Straightforward improvement stimulating honest responding can be achieved through careful selection of the questionnaire format, e.g. question order, wording of questions, and response format. These features can greatly influence the results of a survey based on self-reports. Forgiving wording is sometimes used in formulation of sensitive questions. This aims at alleviation of the normative pressure, which a respondent might experience, by indicating that more people possess the sensitive characteristic. A method of implicit goal priming uses a verbal goal activation principle. It exposes respondents to words related to the goal of achievement in an indirect way. This is done by asking respondents to complete a task containing words like strive, achieve, an succeed, prior to the sensitive question adminstration. Research showed that in this case the respondents have an increased disclosure level of sensitive personal information (Rasinski, Visser, Zagatsky, & Rickett, 2005). Similar methods were described by Bargh, Gollwitzer, Lee–Chai, Barndollar, and Trtschel (2001) and Chartrand and Bargh (1996), among others. Response bias can also be diminished by stressing out the importance of the study, where for example respondents are persuaded to respond truthfully since they provide highly valuable information.

The stage of response revealing can also be adapted. It is, at this stage, that a sensitive nature of a question is taken into account by a respondent. For example, enhancement of truthful responding can be achieved with methods using self-administration with or without use of computers. Another strategy in the data collection process is the bogus pipeline technique. It works on a principle that a respondent is convinced that, regardless of the reported answer, the interviewer will be able to discover the true status of respondent on the variable in question (Jones & Sigall, 1971). Means of convincing ranged from introducing fake polygraph-like devices to taking saliva or breath tests. Respondents will be inclined to respond more honestly to avoid embarrassment of being caught lying. However, due to the element of deceiving, many researchers refrain from using this technique. Discussion of these and other methods to diminish response bias can be found in Sudman, Bradburn, and Schwarz (1996) and Tourangeau et al. (2000).

Most methods will provide more explicit assurances of confidentiality. When confidentiality of individual responses is assured, participants are more willing to reveal their honest responses. In some cases, however, ultimate efforts of a researcher to maintain confidentiality of responses can be destroyed, for example, by an imposed legislative disclosure of research data (Boruch & Cecil, 1979). To further improve the confidentiality of the individual responses, alternative data collection strategies have been developed, which are known as randomized response data collection methods (Fox & Tracy, 1986).

The randomized response technique is a data collection method called to neutralize response bias. Responses are randomly misclassified at the stage of response revealing. In the right-hand track of Figure 1.2, a schematic randomized response

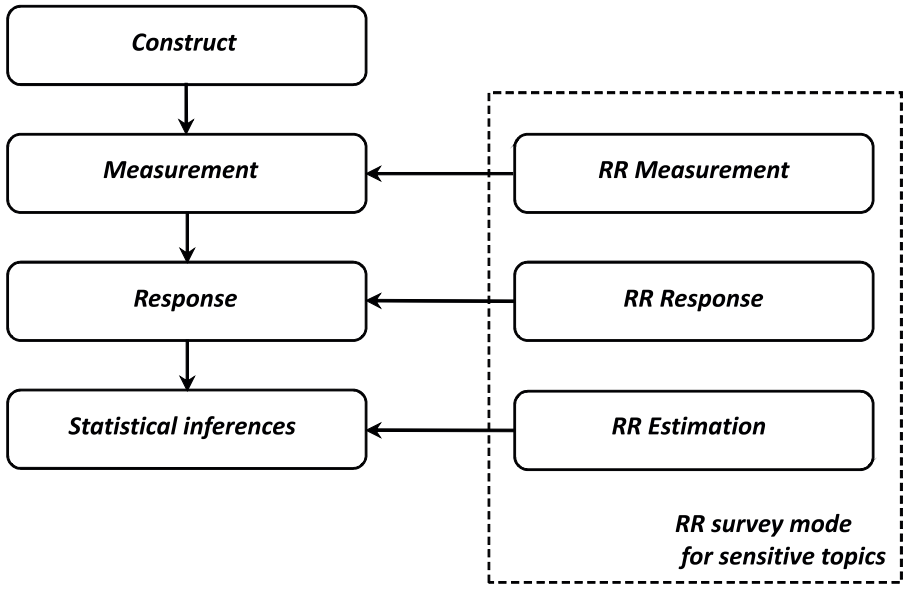

Figure 1.2: Randomized response extension of direct questioning mode for sensitive topics.

extension of the direct questioning mode is presented.

The misclassification due to randomization, is based on a known probability distribution with the purpose to hide individual responses. As a result, individual responses cannot be linked to the identifying information. Free from self-representational concerns respondents are more willing to report truthfully on sensitive behaviors. However, observed data contain randomized responses, which cannot be analyzed with standard statistical procedures. Modified statistical procedures have to relate observed randomized responses to unobserved masked responses taking account of probabilities governing the randomization process at the data collection step. Many authors reported that the randomized response data collection method can have pronounced effects on the statistical inferences. Lensvelt-Mulders, Hox, van der Heijden, and Maas (2005) in their meta-analysis of this topic have shown that RR produces better prevalence estimates than other survey methods.

Originally proposed by Warner (1965), the randomized response technique have been modified in various ways forming a class of randomized response techniques. In the present work, a forced response modification of the RR method is adopted. An extensive overview of RR techniques including nonrandomized response techniques, is given in Chapter 5.

1.3 Randomized Item Response Theory Models

Although initially designed to operate on a single-item, the randomized response techniques can be applied to multi-item scales. Multi-scale measurement instruments, such as tests and questionnaires, are frequently used tools in social and behavioral research. If a scale is designed to measure a sensitive behaviour or attitude, a randomized data collection mode can help to obtain more truthful self-reports. Lack of adequate means to analyse multivariate randomized response data hampered application of RR in a wide range of research areas. Recently Fox (2005b) proposed a class of Bayesian randomized item response theory models. With a comparable purpose, item randomized-response models in a frequentist framework were developed by Böckenholt and van der Heijden (2007).

Item response theory (IRT) is the psychometric theory used for design, analysis, and scoring of tests administered in large-scale study settings. IRT models are used to measure latent constructs such as attitudes, behaviours, and abilities, given manifest responses to test items. The core feature of an IRT model is that item parameters, or item characteristics, and the person parameters, or latent traits, are modeled as disjunct sets of parameters and are placed on the same metric or latent trait continuum. IRT models relate the probability of a manifest item response to person and item parameters. The superiority of IRT models compared to classical test theory models (CTT) is partly expressed in that IRT models are taking account of the differences between items. IRT models are used for major tests such as the Test Of English as a Foreign Language (TOEFL), the Graduate Record Examination (GRE), and the Graduate Management Admission Test (GMAT).

The main assumption of commonly used IRT methods is local independence,

which states that item responses are conditionally independently distributed given the latent trait value. That is, the probability of endorsing an item is strictly determined by the level of individuals' latent trait and not by the responses to other items.

IRT models can be applied to various response formats; that is, dichotomous, polytomous, as well as to scales with mixed response formats. The Rasch model is the most commonly used and is characterized by expressing the success probability of a response by a function of only one item parameter, namely the item difficulty, and a person parameter. The Rasch model can be extended to take account of item discrimination parameters as well as guessing parameters (Embretson & Reise, 2000; Lord & Novick, 1968).

The combination of item response theory and a randomized response data collection procedure can be used to model RR multi-item data. Such a model consists of two modeling stages. First, item response theory is used to model the true unobserved responses. Second, the true responses are linked to observed randomized responses via a randomized response technique, which was used at the data collection step.

To motivate the use of the randomized item response model, the following validation study is considered, which is described in detail in Chapter 3. In this study, a multi-item questionnaire (Appendix D) was used to collect self-report data from lung patients. Patients were randomly assigned to treatment and control groups. Respondents in the control group filled in the questionnaire in a direct questioning manner. In the treatment group, self-report data were obtained using a randomized response data collection mode. The smoking status of each patient was available via a breath test. In Figure 1.3, smokers are represented by filled marks, while empty marks denote non-smokers given the breath test outcomes. Patient's score on the smoking questionnaire determines his position on a vertical, smoking behavior axis. The higher the value of the smoking behavior score, the more likely it is that the patient is a smoker, when making inferences solely from response data.

It can be seen that patients differ in their smoking behavior. Smoking behavior of non-smokers in both groups does not differ much. However, there is a substantial difference between the scores of smokers in the RR group compared to smokers in the DQ group. Patients in the RR group, experiencing the smoking behavior questions as sensitive, were more likely to give an honest response than those in the DQ group. This supports the assumption that RR can improve the quality of self-report data.

Educational and psychological tests are predominantly multidimensional in nature. That is, more than one latent traits are involved in producing the manifest response. The extension of IRT models to multiple latent traits goes under the name of multidimensional item response theory (MIRT, e.g., van der Linden & Hambleton, 1997; Reckase, 2009). Two types of MIRT models can be distinguished, namely the compensatory and noncompensatory models (e.g., Bolt & Lall, 2003). Compensatory multidimensionality assumes that items are characterized by a disjunctive component processes underlying the item response (Maris, 1999). This implies that a deficiency on one trait can be compensated by a proficiency on the other. Noncompensatory models are based on conjunctive component processes;

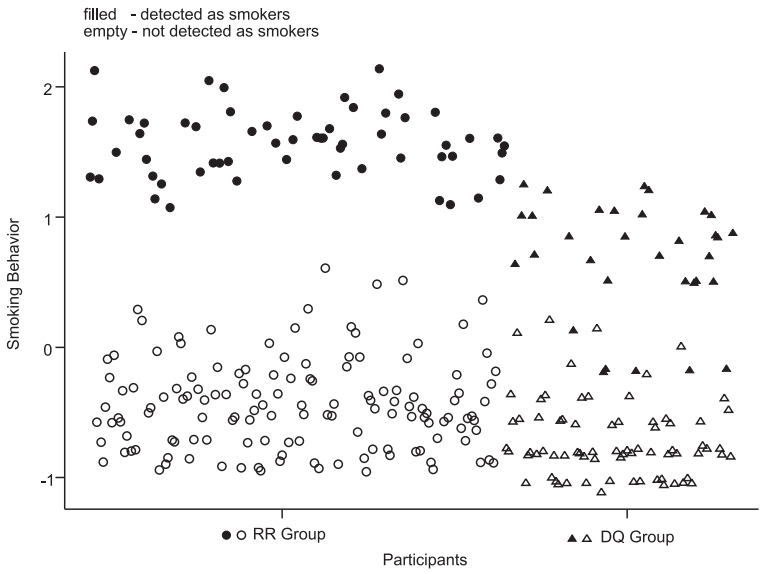

Figure 1.3: Scores of smokers and non-smokers in the DQ and RR group.

that is, when a deficiency on one trait can influence performance on the other.

In this thesis, the compensatory MIRT modeling framework is extended to model the relation between questionnaire scores and dichotomous and polytomous randomized item scores. A two-parameter variant of the multidimensional randomized IRT model (MRIRT) is developed to describe the probability of correct responses in a dichotomous response format. The probability that a respondent scores in a certain category is modeled by the graded response model of Samejima (1969).

1.4 Bayesian Approach to IRT Modeling

The acceptance of the Bayesian methodology was hampered by intractabilities involved in the calculation of posterior distributions. With the introduction of Markov Chain Monte Carlo (MCMC) methods (Gelfand & Smith, 1990; Gelfand, Hills, Racine-Poon, & Smith, 1990; Gelfand & Smith, 1984) Bayesian statistics became feasible in practice.

The Bayesian framework has several advantages. All unknown parameters are defined as random parameters. Each parameter gets a prior distribution, which makes it possible to reflect initial knowledge available. After the data have been observed, beliefs concerning parameters are modified and comprise data and prior information leading to posterior beliefs. Bayesian theory provides a straightforward mechanism of prior knowledge updating. When new data are available, the updated, or posterior, knowledge can be used as prior input in subsequent analysis. Bayesian models are flexible. For instance, it is rather simple to extend Bayesian

models with explanatory information on person parameters. Furthermore, MCMC estimation methods, used for computations involving high-dimensional integration, remain straightforward as model complexity increases.

The Bayesian version of IRT models are discussed by Albert (1992), Junker (2001), Patz and Junker (1999a, 1999b), among others. It is argued, that IRT models can get very complex, depending on the situation to which they are applied. In that case, a large number of parameters have to be estimated. However, MCMC implementations of Bayesian IRT model parameter are often defined in a straightforward way.

WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000) is the free statistical software for Bayesian analysis using Markov Chain Monte Carlo methods. It can be used for full Bayesian estimation of relatively complex problems. Throughout this thesis, a number of models were fitted using WinBUGS. The corresponding code is given in listings, which are presented in appendices.

1.5 Outline

In this thesis, Bayesian models for sensitive measurements, where observations are collected using randomized response methods, are extended in different ways. With respect to sample size, two approaches can be recognized. For small samples, when it is not possible to generate stable estimates for IRT-based models, a Dirichlet-multinomial RR modeling approach is proposed for ordinal data. It is shown that this modeling framework is a generalization of the beta-binomial RR modeling approach for binary data. For large scale data, the randomized IRT model is presented and validated in a unique experimental study. With respect to the dimension of the sensitive measurements, the class of Bayesian randomized IRT models is extended to support the measurement of unidimensional and multidimensional constructs in a compensatory and non-compensatory ways.

An overview of the chapters will be given. In Chapter 2, a Dirichlet-multinomial model for categorical multivariate RR data is proposed. Individual unobserved categorical-response rates are estimated in a straightforward way using a linear transformation with categorical-response rates based on observed RR data. The empirical Bayes estimates are compared to the full Bayes estimates. The full Bayes procedure is implemented in WinBUGS using the collapsing property of the Dirichlet distribution by expressing the Dirichlet as series of beta-binomial distributions. The model is extended to a constrained-Dirichlet-multinomial such that the homogeneity of category-response rates across individuals, or groups of individuals, can be explicitly tested. In the second part of this chapter, a simulation study is presented that shows the recovery of simulated parameters for the full Bayes method, as well as the sensitivity of the parameter estimates to various design conditions. The influence of prior settings is assessed by comparing MSEs. The College Alcohol Problem Scale (CAPS) is used to illustrate the performance of the full Bayes model, which includes an investigation of the group-specific population proportions.

Chapter 3 presents a validation study of the randomized item response technique, where a multi-item measure is developed to assess smoking behavior. A

clinical breath test is also used to determine the smoking status of each patient. Individual smoking behavior is assessed using the mixture randomized item response model given binary and ordinal item response data. Data are collected using the randomized response and the direct questioning technique. For each questioning mode, the outcome of the clinical diagnosis using the breath test is compared to the latent smoking behavior estimate using the multi-item response data. It is shown that the randomized response test data are more accurate than the direct questioning data when comparing the outcomes with the breath test outcomes. Further, a Bayesian latent variable method for diagnostic test accuracy is proposed, which supports the Bayesian diagnostic evaluation of the proposed multi-item smoking behavior test. It allows computation of posterior classification probabilities such as the true positive fraction (sensitivity) and the true negative fraction (specificity). The quantities are used to assess the diagnostic accuracy of the smoking behavior questionnaire. The randomized response technique is further validated using the positive and negative predictive values of the test.

In Chapter 4, a multidimensional randomized item response theory model (MRIRT) is proposed to measure multiple sensitive factors underlying multi-item randomized responses. The MRIRT modeling structure comprehends three stages. First, randomized response scale data are related to individual response probabilities. Second, the response process is described by a multidimensional IRT model. Third, latent sensitive characteristics are considered to be outcomes of a multivariate regression model. An MCMC algorithm with a double data augmentation step is developed for simultaneous estimation of all model parameters. After a simulation study for parameter recovery, an MRIRT analysis of data from the College Alcohol Problem Scale (CAPS), measuring alcohol-related socio-emotional and community problems, and the Alcohol Expectancy Questionnaire (AEQ), measuring alcohol-related sexual enhancement expectancies is presented.

A comprehensive review of RR techniques is presented in Chapter 5, where the distinction is made between traditional and recently developed techniques. Traditional RR methods are described together with a reasoning behind extensions. Various types of randomized response data collection strategies are discussed in detail. Different parameter estimation approaches are presented. More recent, nonrandomized response techniques are summarized and compared to standard procedures. The issue of level of inferences given randomized response data for different measurement formats is addressed. This includes individual-level inferences when multiple individual observations are available, and population-level inferences when dealing with single-item measurements. A few randomized response techniques are illustrated with examples, where parameter estimates are obtained using maximum likelihood and full Bayesian estimation methods.

Chapter 2

The Dirichlet-Multinomial Model for Multivariate Randomized Response Data and Small Samples

Abstract

In survey sampling the randomized response (RR) technique can be used to obtain truthful answers to sensitive questions. Although the individual answers are masked due to the RR technique, individual (sensitive) response rates can be estimated when observing multivariate response data. The beta-binomial model for binary RR data will be generalized to handle multivariate categorical RR data. The Dirichlet-multinomial model for categorical RR data is extended with a linear transformation of the masked individual categorical-response rates to correct for the RR design and to retrieve the sensitive categorical-response rates even for small data samples. This specification of the Dirichlet-multinomial model enables a straightforward empirical Bayes estimation of the model parameters. A constrained-Dirichlet prior will be introduced to identify homogeneity restrictions in response rates across persons and categories. The performance of the full Bayes parameter estimation method is verified using simulated data. The proposed model will be applied to the college alcohol problem scale study, where students were interviewed directly or interviewed via the randomized response technique about negative consequences from drinking.

Key words: randomized response data, beta-binomial distribution, Dirichlet-multinomial, constrained-Dirichlet prior, sensitive-item survey, small data sample

2.1 Introduction

The data collection through surveys based on direct-questioning methods has been the most common way. The direct-questioning techniques are usually assumed to provide the necessary level of reliability when measuring opinions, attitudes, and behaviors. However, individuals with different types of response behavior who are confronted with items about sensitive issues of human life regarding ethical (stigmatizing) and legal (prosecution) implications are reluctant to supply truthful answers. Tourangeau et al. (2000), and Tourangeau and Yan (2007) argued that socially desirable answers and refusals are to be expected when asking sensitive questions directly.

Warner (1965), and Greenberg, Abul-Ela, Simmons, and Horvitz (1969) developed RR techniques to obtain truthful answers to sensitive questions in such a way that the individual answers are protected but population characteristics can be estimated. These techniques are based on univariate RR data. Recently, RR models have been developed to analyze multivariate response data, where the item responses are nested within the individual. Although the individual answers are masked due to the RR technique, individual (sensitive) characteristics can be estimated when observing multivariate RR data. Fox (2005b) and Böckenholt and van der Heijden (2007) introduced item response models for binary RR data. The applications are focusing on surveys where the items measure an underlying sensitive construct. The so-called randomized item response models have been extended to handle categorical RR data by Fox and Wyrick (2008) and De Jong et al. (2010).

The class of randomized item response models are meant for large-scale survey data, since person as well as item parameters need to be estimated (Fox & Wyrick, 2008). For categorical item response data, more than 500 respondents are often needed to obtain stable parameter estimates. Furthermore, the randomized item response data are less informative than the direct-questioning data, since the RR technique engenders additional random noise to the data. Fox (2010) proposed a beta-binomial model for analyzing multivariate binary RR data, which enables the computation of individual response estimates without requiring a large-scale data set. The beta-binomial model has several advantages like a simple interpretation of the model parameters, stable parameter estimates for relatively small data sets, and a straightforward empirical Bayes estimation method.

Here, a Dirichlet-multinomial model is proposed for handling multivariate categorical RR data such that individual category-response rates can be estimated. The individual observed RR data consist of a number of randomized responses per category. Each individual set of observed numbers are assumed to be multinomially distributed given the individual category-response rates. The individual category-response rates are assumed to follow a Dirichlet distribution. The individual response rates are related to the observed randomized responses, which make them not useful for the inferences basing on regular statistical approaches. However, it will be shown that the individual category-response rates are linearly related to the model-based (true) category-response rates. The latter one relates to the latent responses, which are expected under the model when the responses are not masked due to the randomized response technique. The parameters of

the linear transformation are design parameters and are known characteristics of the randomizing device that is used to mask the individual answers. The transformed categorical-response rates will provide information about the latent individual characteristic that is measured by the survey items. Analytical expressions of the posterior mean and standard deviation of the true individual categorical-response rates will be given. The expressions can be used for estimation given prior knowledge or empirical Bayes estimates of the population response rates. Furthermore, a WinBUGS implementation is given for a full Bayes estimation of the model parameters.

To model and to identify constraints of homogeneity in category-response rates, the restricted-Dirichlet prior (Schafer, 1997) is used. The restriction on the Dirichlet prior can be used to identify effects of the randomized response mechanism across individuals, groups of individuals, and response categories.

In the next section, the randomized response technique is described in a multiple-item setting. The beta-binomial model is described for multivariate binary RR outcomes. Then, as a generalization, the Dirichlet-multinomial model is presented for multivariate categorical RR data. Properties of the conditional posterior distribution of the true individual categorical-response rates are derived given observed randomized response data. Then, empirical and full Bayes methods are proposed to estimate all model parameters. A simulation study is given, where the properties of the estimation methods are examined. Finally, the model will be used to analyze data from a college alcohol problem scale survey, where U.S. college students were asked about their alcohol drinking behavior with and without using the randomized response technique. The restricted-Dirichlet prior will be used to test assumptions of homogeneity over persons and response categories. In particular, it will be shown that the effect of the RR method varies over response categories, where the RR effect will be the highest for the most sensitive response option.

2.2 Multivariate Randomized Response Techniques

In Warner's RR technique (Warner, 1965) for univariate binary response data, in the data collection procedure a randomizing device (RD) is introduced. For each respondent the RD directs the choice of one of two logically opposite questions. This sampling design guarantees the confidentiality of the individual answers, since they cannot be related directly to one of the opposite questions.

Greenberg et al. (1969) proposed the unrelated question technique, where the outcome of the RD refers to the study-related sensitive question or an irrelevant unrelated question. The RD is specified in such a way that the sensitive question is selected with probability ϕ_1 and the unrelated question with probability $1 - \phi_1$. This RR method is extended to a forced response method (Edgell, Himmelfarb, & Duchan, 1982), where the unrelated question is not specified but an additional RD is used to generate a forced answer. Each observed individual answer is protected, since it cannot be retrieved whether it is a true answer to the sensitive question or a forced answer generated by the RD. As a result, the observed RR answers are polluted by forced responses.

Let $RD = 1$ denote the event that an answer to the sensitive question is re-

quired and $P(RD = 1) = \phi_{1k}$ and $RD = 0$ otherwise. A forced positive response to item k is generated with probability ϕ_{2k}. For a multiple-item survey, the probability of a positive RR of respondent i, given a forced response sampling design, can be stated as

$$P\left(Y_{ik} = 1 \mid \boldsymbol{\phi}, p_{ik}\right) = P(RD = 1)p_{ik} + (1 - P(RD = 1))\phi_{2k}, \tag{2.1}$$

where the true response rate of person i to item k is denoted as p_{ik}. Note that the response model for the RR data is a two-component mixture model. For the first component the sensitive question needs to be answered and for the second component a forced response needs to be generated. Thus, the randomized response probability equals the true or the forced response probability depending on the RD outcome. With $\phi_{1k} > 1/2$, for all k, the data contain sufficient information to make inferences about the true response rates.

The multiple items will be assumed to measure an underlying individual response rate (e.g., alcohol dependence, academic fraud) such that $p_{ik} = p_i$ for all k. This individual response rate can be estimated from the multivariate RR data. Note that in a multivariate setting the RD characteristics are allowed to vary over items such that the proportion of forced responses can vary over items. In practice, the sensitivity of the items may vary although they relate to the same sensitive latent characteristic. This variation in sensitivity can be controlled by adjusting the RD characteristics, which are under the control of the interviewer.

The forced response model in Equation 2.1 can be extended to handle polytomous multivariate RR data. Let $\phi_{2k}(c)$ denote the probability of a forced response in category c for $c = 1, \ldots, C_k$ such that the number of response categories may vary over items. The categorical-response rates of individual i are denoted as $p_i(1), \ldots, p_i(C_k)$, which represent the probabilities of honest (true) responses corresponding to the response categories of item k. The probability of an observed randomized response of individual i in category c of item k can be stated as,

$$P\left(Y_{ik} = c \mid \boldsymbol{\phi}, p_{ik}\right) = \phi_{1k}p_i(c) + (1 - \phi_{1k})\phi_{2k}(c). \tag{2.2}$$

This forced RR model for categorical data can be used to measure individual categorical response rates related to a sensitive characteristic. The individual answers are not known but the multivariate data make it possible to retrieve information about latent individual characteristics.

2.3 The Beta-Binomial Model for Multivariate Binary RR data

Let each participant $i = 1, \ldots, N$ respond to $k = 1, \ldots, K$ binary items. The observations $u_{i1}, \ldots, u_{iK}$ represent the answers of the i^{th} participant to the K items. The response observations are assumed to be Bernoulli distributed given response rate p_i for individual i. The observations are assumed to be independently distributed given the response rate. Therefore, the sum of individual response observations is binomially distributed with parameters K and p_i.

It is to be expected that the response rates vary over participants. This variation is modeled by means of a beta distribution with parameters $\tilde{\alpha}$ and $\tilde{\beta}$, which

specify the distribution of the response rates. This leads to the following hierarchical model for the multivariate binary response observations,

$$\begin{aligned} U_{i.} \mid p_i &\sim \mathcal{BIN}(K, p_i), \\ p_i \mid \tilde{\alpha}, \tilde{\beta} &\sim \mathcal{B}(\tilde{\alpha}, \tilde{\beta}), \end{aligned}$$

where $U_{i.} = \sum_k U_{ik}$.

Within a Bayesian modeling approach, the beta prior distribution for parameter p_i is a conjugated prior when the data are binomially distributed given the response rate. In that case, the posterior distribution of the response rate is also a beta distribution. That is,

$$\begin{aligned} p\left(p_i \mid u_{i\cdot}, \tilde{\alpha}, \tilde{\beta}\right) &= \frac{f(u_{i\cdot} \mid p_i)\pi(p_i \mid \tilde{\alpha}, \tilde{\beta})}{\int f(u_{i\cdot} \mid p_i)\pi(p_i \mid \tilde{\alpha}, \tilde{\beta})dp_i} \\ &= \frac{\Gamma(K + \tilde{\alpha} + \tilde{\beta})}{\Gamma(u_{i\cdot} + \tilde{\alpha})\Gamma(K - u_{i\cdot} + \tilde{\beta})} p_i^{u_{i\cdot}+\tilde{\alpha}-1}(1 - p_i)^{K-u_{i\cdot}+\tilde{\beta}-1}, \end{aligned}$$

which can be recognized as a beta density with parameters $u_{i\cdot} + \tilde{\alpha}$ and $K - u_{i\cdot} + \tilde{\beta}$. The posterior mean and the variance are

$$E(p_i \mid u_{i\cdot}, \tilde{\alpha}, \tilde{\beta}) = \frac{u_{i\cdot} + \tilde{\alpha}}{K + \tilde{\alpha} + \tilde{\beta}},$$

$$Var(p_i \mid u_{i\cdot}, \tilde{\alpha}, \tilde{\beta}) = \frac{(u_{i\cdot} + \tilde{\alpha})(K - u_{i\cdot} + \tilde{\beta})}{(K + \tilde{\alpha} + \tilde{\beta} + 1)(K + \tilde{\alpha} + \tilde{\beta})^2},$$

respectively. It follows that posterior inferences can be directly made when knowing the population parameters $\tilde{\alpha}$ and $\tilde{\beta}$.

In a forced response design, the observations $\mathbf{u}$ are masked and randomized responses $\mathbf{y}$ are observed. The RD specifies the probabilities governing this randomization process such that an honest response is to be given with probability ϕ_1 and a positive forced response with probability $(1 - \phi_1)\phi_2$. The probability of observing a positive response from participant i to item k is related to the true response by the following expression:

$$\begin{aligned} P\left(Y_{ik} = 1 \mid p_i\right) &= \phi_1 f(u_{ik} \mid p_i) + (1 - \phi_1)\phi_2 \\ &= \phi_1 p_i + (1 - \phi_1)\phi_2 = \Delta(p_i). \end{aligned}$$

It can be seen that the forced response design corresponds with a linear transformation of the response rate. This linear transformation function, $\Delta(.)$, operates on the individual response rate of the true responses. Therefore, the beta-binomial model accommodates the forced response sampling mechanism by modeling the linearly transformed response rates; that is,

$$\begin{aligned} Y_{i.} \mid p_i &\sim \mathcal{BIN}\left(K, \Delta(p_i)\right), \\ \Delta(p_i) &\sim \mathcal{B}(\alpha, \beta), \end{aligned}$$

where the transformation parameters ϕ_1 and ϕ_2 are characteristics of the RD and are known a priori.

A population distribution is specified for the transformed response rates. The transformed response rates are a priori beta distributed, which is the conjugated prior for the binomially distributed likelihood. As a result, the posterior distribution of the transformed response rates is beta distributed with parameters $y_{i\cdot} + \alpha$ and $K - y_{i\cdot} + \beta$.

The posterior expected response rate given the randomized responses can be expressed as

$$\begin{aligned} E\left(\Delta(p_i) \mid y_{i\cdot}, \alpha, \beta\right) = \frac{y_{i\cdot} + \alpha}{K + \alpha + \beta} &= \Delta\left(E(p_i) \mid y_{i\cdot}, \alpha, \beta\right) \\ &= \phi_1 E\left(p_i \mid y_{i\cdot}, \alpha, \beta\right) + (1 - \phi_1)\phi_2, \end{aligned}$$

using that the expected value of the linearly transformed response rate equals the linearly transformed expected response rate. As a result, the posterior expected value of the (true) response rate can be expressed as

$$E\left(p_i \mid y_{i\cdot}, \alpha, \beta\right) = \phi_1^{-1}\left(\frac{y_{i\cdot} + \alpha}{K + \alpha + \beta}\right) + (1 - \phi_1^{-1})\phi_2. \tag{2.3}$$

In the same way, an expression can be found for the posterior variance of the true response rate,

$$Var\left(p_i \mid y_{i\cdot}, \alpha, \beta\right) = \frac{(y_{i\cdot} + \alpha)(K - y_{i\cdot} + \beta)}{\phi_1^2(K + \alpha + \beta + 1)(K + \alpha + \beta)^2}.$$

There are two straightforward methods for estimating the hyperparameters α and β. The method of moments and the method of maximizing the marginal likelihood. Given the estimated hyperparameters, empirical Bayes estimates of the response rates can be derived by inserting the hyperparameter estimates into Equation 2.3. Furthermore, the estimation of confidence intervals and Bayes factors is described in Fox (2008).

2.4 The Dirichlet-Multinomial Model for Multivariate Categorical RR Data

The number of responses per response category over items for person i are stored in a vector $\mathbf{u}_{i\cdot} = (u_{i\cdot 1}, \ldots, u_{i\cdot C})^t$, where $u_{i\cdot c} = \sum_k u_{ikc}$ for $c = 1, \ldots, C$. They represent the number of choices per response category over items. In the college alcohol study we will present in Section 2.7.2, the data represent the frequency of alcohol-related negative consequences. In marketing research, Goodhardtn, Ehrenberg, and Chatfield (1984) considered data about individual number of purchases per brand in a time period. In social research, Wilson and Chen (2007) considered frequencies to television viewing questions from the High School and Beyond survey study in the United States. Their item-based test is assumed to measure the daily television viewing habit and interest is focused on time-specific population response rates.

The number of responses per category given the category response rates are assumed to be independently distributed. They can be modeled by a multinomial

distribution with parameters K and category response rates $p_{i1}, \ldots, p_{iC}$. For respondent i, the contribution to the likelihood is

$$f(\mathbf{u}_{i\cdot} \mid \mathbf{p}_i) = \frac{K!}{\prod_c u_{i\cdot c}!} \prod_c p_{ic}^{u_{i\cdot c}}.$$

The variability in the vectors of response counts is often higher than can be accommodated by the multinomial distribution. Therefore, individual variation in category response rates is modeled by a Dirichlet distribution with parameters $\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_C)$, which is represented by

$$\pi(\mathbf{p}_i \mid \tilde{\boldsymbol{\alpha}}) = \frac{\Gamma(\tilde{\alpha}_0)}{\prod_c \Gamma(\tilde{\alpha}_c)} \prod_c p_{ic}^{\tilde{\alpha}_c - 1}.$$

where $\tilde{\alpha}_0 = \sum_c \tilde{\alpha}_c$. The within-individual and between-individual variability in response rates is described by a Dirichlet-multinomial model; that is,

$$\begin{array}{rcl} U_{i\cdot 1}, \ldots, U_{i\cdot C} \mid p_{i1}, \ldots, p_{iC} & \sim & \mathit{Mult}(K, p_{i1}, \ldots, p_{iC}), \\ p_{i1}, \ldots, p_{iC} & \sim & \mathcal{D}(\tilde{\alpha}_1, \ldots, \tilde{\alpha}_C), \end{array}$$

where $U_{i\cdot c} = \sum_k U_{ikc}$ for $c = 1, \ldots, C$. The compact form of this expression can be written in terms of vector notation

$$\begin{array}{rcl} \mathbf{U}_{i\cdot} \mid \mathbf{p}_i & \sim & \mathit{Mult}\left(K, \mathbf{p}_i\right), \\ \mathbf{p}_i & \sim & \mathcal{D}(\tilde{\boldsymbol{\alpha}}), \end{array}$$

where $\mathbf{U}_{i\cdot} = (U_{i\cdot 1}, \ldots, U_{i\cdot C})^t$.

The Dirichlet distribution is a conjugate prior for the parameters of the multinomially distributed responses. Therefore, the conditional posterior distribution of the category response rates is a Dirichlet distribution, which is represented by

$$\begin{array}{rcl} p\left(\mathbf{p}_i \mid \mathbf{u}_{i\cdot}, \tilde{\boldsymbol{\alpha}}\right) & = & \dfrac{f(\mathbf{u}_{i\cdot} \mid \mathbf{p}_i)\pi(\mathbf{p}_i \mid \tilde{\boldsymbol{\alpha}})}{\int f(\mathbf{u}_{i\cdot} \mid \mathbf{p}_i)\pi(\mathbf{p}_i \mid \tilde{\boldsymbol{\alpha}}) d\mathbf{p}_i} \\ & = & \dfrac{\Gamma(K + \tilde{\alpha}_0)}{\prod_c \Gamma(u_{i\cdot c} + \tilde{\alpha}_c)} \prod_c p_{ic}^{u_{i\cdot c} + \tilde{\alpha}_c - 1}. \end{array}$$

The posterior mean and the variance of the category response rates of individual i equals

$$E(p_{ic} \mid \mathbf{u}_{i\cdot}, \tilde{\boldsymbol{\alpha}}) = \frac{u_{i\cdot c} + \tilde{\alpha}_c}{K + \tilde{\alpha}_0}$$

and

$$Var(p_{ic} \mid \mathbf{u}_{i\cdot}, \tilde{\boldsymbol{\alpha}}) = \frac{(u_{i\cdot c} + \tilde{\alpha}_c)(K + \tilde{\alpha}_0 - (u_{i\cdot c} + \tilde{\alpha}_c))}{(K + \tilde{\alpha}_0 + 1)(K + \tilde{\alpha}_0)^2},$$

respectively, where the prior parameters $\tilde{\boldsymbol{\alpha}}$ are unknown.

According to Equation 2.2, the probability of an observed randomized response in category c for item k can be expressed as

$$\begin{array}{rcl} P\left(Y_{ik} = c \mid p_{ic}\right) & = & \phi_1 p_{ic} + (1 - \phi_1)\phi_2(c) \\ & = & \Delta(p_{ic}), \end{array}$$

where $\Delta(p_{ic})$ is the linearly transformed category-response rate of person i, which depends on the parameters of the forced randomized response design. Let $\mathbf{y}_{i.} = (y_{i.1}, \ldots, y_{i.C})^t$ denote the vector of observed randomized count data per response category across items for subject i. The Dirichlet-multinomial model for the observed randomized count data per category takes the form

$$\begin{array}{rcl} \mathbf{Y}_{i.} \mid \mathbf{p}_i & \sim & \mathit{Mult}(K, \Delta(\mathbf{p}_i)), \\ \Delta(\mathbf{p}_i) & \sim & \mathcal{D}(\boldsymbol{\alpha}), \end{array} \tag{2.4}$$

where $\mathbf{Y}_{i.} = (Y_{i.1}, \ldots, Y_{i.C})^t$ and $\Delta(\mathbf{p}_i) = (\Delta(p_{i1}), \ldots, \Delta(p_{iC}))^t$.

The conditional posterior distribution of the transformed category-response rate can now be stated as

$$p\left(\Delta(\mathbf{p}_i) \mid \mathbf{y}_{i.}, \boldsymbol{\alpha}\right) = \frac{\Gamma(K+\alpha_0)}{\prod_c \Gamma(y_{i.c}+\alpha_c)} \prod_c (\Delta(p_{ic}))^{y_{i.c}+\alpha_c-1}.$$

Subsequently, the posterior expected (true) category-response rate can be obtained through a linear transformation. That is,

$$\begin{array}{rcll} E\left(\Delta(p_{ic}) \mid \mathbf{y}_{i.}, \boldsymbol{\alpha}\right) = \dfrac{y_{i.c}+\alpha_c}{K+\alpha_0} & = & \Delta\left(E\left(p_{ic}\right) \mid \mathbf{y}_{i.}, \boldsymbol{\alpha}\right) & (2.5) \\ & = & \phi_1 E\left(p_{ic} \mid \mathbf{y}_{i.}, \boldsymbol{\alpha}\right) + (1-\phi_1)\phi_2(c). & \end{array}$$

Applying the inverse of the linear transformation on $E(\Delta(p_{ic}) \mid \mathbf{y}_{i.}, \boldsymbol{\alpha})$, the conditional posterior expected value can be obtained as

$$E(p_{ic} \mid \mathbf{y}_{i.}, \boldsymbol{\alpha}) = \phi_1^{-1}\left(\frac{y_{i.c}+\alpha_c}{K+\alpha_0}\right) + (1-\phi_1^{-1})\phi_2(c).$$

The expression for the conditional posterior variance can be derived in a similar way and is equal to

$$Var\left(p_{ic} \mid \mathbf{y}_{i.}, \boldsymbol{\alpha}\right) = \frac{\left(y_{i.c}+\alpha_c\right)\left(K+\alpha_0-\left(y_{i.c}+\alpha_c\right)\right)}{\phi_1^2\left(K+\alpha_0+1\right)\left(K+\alpha_0\right)^2}.$$

2.5 Empirical Bayes and Full Bayes Estimation

There are two major approaches for estimating the model parameters when the prior parameters are unknown. An empirical Bayes approach, where the prior parameters are estimated from the marginal likelihood of the data and a full Bayes approach where hyperpriors are defined for the prior parameters and all model parameters are simultaneously estimated.

2.5.1 Empirical Bayes Estimation

The marginal distribution of the data given the prior parameters is obtained by integrating out the category-response rates. In Appendix A, a derivation is given of

the marginal likelihood of the randomized response data given the prior parameters $\boldsymbol{\alpha}$. This conditional distribution is given by

$$\begin{aligned} p(\mathbf{y} \mid \boldsymbol{\alpha}) &= \prod_i \int_{\Delta(\mathbf{p}_i)} p(\mathbf{y}_{i\cdot} \mid \Delta(\mathbf{p}_i))\, p(\Delta(\mathbf{p}_i) \mid \boldsymbol{\alpha})\, d(\Delta(\mathbf{p}_i)) \\ &= \prod_i \frac{K!}{\prod_c y_{i\cdot c}!} \frac{\Gamma(\alpha_0)}{\prod_c \Gamma(\alpha_c)} \frac{\prod_c \Gamma(\alpha_c + y_{i\cdot c})}{\Gamma(\alpha_0 + K)}. \end{aligned}$$

There are two ways of obtaining empirical Bayes estimates from this marginal likelihood. The most straightforward way is using the method of moments (Brier, 1980; Danaher, 1988; Mosimann, 1962). The second way is the method of marginal maximum likelihood (Paul, Balasooriya, & Banerjee, 2005).

Method of Moments

Let the sum of the prior parameters be α_0 and the fraction $\frac{\alpha_c}{\alpha_0}$ for each c be greater than zero. Now, the observed proportion of category responses is used to estimate the fraction $\frac{\alpha_c}{\alpha_0}$; that is,

$$N^{-1} \sum_{i=1}^{N} y_{i\cdot c}/K = \widehat{\frac{\alpha_c}{\alpha_0}},$$

for $c = 1, \ldots, C$. The sum of the prior parameters α_0 is estimated using a relationship between the covariance matrix of the observed data, denoted as $\boldsymbol{\Sigma}_y$ of dimension $(C-1)(C-1)$, and of the category response rates, denoted as $\boldsymbol{\Sigma}_{\Delta(p)}$ of dimension $(C-1)(C-1)$. Mosimann (1962) showed that

$$(1 + \alpha_0)\boldsymbol{\Sigma}_y = (K + \alpha_0)\boldsymbol{\Sigma}_{\Delta(p)}. \tag{2.6}$$

The observed data can be used to estimate the covariance matrices; that is,

$$\widehat{\boldsymbol{\Sigma}}_y = \begin{cases} (N-1)^{-1} \sum_{i=1}^{N} (y_{i\cdot c} - \overline{y}_{\cdot\cdot c})^2 & \text{diagonal terms,} \\ (N-1)^{-1} \sum_{i=1}^{N} (y_{i\cdot c} - \overline{y}_{\cdot\cdot c})(y_{i\cdot c'} - \overline{y}_{\cdot\cdot c'}) & \text{off-diagonal terms, } c \neq c' \end{cases}$$

and

$$\widehat{\boldsymbol{\Sigma}}_{\Delta(p)} = \begin{cases} \overline{y}_{\cdot\cdot c}(K - \overline{y}_{\cdot\cdot c})/K & \text{diagonal terms,} \\ -\overline{y}_{\cdot\cdot c}\overline{y}_{\cdot\cdot c'}/K & \text{off-diagonal terms, } c \neq c', \end{cases}$$

where $\overline{y}_{\cdot\cdot c} = \sum_i y_{i\cdot c}/N$. The relationship in Equation 2.6 can be transformed to specify a relationship between the determinants of both covariance matrices, which can be used to estimate the α_0. In this way, the estimate $\hat{\alpha}_0$ can be obtained from

$$\left(\frac{\left| \widehat{\boldsymbol{\Sigma}}_y \right|}{\left| \widehat{\boldsymbol{\Sigma}}_{\Delta(p)} \right|} \right)^{1/(C-1)} = \frac{K + \hat{\alpha}_0}{1 + \hat{\alpha}_0}.$$

Method of Marginal Maximum Likelihood

The Dirichlet prior parameters can also be estimated from the marginal likelihood given the observed randomized response data. The so-called marginal maximum likelihood estimates are the values for the parameters that maximize the marginal (log-)likelihood function. To facilitate the computation of marginal maximum likelihood estimates, an analytical expression is required of the marginal log-likelihood of the Dirichlet parameters given the randomized response data. The derivation of this marginal log-likelihood function is given in Appendix A. The terms not including any parameters can be ignored, which leads to the following expression

$$l(\boldsymbol{\alpha} \mid \mathbf{y}) \propto \sum_{i=1}^{N} \left[\sum_{j=0}^{y_{i.1}-1} log\left(\alpha_1 + j\right) + \ldots + \sum_{j=0}^{y_{i.C}-1} log\left(\alpha_C + j\right) - \right. \\ \left. - \sum_{j=0}^{K-1} log\left(\alpha_0 + j\right) \right]. \tag{2.7}$$

The marginal maximum likelihood estimates can be obtained using the Newton-Raphson algorithm. Convergence problems of the latter are often associated with the parameter initialization step. Dishon and Weiss (1980) suggested using moment estimates as initial parameter values for the Newton-Raphson procedure.

2.5.2 Full Bayes Estimation

The model in Equation 2.4, can be extended with a hyperprior for the prior parameters. Then, the model consists of three levels, where level 1 defines the distribution of the randomized response data, level 2 the prior distribution for the level-1 parameters, and level 3 the distribution of the prior parameters. In such an hierarchical modeling approach, uncertainties are defined at different hierarchical levels. In the empirical Bayes estimation approach, the prior parameters are estimated using only the observed data, but in a full Bayes estimation approach the (hyper) prior information as well as the data are used.

In a full Bayes estimation approach all defined uncertainties can be taken into account. Therefore, a Markov Chain Monte Carlo (MCMC) method will be used to estimate the posterior densities of all model parameters, which includes the transformed category response rates and the population parameters $\boldsymbol{\alpha}$.

To implement an MCMC procedure the collapsing property of the multinomial and Dirichlet distribution can be used. Assume that for each respondent the cells $2, \ldots, C$ are collapsed and that in total two cells are observed with $y^*_{i.2} = y_{i.2} + \ldots + y_{i.C}$. The distribution of the collapsed data are binomially distributed given the category response rate; that is,

$$p\left(y_{i.1}, y^*_{i.2} \mid \Delta(\mathbf{p}_i)\right) \propto \left(\Delta(p_{i1})\right)^{y_{i.1}} \left(1 - \Delta(p_{i1})\right)^{y^*_{i.2}}. \tag{2.8}$$

In the same way, the collapsing property of the Dirichlet distribution can be used. The collapsed Dirichlet prior for the transformed category response rate, $\Delta(p_{i1})$, is a beta distribution with parameters α_1 and $\alpha_0 - \alpha_1$, which leads to a beta-binomial model for the first transformed category response rate.

This procedure can also be applied to the second response category. Let $y^*_{i.3} = y_{i.3} + \ldots + y_{i.C}$ denote the collapsed data. The observed data of respondent i in category two are binomially distributed, where the responses to category one are excluded. Therefore, consider $\Delta(p_{i2})/(1 - \Delta(p_{i1}))$ as the correctly scaled success probability such that the collapsed randomized response data are binomially distributed,

$$p\left(y_{i.2}, y^*_{i.3} \mid \Delta(\mathbf{p}_i)\right) \propto \left(\frac{\Delta(p_{i2})}{1 - \Delta(p_{i1})}\right)^{y_{i.2}} \left(1 - \frac{\Delta(p_{i2})}{1 - \Delta(p_{i1})}\right)^{y^*_{i.3}}. \tag{2.9}$$

Subsequently, the induced beta prior has parameters α_2 and $(\alpha_0 - \alpha_1 - \alpha_2)$.

Now, the distribution of the observed data according to the multinomial distribution can be factorized as a product of binomial distributions. Let the data consist of three cells such that $K = y_{i.1} + y_{i.2} + y_{i.3}$, and let Equations 2.8 and 2.9 define the distribution of the collapsed data sets. Then, the conditional distribution of the observed data can be given as

$$\begin{aligned}
&p\left(\mathbf{y} \mid \Delta(\mathbf{p})\right) \\
&\propto \Delta(p_{i1})^{y_{i.1}} (1 - \Delta(p_{i1}))^{K - y_{i.1}} \left(\frac{\Delta(p_{i2})}{1 - \Delta(p_{i1})}\right)^{y_{i.2}} \left(1 - \frac{\Delta(p_{i2})}{1 - \Delta(p_{i1})}\right)^{y_{i.3}} \\
&\propto \Delta(p_{i1})^{y_{i.1}} (1 - \Delta(p_{i1}))^{y_{i.2} + y_{i.3}} \left(\frac{\Delta(p_{i2})}{1 - \Delta(p_{i1})}\right)^{y_{i.2}} \left(\frac{\Delta(p_{i3})}{1 - \Delta(p_{i1})}\right)^{y_{i.3}} \\
&\propto \Delta(p_{i1})^{y_{i.1}} \Delta(p_{i2})^{y_{i.2}} \Delta(p_{i3})^{y_{i.3}},
\end{aligned}$$

which equals the unnormalized multinomial density. It can be shown in a similar way that the product of beta distributions defines the Dirichlet prior due to the collapsing property of the latter one.

This factoring of the Dirichlet-multinomial in components of beta-binomials is used in the WinBUGS (Lunn et al., 2000) implementation given in Appendix B. The implementation is given for N persons, K items, and five response categories, where the randomized response data are specified as multinomially distributed. Then, the individual category-response probabilities are specified as beta distributed, where the beta prior parameters are derived from the Dirichlet parameters.

The implementation requires the specification of a hyperprior for the Dirichlet parameters. There is often little information available about the category-response rates in the population. When a substantial number of cells does not contain observations, the parameters might not be estimable or the estimates are located on the boundary of the parameter space. A flattening prior that smooths the estimates toward a unique mode located in the interior of the parameter space is preferred when the data are sparse. The prior that assigns a common value of one or greater(say, e.g., $\alpha_c = 1$ for $c = 1, \ldots, C$) will have this smoothing or flattening property. Therefore, it might seem reasonable to restrict the prior parameters to a common value but this uninformative proper hyperprior also fixes the influence of the prior, which might be too weak for small sample sizes. It is also difficult to determine the amount of prior information given the sample information. A uniform prior, $\boldsymbol{\alpha} \sim U(0, 10)$, will also have this flattening property

but the data will be used to estimate the prior parameters. The influence of the prior is estimated from the data. When the data are sparse, a more informative prior is needed to obtain stable parameter estimates but the data will be used to estimate the amount of prior information. Furthermore, the estimated prior parameter estimates will reveal whether the observed data do not support the model. In that case, a substantial amount of prior information is needed, more than 20% of the sample data, to obtain stable parameter estimates.

2.6 Restricted Dirichlet-Multinomial Modeling

The Dirichlet-multinomial model in Equation 2.4 is a saturated model in the sense that the category-response rates are freely estimated over individuals. The Dirichlet prior does not impose any restrictions that are typically present in a cross-classified data structure.

Schafer (1997) proposed a constrained Dirichlet prior to impose a loglinear model on the individual response rates. This constrained prior forms a conjugate class since it has the same functional form as the multinomial likelihood. The constrained Dirichlet prior is represented by

$$\begin{aligned} \Delta(\mathbf{p}_i) &\propto \prod_c \Delta(p_{ic})^{\alpha_c - 1} \\ \log\left(\Delta\left(\mathbf{p}_i\right)\right) &= \mathbf{M}\boldsymbol{\lambda}, \end{aligned}$$

where $\mathbf{M}$ is the design matrix that defines a restriction on the transformed response rates.

In the same way, a restriction can be defined on the (true) category-response rates instead of the transformed category-response rates. It will restrict the posterior solution to that area where the loglinear model on the category response rates is true; that is, $\log\left(p_{ic}\right) = \mathbf{M}_c^t \boldsymbol{\lambda}_c$, for $c = 1, \ldots, C$. Such a constrained prior makes the strong assumption that the category-response rates can be partitioned according to the implied structure. Here, such a model restriction will be particularly used to test alternative models that assume a certain homogeneity in category-response rates over individuals or groups of individuals.

2.7 Application of the Dirichlet-Multinomial Model

A simulation study is performed to evaluate the performance of the full Bayes method for estimating the population proportions. Furthermore, the full and empirical Bayes estimates of the true individual response rates are compared under different conditions given categorical randomized response data. Then, the model is used to analyze randomized response data from the college alcohol problem scale (CAPS, O'Hare, 1997).

2.7.1 Simulation Study

In order to investigate the performance of the full Bayes estimation method, data were simulated under various conditions. The number of persons (N equaled 100 or

500), items (K equaled ten or fifteen), response categories (C equaled three or five), and randomizing device characteristics (ϕ_1 equaled .6 or .8) were varied. The data generation procedure comprised the following. For each respondent C category-response rates were simulated from a Dirichlet distribution given prior parameters $\boldsymbol{\alpha}$. The prior parameters were constant or varied over response categories. For the constant case, the sum of the prior parameters equaled C and the prior parameters equaled one such that the population proportions equaled $1/C$. For the non-constant case, the sum of the prior parameters was not equal to C and the prior parameters $(\alpha_1, \alpha_2, \alpha_3)$ equaled $(1, 2, 1)$ for C = 3 and $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5)$ equaled $(1, 2, 4, 2, 1)$ for C=5. The simulated category-response rates were used to generate true response patterns, which were randomized using the forced response design with randomizing device probabilities ϕ_1 and $\phi_2 = 1/C$. Ten independent samples were generated for each condition.

The parameters were re-estimated using WinBUGS. The WinBUGS code of the Dirichlet-multinomial model for RR data is given in Appendix B. For each data set, 15,000 iterations were made with a burn-in period of 5,000 iterations. Each model parameter was estimated by the average of the corresponding sampled values, which is an estimate of the posterior mean.

The method was successful in model parameter estimation. The point estimates are close to the true values and the standard deviations become smaller when increasing the number of respondents. Similar trends were found for the cases of three and five response categories. However, for the C=5 case, the reduction in the estimated prior weights is better visible when increasing the number of items and/or decreasing the percentage of forced responses. This follows from the fact that more parameters need to be estimated with the same amount of observed data.

In Table 2.1, for C=5, the estimated population proportions per category are presented. The prior parameters were divided by the sum of the prior parameters such that they were scaled in the same way as the true generating values. Note that each estimate is an average of the estimates corresponding to the ten independently generated data sets. It can be seen that the prior parameter estimates resemble the true values quite well for the constant and non-constant case. Increasing the number of persons leads to more accurate results, since the estimated standard deviations become smaller.

When decreasing the percentage of forced responses, the standard deviations remain constant for the case of ten and fifteen items, and 100 and 500 persons. The actual amount of information will increase when the amount of forced responses is reduced, since the forced responses are just random noise to mask the individual answers. From Equation 2.5 it can be seen that the number of items K as well as α_0 determine the prior weight in the computation of the individual expected posterior category-response rate. It is clear that particularly for these situations the prior weights reduce since the sum of the prior parameters become smaller. That is, the influence of the population prior on the posterior mean category-response rates becomes smaller when decreasing the amount of forced responses. The observed RR data will contain more information about the individual category-response rates when less forced responses are observed and less prior information will be used to estimate the response rates. Note that the standard deviations of the sum

Table 2.1: Full Bayes estimates of population proportions for 5 response categories, 100 and 500 respondents, and 10 and 15 items.

		K = 10				K = 15			
		Const.		Non-Const.		Const.		Non-Const.	
	Parameter	Mean	SD	Mean	SD	Mean	SD	Mean	SD
N = 100									
$\phi_1 = 0.6$	α_1/α_0	.209	.016	.144	.013	.207	.014	.145	.012
	α_2/α_0	.203	.016	.200	.015	.205	.014	.205	.013
	α_3/α_0	.197	.016	.318	.018	.197	.014	.308	.015
	α_4/α_0	.195	.016	.198	.015	.195	.014	.195	.013
	α_5/α_0	.196	.016	.140	.013	.195	.014	.147	.012
	α_0	12.003	1.642	15.615	2.225	12.082	1.432	17.445	2.204
$\phi_1 = 0.8$	α_1/α_0	.199	.017	.123	.013	.206	.016	.114	.011
	α_2/α_0	.207	.017	.208	.016	.194	.015	.201	.014
	α_3/α_0	.190	.017	.350	.019	.202	.016	.357	.017
	α_4/α_0	.205	.017	.204	.016	.201	.016	.197	.014
	α_5/α_0	.200	.017	.116	.012	.197	.016	.119	.011
	α_0	7.785	1.008	12.217	1.679	7.984	.861	14.183	1.781
N = 500									
$\phi_1 = 0.6$	α_1/α_0	.196	.007	.140	.006	.204	.006	.141	.005
	α_2/α_0	.197	.007	.203	.007	.201	.006	.201	.006
	α_3/α_0	.202	.007	.321	.008	.201	.006	.319	.007
	α_4/α_0	.201	.007	.195	.007	.196	.006	.200	.006
	α_5/α_0	.203	.007	.141	.006	.198	.006	.139	.005
	α_0	14.276	1.116	22.994	2.083	15.534	1.001	26.183	2.057
$\phi_1 = 0.8$	α_1/α_0	.197	.008	.122	.006	.200	.007	.122	.005
	α_2/α_0	.202	.008	.201	.007	.202	.007	.198	.006
	α_3/α_0	.203	.008	.357	.008	.200	.007	.357	.007
	α_4/α_0	.198	.008	.201	.007	.200	.007	.201	.006
	α_5/α_0	.200	.008	.119	.005	.198	.007	.122	.005
	α_0	8.351	.526	15.232	1.265	8.672	.454	16.511	1.110

of the prior parameters (α_0) become smaller when decreasing the number of forced responses. The typical advantage of the full Bayes estimation method applies here, where the prior weights are also estimated from the data. Note that for $\phi_1 = .6$, 40% of the data are forced responses. So, the actual amount of information in the data is rather limited but the population parameters can still be recovered. The decrease in the amount of forced responses leads to more accurate results.

When increasing the number of items, the standard deviations only become slightly smaller. The additional amount of five RR observations did not led to a substantial increase in the precision of the posterior mean estimates. The increase in items led to a higher estimate of α_0 with a smaller standard deviation. As a result, the posterior mean response rates will be more influenced by the data than the prior information due to the higher number of items and the higher posterior mean estimate of α_0.

In this simulation study, the posterior mean response rates were also compared for different number of respondents (100 and 500) and different prior values. To evaluate the prior influence, individual category-response rates were estimated using the simulated and estimated prior parameters, and through a full Bayes estimation method. Per response category, the mean squared error (MSE) of the estimated response rates was calculated. The MSE comprises a comparison of the estimate of the individual response rate with the true value. For category c, the MSE can be stated as

$$MSE(\hat{\mathbf{p}}_c \mid \mathbf{y}) = \sum_{i=1|c}^{N} E(p_{ic} - \hat{p}_{ic})^2 + \sum_{i=1|c}^{N} Var(\hat{p}_{ic})^2,$$

where the first term is the cumulative bias between the true value and its estimate and the second term represents the cumulative variance of the estimate.

In Table 2.2, the MSEs are presented for five response categories, where the cumulative bias and variance terms are given in parenthesis. Two empirical Bayes estimates are considered, the $MSE(\hat{p}_{EB}, \alpha_c = 1)$ and the $MSE(\hat{p}_{EB}, \alpha_c)$, which is based on prior parameters that are set to one, denoted as the homogenous prior, and the simulated prior parameters, respectively. The $MSE(\hat{p}_{FB})$ is based on full Bayes estimates using the WinBUGS program. The simulated prior parameters correspond to the non-constant case with $\boldsymbol{\alpha} = (1, 2, 4, 2, 1)$, which differs from the homogenous prior with prior parameters equal to one.

For 100 and 500 persons, the bias is smallest for the full Bayes estimates and they are slightly better than empirical Bayes estimates given the true prior parameters. The full Bayes estimates are accurate with respect to bias since the prior weights are also estimated from the sampled data. The estimates using the homogenous prior, with all prior parameters equal to one, have the highest bias but smaller MSEs than those based on the full Bayes estimates. For the latter one, the estimated variances are much higher due to the fact that the prior parameters also need to be estimated. Therefore, the homogenous prior leads to quite accurately estimated category-response rates and performs better than the full Bayes estimates given the MSEs.

Table 2.2: Estimated MSEs concerning response-rate estimates for five response categories, and 100 and 500 respondents. The bias and variance components of the MSE are given in parenthesis, respectively.

	Category	$MSE(\hat{p}_{EB}, \alpha_c)$	$MSE(\hat{p}_{EB}, \alpha_c = 1)$	$MSE(\hat{p}_{FB})$
N=100				
	c = 1	.45 (.44, .00)	.52 (.51, .01)	.90 (.42, .48)
	c = 2	.66 (.66, .01)	.85 (.84, .01)	1.33 (.59, .74)
	c = 3	1.18 (1.16, .02)	1.53 (1.50, .02)	2.11 (1.01, 1.09)
	c = 4	.90 (.89, .01)	1.05 (1.03, .02)	1.67 (.88, .79)
	c = 5	.51 (.51, .00)	.56 (.55, .01)	.98 (.47, .50)
N = 500				
	c = 1	2.22 (2.20, .02)	3.12 (3.08, .04)	4.27 (1.95, 2.32)
	c = 2	4.29 (4.24, .05)	5.32 (5.25, .07)	7.66 (4.00, 3.66)
	c = 3	6.24 (6.15, .09)	7.91 (7.79, .12)	10.78 (5.62, 5.16)
	c = 4	4.34 (4.30, .05)	5.22 (5.15, .07)	7.63 (4.14, 3.49)
	c = 5	2.51 (2.49, .02)	3.18 (3.14, .40)	4.54 (2.26, 2.27)

2.7.2 Response Rates of Alcohol-Related Negative Consequences

The college alcohol problem scale (CAPS; O'Hare, 1997) was used to measure frequencies of alcohol-related negative consequences among college students. The first thirteen items of the CAPS scale (Appendix C) were used that covered socio-emotional problems (hangovers, memory loss, nervousness, depression) and community problems (drove under the influence, engaged in activities related to illegal drugs, problems with the law). Each item has a five-point scale (one = never/almost never, five =almost always).

A total of 793 US college student were at random divided in two groups. One group of 351 participants answered the questionnaire directly without using a randomizing device, denoted as the direct-questioning (DQ) group. The other group, denoted as the RR group, consisted of 442 participants and they responded to the questionnaire according to a forced randomized response design, where $\phi_1 = .60$ and $\phi_2(c) = .20$ for $c = 1, \ldots, 5$. The RR group used a spinner to answer the questions. The spinner was developed such that 60% of the area was comprised of 'answer honestly' space, and 40% of the area was divided into equal sections to represent the five possible answer choices.

The main focus of the study was to investigate whether the RR technique improved the accuracy of self-reports. The sensitivity of the response categories was evaluated, where it was expected that a strong confirmation to an item is more sensitive than a negative confirmation. The Dirichlet-multinomial model with a restricted Dirichlet prior was used to evaluate the effect of the RR technique per category, where the between-group differences in mean category-response rates were investigated.

Group-Specific Population Proportions

The Dirichlet-multinomial model was estimated using group-specific population proportions such that the DQ and the RR group each had their own prior parameters. In Table 2.3, the full Bayes mean estimates of the population proportions and standard deviations are presented. Note that the estimated population proportions of the RR group are transformed such that they estimate the true category proportions in this group. The estimated population proportions corresponding to the observed response rates are given in parenthesis, which also take the forced responses into account. The standard deviations are given in the next column, where the standard deviations of the non-transformed proportions are given in parenthesis.

It can be seen that the estimated population proportion of category one of the DQ group is higher and that the other category-mean proportions are smaller compared to the RR group. This indicates that the respondents in the DQ group are less likely to confirm experiences with alcohol-related negative consequences. However, the respondents were randomly assigned to the RR group, which suggests serious underreporting in the DQ group. Although the RR group contains more respondents than the DQ group, the standard deviations of the estimated population proportions of the RR group are higher, since 40% of the data are forced responses. The standard deviations of the non-transformed population proportions are comparable since they are based on all data. The RR group has a higher estimate of the prior weight (α_0) compared to the DQ group due to the number of forced responses. As a result, the posterior means of the individual category-response rates are more influenced by the prior in the RR group than in the DQ group.

Table 2.3: CAPS: Estimated population proportions per category for the DQ and RR group.

	DQ (351)		RR (442)	
	Mean	SD	Mean	SD
α_1/α_0	.590	.010	.471(.363)	.012(.007)
α_2/α_0	.161	.007	.170(.182)	.010(.006)
α_3/α_0	.129	.006	.153(.172)	.009(.006)
α_4/α_0	.069	.005	.131(.159)	.009(.005)
α_5/α_0	.051	.003	.075(.125)	.008(.005)
α_0	9.872	.648	20.360	1.672

Linear Restricted Category Response Rates

To investigate the category-specific effect of the RR method, a loglinear model was defined for the category-response rates. For each category, the logarithm of the true category-response rates were explained by a constant and a category-specific

RR effect. This restriction of the Dirichlet prior is given by

$$log(p_{ic}) = \lambda_{0c} + \lambda_{1c} RR_i,$$

for $c = 1, \ldots, C$, where RR_i equals one when respondent i belongs to the RR group and zero otherwise. Note that the loglinear representation was only used to evaluate the category-specific RR effect.

In Table 2.4, the intercept and RR effect are presented per category for the DQ and RR group. The 95% highest posterior density (HPD) region is also given for each parameter. The estimates can be used to compute the posterior expected population proportion per category in each group. It follows that about $\exp(-.62) = 54\%$ and $\exp(-.62 - .27) = 41\%$ in the DQ group and the RR group, respectively, say that they almost never experience negative consequences of drinking. The other posterior mean percentages can be computed in a similar way.

The RR effect is negative for the first category (labeled almost never) and positive for all other categories. By investigation of the HPD regions, it follows that all RR effects are significantly different from zero. It can be concluded that in the DQ group, respondents underreported any experiences of negative consequences and, subsequently, overreported that almost never negative consequences were experienced. Furthermore, the estimated RR effects increase with an increase in the number of negative experiences, where the fifth category has the highest RR effect. Thus, an increase in the number of experiences of alcohol-related negative consequences leads to a more sensitive response option. The difference between the groups with respect to the posterior expected proportion of respondents that admit experiencing negative consequences is highest for the fifth response option. In that case, around 1% of the DQ group admits to have almost always alcohol-related negative consequences, which is around 13% in the RR group. It can be concluded that the RR technique led to a higher degree of cooperation and more accurate data, especially when the response options become more sensitive.

Table 2.4: CAPS: Category-specific intercepts and RR effects.

	Intercept			RR Effect		
	Mean	SD	HPD	Mean	SD	HPD
1 Almost never	-.62	.03	-.68, -.56	-.27	.03	-.33, -.22
2 Seldom	-2.13	.07	-2.28,-2.00	.43	.08	.28, .57
3 Sometimes	-2.39	.08	-2.56,-2.23	.65	.08	.49, .81
4 Often	-3.61	.16	-3.93,-3.33	1.65	.13	1.39, 1.91
5 Almost always	-4.71	.18	-5.09,-4.37	2.70	.13	2.45, 2.94

The response data from the RR group were used to explore ethnic differences in experiencing alcohol-related negative consequences. The responses from the DQ group were shown to be biased, since effects of under- and overreporting were found. In this study, the racial origin of the respondents was administered. The RR

group consisted of 2% Asians, 83% white Americans, 11% African Americans, and 12% belonged to another ethnicity. An indicator variable, labeled Ethnicity, was used in the log-linear model that represented the racial origin of each respondent.

In Table 2.5, the estimated category-specific effects of ethnicity on the individual category response rates are given. Each posterior mean effect is accompanied

Table 2.5: CAPS: Ethnic differences per response category.

		Effect			Scaled Effect		
Cat.	Ethnicity	Mean	SD	HPD	Mean	SD	HPD
1							
	Asian	-.04	.12	-.29, .17	-.18	.16	-.51, .11
	White American	-.03	.05	-.12, .07	-.16	.05	-.26,-.06
	African American	.13	.06	.02, .24	.00		
	Other	-.06	.08	-.22, .08	-.20	.10	-.40,-.02
2							
	Asian	.25	.19	-.14, .60	.46	.26	-.08,1.00
	White American	-.02	.08	-.17, .13	.19	.09	.00, .37
	African American	-.21	.10	-.40,-.03	.00		
	Other	-.01	.12	-.26, .22	.20	.16	-.14, .50
3							
	Asian	-.23	.20	-.60, .17	-.28	.27	-.80, .25
	White American	.07	.08	-.09, .22	.03	.09	-.15, .20
	African American	.04	.10	-.15, .23	.00		
	Other	.11	.12	-.12, .36	.08	.16	-.23, .39
4							
	Asian	.13	.20	-.27, .50	.22	.27	-.30, .76
	White American	.06	.08	-.10, .21	.15	.09	-.02, .32
	African American	-.09	.10	-.29, .09	.00		
	Other	-.10	.12	-.34, .13	-.01	.16	-.33, .28
5							
	Asian	-.01	.19	-.37, .35	.07	.25	-.43, .55
	White American	-.03	.07	-.18, .10	.04	.08	-.12, .19
	African American	-.08	.09	-.25, .09	.00		
	Other	.12	.11	-.11, .33	.19	.14	-.09, .46

with a posterior standard deviation (SD) and a 95% highest posterior density (HPD) interval. Under the column labeled Effect, the estimated effects are presented, where each category-specific intercept represents the average population level on the logarithmic scale. Under the column labeled Scaled Effect, the intercept represents the average population level of the African Americans on the logarithmic scale. The scaled ethnic effect of the this group is in that case equal to zero.

For the first category, the African Americans score significantly higher than the other groups. Furthermore, the estimated scaled effects of the other groups

are significantly smaller. This means that the percentage of African-Americans experiencing almost never negative consequences is much higher compared to the other groups. In the same way, it follows that the percentage of African-Americans experiencing negative consequences seldom to almost always is lower than that of other groups. In the second category, the white Americans also score significantly higher than the African Americans, which follows from the scaled effects. The third to fifth response categories did not show any significant ethnic differences.

2.8 Discussion

For small data samples, the Dirichlet-multinomial model is proposed for categorical RR data, where a linear transformation of the response rates is implemented to adjust for the RR sampling design. This model is a generalization of the beta-binomial model for binary RR data. Both models are suitable for sensitive-survey studies and small data samples. The individual category-response rates are related to the observed data, but a linear transformation can be used to derive the true categorical-response rates. The parameters of this linear transformation are the characteristics of the randomizing device and they are usually known. The derived expressions of the posterior expectation and variance of the category-response rates are useful in case of empirical Bayes estimation or explicit prior knowledge about response rate population parameters.

The idea of full Bayes parameter estimation was elaborated using the synthetic data set. The simulation study has shown that full Bayes model was able to rather accurately estimate values of parameters for different number of response categories. The method was equally successful in retrieving the parameters for the constant case of homogenous prior parameters as well as for the case of non-constant prior parameters. Moreover, the simulation study concluded that increasing the number of persons leads to more accurate results, while the variation of the percentage of forced responses does not influence the accuracy.

A constrained-Dirichlet prior is used to identify homogeneity in response rates over items and persons. Therefore, the WinBUGS program was extended to define a constrained-Dirichlet prior, where a loglinear model was defined on the true category-response rates.

An important effect was identified in the real data study, which showed that the effect of the RR method varied over response categories. A priori it was assumed that the response options varied in their sensitivity, where a higher degree of accordance with the sensitive item implied a higher sensitivity, but this hypothesis was never tested in the literature. The analysis showed a substantial increase in agreement with more sensitive response options under the randomized response condition.

In RR studies the topic of compliance is often an issue. Respondents are instructed to follow the RR instructions but may for different reasons act differently. In large-scale sample studies, a latent class structure can be integrated in the model to identify non-compliant behavior. The responses from the non-compliant subjects are modeled differently. The Dirichlet-multinomial model can also be extended with a two-component latent-class structure to allow for non-compliance,

but that would require more response data to obtain stable parameter estimates.

Chapter 3

Mixture Randomized Item Response Modeling: A Smoking Behavior Validation Study

Abstract

Misleading response behavior is expected in medical settings where incriminating behavior is negatively related to the recovery from a disease. In the present study, lung patients feel social and professional pressure concerning smoking and experience questions about smoking behavior as sensitive, and tend to conceal embarrassing or threatening information. The randomized item-response survey method is expected to improve the accuracy of self-reports since individual item responses are masked and only randomized item responses are observed.

The validation of the randomized item-response technique is explored in a unique experimental study. Therefore, a new multi-item measure assessing smoking behavior was administered using a treatment-control design (randomized response or direct questioning). After the questionnaire, a breath test using a carbon monoxide (CO) monitor was administered to determine the smoking status of the patient. The response data were used to measure the individual latent smoking behavior using a mixture item response model. It is shown that the detected smokers scored significantly higher in the RR condition compared to the DQ condition. A Bayesian latent variable framework is proposed to evaluate the diagnostic test accuracy of the questionnaire using the randomized response technique, which is based on the posterior densities of the subject latent scores together with the breath test measurements. For different thresholds, moderate posterior mean estimates of sensitivity and specificity were obtained due to observing a limited number of discrete randomized item responses.

Key words: randomized response, validation, mixture item response theory,

diagnostic test accuracy, classification probabilities.

3.1 Introduction

The research on smoking status assessment often rests on information available from self-reports. Many researchers indicate that denial and underreporting of the extent of smoking are quite usual in certain populations, such as adolescents and announced quitters (Akers, Massey, Clarke, & Lauer, 1983; Monninkhof et al., 2004). Self-reports are also known to be less reliable for special subgroups such as individuals with a coronary diagnosis or other smoking-related diseases (Attebring, Herlitz, Berndt, Karlsson, & Hjalmarson, 2001). This follows from the fact that these smokers fail to discontinue smoking, but feel a strong pressure to do so. However, for a wide variety of medical conditions and particularly for diseases of the respiratory system, accurate information about the smoking status of patients is crucial for the choice of suitable treatments. In addition, accurate information is also essential for health care professionals during smoking cessation programs. In general, survey research techniques have several limitations that can undermine its usefulness in health research. One of the limitations is that individuals are reluctant to provide personal information about sensitive attitudes or behaviors (Tourangeau et al., 2000). Respondents tend to under- or overreport or avoid questions which are perceived as threatening or sensitive. Threatening or sensitive questions can concern personal behavior (such as sexual practices), illegal behavior (such as drug use or alcohol consumption), and other health-related behaviors (such as smoking).

Many agree on the fact that a validation of self-reported levels of smoking is necessary (Daly & Blann, 1996; Hill, Haley, & Wynder, 1983). A broad spectrum of biochemical validation techniques that differentiate between smokers and non-smokers is available in clinical settings. Many methods are based on the determination of the levels of nicotine and its metabolites in blood or urine, as well as the traces of carbon monoxide in blood and saliva (Jarvis, Tunstall-Pedoe, Feyerabend, Vesey, & Saloojee, 1987). Most of these tests are rather invasive, expensive or time-consuming. An alternative less expensive method offering immediate result, not requiring specialized training and that is suitable for situations where nicotine replacement strategies are used is the expired air carbon monoxide (CO) test.

Here, attention will be focused on the improvement of self-reports on a sensitive attribute by means of alternative survey techniques ensuring the privacy of the respondents' answers. In particular, Warner (1965) proposed a univariate randomized response data collection method that led to a whole family of techniques, which insure confidentiality of responses due to randomization. For example, in the so-called forced RR design, a randomizing device is used before an answer is given and the outcome decides whether a truthful or forced (simulated) response is requested. The answer is protected since the outcome of the randomizing device is only known to the respondent. Although Soeken (1987), Rittenhouse (1996a, 1996b), and Williams, Suen, and Baffi (1993) explicitly recommended the use of the RR technique (RRT) in health research, the RR survey method did not re-

ceive much attention. Only recently, Cross, Edwards-Jones, Omed, and Williams (2010) used the RRT to estimate the sheep scab prevalence in Welsh flocks by interviewing farmers in Wales, and Ostapczuk, Musch, and Moshagen (2011) to investigate lifetime prevalence of medication non-adherence in Germany.

The little attention toward the univariate RR techniques in health research may be attributed to the limitation of inferences to the population level. However, it is possible to make individual-level inferences given multivariate randomized response data (Fox, 2005b). In other research fields, the RRT has obtained more attention and various studies have been performed successfully to measure the prevalence of a sensitive behavior. De Jong et al. (2010) reported on desires for products and services in the domain of adult entertainment. Fox and Meijer (2008) and Fox (2005b) analyzed students' academic cheating behavior at a Dutch university. Fox and Wyrick (2008) measured the prevalence of alcohol use, and alcohol-related problems among college students. Rates of academic cheating and rates of criminal behavior were investigated by Tracy and Fox (1981) and Scheers and Dayton (1988), respectively. The prevalence of academic cheating, tax evasion, and software piracy, among others, have been discussed by Lensvelt-Mulders, Hox, van der Heijden, and Maas (2005); Lensvelt-Mulders, Hox, and van der Heijden (2005) and van der Heijden et al. (2000).

An item response theory model (Lord & Novick, 1968) combined with an RR model will be used to measure the sensitive latent trait smoking behavior given multivariate RR data. The so-called randomized item response model enables the measurement of item and person characteristics while accounting for the RR nature of the data being analyzed. In practice, the RR design might not always be followed. Non-compliant respondents (cheaters) are expected to elicit the least incriminating response to improve their self-report. This leads to a surplus of such responses. Therefore, Clark and Desharnais (1998), Böckenholt and van der Heijden (2007), De Jong et al. (2010), and Ostapczuk et al. (2011), defined a mixture randomized response model to account for respondents that do not follow the RR instructions. Here, a generalized mixture model is proposed using response-type specific (binary and ordinal) latent classes to capture extreme use of the less incriminating response category. In this way it is possible to control for response-type specific positive self-presentation biases, where it is expected that the binary response format will induce more (positive self-presentation) bias than the ordinal response format.

In experimental studies using RRT, a treatment-control design is often used for validation of the results, where the randomly selected members of the control and the randomized-response group are interviewed via direct or randomized response questioning, respectively. Without knowing the truth, a difference between the mean response of the DQ and the RR group cannot be interpreted as a validation of the item randomized response technique. Umesh and Peterson (1991) commented that a legitimate validation of randomized response estimates will rest on true response values, preferably at the individual level. This is done in the present smoking behavior survey study. The true smoking status (smoker/non-smoker) of every participant was determined after the interview by the expired air carbon monoxide test. The participants were randomly selected into two groups, where one group was questioned using the forced randomized response technique and the

other group was questioned in a direct mode. Response data were collected using a new multi-item smoking scale questionnaire, comprising ordinal and dichotomous items.

Within a Bayesian framework, individual latent scores on the smoking scale are estimated using a mixture randomized item response model for mixed response types (ordinal and binary), which generalizes the models of van der Heijden et al. (2000), De Jong et al. (2010), and Fox (2005b, 2010). Furthermore, more general response-type specific latent class components are defined for non-compliance, since it is expected that the willingness of subjects to cooperate is influenced by the variety in response options. At the level of the individual, a mean structure is defined for the latent behavior variable that includes the random assignment of subjects to treatment and control groups and various explanatory background variables. In this way, latent score differences can be attributed by individual, group, and questioning technique differences, where also the true smoking status from the CO measurement will be used to validate the RRT.

The continuous latent level of smoking behavior can be translated to a categorical latent smoking status (referred to as a diagnosis) given a selected decision threshold on the continuous latent scale. Together with the true smoking status as a gold standard posterior densities of the individual latent scores are used to compute posterior classification probabilities such as the true positive fraction (sensitivity) and the true negative fraction (specificity). Both quantities will be used to assess the diagnostic test accuracy of the smoking questionnaire under randomized response questioning. The RRT is further validated using the predictive properties of the smoking behavior questionnaire. That is, the positive and negative predictive value of the test under the randomized item-response technique will be analyzed.

This paper is organized as follows. First, the forced randomized response method and the study design are presented. Thereafter, the mixture randomized item response model for mixed responses is discussed. The Bayesian latent variable method for diagnostic accuracy is described together with posterior classification probabilities. Then, the model and the Bayesian test diagnostics are used to validate the RRT and to examine the properties of the smoking behavior test in an experimental-clinical smoking study. Finally, some concluding comments and suggestions for further research are given.

3.2 Method

This study involved outpatients of a pulmonary department of a hospital in the Netherlands. Health care providers strongly recommend lung patients to quit smoking, therefore the patients that are not complying with the medical advice often experience questions about smoking as very sensitive. A smoking questionnaire was developed to assess subject's smoking behavior. The developed smoking questionnaire to assess smoking behavior is a multi-item screening instrument developed in this study for measuring the latent smoking behavior of each lung patient toward smoking. The smoking scale comprises nine dichotomous and three polytomous items, see Appendix D. The list of items is constructed from questions

that are usually asked by medical personal during the routine visits to the treating pulmonologist.

According to a randomized control trial, subjects were randomly assigned to an RR or DQ group. In the DQ group, subjects completed the questionnaire in a conventional way. In the RR group, subjects completed the questionnaire using a randomizing spinner-device for the items with two and three response categories.

The demographic characteristics such as gender, age, and educational level of the patient were recorded as background characteristics, which are known to be related to smoking behavior (Wetter et al., 2004). In addition, information on medical condition and treating pulmonologist were noted as well.

After completion of the questionnaire the smoking status of the participants was assessed by means of physical carbon monoxide level measurement in the expired air (Middleton & Morice, 2000). The Bedfont Micro 4 Smokerlyzer portable carbon monoxide monitor was used for measuring expired-air carbon monoxide level. The carbon monoxide (CO) measurement, which in the following will be referred to as the true smoking status of the patient, was used to distinguish smokers from non-smokers. The response data was used to estimate patient's smoking behavior, as a continuous latent trait, and patient's smoking status, as a discrete latent trait. The true smoking status was used to validate RR as a self-report improving questioning technique.

3.2.1 Mixture Randomized Item Response Model

In the RR group, the outcome of the randomizing device determines whether the respondent is requested to answer honestly or to give a (simulated) forced response. The respondent's answer is protected, since the outcome of the randomizing device is only known to the respondent. Each individual response observed might be a forced response, which makes it much easier for the individuals to give truthful responses to sensitive questions when an honest response is requested by the randomizing device. A forced randomized response design will be utilized, where an honest response is requested with probability ϕ_1 and a forced response with probability $(1 - \phi_1)$. Subsequently, a forced positive response is prompted with probability ϕ_2. The a priori known characteristics of the randomizing device determine the probabilities ϕ_1 and ϕ_2.

Let Y_{ik} denote the randomized response of subject i to item k. Subsequently, let $\tilde{Y}_{ik}$ denote the response that will be observed when an honest response is requested, which will be referred to as the nonrandomized response. The probability of an observed randomized response given by participant i to item k can be expressed as

$$P\left(Y_{ik} = c\right) = \phi_1 P\left(\tilde{Y}_{ik} = c\right) + (1 - \phi_1)\phi_2(c) \quad c \in 1, \ldots, C. \tag{3.1}$$

The observed RR data are modeled using a mixture randomized item response model for mixed responses (binary and ordinal). For each response-type format, it will be assumed that the mixture component is represented by two unobserved groups in the population. The subjects belong to the compliance group who follow the randomization scheme or to the non-compliance group who always choose the least stigmatizing category. This non-compliance group is characterized by

subjects that respond negatively (the least self-incriminating answer) with probability one to all questions and ignore the instructions of the randomized response design. The randomized item responses of subjects in the compliance group are assumed to be distributed according to an item response model when an honest answer is requested and a generated (forced) response when a forced response is prompted by the randomization device. The responses from the DQ group are assumed to be distributed according an item response model. Furthermore, the item characteristics are assumed to be invariant over questioning techniques.

Let θ_i denote the latent smoking behavior of subject i and β_k the threshold parameter of item k. For subject i $(i = 1, \ldots, N)$, the probability of a positive response to item k depends on the random assignment to the RR or DQ group and its membership to the compliance or non-compliance group. When item k has two response categories, the response probabilities of subject i are given by,

$$\begin{aligned} P\left(Y_{ik} = 0 \mid \theta_i, b_k\right) &= \begin{cases} 1 - \Phi\left(\theta_i - b_k\right) & \text{for DQ} \\ \pi_i^b + \left(1 - \pi_i^b\right)\left(1 - \left(\phi_1 \Phi\left(\theta_i - b_k\right)\right.\right. \right. \\ \qquad \left.\left.\left. + \left(1 - \phi_1\right)\phi_2\right)\right) & \text{for RR,} \end{cases} \\ P\left(Y_{ik} = 1 \mid \theta_i, b_k\right) &= \begin{cases} \Phi\left(\theta_i - b_k\right) & \text{for DQ} \\ \left(1 - \pi_i^b\right)\left(\phi_1 \Phi\left(\theta_i - b_k\right)\right. \right. \\ \qquad \left.\left. + \left(1 - \phi_1\right)\phi_2\right) & \text{for RR,} \end{cases} \end{aligned} \tag{3.2}$$

where $\Phi(.)$ denotes the cumulative normal distribution function and $\pi_i^b = 1$ when subject i belongs to the non-compliance group for the binary items.

When item k has three response categories, the response probabilities of subject i are given by,

$$\begin{aligned} P\left(Y_{ik} = 0 \mid \theta_i, \kappa_{k1}\right) &= \begin{cases} \Phi\left(\kappa_{k1} - \theta_i\right) & \text{for DQ} \\ \pi_i^o + \left(1 - \pi_i^o\right)\left(\phi_1 \Phi\left(\kappa_{k1} - \theta_i\right)\right. \right. \\ \qquad \left.\left. + \left(1 - \phi_1\right)\phi_2(0)\right) & \text{for RR} \end{cases} \\ P\left(Y_{ik} = 1 \mid \theta_i, \boldsymbol{\kappa}_k\right) &= \begin{cases} \Phi\left(\theta_i - \kappa_{k1}\right) - \Phi\left(\theta_i - \kappa_{k2}\right) & \text{for DQ} \\ \left(1 - \pi_i^o\right)\left(\phi_1\left[\Phi\left(\theta_i - \kappa_{k1}\right) - \Phi\left(\theta_i - \kappa_{k2}\right)\right]\right. \right. \\ \qquad \left.\left. + (1 - \phi_1)\phi_2(1)\right) & \text{for RR} \end{cases} \\ P\left(Y_{ik} = 2 \mid \theta_i, \kappa_{k2}\right) &= \begin{cases} \Phi\left(\theta_i - \kappa_{k2}\right) & \text{for DQ} \\ \left(1 - \pi_i^o\right)\left(\phi_1 \Phi\left(\theta_i - \kappa_{k2}\right)\right. \right. \\ \qquad \left.\left. + \left(1 - \phi_1\right)\phi_2(2)\right) & \text{for RR,} \end{cases} \end{aligned} \tag{3.3}$$

where $\pi_i^o = 1$ when subject i belongs to the non-compliance group for the ordinal items.

At a higher level, general priors for the model parameters are specified as follows;

$$\begin{aligned} \theta_i &\sim \mathcal{N}\left(\mu_\theta, \sigma_\theta^2\right) \\ b_k &\sim \mathcal{N}\left(\mu_b, \sigma_b^2\right) \\ \pi_i^b &\sim \mathcal{B}\left(\pi_{01}\right) \\ \pi_i^o &\sim \mathcal{B}\left(\pi_{02}\right). \end{aligned}$$

Furthermore, restricted normally distributed priors are specified for the thresholds

κ_{k1} and κ_{k2} such that $\kappa_{k1} < \kappa_{k2}$. The hyperpriors are specified as follows,

$$\begin{aligned} p\left(\mu_\theta \mid \sigma_\theta^2\right) &= \mathcal{N}(0, \sigma_\theta^2/n_0) \\ p\left(\mu_b \mid \sigma_b^2\right) &= \mathcal{N}(0, \sigma_b^2/n_0) \\ p\left(\pi_{01}\right) = p\left(\pi_{02}\right) &= \mathcal{B}e\left(g_1, g_2\right) \end{aligned}$$

and inverse-gamma priors for the variances σ_b^2 and σ_θ^2.

The population prior for the latent variable can be changed to model mean latent differences between questioning techniques or to include effects of background information, and/or to analyze the relationship between the self-report measure of smoking behavior and the CO measure of smoking.

The model parameters can be estimated simultaneously using MCMC. The WinBUGS program-code is given in the Appendix E. The model is identified by restricting the sum of the difficulty parameters or by restricting the mean of the latent variable. In both cases, the mean of the latent scale is identified. The direct questioning and randomized response data are analyzed simultaneously. As a result, a common scale is defined for the subjects in the RR and DQ group.

3.2.2 Bayesian Latent Variable Methods for Diagnostic Accuracy

One of the objectives is to assess the smoking status of each subject from the self-reported responses to the questionnaire. The multiple-item questionnaire ensures that a more reliable measure of smoking behavior is obtained but still serves as an imperfect measure of the true smoking status. The unobserved smoking status is therefore considered to be a latent variable, and the response data will be used to estimate the smoking status. A positive diagnosis of smoking status of subject i will be denoted as $X_i = 1$, where the true smoking status will be denoted as $D_i = 1$.

The posterior density of the continuous latent variable representing smoking behavior will be used to assess the smoking status. Therefore, a threshold value on the latent scale will be considered, say θ_c. The smoking status will be positively diagnosed when the posterior probability that the smoking behavior is above this threshold is significantly high. Thus, the posterior probability of a positive diagnosis $X_i = 1$, conditional on the response pattern, is given by

$$\begin{aligned} p_{ic} = P\left(X_i = 1 \mid \mathbf{y}_i\right) &= P\left(\theta_i > \theta_c \mid \mathbf{y}\right) \\ &= \int_{-\infty}^{\theta_c} p\left(\theta_i \mid \mathbf{y}\right) d\theta_i. \end{aligned} \tag{3.4}$$

Garrett, Eaton, and Zeger (2002) and Formann and Kohlmann (1996) evaluated subject-specific medical diagnosis via a latent class analysis. In their approach, multiple test indicators are observed and the posterior probability of having the disease given the test results is computed via a latent class analysis. In the present approach, through a more advanced item response modeling approach the multiple indicators are used to measure smoking behavior on a continuous latent scale. The posterior distribution of this continuous latent variable is used

to assess the smoking status (see Equation 3.4). In a two-component latent class approach, subject-specific latent class probabilities can be computed, which will depend on the other members in the group. The observed response patterns of the subjects are used to classify each subject to one of the two groups (smokers and non-smokers). In this continuous latent variable approach, the subject-specific response data are used to measure smoking behavior, which provides accurate information about the smoking status. A subject's smoking status can be seen as the dichotomous observation of the underlying measurement of smoking behavior.

The probabilities of diagnoses are not definitive indicators of the true disease status. Each probability of diagnosis depends on the corresponding response pattern and the chosen threshold value. For making accurate decisions, it is important to evaluate the operating characteristics of the test. However, the computation of associated classification probabilities is complicated since response data are observed and not the test diagnosis of smoking behavior. Therefore, expected posterior classification probabilities are considered, where the expectation is taken over the posterior distribution of the smoking diagnosis given the observed response data.

The joint distribution of the smoking diagnoses $(X_1, \ldots, X_N)$ is represented by

$$P\left(x_1, x_2, \ldots, x_N \mid \mathbf{y}\right) = \prod_{i=1}^{N} P\left(x_i \mid \mathbf{y}_i\right) = \prod_{i=1}^{N} p_{ic}^{x_i}(1-p_{ic})^{(1-x_i)}, \quad (3.5)$$

since the smoking diagnoses are conditionally independently distributed given the individual response patterns. The joint distribution can be recognized as the product of independent Bernoulli distributions. Let M denote the random variable that represents the number of subjects with test diagnosis $X_i = 0, 1$ and smoking status $D_i = 0, 1$. Subsequently, let $M = M_{11}$ denote the number of true smokers using the reference test and diagnosed as smokers using the response data. The posterior probability of $M = M_{11}$ equals

$$P\left(M = M_{11} \mid \mathbf{y}, \theta_c\right) = \sum_{\mathcal{M}_{11}} \prod_{j \in M_{11}} p_{jc} \prod_{h \notin M_{11}} \left(1 - p_{hc}\right), \quad (3.6)$$

where $\mathcal{M}_{11}$ denotes the total set of combinations with $M = M_{11}$.

Following Broemeling (2007), the conditional probability of being positively diagnosed (conditional sensitivity), and the conditional probability of being negatively diagnosed (conditional specificity), can be computed given the diagnoses $\mathbf{x}$. The conditional posterior sensitivity (true positive fraction) of the self-report test given the diagnoses equals

$$SE\left(\theta_c \mid \mathbf{y}, \mathbf{x}\right) = \frac{P\left(M = M_{11} \mid \mathbf{y}\right)}{P\left(M = M_{11} \mid \mathbf{y}\right) + P\left(M = M_{01} \mid \mathbf{y}\right)}, \quad (3.7)$$

where M_{01} denotes the number of smokers according to the reference test but diagnosed as non-smokers using the response data. The conditional posterior specificity given the diagnoses equals

$$SP\left(\theta_c \mid \mathbf{y}, \mathbf{x}\right) = \frac{P\left(M = M_{00} \mid \mathbf{y}\right)}{P\left(M = M_{00} \mid \mathbf{y}\right) + P\left(M = M_{10} \mid \mathbf{y}\right)}, \quad (3.8)$$

where M_{10} is the number of non-smokers that are positively diagnosed. Subsequently, the expected posterior sensitivity and specificity are expressed as

$$SE(\theta_c \mid \mathbf{y}) = E[SE(\theta_c \mid \mathbf{y}, \mathbf{x}) \mid \mathbf{y}] = \sum_{\mathbf{x} \in \mathcal{X}} SE(\theta_c \mid \mathbf{y}, \mathbf{x}) P(\mathbf{x} \mid \mathbf{y}, \theta_c) \quad (3.9)$$

and

$$SP(\theta_c \mid \mathbf{y}) = E[SP(\theta_c \mid \mathbf{y}, \mathbf{x}) \mid \mathbf{y}] = \sum_{\mathbf{x} \in \mathcal{X}} SP(\theta_c \mid \mathbf{y}, \mathbf{x}) P(\mathbf{x} \mid \mathbf{y}, \theta_c) \quad (3.10)$$

respectively, where $\mathcal{X}$ denote the set of all possible diagnoses for the N subjects.

Finally, an expected posterior predictive value (PPV) can be defined as

$$PPV(\theta_c \mid \mathbf{y}) = \sum_{\mathbf{x} \in \mathcal{X}} \frac{P(M = M_{11} \mid \mathbf{y})}{P(M = M_{11} \mid \mathbf{y}) + P(M = M_{10} \mid \mathbf{y})} P(\mathbf{x} \mid \mathbf{y}, \theta_c),$$

an expected posterior negative predictive value (NPV) as

$$NPV(\theta_c \mid \mathbf{y}) = \sum_{\mathbf{x} \in \mathcal{X}} \frac{P(M = M_{00} \mid \mathbf{y})}{P(M = M_{00} \mid \mathbf{y}) + P(M = M_{01} \mid \mathbf{y})} P(\mathbf{x} \mid \mathbf{y}, \theta_c).$$

The posterior expected classification probabilities can be computed as by-products of the MCMC algorithm. In the first step, smoking diagnoses are sampled given a threshold θ_c from the joint distribution represented in Equation 3.5. In the second step, the conditional posterior sensitivity and conditional posterior specificity probabilities in Equation 3.7 and 3.8 are computed in each MCMC iteration. The posterior expected specificity and expected sensitivity, Equations 3.9 and 3.10, are estimated by the average values over MCMC iterations.

3.3 Results

The 37-day survey was conducted in cooperation with ten pulmonologists. Lung patients over 16 years of age were asked to voluntarily participate in the survey on smoking. None of the randomly selected patients refused to cooperate. A total of 305 patients were assessed and 198 completed the test using the RR technique. The RR group was significantly larger than the DQ group to account for the generated forced responses. For the items with two response categories, the developed random device requested an honest answer with probability .778 and simulated each forced response ('Yes', 'No') with probability .5. For the items with three response categories, an honest response was requested with probability .611, and the forced category response probabilities were .25 for the first and third category and .5 for the second category. Carbon monoxide was used to detect smokers, and a cut-off value of 6 parts per million (ppm) was taken as the cutoff between smokers (D=1) and non-smokers (D=0).

From the classical test procedure, for the 12 item direct-questioning data, Cronbach's alpha equals .78, which indicates that the scale has a good reliability. Cronbach's alpha cannot be directly computed from the RR data due to

Table 3.1: Assessing smoking behavior: Estimated model parameters for RR and DQ data.

	Item	Mean	SD	95% CI	
Threshold Parameters					
b_1	1	.693	.124	.444	.935
b_2	2	.645	.119	.407	.873
b_3	3	1.094	.138	.841	1.366
b_4	4	-.321	.114	-.556	-.094
b_5	5	2.143	.206	1.768	2.563
b_6	6	.513	.121	.284	.762
b_7	7	1.328	.154	1.046	1.634
b_8	8	.299	.120	.062	.543
b_9	9	-.132	.117	-.366	.102
κ_1	10	-1.498	.151	-1.803	-1.203
κ_2	10	-.749	.131	-1.008	-.497
κ_1	11	.656	.132	.402	.924
κ_2	11	1.557	.164	1.245	1.888
κ_1	12	.625	.131	.359	.884
κ_2	12	.817	.133	.555	1.077
Population Parameters					
σ^2_θ		.953	.133	.721	1.239
σ^2_b		.836	.498	.319	2.066
μ_b		.691	.312	.073	1.310
π_{01}		.037	.022	.004	.086
π_{02}		.017	.013	.001	.049

the forced responses. Following Himmelfarb (2008), that proposed a correction method, Cronbach's alpha equals .86, which shows that the items slightly better correlate under the RR questioning technique.

The mixture randomized item response model was fitted using WinBUGS (Lunn et al., 2000), with a burn-in period of 5,000 followed by 10,000 iterations. The computer code is given in Appendix E.

Table 3.1, represents the model parameter estimates. The estimated item characteristics span a wide range, which is useful for assessing accurately different smoking behaviors. For the binary items, the item population distribution has a mean of .691 and a variance of .953, which indicates the relatively large variation in item difficulties. The smoking behavior population distribution has a mean of zero, to identify the latent scale, and a variance of .953. Participants endorsing the items with high values of threshold parameters score higher on the latent scale, and show a stronger positive smoking behavior. The low estimated thresholds of item 10 reveal that most subjects confirmed to have smoked at least a few years, where most of them (around 80%) smoked more than 15 years. Most of the patients are aware that smoking is an unhealthy habit, which follows from the scores on item five.

Mixture components (compliant and non-compliant class) were implemented for the binary and polytomous items to control for positive self-report behavior of subjects not following the RR instructions. It follows that around 3.7% (π_{01}) and 1.7% (π_{02}) of the patients were detected as non-complying given the binary and polytomous response patterns, respectively.

3.3.1 RRT Validation

Participants were randomly assigned to the RR group or the DQ group. Therefore, it is to be expected that mean smoking behavior differences are only attributable to the questioning technique. Furthermore, the guarantee of confidentiality inherent to the RR technique will induce subjects in the RR group to score more honestly and therefore more strongly correlate with the (physical) CO measurement than those in the DQ group.

The mean structure of population distribution of the latent smoking behaviors is partitioned to identify between-group differences in latent smoking behaviors. The following structural population model is defined,

$$\theta_i = \beta_0 + \beta_1 RR_i + \beta_2 CO_i + \beta_3 RR_i CO_i + \epsilon_i, \tag{3.11}$$

where the intercept, β_0, represents the mean smoking behavior of the DQ group, β_1 and β_2 represent the effect of RRT and of CO-measured smokers, respectively. The interaction effect, β_3, denotes the effect of CO-measured smokers in the RR group.

Model 2 contains the population distribution defined in Equation 3.11, and model 1 only contains the main effects. Model 3 contains a random day effect to account for unexplained additional dependencies between patients, which were assessed the same day. The WinBUGS program in Appendix E was extended to estimate all structural effects, which are reported in Table 3.2.

Model 1 estimates show that the average smoking behavior is around -.83 for the DQ group, and significantly higher for the RR group with an increase of around .47. Patients diagnosed as smokers score on average 1.75 higher than those diagnosed as non-smokers according to the CO measurement. This strong relationship between the latent smoking behavior scores and the CO scores indicates that the response data can be used to classify patients as smokers and non-smokers.

Model 2 contains an interaction effect to evaluate a differential effect of CO measurements over questioning techniques. It follows that the average latent score is -.69 for the DQ group. Patients in the DQ group detected as smokers (CO=1) score on average 1.39 higher compared to the non-smokers in the DQ group. The latent mean of subjects in the RR group is .24 higher. Furthermore, in the RR group smokers (CO=1) score on average 1.39 plus .62 higher than the non-smokers, which is significantly higher compared to the DQ group.

It follows that a significant RR effect is detected for the non-smokers. The non-smokers in the RR group show on average a higher smoking behavior than the non-smokers in the DQ group. The mean smoking behavior difference between questioning techniques is relatively small since the items are less sensitive for non-smokers. The CO measure might be negative when a smoker abstains for a period of 12 hours and, in that case, the self-report under RR may lead to a relatively

Table 3.2: Smoking behavior: Influence of RRT and CO-measurement and accounting for random day effects.

	Model 1			Model 2			Model 3		
Parameter	Mean	SD	HPD	Mean	SD	HPD	Mean	SD	HPD
β_0 (Intercept)	-.83	.07	-.97,-.70	-.69	.08	-.84,-.54	-.70	.10	-.92,-.55
β_1 (RR)	.47	.09	.28, .64	.24	.11	.04, .43	.27	.14	.04, .56
β_2 (CO)	1.75	.08	1.58,1.91	1.39	.14	1.10,1.63	1.40	.13	1.14,1.66
β_3 (RR x CO)				.62	.19	.21, .93	.52	.18	.13, .85
Residual Variance									
σ^2	.31	.04	.28, .38	.28	.05	.17, .37	.29	.04	.20, .36
Random Day effect									
τ^2							.05	.02	.01, .09

Note: HPD = 95% highest posterior density

high smoking behavior score. The more interesting case concerns the smokers in both groups. The estimated interaction effect is around 2.5 times higher than the main RR effect, which means that exactly smokers score significantly higher in the RR condition. This result validates the RR technique. The smokers experience the items as more sensitive than the non-smokers, and they score significantly higher in the RR condition, which corresponds to the true smoking status according to the CO measurements.

In Figure 3.1, the EAP latent scores are plotted for the smokers (CO=1) and non-smokers in the DQ an RR group. It follows that the estimated latent scores of the non-smokers in the DQ and RR group do not differ much, but they differ for the smokers. Furthermore, the estimated latent scores differ more between smokers and non-smokers in the RR group than in the DQ group. This illustrates that the RR technique improves the quality of the self-reports for those that experience the items as sensitive.

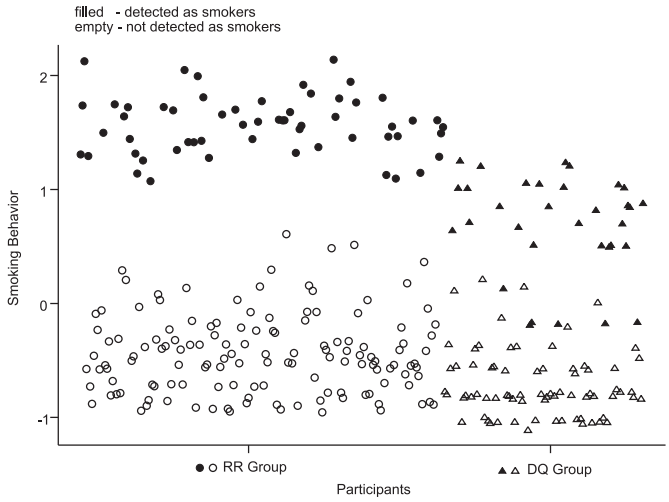

Figure 3.1: Estimated latent behaviors of smokers and non-smokers, diagnosed with the CO measurement, in the DQ and RR group.

In model 3, a random day effect is introduced that account for dependencies between latent scores given the mean structure of model 2. It can be seen that the latent mean scores differ over days, with a variance of .05. Around 15% of the variation in latent scores is explained by the nesting of patients in days. This might be caused by synchronizing hospital visits of patients with comparable background characteristics. This relates to the fact that pulmonologists have different specializations and are working on a fixed-shift basis, which supports the clustering of comparable patients in days. Accounting for random variation across days hardly influences the group effects, since the estimates of model 2 and 3 do not differ much.

Different background information was collected to explain variation in latent smoking behaviors. Therefore, the population distribution in Equation 3.11 was

Table 3.3: Smoking behavior: Influence of age and type of disease.

	Model 4		Model 5	
Parameter	M	SD	M	SD
Intercept	-.72	.11	-.52	.11
RR	.22	.15	.21	.15
CO status	1.43	.14	1.34	.14
RR x CO status	.52	.19	.54	.19
Age 17-40	.08	.17		
Age 41-60	-.16	.12		
Age 17-40 x RR	-.32	.23		
Age 41-60 x RR	.31	.15		
Asthma			-.25	.17
Lung Cancer			-.19	.41
Bronchitis			-.41	.38
Other			-.43	.17
Asthma x RR			-.18	.24
Lung Cancer x RR			.38	.47
Bronchitis x RR			.03	.46
Other x RR			.48	.26
σ^2	.27	.04	.25	.04

extended with explanatory variables: gender, education, age, and type of disease. It was not possible to simultaneously investigate all main effects and (higher level) interactions due to the relatively small data set. The random day effect was removed from the model since background variable(s) explained significant variation between days. There were no significant differences detected between males and females in latent scores conditional on the mean structure in Equation 3.11. A significant effect of education was also not found.

Three age groups were considered: 17-40, 41-60, and above 60, where the last group was the baseline. In Table 3.3, the parameter estimates are given of the main and interaction effects with RR under model 4. It can be seen that the middle group scored lower but the effect is not significantly different from zero. The positive significant interaction effect of the middle age group with RRT reveals that subjects in this age group score significantly higher under protection of the RR technique. The RR technique prove to be more powerful for this age group.

The different diseases were categorized as; asthma, lung cancer, bronchitis, and a group other. It follows that patients classified in group other (group effect -.43) score significantly lower under direct questioning. However, they score significantly higher in the RR group. Although not significant this pattern is apparent for all groups except the disease group asthma.

3.3.2 Bayesian Diagnostic Evaluation of Randomized Response Testing

To compute basic measures of diagnostic test accuracy of the questionnaire under the randomized response design the CO measure was excluded from the measurement model. The expected number of diagnosed smokers in the sample was computed given the response data but independently of the observed CO measures. Subsequently, the CO measures were used to validate the quality of the model predictions based on the test data.

According to Equation 3.4, the posterior probability of smoking per subject given the response data was computed for threshold values ranging from zero to two. This interval covers the posterior smoking behavior estimates above the mean since the population mean and standard deviation of the smoking behaviors in the RR group equals .33 and .78, respectively. Subsequently, for each subject and threshold value posterior predictive diagnoses were simulated under the model according to Equation 3.5. Each posterior predictive sample of diagnoses, $\mathbf{x}$, based on the response data was compared to the true diagnoses based on the CO measure.

In Figure 3.2, the estimated expected posterior sensitivity (true positive fraction) is given per threshold value together with 95% highest posterior density (HPD) intervals. It can be seen that for small threshold values, around the la-

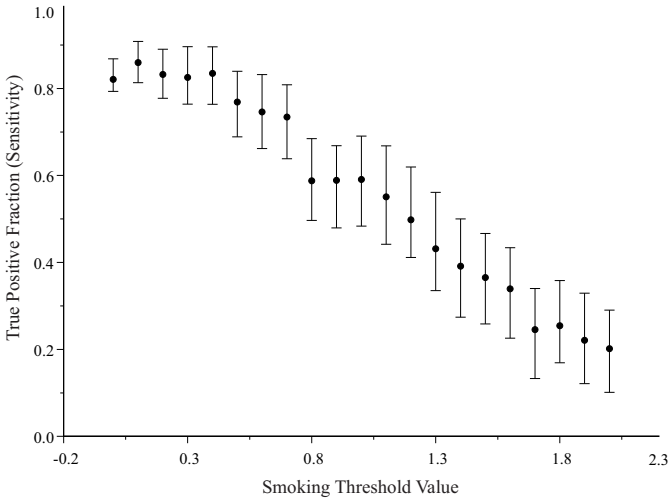

Figure 3.2: The posterior sensitivity as a function of threshold values from zero to two based using response data from subjects in the RR group.

tent population mean, the posterior probability of diagnosing smokers correctly on basis of the response data is around 85%, and for lower threshold values (not plotted) this probability goes to one. When increasing the higher threshold value, the true positive fraction decreases and the uncertainty increases, which follows from the increase of the HPD interval. The posterior latent mean of subjects in the RR group detected as smokers equals 1.31. When the threshold value equals this posterior latent mean ($\theta_c = 1.3$), the sensitivity is around .43. Each sensitivity

value is based on smoking diagnoses simulated from the smoking behavior posterior distributions. The additional posterior uncertainty of each subject-specific latent smoking behavior causes that the sensitivity decreases more slowly towards zero as the threshold value increases.

In Figure 3.3, the estimated posterior specificity (one minus false positive fraction) and 95% HPD interval is given per threshold value using the randomized response data. For high threshold values the posterior probability of smoking goes to zero since all subjects are classified as non-smokers. The HPD intervals are relatively small since the posterior latent smoking behavior distributions support the decision of classifying subjects as non-smokers for such high threshold values. The probability of correctly classifying non-smokers decreases when the threshold value decreases, where the uncertainty increases. When the threshold value equals 1.3, the posterior specificity equals .90. When the threshold value equals .70, both the sensitivity and specificity are around .73. The posterior latent smoking behavior estimates are subject to considerable uncertainty due to the limited number of items and noise introduced by the randomized response mechanism. Together with the limited number of subjects in the study, the optimal posterior classification probabilities (specificity and sensitivity) are around .73.

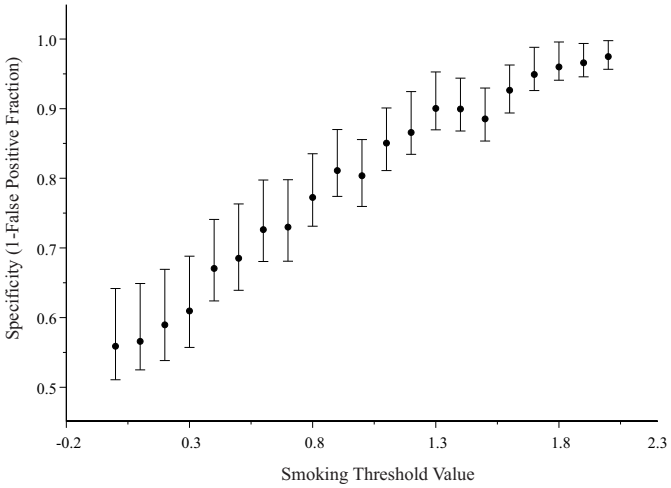

Figure 3.3: The posterior specificity as a function of threshold values from zero to two using response data from subjects in the RR group.

For $\theta_c = .7$, the posterior mean PPV equals .76 for the RR data and .60 for the DQ data. In the same way, the negative posterior predictive value (NPV) equals .89 and .84 for the RR and DQ group, respectively. For a perfect test, both probabilities are equal to one, but the mentioned different types of uncertainty reduce both predictive probability estimates. Furthermore, the posterior distributions of the PPV and NPV are skewed to the left due to the natural upper bound of one. The sampling-based algorithm renders posterior means, which are most likely conservative underestimates of the true values.

A comparison between the DQ and RR group of posterior classification prob-

abilities is complicated due to scale differences, which makes it problematic to define one common set of threshold values over groups. Furthermore, the DQ group shows serious underreporting, which increases the probability and accuracy of classifying non-smokers correctly. The consistent low scores in the DQ group significantly improves the test accuracy when it also concerns non-smokers, which are 71% of the respondents in the DQ group. From that perspective, the test accuracy only increases using the RR technique when it concerns sensitive information. In Figure 3.4, the estimated smoking behavior of smokers (D=1) and non-smokers (D=0) in the DQ and RR group are plotted against the log-likelihood of the corresponding response patterns. For the non-smokers (D=0), it can be seen that a relatively large group of subjects in the RR condition have relatively low log-likelihood values, which indicates that on average the response patterns in the DQ condition fit better. However, for the smokers (D=1), the subjects in the RR conditions have higher log-likelihood values. On average the response patterns of smokers in the RR condition fit better and correspond to higher smoking behavior values than those in the DQ condition.

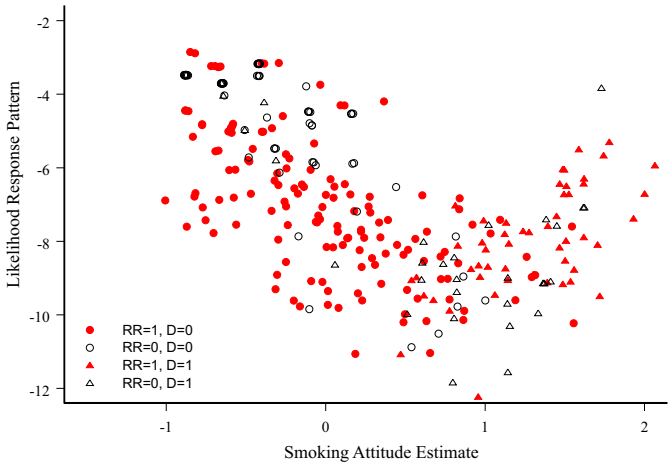

Figure 3.4: For each subject, the estimated latent smoking behavior against the log-likelihood of the response pattern

3.4 Discussion and Conclusions

In the present sensitive survey study, smoking behaviors of outpatients of a Dutch pulmonary department were assessed using different questioning techniques. It was shown that the randomized response technique led to more accurate responses in comparison to the direct-questioning technique. Therefore, CO measurements were used as a gold standard to evaluate test diagnoses based on (randomized) item response data. Participants detected as smokers scored significantly higher in the RR condition compared to smokers in the DQ condition. Such an apparent difference in scoring was not detected for the non-smokers, although non-smokers

scored slightly higher in the RR condition. Non-smokers did not perceive the smoking questionnaire as sensitive and the questioning technique hardly influenced the test results. The randomized response technique positively influenced the quality of smokers' response data according to the gold standard, which validates the multi-item randomized response technique.

Although not for the item randomized-response technique, a few studies did use true individual information about the sensitive behavior to validate the (univariate) randomized response technique. van der Heijden et al. (2000) reported about respondents caught for welfare or unemployment benefit fraud and investigated the percentage of underreporting without having a control group. Akers et al. (1983) used a physiological measure, salivary thiocyanate, to identify a respondent's smoking behavior. They investigated smoking behavior among adolescents and found approximately similar results with both the randomized-response and direct-questioning method. They concluded that the participants might have experienced smoking as non-sensitive behavior. However, the participants were informed that saliva samples were being collected after the survey, which was to convince the subjects that their self-reports could be verified. As a result, this feature of the study design stimulated truthful reporting in both RR and DQ groups. This so-called bogus pipeline (Jones & Sigall, 1971) probably diminished the effect of the RRT.

The developed mixture randomized item response model enabled the simultaneous analysis of mixed item response data (binary and ordinal) that were obtained using the RR or DQ technique. A mixture modeling approach was pursued to account for subjects that ignore the RR instruction and consistently scored in the least incriminating response category. Therefore, a non-compliance group was defined as a separate latent class for the ordinal and binary response data.

The subject-specific posterior probability that the (continuous) latent smoking behavior variable is greater than a specified threshold value defined the posterior probability of smoking given the response data. This posterior probability distribution was used to simulate smoking diagnoses to evaluate the diagnostic accuracy of the self-report smoking questionnaire in the RR condition. This innovative nested latent variable approach enabled the computation of the posterior sensitivity and specificity given self-reported randomized response data and CO measurements. Further research in this area is needed to investigate the influence of different randomized response techniques, test length, and item sensitivity on the test accuracy. The presented Bayesian latent variable framework for evaluating the test accuracy can also be extended to ordinal diagnostic measurements.

Chapter 4

A Multidimensional Randomized Item Response Model

Abstract

Randomized response (RR) models are often used for analyzing univariate randomized response data and measuring population prevalence of sensitive behaviors. There is much empirical support that RR methods improve the cooperation of the respondents. Recently, RR models have been extended to measure individual unidimensional behavior. An extension of this modeling framework is proposed to measure multiple sensitive factors underlying the randomized item response process. A multidimensional randomized item response theory model (MRIRT) is developed for the analysis of multivariate RR data by modeling the response process and specifying structural relationships between sensitive behaviors and background information. An MCMC algorithm is developed to estimate simultaneously the parameters of the MRIRT model. The model extension enables the computation of individual true item response probabilities, estimates of individuals' sensitive behavior on different domains, and their relationships with background variables. An MRIRT analysis is presented of data from a college alcohol problem scale (CAPS), measuring alcohol-related socio-emotional and community problems, and alcohol expectancy questionnaire (AEQ), measuring alcohol-related sexual enhancement expectancies. Students were interviewed via direct or randomized-response questioning. Scores of alcohol-related problems and expectancies are significantly higher for the group of students that were questioned using the randomized response technique. Alcohol-related problems and sexual enhancement expectancies are positively moderately correlated and vary differently across gender and universities.

Key words: multidimensional item response theory model, MCMC, randomized response data.

4.1 Introduction

In sample surveys, it can be difficult to obtain reliable information on stigmatizing or socially undesirable/unacceptable matters using the common direct questioning procedure. Direct questioning of sensitive questions often leads to refusals, non-responses, or socially desirable answers. Warner (1965) developed the randomized response (RR) technique to gather information on such sensitive matters by protecting the privacy of the respondents. It is shown in several studies (e.g., Lensvelt-Mulders, Hox, van der Heijden, and Maas, 2005) that the cooperation of respondents improved due to the RR technique. Besides the evident usefulness of the RR technique, inferences from applications utilizing them are limited to estimating population proportions. Further, the traditional RR models (e.g., Greenberg et al., 1969; Warner, 1965) are only appropriate for the analysis of univariate RR data, they do not account for individual response probabilities, and they do not allow for heterogeneity across respondents. In many cases, outcome data are multivariate or correlated, and it is appealing to model the individual outcomes while taking account of the dependency structure.

To motivate the problem, questionnaire data to measure different psychosocial dimensions of problem drinking among college students were collected on 793 students from four colleges/universities in North Carolina. Responses to alcohol expectancy questionnaire items were collected, which measure alcohol-related sexual enhancement expectancies. A part of the participants was questioned via a randomized response technique to investigate whether this technique increases the accuracy of self-reports of sensitive information on different latent dimensions. Further, interest is focused on relationships between latent factors, and their relationship with background variables (e.g., age, gender, racial origin).

Several attempts have been made to extend the class of RR models by modeling the item response process and/or by including various sources of information such as ancillary variables. Scheers and Dayton (1988) modeled the relation between the population proportion with the sensitive characteristic and covariate information. They showed that the use of relevant covariate information improved the estimation of the population proportion with the sensitive characteristic. Böckenholt and van der Heijden (2007) and Böckenholt, Barlas, and van der Heijden (2009) proposed an item randomized response model for multivariate dichotomous RR data that accounts for the possibility (1) that not all respondents may follow the randomization instructions (say "No" regardless of the question), and (2) that respondents provide intentionally misleading answers to conceal socially undesirable behavior. Fox (2005a) and Fox and Wyrick (2008) modeled multivariate dichotomous and polytomous data with a randomized IRT model that accounts for individual differences in the response process, and that enables the computation of individual response probabilities, and the measurement of an unidimensional underlying sensitive characteristic. De Jong et al. (2010) developed a mixture randomized item response theory model for polytomous randomized responses.

The IRT modeling approach of RR data is generalized to measure multiple latent sensitive characteristics together with relationships with background variables using a structural multivariate regression model, given dichotomous or polytomous randomized responses. The situation is considered in which multiple

latent factors underly the manifest randomized responses in a compensatory or non-compensatory way. This study extends the work of Böckenholt and van der Heijden (2007), who proposed a between-item multidimensional item randomized-response model for binary response data. In their study, multiple item bundles are considered, where each item bundle is used to measure a specific construct using the Rasch model in a non-compensatory way. At the level of observations, responses are assumed to be conditionally independently distributed given one of the factors.

Here, in the spirit of multidimensional confirmatory item-factor models, a Bayesian multidimensional confirmatory IRT model is proposed for dichotomous and polytomous data to measure and relate factors underlying the individual sensitive characteristics given randomized response observations. The unobservable factors can be interpreted as a combination of sub-scale components, or as compensatory or non-compensatory factors that influence the item probabilities in a combined way. This modeling approach connects with recent developments in multidimensional IRT research showing the computational feasibility and increasing attention in the methodology (e.g., Chambers, 2010; Edwards, 2010; Reckase, 2009; Wirth & Edwards, 2007).

The proposed model consists of three components. At the first stage the multivariate RR data are related to individuals response probabilities via an RR model. At a second stage, the response process is modeled by assuming a multidimensional IRT model for the underlying true responses, that would have been observed when the responses were not randomized. This enables the measurement of individual response probabilities and individual latent sensitive characteristics. At a third stage, the latent sensitive characteristics are considered to be outcomes of multivariate regression model. This enables a marginal interpretation for the individual outcomes while appropriately accounting for the dependency structure. The multivariate model has the advantage that the dependency structure can be described parsimoniously in terms of correlation coefficients of the underlying latent characteristics. That is, heterogeneity across respondents and across groups can be properly modeled.

An MCMC algorithm is developed for estimating simultaneously all parameters. It is shown that a posterior computation can proceed through a Gibbs sampling algorithm using auxiliary variables. Two augmentation steps facilitate a sampling-based approach for estimating simultaneously all model parameters. First, discrete variables are defined that represent true item responses that would have been observed without randomizing responses. A conditional distribution of latent true item responses given randomized responses are derived via Bayes' Theorem. Second, normally distributed latent variables are defined that are manifested as discrete true item responses through a threshold specification. The developed algorithm generalizes the procedure of Fox and Wyrick (2008) and De Jong et al. (2010) to deal with the multidimensional IRT model and the structural multivariate latent variable component.

In the following the three-stage MRIRT model is presented. Next, a general MCMC algorithm is described for dichotomous and polytomous randomized response data. Different prior choices are discussed that lead to proper posterior distributions. Then, the posterior computations are illustrated with a simulation

study. Subsequently, a description is given of the joint Bayesian multidimensional IRT analysis of the CAPS and AEQ randomized response data with careful attention to the underlying factor structure. Finally, the pros and cons of the new model are discussed and briefly compared with other approaches in the literature.

4.2 Modeling Individual Response Probabilities

In general, the RR technique is used to estimate the proportion, π, of respondents belonging to a sensitive class in the population. Horvitz, Shah, and Simmons (1967) proposed an unrelated question RR design that is based on two questions. The provocative question relates to the sensitive characteristic, and the other, unrelated question, refers to a non-sensitive innocuous attribute. Each respondent randomly selects, by means of a randomizing device such as a die or spinner, one of the two questions and answers it truthfully. The respondent does not tell the interviewer which question he has selected to answer. If the population proportion of the non-sensitive characteristic is known, and this is build into the randomizing device (see, Boruch, 1971b); the probability of a positive response will be

$$n_1/n = p_1\pi + (1 - p_1)p_2 \tag{4.1}$$

where n_1 (n_1/n) is the number (proportion) of "yes" answers reported by n individuals, and p_1 is the probability that the randomizing device selects the question related to the sensitive characteristic. Subsequently, with probability $(1 - p_1)$ the unrelated question is selected and with probability p_2 a positive response is given. Note that parameters p_1 and p_2 are known since they are specified by the randomizing device. Using Equation 4.1 the proportion of persons affirming or denying the intrusive item can be accurately estimated. De Schrijver (2012) compared the forced response technique with the unrelated questioning technique and concluded that the forced response technique was better understood.

In case of multivariate randomized response data, let random variable $\mathbf{Y}_i$ denote the observed randomized responses and random variable $\mathbf{U}_i$ the true responses that would have been observed when these responses were not randomized. The distribution of observed randomized responses relates to the distribution of true item responses according to the RR sampling design; that is,

$$P(\mathbf{Y}_i) = P(Y_{i1}, \ldots, Y_{iK}) \quad = \quad \prod_k \big[p_1 P(U_{ik}) + (1 - p_1)p_2\big]. \tag{4.2}$$

The true individual response probabilities $P(U_{ik})$ will be modeled, which can improve the estimate of the proportion with the sensitive characteristic in the population depending on the available explanatory information. It will provide information at the individual level since estimates of individual response probabilities can be obtained as well as relationships with explanatory variables.

4.3 The Model

4.3.1 Probit Response Functions

It will be assumed that a set of K items are composed to measure a multidimensional latent trait, $\boldsymbol{\theta}_i$, where $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iQ})^t$. Let, $\boldsymbol{\theta}_q$, $q = 1, \ldots, Q$, contain the traits of respondents $i = 1, \ldots, N$ on dimension q. The true categorical outcome, U_{ik}, represents the item response of person i on item k ($k = 1, \ldots, K$). These item responses may be dichotomous or polytomous. For dichotomous item responses a two-parameter Q-dimensional IRT model is used for specifying the relation between the level of a latent trait and the probability of a particular item response; that is,

$$\begin{aligned} P\big(U_{ik} = 1 \mid \boldsymbol{\theta}_i, \boldsymbol{A}_k, b_k\big) &= \Phi\Big(\boldsymbol{A}_k^t \boldsymbol{\theta}_i - b_k\Big) \\ &= \Phi\Big(\sum_q A_{kq}\theta_{iq} - b_k\Big) \end{aligned} \tag{4.3}$$

where $\boldsymbol{A}_k$ is the vector of item discrimination parameters or factor loadings, and b_k is the item difficulty parameter. The item parameters will also be denoted by $\boldsymbol{\xi}_k$, with $\boldsymbol{\xi}_k = (\boldsymbol{A}_k, b_k)$. For polytomous item responses, the probability that an individual obtains a grade c ($c = 1, \ldots, C$) on item k is defined by a graded response model

$$\begin{aligned} P\big(U_{ik} = c \quad \mid \quad & \boldsymbol{\theta}_i, \boldsymbol{A}_k, \boldsymbol{\kappa}_k\big) \\ &= \Phi\Big(\boldsymbol{A}_k^t \boldsymbol{\theta}_i - \kappa_{kc-1}\Big) - \Phi\Big(\boldsymbol{A}_k^t \boldsymbol{\theta}_i - \kappa_{kc}\Big) \\ &= \Phi\Big(\sum_q A_{kq}\theta_{iq} - \kappa_{kc-1}\Big) - \Phi\Big(\sum_q A_{kq}\theta_{iq} - \kappa_{kc}\Big) \end{aligned} \tag{4.4}$$

where the boundaries between the response categories are represented by an ordered vector of thresholds $\boldsymbol{\kappa}$ and let $\boldsymbol{\xi}_k = (\boldsymbol{A}_k, \boldsymbol{\kappa}_k)$. There are a total of $C - 1$ threshold parameters and Q discrimination parameter for each item. For the logistic IRT model replace $\Phi(.)$ with $\mathcal{L}(.)$. Although the polytomous IRT model in Equation 4.4 also comprehends the two-parameter IRT model, the two-parameter IRT model is presented separately according to Equation 4.3. Reckase (2009) gives an overview of multidimensional IRT models.

4.3.2 Forced Randomized Response Design

According to the forced randomized response design, two probabilities are specified by the randomizing device, probability p_1 that the respondent has to answer the sensitive question, probability p_2 a forced positive response is given. Using Equations 4.2 and 4.3, the probability of a positive randomized response equals,

$$\begin{aligned} P(Y_{ik} = 1 \mid \boldsymbol{\theta}_i, \boldsymbol{A}_k, b_k) &= p_1 P\big(U_{ik} = 1 \mid \boldsymbol{\theta}_i, \boldsymbol{A}_k, b_k\big) + (1 - p_1)p_2 \\ &= p_1 \Phi\Big(\boldsymbol{A}_k^t \boldsymbol{\theta}_i - b_k\Big) + (1 - p_1)p_2. \end{aligned} \tag{4.5}$$

This is easily extended to multiple, say $c = 1, \ldots, C$, response categories. The randomizing device determines if the item is to be answered honestly with probability p_1 or a forced response is scored in category c with probability $(1-p_1)p_2(c)$. Using Equations 4.2 and 4.4, the probability of scoring in category c equals

$$\begin{aligned} P(Y_{ik} = c \mid \boldsymbol{\theta}_i, \boldsymbol{A}_k, \boldsymbol{\kappa}_k) &= p_1 P\big(U_{ik} = c \mid \boldsymbol{\theta}_i, \boldsymbol{A}_k, \boldsymbol{\kappa}_k\big) + (1-p_1)p_2(c) \quad (4.6) \\ &= p_1\Big[\Phi\Big(\boldsymbol{A}_k^t\boldsymbol{\theta}_i - \kappa_{kc-1}\Big) - \Phi\Big(\boldsymbol{A}_k^t\boldsymbol{\theta}_i - \kappa_{kc}\Big)\Big] \\ &+ (1-p_1)p_2(c). \end{aligned}$$

4.3.3 Structural Multivariate Latent Model

In multidimensional IRT, it is often assumed that the latent traits $\boldsymbol{\theta}_i$ for $i = 1, \ldots, N$ are exchangeable and multivariate normally distributed. That is, $\boldsymbol{\theta}_i \sim \mathcal{N}_Q\big(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta\big)$ where the population means of latent traits and variations and dependencies among latent trait dimensions are not depending on any individual information. This model specification can be extended to include clustering of background information of respondents. In general, a multivariate model makes it possible to investigate correlations between latent traits, its dependence on the individual and the group level, to test simultaneously effects of an explanatory variable on several latent traits, and to test for differential effects of explanatory variables on various latent traits.

Let S predictor variables for respondent i be stored in $\boldsymbol{X}_i^t = (X_{i1}, \ldots, X_{iS})$. A structural multivariate latent model for the latent traits is expressed by

$$\begin{pmatrix} \theta_{11} & \cdots & \theta_{1Q} \\ \theta_{21} & \cdots & \theta_{2Q} \\ \vdots & \ddots & \vdots \\ \theta_{N1} & \ldots & \theta_{NQ} \end{pmatrix} = \begin{pmatrix} X_{11} & \cdots & X_{1S} \\ X_{21} & \cdots & X_{2S} \\ \vdots & \ddots & \vdots \\ X_{N1} & \ldots & X_{NS} \end{pmatrix} \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1Q} \\ \gamma_{21} & \cdots & \gamma_{2Q} \\ \vdots & \ddots & \vdots \\ \gamma_{S1} & \cdots & \gamma_{SQ} \end{pmatrix}$$

$$+ \begin{pmatrix} e_{11} & \cdots & e_{1Q} \\ e_{21} & \cdots & e_{2Q} \\ \vdots & \ddots & \vdots \\ e_{N1} & \ldots & e_{NQ} \end{pmatrix},$$

that is,

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{e}, \tag{4.7}$$

where $\mathbf{e}$ is an $N \times Q$ matrix whose rows are independently distributed, $\mathbf{e}_i \sim \mathcal{N}_Q\big(\mathbf{0}, \boldsymbol{\Sigma}_\theta\big)$. The $S \times Q$ matrix $\boldsymbol{\gamma}$ contains unknown fixed effects parameters. The latent traits on the Q-dimensions for each respondent have a covariance matrix $\boldsymbol{\Sigma}_\theta$ but the latent traits are uncorrelated across individuals. A complete overview of multivariate multiple regression models can be found in Johnson and Wichern (2002).

The within-subject covariance matrix $\boldsymbol{\Sigma}_\theta$ can be modeled as functions of unknown covariance parameters, where each structure for $\boldsymbol{\Sigma}_\theta$ can have important

subject-specific interpretations. The efficiency of the regression parameter estimates may be improved by modeling the covariance matrix parsimoniously as independent, with a constant variance parameter across subject's latent trait dimensions,

$$\boldsymbol{\Sigma}_\theta = \sigma^2 \mathbf{I}_Q, \tag{4.8}$$

where $\mathbf{I}_Q$ is a $Q \times Q$ identity matrix.

Other cases arise by allowing the variance σ^2 to vary across latent trait dimensions, or by allowing the variance to vary across groups of subjects. In both cases $\boldsymbol{\Sigma}_\theta$ remains a diagonal matrix. The general extension is the unstructured covariance matrix with $Q(Q+1)/2$ parameters, and possibly with covariance parameters that vary across groups.

Finally, in a multivariate mixed effects modeling approach, variation can be modeled in latent trait dimensions within and between groups of subjects, for example, due to treatment effects. Following Schafer and Yucel (2002), let n_j $(j = 1, \ldots, J)$ denote the number of subjects in cluster j, and let $\mathbf{W}_j$ be an $n_j \times R$ design matrix linking to the random effect parameters $\boldsymbol{\zeta}_j$, an $(R \times Q)$ matrix of coefficients specific to subjects in cluster j. With these random effects, Equation 4.7 is generalized to

$$\boldsymbol{\theta}_j = \mathbf{X}_j \boldsymbol{\gamma} + \mathbf{W}_j \boldsymbol{\zeta}_j + \mathbf{e}_j, \tag{4.9}$$

where $\mathbf{X}_j$ is a $n_j \times S$ matrix of explanatory variables. The n_j rows of $\mathbf{e}_j$ are assumed to be independently distributed as $\mathcal{N}_Q\big(\mathbf{0}, \boldsymbol{\Sigma}_\theta\big)$. The random effects are independently distributed as $vec(\boldsymbol{\zeta}_j) \sim \mathcal{N}\big(\mathbf{0}, T\big)$, where the $vec(.)$ operator creates a column vector from the columns of $\boldsymbol{\zeta}_j$. It is assumed that for clusters j and j', $cov(\mathbf{e}_j, \mathbf{e}_{j'}) = 0$, $cov(\boldsymbol{\zeta}_j, \boldsymbol{\zeta}_{j'}) = 0$, and $cov(\mathbf{e}_j, \boldsymbol{\zeta}_{j'}) = 0$. Then, the covariance matrix for $vec(\boldsymbol{\theta}_j)$ equals

$$\big(\mathbf{I}_Q \otimes \mathbf{W}_j\big) T \big(\mathbf{I}_Q \otimes \mathbf{W}_j\big)^t + \boldsymbol{\Sigma}_\theta \otimes \mathbf{I}_{n_j}. \tag{4.10}$$

A well-known structure is the case where each $\mathbf{W}_j$ is a vector of ones that gives the between-within mixed model structure called the compound symmetry. In case of modeling a large number of latent traits it may be advantageous to restrict T to a block-diagonal structure indicating that there are no a priori associations between the columns of $\boldsymbol{\zeta}_j$.

4.3.4 Identification Issues

The multidimensional RIRT model can be identified by imposing restrictions on some of the parameters. Following Béguin and Glas (2001), one approach is to fix the mean and covariance matrix of $\boldsymbol{\theta}$ to zero and the identity matrix, respectively. To avoid the rotational invariance constraints are necessary for some elements of $\mathbf{A}$, for example, $A_{kq} = 0$ for $k = 1, \ldots, Q-1$ and $q = 2, \ldots, Q$. Furthermore, at least one element in each column of $\mathbf{A}$ is constrained to be positive (see also Lopes and West (2004)).

It is also possible to consider the covariance parameters as unknowns by imposing the restrictions; $A_{kq} = 1$ if $k = q$, and $A_{kq} = 0$ if $k \neq q$ for $k = 1, \ldots, Q$ and $q = 1, \ldots, Q$, and setting the mean equal to zero. Instead of restricting the

upper-diagonal elements of the matrix of factor loadings, the diagonal elements of $\boldsymbol{\Sigma}_\theta$ can also be restricted to one such that only the non-diagonal elements are free parameters.

4.4 Bayesian Inference

A Markov Chain Monte Carlo (MCMC) procedure is proposed for model estimation. This MCMC algorithm is based on developed MCMC methods for multidimensional IRT and factor analytic models (e.g., Béguin and Glas, 2001; Bolt and Lall, 2003; Jackman, 2001; Lopes and West, 2004; Sheng and Wikle, 2007; Shi and Lee, 1998; Song and Lee, 2001; Yao and Boughton, 2007). A straightforward MCMC implementation is not possible, since discrete randomized response data are observed. Therefore, following the MCMC method of Fox and Wyrick (2008) for unidimensional IRT, a double augmentation step is proposed to sample the discrete true item response data and continuous latent response data given the randomized response data. As a result, the joint distribution of the parameters and augmented data are considered to circumvent a direct evaluation of the likelihood function, which is computationally intensive.

The joint posterior distribution can be expressed as

$$\begin{aligned} p\left(\mathbf{u}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \mathbf{A}, \boldsymbol{\kappa}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}_\theta, T \mid \mathbf{y}\right) &\propto p\left(\mathbf{y} \mid \mathbf{u}\right) p\left(\mathbf{u} \mid \boldsymbol{\theta}, \mathbf{A}, \boldsymbol{\kappa}\right) p\left(\boldsymbol{\theta} \mid \boldsymbol{\gamma}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\zeta}\right) p\left(\boldsymbol{\zeta} \mid T\right) \\ &\quad p(\mathbf{A}) p(\boldsymbol{\kappa}) p(T) p(\boldsymbol{\Sigma}_\theta) p(\boldsymbol{\gamma}) \\ &\propto \prod_j \left[\prod_i \left[\prod_k p\left(y_{ijk} \mid u_{ijk}\right) p\left(u_{ijk} \mid \boldsymbol{\theta}_i, \mathbf{A}_k, \boldsymbol{\kappa}_k\right) \right] \right. \\ &\quad \left. p\left(\boldsymbol{\theta}_i \mid \boldsymbol{\gamma}, \boldsymbol{\zeta}_j, \boldsymbol{\Sigma}_\theta\right) \right] p\left(\boldsymbol{\zeta}_j \mid \mathbf{T}\right) p(\mathbf{A}) p(\boldsymbol{\kappa}) p(T) p(\boldsymbol{\Sigma}_\theta) p(\boldsymbol{\gamma}), \end{aligned}$$

where $p\left(y_{ijk} \mid u_{ijk}\right)$ defines the randomized response process given the characteristics of randomizing device. The term $p\left(u_{ijk} \mid \boldsymbol{\theta}_i, \mathbf{A}_k, \boldsymbol{\kappa}_k\right)$ defines the multidimensional IRT component for the true response data, and $p\left(\boldsymbol{\theta}_i \mid \boldsymbol{\gamma}, \boldsymbol{\zeta}_j, \boldsymbol{\Sigma}_\theta\right)$ the multivariate model component for the factors.

The structural multivariate parameters and the multidimensional IRT parameters are sampled from the full conditionals, where only the threshold parameters $\boldsymbol{\kappa}$ are sampled using a Metropolis-Hastings step. The full conditionals are easily derived given the augmented data. Therefore, the randomized response data $\mathbf{Y}$ will be augmented with latent data $\mathbf{U}$ and $\mathbf{Z}$. The random variable U_{ijk} represents the true response to item k of a person indexed ij. The conditional probability of a true response U_{ijk} given a randomized response Y_{ijk} is considered. Let $\pi_{ijk}(c)$ denote the probability of responding in category c according to the IRT model in Equations 4.5 or 4.6, it follows that

$$\begin{aligned} P\left(U_{ijk} = c \mid Y_{ijk} = c; \pi_{ijk}(c)\right) &= \frac{P\left(U_{ijk} = c, Y_{ijk} = c; \pi_{ijk}(c)\right)}{P\left(Y_{ijk} = c; \pi_{ijk}(c)\right)} \\ &= \frac{\pi_{ijk}(c)[p_1 + (1 - p_1)p_2]}{\pi_{ijk}(c)p_1 + (1 - p_1)p_2}, \end{aligned}$$

where p_1 and p_2 are the known randomizing device characteristics. In general, the conditional distribution of a true item response given a randomized item response is a multinomial with cell probabilities

$$\Delta(c) = \frac{\pi_{ijk}(c)p_1 I(c = c') + \pi_{ijk}(c)(1 - p_1)p_2}{\pi_{ijk}(c')p_1 + (1 - p_1)p_2}, \tag{4.11}$$

for $c, c' = 1, \ldots, C_k$. For binary response data, the conditional distribution is Bernoulli with the success probability defined in (4.11) and $c = 1$.

Subsequently, the full conditional of the augmented data $\mathbf{Z}$ is a normal distribution,

$$Z_{ijk} \mid U_{ijk}, \boldsymbol{\theta}_i, \boldsymbol{\xi}_k \sim \begin{cases} N\left(\mathbf{A}_k^t \boldsymbol{\theta}_i - b_k, 1\right) \text{ for binary data} \\ N\left(\mathbf{A}_k^t \boldsymbol{\theta}_i, 1\right) \text{ for polytomous data,} \end{cases} \tag{4.12}$$

with U_{ijk} the indicator of Z_{ijk} being positive for binary response data, and $U_{ijk} = c$ if $\kappa_{kc-1} \leq Z_{ijk} \leq \kappa_{kc}$ in case of polytomous response data.

The prior for the non-fixed factor loadings is standard normal. The full conditional for $\mathbf{A}_k$ is normal with mean $\left(\boldsymbol{\theta}^t\boldsymbol{\theta} + \mathbf{I}_Q\right)^{-1}\left(\boldsymbol{\theta}^t\left(\mathbf{Z}_k + b_k\right)\right)$ and variance $\left(\boldsymbol{\theta}^t\boldsymbol{\theta} + \mathbf{I}_Q\right)^{-1}$. If $\mathbf{A}_k$ contains fixed elements, the vector of free random loadings are sampled given the other vector of fixed loadings. The difficulty parameters are not included in the mean term when dealing with polytomous data but thresholds are introduced as additional restrictions on the space spanned by the augmented data vectors.

Noninformative proper priors for the difficulty or threshold parameters are specified. The threshold parameters in Equation 4.4 are assumed to be independent uniformly distributed subject to the condition $\kappa_{k0} < \kappa_{k1} < \ldots < \kappa_{kC_k}$ with $\kappa_{k0} = -\infty$ and $\kappa_{kC_k} = \infty$. Sampled values from the conditional distribution of the threshold parameters can be obtained using the Metropolis-Hastings algorithm, see Fox (2005a), (Fox, 2010) for specific details. For binary data, the full conditional posterior of the difficulty parameter is normal with mean

$$\frac{-\sum_i \left(Z_{ik} - \mathbf{A}_k^t \boldsymbol{\theta}_i\right)/N + \mu_b/\sigma_b}{N^{-1} + \sigma_b^{-1}} \tag{4.13}$$

and variance $(N^{-1} + \sigma_b^{-1})^{-1}$ using a normal prior with mean μ_b and variance σ_b. The mean and variance parameters are sampled from a normal and an inverse-gamma distribution, respectively, given the difficulty parameters.

Consider the multivariate regression model in Equation 4.7 for the latent variables $\boldsymbol{\theta}$ and the multidimensional IRT model in Equation 4.3. Subsequently, the full conditional distribution is normal with mean

$$\boldsymbol{\theta}_{ij} \mid \mathbf{Z}_{ij}, \mathbf{A}, \mathbf{b}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}_\theta \sim N\left(\left(\mathbf{A}^t\mathbf{A} + \boldsymbol{\Sigma}_\theta^{-1}\right)^{-1}\hat{\boldsymbol{\theta}}_{ij}, \left(\mathbf{A}^t\mathbf{A} + \boldsymbol{\Sigma}_\theta^{-1}\right)^{-1}\right) \tag{4.14}$$

where $\hat{\boldsymbol{\theta}}_{ij} = \left(\mathbf{A}^t\left(\mathbf{Z}_{ij} + \mathbf{b}\right) + \boldsymbol{\Sigma}_\theta^{-1}\boldsymbol{\gamma}^t\mathbf{X}_{ij}\right)$.

Structural Multivariate Parameters

The posterior distributions of the fixed effects parameters $\boldsymbol{\gamma}$ in Equation 4.7 follow from standard Bayesian linear model results. The prior distribution for the

multivariate regression parameter is assumed to be normal with fixed mean and variance parameter $\boldsymbol{\gamma}_0$ and Γ, respectively, it follows that,

$$\boldsymbol{\gamma} \mid \boldsymbol{\theta}, \boldsymbol{\gamma}_0, \boldsymbol{\Sigma}_\theta, \Gamma \sim N\big(\tilde{\gamma}, \boldsymbol{\Sigma}_\gamma\big)$$

with

$$\begin{aligned} \tilde{\gamma} &= \boldsymbol{\Sigma}_\gamma \left(\left(\boldsymbol{\Sigma}_\theta^{-1} \otimes \mathbf{X}^t\mathbf{X} \right) \hat{\gamma} + \Gamma^{-1}\boldsymbol{\gamma}_0 \right) \\ \boldsymbol{\Sigma}_\gamma &= \left(\left(\boldsymbol{\Sigma}_\theta^{-1} \otimes \mathbf{X}^t\mathbf{X} \right) + \Gamma^{-1} \right)^{-1}. \end{aligned}$$

Subsequently, the covariance matrix $\boldsymbol{\Sigma}_\theta$ has an inverse-Wishart distribution with $N + n_0$ degrees of freedom and scale matrix S_θ^{-1} with $S_\theta = \sum_j \big(\boldsymbol{\theta}_j - \mathbf{X}_j\boldsymbol{\gamma}\big)\big(\boldsymbol{\theta}_j - \mathbf{X}_j\boldsymbol{\gamma}\big)^t + I_Q$ where $n_0 \geq Q$ to specify a noninformative proper prior distribution. Note that all parameters of the covariance matrix $\boldsymbol{\Sigma}_\theta$ are free parameters when using identifying restrictions on the item parameters and factor loadings.

The posterior distribution of the fixed effects parameters and the within-subject covariance matrix given a multivariate mixed effects model according to (4.9) are slightly different due to the fact that the error terms are defined as $\mathbf{e}_j = \boldsymbol{\theta}_j - \mathbf{X}_j\boldsymbol{\gamma} - \mathbf{W}_j\boldsymbol{\zeta}_j$, and by allowing variation across clusters. According to Equation 4.9, it follows that the posterior distribution of the random effect parameters is a product of normal distributions. As a result,

$$vec\big(\boldsymbol{\zeta}_j\big) \mid \boldsymbol{\theta}_j, \boldsymbol{\Sigma}_\theta, \boldsymbol{\gamma}, T \sim N\big(vec\big(\hat{\boldsymbol{\zeta}}_j\big), \Omega_\zeta\big), \tag{4.15}$$

where

$$\begin{aligned} vec\big(\hat{\boldsymbol{\zeta}}_j\big) &= \Omega_\zeta\big(\boldsymbol{\Sigma}_\theta^{-1} \otimes \mathbf{W}_j^t\big) vec(\boldsymbol{\theta}_j - \mathbf{X}_j\boldsymbol{\gamma}) \\ \Omega_\zeta &= \big(T^{-1} + \boldsymbol{\Sigma}_\theta^{-1} \otimes \mathbf{W}_j^t\mathbf{W}_j\big)^{-1}. \end{aligned}$$

Subsequently, when T is block diagonal, independent draws from a inverse-Wishart distribution with degrees of freedom $\nu + J$ and scale parameter $\Lambda_\zeta + \sum_j \boldsymbol{\zeta}_{jq}\boldsymbol{\zeta}_{jq}^t$, where $\boldsymbol{\zeta}_{jq}$ denotes the qth column of $\boldsymbol{\zeta}_j$. The hyperparameter ν is usually set to R to specify a disperse prior and a priori independence is assumed among the random effects, that is, $\Lambda_\zeta = \mathbf{I}_R$.

4.4.1 Implementation Issues

The MCMC algorithm can handle randomized response data as well as direct-questioning data since the properties of the randomizing device are known and corresponding parameters can be set to specific values. For direct-questioning data, p_1 is set to one. This corresponds with the approach of Chaudhuri and Mukerjee (1988), they permitted an option for direct questioning to those who volunteer to reveal the truth viewing the attribute not stigmatizing enough.

Ignorable missing response values are handled by sampling latent augmented data Z_{ijk} without truncating the values to a specific domain but based on given values of the item parameters and the latent variable. This imputation-based procedure creates a complete data set and the procedure is easily implemented

in the MCMC algorithm. The imputed augmented values have larger standard deviations since they are not restricted to a specific domain such that uncertainty due to missing values is taken into account.

The convergence of the MCMC algorithm is depending on several factors. Convergence can be slow when the amount of missing information is high. In that case the latent person parameters and item parameters are poorly estimated with large variances. Also more iterations might be needed for obtaining stable parameter estimates. Convergence can also be slow when the number of clusters is large, because for large J the posterior distribution for T given $\boldsymbol{\zeta}$ becomes very tight, and, as a result, a drawn value of the covariance matrix T remain close to its previous value. Convergence can be informally assessed by examining trace plots, time-series plots, plots of the average of each parameter across multiple chains, and plots of the running average. Formal and informal convergence diagnostics can be found in Brooks and Gelman (1998), and Gilks, Richardson, and Spiegelhalter (1996). Starting values for the MCMC algorithm can be obtained by fitting a multidimensional IRT model to the data but ignoring the randomized response character of the discrete response data using the MCMC algorithm of Béguin and Glas (2001). However, it will be shown in a simulation study that convergence properties of the proposed MCMC algorithm are good and independent of chosen starting values.

4.5 Simulation Study

In this section, results are reported from a simulation study for parameter recovery based on the RIRT model for randomized item response data.

Data were simulated for two orthogonal ability dimensions, $Q = 2$, for $N = 750$ subjects who responded to $K = 20$ items. For the two ability dimensions a structural multivariate model was assumed with two covariates. The covariate values, $\mathbf{X}$, and regression coefficients, $\boldsymbol{\gamma}$, were generated randomly from $N(0, 1)$. The between-subject variation was .25 for each dimension. The intercept of both dimensions was fixed at zero for identification. Furthermore, the first two items were loading only on the first dimension, while items 3 and 4 only on the second dimension, where the other loadings were randomly generated from $N(0, .25)$ and were allowed to load on both dimensions. Four response categories were assumed, with the threshold parameters for all items set at $\boldsymbol{\kappa}_C = (-1, 0, 1)$. For the forced randomized response part, the probabilities were set at $p_1 = 0.667$ and $p_2 = 0.25$.

Using randomly generated starting values, two MCMC chains of 6,000 iterations each were simulated, and the first 1,000 iterations were discarded for the burn-in. Using the CODA package in R, convergence of the algorithm was assessed using several statistics as well as the visual inspection of trace plots. Trace plots for the regression coefficients for the two chains are shown in Figure 4.1.

For the MCMC chains of the regression parameters, Gelman's R statistic was estimated to be 1.05, which is below the recommended threshold of 1.10. Geweke's test for equality of means for the first and the last part of the MCMC chains did not reject the null hypothesis of equal means. Also for other parameters those statistics suggested convergence of the algorithm. The partial autocorrelation

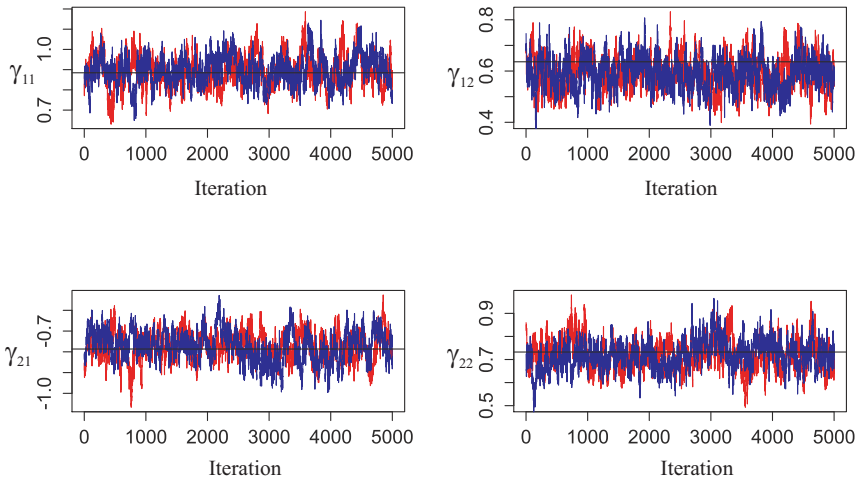

Figure 4.1: Trace plots of estimated regression coefficients showing two MCMC runs, where the solid black line represents the simulated true value.

function for several chains suggested a first-order Markov process, only for some chains showing a statistically significant, but minor, autocorrelation of around 0.20 at lag 2.

Recovery of the loadings and the ability parameters was assessed graphically (not shown) and showed that the true against the re-estimated parameters closely followed the identity line. Finally, Table 4.1 shows the true and re-estimated parameters for the structural multivariate model on the person parameters, indicating that the developed MCMC algorithm worked well for estimation of the model.

Table 4.1: Results of simulation study. Generating values, means and standard errors of recovered values.

	Gen.	RIRT Model		
Fixed	Coeff.	Mean	SD	95% HPD
First Dimension				
γ_{10}	0	-	-	-
γ_{11}	.88	.89	.08	$[.74, 1.05]$
γ_{12}	.63	.59	.06	$[.47, .72]$
Second Dimension				
γ_{20}	0	-	-	-
γ_{21}	$-.79$	-.77	.07	$[-.91, -.65]$
γ_{22}	.73	.72	.07	$[.60, .87]$
Random	Coeff.	Mean	SD	95% HPD
$\Sigma_{\theta_{11}}$	.25	.30	.07	$[.19, .46]$
$\Sigma_{\theta_{12}}$	.00	.00	.03	$[-.05, .06]$
$\Sigma_{\theta_{22}}$	.25	.24	.05	$[.15, .34]$

4.6 Measuring Drinking Problems and Alcohol-Related Expectancies among College Students

Thirteen items of the College Alcohol Problem Scale (CAPS; O'Hare, 1997) and four items of the Alcohol Expectancy questionnaire (AEQ; Brown, Christiansen, and Goldman, 1987) were used to assess alcohol problems and alcohol-related expectancies among college students. The questionnaire items are given in Appendix C. The sensitive nature of the study supports a randomized response questioning technique to avoid refusals and misleading responses to conceal socially undesirable behavior. Any self-reported information about negative consequences of drinking is likely to be biased due to socially desirable responding. It is investigated whether the randomized response technique improved the cooperation of the respondents and the quality of the data by comparing the randomized response outcomes with the direct-questioning outcomes. Furthermore, using a the joint multidimensional RIRT modeling approach, multiple sensitive constructs underlying both scales (CAPS and AEQ) are measured and their relationships with background information analyzed.

The AE questionnaire is used to measure the degree of expectancies associated with drinking alcohol. Expectancies related to alcohol use are are known to influence alcohol use and behavior while drinking (e.g., Werner, Walker, and Greene, 1995). The entire test consists of 90 items and covers six dimensions (see Brown et al., 1987). In the present study, attention was focused on alcohol use-related sexual enhancement expectancies using four items covering sexual enhancement expectancies, which are given in Appendix C.

The CAPS instrument is one of the major self-report measures used to asses drinking problems. The items cover socio-emotional (e.g., hangovers, memory loss, depression) as well as community problems (e.g., drove under the influence, engaged in activities related to illegal drugs, problems with the law). O'Hare (1997) developed the CAPS instrument to measure different psychosocial dimensions of problem drinking among college students. The two factors, socio-emotional and community problems, were identified from a factor analysis, which explained more than 60% of the total variance. Fox and Wyrick (2008) analyzed the CAPS data using a unidimensional RIRT model to measure a general construct alcohol dependence. Although the model described the data well, a multidimensional approach provides insight in the different factors related to problem drinking, factor-specific relationships with background variables, and supports the multidimensional nature of the CAPS.

Data were collected through a survey study in 2002 at four local colleges/universities, Elon University (N=495), Guilford Technical Community College (N=66), University of North Carolina (N=166), and Wake Forest University (N=66). A total of 351 students was assigned to the direct-questioning (DQ) condition and 442 to the randomized response (RR) condition. Students in the DQ group served as the study's control group and were instructed to answer the questionnaire as they normally would. Students in the RR condition used a spinner to assist them in completing the questionnaire. For each item of the CAPS and AEQ, the spinner was used as a randomizing device and the outcome determined whether to answer

honestly or to register a forced response. The properties of the spinner were set such that an honest answer was requested with a probability of 60% and with a probability of 40% a forced response was dictated. When a forced response was to be generated, each response was given an equal probability of 20%. No identifying information was collected but age, gender, and ethnicity were also registered. Each class of students (5-10 participants) was randomly assigned to the DQ group or the RR group and were selected from the same population. It was not possible to randomly assign students.

The following multidimensional RIRT model was used to analyze the response data,

$$\begin{aligned} P\left(Y_{ik} = c \mid \boldsymbol{\theta}_i, \mathbf{A}_k, \boldsymbol{\kappa}_k\right) &= p_1 \pi_{ik} + (1 - p_1) p_2(c) \\ \pi_{ik} &= \Phi\left(\mathbf{A}_k^t \boldsymbol{\theta}_i - \kappa_{i,(c-1)}\right) - \Phi\left(\mathbf{A}_k^t \boldsymbol{\theta}_i - \kappa_{i,c}\right), \quad (4.16) \\ \boldsymbol{\theta}_i &= \boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_1 RR_i + \mathbf{e}_i, \end{aligned}$$

where $\mathbf{e}_i \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_\theta\right)$ and RR_i equals one for each dimension q when student i was assigned to the RR group and zero otherwise. According to the forced randomized response sampling design, $p_1 = .60$ and $p_2(c) = .20$, for $c = 1, \ldots, 5$, and in the direct-questioning design $p_1 = 1$. The factor loadings $\mathbf{A}$ and item thresholds are assumed to be independent of the questioning technique. The multidimensional RIRT model was identified by fixing the mean in each dimension, such that $\boldsymbol{\gamma}_0 = \mathbf{0}$. The variance was identified by restricting the variance components of each factor to one. The so-called rotational variance was identified by assigning Q items uniquely to Q dimensions. Each model was estimated using 50,000 MCMC iterations using a burn-in period of 10,000 iterations for parameter estimation. The convergence of chains were inspected using plots and various MCMC convergence diagnostics.

4.6.1 Multi-Dimensional Scale Analysis

First, a two-factor RIRT model was estimated, where item 1 and item 14 had just one free non-zero loading to identify two factors. In Figure 4.2, the estimated factor loadings are given for the two-component RIRT model stated in Equation 4.16.

The estimated factor loadings are standardized by dividing each loading with the average item loading. For each component, the sign of the loadings is set in such a way that a higher latent score corresponds to a higher observed score. It can be seen that items 14-17 measure the factor alcohol-related expectancy and that most other items are clearly measuring another factor socio-emotional/community problems, which was labeled alcohol dependence in the Fox and Wyrick (2008) analysis. The loadings are all above .75, which indicates that both factors can be interpreted. Note that expectancies are increasing with alcohol consumption and slightly diminish socio-emotional/community problems given the negative factor loadings for the other component.

Second, a three-factor RIRT model was estimated to investigate whether the CAPS measures the two dimensions socio-emotional and community problems. Besides item 1 and 14, the loadings of item 5 were also restricted to identify the factor community problems as reported in the literature. In Table 4.2, the standardized estimated factor loadings of the three factors are given. It follows that

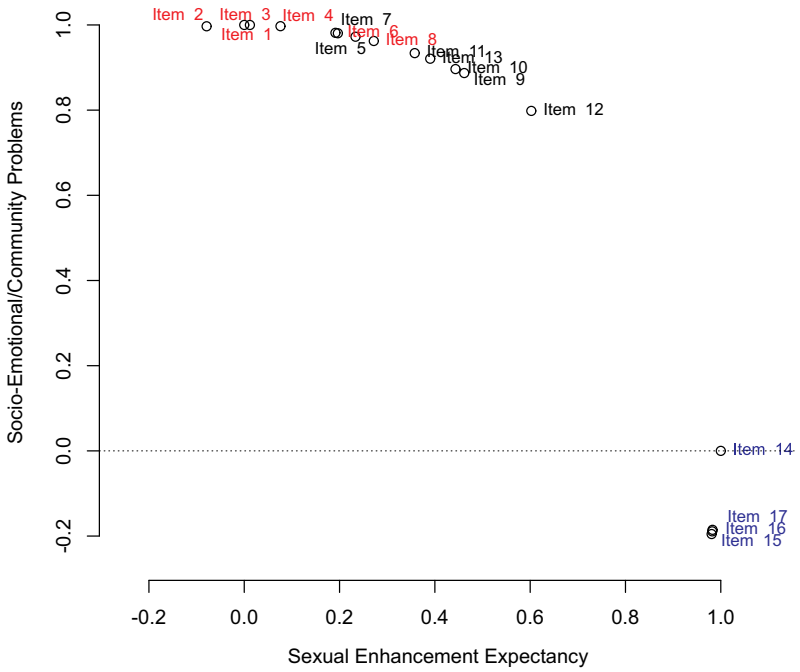

Figure 4.2: Estimated standardized factor loadings for the two-component RIRT model

items 1-4, 6,8, and 9 associate with the first factor, representing socio-emotional problems, and have factor loadings higher than .60. This first component represents drinking-related problems that includes depression, anxiety, and troubles with family, where the problems will increase with alcohol consumption. Some of the items also associate with the two other components. The second component labeled community problems covers items 5,7,and 10-13, with loadings higher than .60, except for item 12. As reported in the literature, this item is associated with the factor community problems but relates also with the other components and most strongly with the third component. The factor community problems covers acute physiological effects of drunkenness together with illegal and potentially dangerous activities (e.g., driving under the influence). For the two-factor RIRT model, the second and third response options were more likely to be endorsed of the AEQ items than of the CAPS items. The corresponding threshold estimates of the AEQ items in comparison to the CAPS items for response category two and three are lower. Except for CAPS item 12 (feeling tired or hung over) on which students scored relatively high. The AEQ items 14-17 and CAPS item 12 can be considered as the less severe items. Items with relatively high thresholds, item 4 (family problems related to drinking), item 5 (spent too much money on drugs), item 7 (hurt another person physically), and 13 (Illegal activities associated with drug use) were severe, where most students responded almost never to seldom. The threshold estimates of the three-factor model were similar, except the threshold estimates of item 5 were not very stable and much higher. This

Table 4.2: CAPS-EA Scale: Estimated weighted factor loadings for the three-component analysis

Subscale Items	Three-Component RIRT Model		
Socio-Emotional	Comp. 1	Comp. 2	Comp. 3
1 Feeling sad, blue or depressed	1.00	.00	.00
2 Nervousness or irritability	.99	.12	-.05
3 Hurt another person emotionally	.94	.33	.08
4 Family problems related to your drinking	.82	.55	.15
6 Badly affected friendship or relationship	.83	.49	.26
8 Caused other to criticize your behavior	.78	.50	.38
9 Nausea or vomiting	.73	.39	.58
Community Problems	Comp. 1	Comp. 2	Comp. 3
5 Spent too much money on drugs	.00	1.00	.00
7 Hurt another person physically	.49	.83	.24
10 Drove under the influence	.44	.73	.52
11 Spent too much money on alcohol	.61	.65	.46
12 Feeling tired or hung over	.59	.41	.69
13 Illegal activities associated with drug use	.08	.95	.30
Alcohol Expectancy Scale	Comp. 1	Comp. 2	Comp. 3
14 I often feel sexier after I've had a couple of drinks	.00	.00	1.00
15 I'm a better lover after a few drinks	-.11	-.06	.99
16 I enjoy having sex more if I've had some alcohol	-.16	-.01	.99
17 After a few drinks, I am more sexually responsive	-.16	-.02	.99

follows from the fact that most responses above one were considered to be forced randomized responses. Item 5 with a very low prevalence did not provide much information to assess drinking problems. A prior restriction on the upper bound led to a numerical stable solution.

4.6.2 Structural Model Analysis

In the two and three-factor RIRT models, the multivariate latent factor model was extended with the explanatory variable RR and an indicator variable female, which equals one when true. For each dimension, both explanatory variables were included.

In Table 4.3, the structural multivariate parameter estimates are given of the three-component and two-component RIRT model. For the two-component model, loadings of item 1 and 14 were fixed to identify two dimensions. One overall factor represents a composite measure of alcohol-related problems (i.e., socio-emotional and community problems) and the other factor alcohol-related sexual enhancement expectancies. It can be seen that there is a moderate positive covariance of .45 between the two factors, where the component variances are slightly smaller than

Table 4.3: CAPS-EA Scale: Parameter estimates of two and three-component RIRT model.

	Two Component		Three Component	
Parameter	Mean	SD	Mean	SD
Dimension				
Socio-Emotional/Community				
γ_{11} (RR)	.20	.09	.21	.10
γ_{21} (Female)	.01	.06	.05	.07
Alcohol Expectancy				
γ_{12} (RR)	.22	.06	.21	.07
γ_{22} (Female)	.03	.04	.06	.05
Community				
γ_{13} (RR)			.32	.10
γ_{23} (Female)			$-.30$	.09
Variance Parameters	Mean	SD	Mean	SD
$\Sigma_{\theta_{11}}$	.96	.05	.98	.05
$\Sigma_{\theta_{12}}$	.45	.07	.55	.06
$\Sigma_{\theta_{13}}$			.38	.08
$\Sigma_{\theta_{22}}$	.98	.05	1.06	.05
$\Sigma_{\theta_{23}}$			.42	.08
$\Sigma_{\theta_{33}}$			.99	.07
Information Criteria				
-2log-likelihood		20622		19625

one.

The students in the RR condition score significantly higher in both dimensions. For the RR group, the average latent scores are .20 and .22 on the composite drinking problem scale and the alcohol-related expectancy scale, respectively, which are both zero in the DQ group. Fox and Wyrick (2008), who performed a unidimensional RIRT analysis using only the CAPS items, reported an RR effect of .23. The present multidimensional approach shows a comparable RR effect, also for the AEQ scores. The females and males show comparable scores on both dimensions.

In the three-factor RIRT model, problems associated with drinking are represented by two factors (i.e., socio-emotional and community problems). The randomized response effects are significantly different from zero for all three factors, where the effect on the factor representing community problems related to alcohol use is around .32 and slightly higher than the effects of the other components, which are around .21. It seems that students were less willing to admit to alcohol-related community problems, which induced more socially desirable responses than the other factors. The relatively high thresholds of items 5,7, and 13 measuring community problems indicated that they were more severe and most likely more

Table 4.4: CAPS-EA Scale: Differences across colleges and universities using the two and three-component RIRT model.

	Two Component		Three Component	
Parameter	Mean	SD	Mean	SD
Dimension				
Socio-Emotional/Community				
γ_{11} (RR)	.29	.08	.29	.09
School variables				
γ_{21} (Elon)	.19	.06	.19	.07
γ_{31} (UNCG)	−.19	.10	−.14	.11
γ_{41} (Wake Forest)	−.23	.12	−.11	.14
γ_{51} (Guilford)	.24	.12	.05	.13
Alcohol Expectancy				
γ_{12} (RR)	.19	.06	.33	.07
School variables				
γ_{22} (Elon)	.04	.05	−.07	.05
γ_{32} (UNCG)	−.11	.07	−.13	.08
γ_{42} (Wake Forest)	−.23	.08	−.15	.09
γ_{52} (Guilford)	.30	.09	.36	.11
Community				
γ_{13} (RR)			.25	.08
School variables				
γ_{23} (Elon)			.16	.06
γ_{33} (UNCG)			−.13	.10
γ_{43} (Wake Forest)			−.43	.13
γ_{53} (Guilford)			.38	.13
Variance Parameters	Mean	SD	Mean	SD
$\Sigma_{\theta_{11}}$	.94	.05	.96	.05
$\Sigma_{\theta_{12}}$	.60	.07	.51	.06
$\Sigma_{\theta_{13}}$			.52	.07
$\Sigma_{\theta_{22}}$	.97	.05	.99	.05
$\Sigma_{\theta_{23}}$			.47	.06
$\Sigma_{\theta_{33}}$			.94	.05
Information Criteria				
-2log-likelihood		20816		19717

sensitive. Male students scored significantly higher in comparison to female students on the dimension representing community problems related to alcohol use. Male students were more likely to experience alcohol-related community problems than females. This gender effect was not found for the other dimensions.

Interest was focused on the average student drinking problems and expectan-

cies of the four selected colleges/universities that took part in the experiment. The clustering of students in colleges/universities was represented using effect coding. In Table 4.4 the parameter estimates are given for the two-component and three-component model. For each dimension, the intercept represents the average score across colleges and universities, which was set to zero. In the two-component model, the average score of the RR group is .29 and .19 and are both significantly higher than zero. It follows that the mean score of the factor alcohol-related problems of Guilford Technical Community College and Elon University are significantly higher than the means of UNCG and Wake Forest. For the factor alcohol-related expectancy, Guilford Technical Community College scored on average higher than the other colleges and universities.

For the three-component model, the mean scores of the three factors of the RR group are significantly higher than zero and comparable when controlling for differences across universities and colleges. It follows that Guilford Technical Community College has the highest average score of alcohol-related community problems and of alcohol-related sexual enhancement expectancies. The results show that alcohol-related sexual enhancement expectancies and community problems are positively correlated, where scores differ across universities and colleges. The estimates of the RR effect indicate that the RR-group scored significantly higher in comparison to the DQ-group on each subscale. Although validation data are not available, it is to be expected that the RR technique led to an improved willingness of students in answering truthfully, given the random assignment to direct-questioning and randomized response questioning.

4.7 Discussion

In educational and psychological measurement, it is often more realistic to assume that multiple constructs influence the performance on test items. The multidimensional item response theory model can be used to asses the underlying latent variable structure given the test results (Reckase, 2009). When surveying sensitive topics, direct questioning may lead to social desirability bias. Therefore, in combination with a randomized response design, a multidimensional randomized item-response model is proposed to measure multiple sensitive constructs given multivariate randomized response data. The presented confirmatory multidimensional model can handle dichotomous and polytomous randomized item responses. The application shows a model belonging to the class of compensatory models. However, when every item measures one construct a non-compensatory multidimensional RIRT model can be stated in a similar way.

MCMC methods have developed to tackle the high dimensional integration problem in confirmatory multidimensional IRT analysis (e.g., Edwards, 2010; Sheng, 2010). Cai (2010) proposed a Metropolis-Hastings Robbins-Monro algorithm for an exploratory analysis given polytomous response data. Béguin and Glas (2001) developed a full Gibbs sampling procedure and proposed several posterior predictive checks. Yao and Boughton (2007) showed MCMC estimation of multidimensional partial credit models and the assessment of subscale scores. The present MCMC algorithm for the estimation of the multidimensional RIRT

model is based on a double augmentation scheme to deal with the categorical randomized response outcomes. Furthermore, the MCMC algorithm can also handle the estimation of the multivariate structural model parameters. This includes the structural regression parameters, which specify the relationship between the multiple sensitive constructs and the explanatory background information, and the correlation structure among the latent variables. The MCMC algorithm has been developed in R and will be made freely available trough the internet.

The present survey study about alcohol-related sexual enhancement expectancies and drinking problems showed that randomized response questioning improved the cooperation of the respondents and reduced domain-specific social desirability bias. The joint analysis results support the alcohol expectancy theory (e.g., Brown et al., 1987), which states that positive expectancies due to alcohol use lead to more positive initial drinking experiences leading to more positive expectancies. Here, it was shown that alcohol-related sexual enhancement expectancy scores were positively correlated with subscale scores for alcohol-related socio-emotional and community problems. In the literature, alcohol-related expectancies have been found to be useful in predicting drinking problems and drinking behavior, and patterns of problematic use (e.g., Werner et al., 1995). The randomized response technique can improve the accuracy of the self-report data and related predictions, while the multidimensional modeling approach can improve the accuracy of subscale scores by using the additional subscale information (e.g., Yao and Boughton, 2007).

The measurement of alcohol expectancies of individuals is important to identify current and predict future problem drinking. The randomized response technique can improve the quality of the diagnostic self-report data, when respondents are inclined to underreport alcohol consumption and the negative effects of alcohol use due to social desirability or potential legal implications. The multifactor modeling approach will support the multifactorial nature of the expectancy questionnaires and individual measurements of expectancy behavior given randomized response data.

Chapter 5

Randomized Response Techniques

Abstract

A general overview of randomized response techniques is given. The randomized response technique is a data collection method used for response bias reduction in sensitive inquiries that in case of direct questioning have the potential of resulting in socially desirable responses. Respondent cooperation and truthful responding are stimulated since each individual response is masked by means of a chance mechanism introduced in the response process. Although the resulting data set are polluted, inferences can be made taking into account the probability distribution of the chance mechanism. Well-known traditional and more recent randomized response techniques operating on binary and categorical ordinal data in a univariate setting will be thoroughly discussed and compared. To give a complete overview, recently developed techniques that avoid the use of a randomizing device will be discussed. For the multi-item setting, models are discussed for individual level inferences where a randomized response technique is used to obtain sensitive responses to scale items. A distinction will be made between models for large-scale assessments and for small sample sizes. Various randomized response technique approaches will be illustrated through a Bayesian analysis.

5.1 Introduction

Surveys including self-reports are usually based on a direct questioning technique for data collection. Generally, it is assumed that this technique provides the necessary level of reliability when measuring respondent characteristics such as attitudes, preferences, opinions, and behaviors. However, when studying sensitive topics respondents are reluctant to supply truthful answers. Self-representational concerns lead to socially desirable answers or non-responses, as reported by Tourangeau et al. (2000), and Tourangeau and Yan (2007). As a result, missing data and inaccurate data can give rise to a substantial bias in estimates of prevalence of

stigmatized behaviors or attitudes. To obtain more accurate estimates based on self-reports in inquiries of a sensitive nature, alternative methods of data collection should potentially relieve the self-representational concerns and stimulate the respondent's cooperation by providing response privacy protection.

The randomized response (RR) technique is a survey technique that can reduce response bias in sensitive survey studies (Clark & Desharnais, 1998; Lamb & Stem, 1978; Lensvelt-Mulders, Hox, van der Heijden, & Maas, 2005; Reinmuth & Geurts, 1975; Warner, 1965). Initially proposed by Warner, the developed randomized response techniques insure confidentiality of individual responses by introducing a chance mechanism in the response process governed by a randomizing device. Depending on the RR design, the outcome of a randomizing device either determines a randomization of the question to be answered or a randomization of the response (honest or preselected) to be given. In either case, individual answers are masked. This guarantees the privacy of responses, alleviates self-representational concerns and stimulates the respondent to provide a truthful answer on the sensitive characteristic. Still, the RR information obtained, together with the known distribution of the randomizing device, is sufficient to compute improved estimates of parameters associated with the sensitive characteristic. Review texts on randomized response techniques can be found in Chaudhuri and Mukerjee (1988) and Fox and Tracy (1986), as well as in Scheers (1992) and Umesh and Peterson (1991).

A Schematic Overview

The RR techniques can be subdivided in different ways. A schematic outline of the techniques is given in Figure 5.1, which provides a thorough insight in the various ways of dealing with randomized responses. First, a basic distinction is made between single-item and multi-item randomized response techniques. Both types of techniques allow aggregate level estimation of a sensitive characteristic. Group estimates of population parameters can be determined, possibly using additional background information. In addition to population parameter estimates, multivariate data format enables inferences on individual level.

Second, two families of single-item randomized response techniques can be recognized: (1) the traditional RR methods making use of a randomizing device, and (2) the so-called nonrandomized response (NRR) models. The NRR models make no use of a randomizing device, but to stimulate respondents' cooperation a non-sensitive question is used. Three traditional RR methods will be discussed in detail; that is, Warner's opposite-question method, the unrelated-question method, and the forced response method. The forced response method will be extended to operate also on ordinal RR data. Various NRR methods will be discussed for univariate analysis, which includes Takahasi's RR technique, the triangular and the crosswise RR method, as well as the hidden sensitivity RR method.

Individual level inferences can be made, when multiple RR observations are recorded for each respondent. Single item or univariate methods are extended to handle multivariate randomized response data. The FRR design is chosen to illustrate extensions and elaborations of RR techniques to a multivariate setting. In this setting, a distinction will be made between models for small and large data samples.

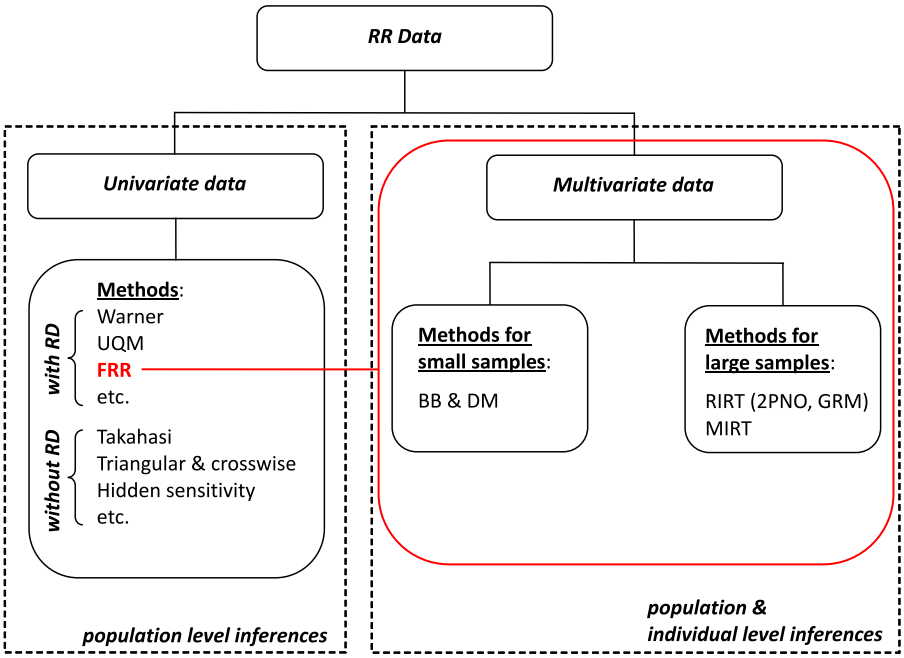

Figure 5.1: Schematic outline of RR techniques for different data collection settings.

Two different modeling approaches for multivariate RR data analysis are described. First, models for small data sizes will be discussed, which includes the beta-binomial and the Dirichlet-multinomial model. The beta-binomial RR model is used for individual response rate estimation given dichotomous RR data. An extension of this model to accommodate polytomous RR data is the Dirichlet-multinomial RR model. Second, for large RR data cases where the object is to measure a single or multiple sensitive constructs (multidimensional) IRT-based RR models (e.g., Lord & Novick, 1968; Reckase, 2009) can be used. Models for the measurement of person and item characteristics taking account of the RR nature of the observed data will be described. An extension is made to mixture IRT models, when respondents who do not trust the protection offered by the randomized response design still give self-protective responses and ignore the design instructions. In that case, a mixture modeling approach can account for the consistent non-compliant response behavior.

This chapter is organized as follows. First, various randomizing devices are described that underly the different RR techniques. After discussing various data formats, univariate randomized response techniques are discussed. Subsequently, the multivariate extension of randomized response models are presented making use of the forced response method. Selected models are illustrated with empirical examples. The inferences stemming from a fully Bayesian analysis are compared to the results from a maximum likelihood approach. Appendix F and G provide

WinBUGS code for presented Bayesian model analysis.

5.2 Randomizing Device

The randomizing device or rule of procedure is an essential element of the randomized response technique. It has to be unbiased, easy to use, easy to understand, and has to promise high degree of protection. Examples of randomizing devices are deck of cards, a sealed plastic box with colored beads (Horvitz et al., 1967), a flask containing different colored balls (the Hopkins' randomizing device III; Liu, Chow, & Mosley, 1975), coins, dice, spinners, final digits on the respondent's social security number, or those of the serial number on a dollar bill in the respondent's possession, etc. Due to their problematic accessibility and questionable randomness, the last two randomizing device alternatives did not gain much popularity.

Most randomizing devices are not without problems (Tracy & Fox, 1981). For instance, to reproduce the required design probabilities a deck has to be adequately shuffled. Spinner devices have to be perfectly horizontal and not bent or damaged. In addition, rules of procedure have to be defined to avoid confusion when the arrow stops on a line separating two adjacent segments.

In general, unsophisticated randomizing devices such as coins and dice are preferred over elaborate randomizing devices since the related procedures are easier to explain and make it easier to convince respondents of their protection.

On top of the usual critical features of question wording and placement in the design of any conventional survey instrument, significant considerations related to application of randomized response come into play. The choice of the type and parameters of the randomizing procedure also influences the level of respondent protection and efficiency in estimation. The lower the probability of selection of the sensitive question the more data are masked and the more noise is allowed into the data set, the higher the protection of responses, and the greater the sampling variance of the estimate. Depending on the level of sensitivity of the inquiry, a careful consideration of the randomizing procedure is necessary to offer high privacy protection, while staying in control of the accuracy of the estimates.

In addition, the subjective assessment of the degree of protection largely influences the extent of cooperation. Randomizing devices that are believed to offer higher response protection are very desirable in RR surveys.

5.3 The Type of Data

Test items comprising a survey instrument are acting as indicators of a latent construct. In surveys the nature of the items dictates the type of randomizing device used to diffuse sensitivity. The procedure where responses are scored in categories leads automatically to categorical data. For categorical data (e.g. dichotomous, ordinal, nominal, etc.) the device used to randomize the responses usually generates discrete values. For instance, given quantitative data, Greenberg, Kuebler, Abernathy, and Horvitz (1971) estimated the mean number of abortions and mean income of heads of household in North Carolina. Franklin (1989) considered a different approach for dichotomous populations using a normal distribution for the

process of randomization. Randomizing devices and procedures can be adapted to handle continuous data, but this type of outcome data will not be discussed in this thesis.

Dichotomous responses can be masked by using a coin (Folsom, Greenberg, Horvitz, & Abernathy, 1973; Fox & Tracy, 1986), a die (Boruch, 1971a), a spinner (Warner, 1965; Stem & Steinhorst, 1984; Scheers & Dayton, 1988; Clark & Desharnais, 1998; Gingerich, 2010), a sealed clear plastic box containing colored beads of two varieties (Horvitz et al., 1967), etc. Boruch (1971a) suggested to use a die to generate masked binary responses in the study on marijuana smoking, where the sensitive question was selected if the outcome of a die was one and nonsensitive question was being answered if the outcome was two or more.

De Jong et al. (2010) described a randomized response procedure making use of a die for masking responses to items that were using a five-point ordinal response format. O'Hare (1997) and Fox and Wyrick (2008) described a spinner that was adapted to accommodate ordinal randomized responses. Another interesting device for discrete quantitative data randomization is a flask containing different colored or numbered balls, known as Hopkins' randomizing device, described by Liu et al. (1975) and Liu and Chow (1976b).

The major drawback of traditional RR models, namely, the restriction of inferences to aggregate-level, diminishes when RR models are adapted for a multivariate survey instrument. Individual answers are still masked, since a randomizing device is used for each item, but at the same time the individual (sensitive) characteristics can be estimated.

5.4 Single-Item Randomized Response Techniques

The different single-item randomized response techniques for self-report data, which will be discussed in this section, are the opposite-question method, also known as Warner's method, the unrelated-question method, with its special case, the forced response method, and the family of randomized response methods which make no use of a randomizing device, frequently referred to in the literature as nonrandomized response methods.

5.4.1 Opposite-Question Method (Warner)

The method developed by Warner (1965) was aimed at controlling response tendencies. In his seminal study, Warner introduced a randomizing device. The device consisted of a spinner, which points to the letter A with probability p and to the letter B with probability $1 - p$. In the data collection procedure the randomizing device directed the choice of one of the two logically opposite questions. The respondent was asked to give an honest answer. The choice was known solely to the respondent. The latter feature guaranteed confidentiality and implied a higher degree of cooperation. An example of two opposed questions is:

Question A: Are you a member of "A"?

Question B: Are you not a member of "A"?

While the questions answered are irretrievable the researcher is still able to make valid inferences given the characteristics of the randomizing device. The

characteristics are incorporated in a probabilistic model relating observed randomized responses to unobserved responses.

Figure 5.2 depicting the response scheme can be explained as follows. The population is split into group A possessing the sensitive attribute, and its compliment $\bar{A}$, not possessing the sensitive attribute. The π denotes the proportion of the population possessing or belonging to A and p denotes the selection probability of Question A. If a survey participant possesses the sensitive attribute then the top branch of the probability tree is followed. With probability πp a 'Yes' response is observed, when a participant is selected that possess the characteristic and is requested by the randomizing device to respond to Question A (Q_A). The probability equals $\pi(1-p)$ and a 'No' response is observed when the complimentary Question B (Q_B) is selected.

For participants belonging to the complimentary class $\bar{A}$ the bottom probability branch is followed. With probability $(1-\pi)p$ a 'No' response is recorded when a participant is selected that does not possess the characteristic and is requested by the randomizing device to respond to Question A (Q_A). A 'Yes' response is recorded with probability $(1-\pi)(1-p)$ when the respondent is confronted with the complimentary Question B (Q_B).

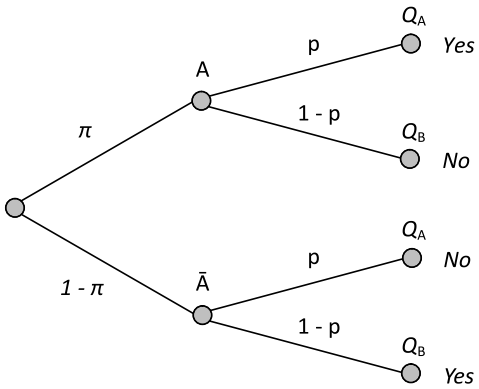

Figure 5.2: Probability tree of the opposite-question model for a dichotomous item.

In a random sample of n respondents, the proportion of affirmative responses $\lambda = \frac{n_1}{n}$ is then expressed in terms of the proportion of respondents possessing the sensitive attribute π by the following expression:

$$\lambda = p\pi + (1-p)(1-\pi), \tag{5.1}$$

where the randomizing device outcome follows a Bernoulli distribution with parameter p (probability of selection of Question A). The maximum likelihood estimate (MLE) of the prevalence of the attribute in the population in terms of the observed affirmative answers equals

$$\hat{\pi} = \frac{\lambda + p - 1}{2p - 1}, \tag{5.2}$$

where $p \neq 0.5$, with variance

$$var(\hat{\pi}) = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2}. \tag{5.3}$$

RR estimators of population proportions have larger standard errors than estimators based on direct questioning data (Lensvelt-Mulders, Hox, & van der Heijden, 2005). For Warner's method, this means that the second term of expression (5.3) represents the additional variance related to the use of a randomizing mechanism. In addition, the closer p is to 0.5, the greater the variance inflation (Tracy & Fox, 1981). Obviously, one way to reduce variance is to increase the sample size.

Warner's technique gave rise to a whole variety of randomized response methods that comprised efficiency improvements. Horvitz et al. (1967) and Liu and Chow (1976a) suggested that by multiple trials, performed on each respondent, a reduction in variance can be achieved. Other developments of Warner's method concerned extensions to multiple categories. For example, Abul-Ela, Greenberg, and Horvitz (1967) discussed an extension to the trichotomous case to estimate proportions of three mutually exclusive groups with at least one nonsensitive group. This method required an extra data sample for each additional parameter. Categorical proportion estimation with only one sample using a modified randomizing device was proposed by Liu et al. (1975) and Eriksson (1973). To address the problem of efficiency, various modifications of the randomized response design were proposed. For instance, Mangat and Singh (1990) proposed an alternative RR procedure where a direct honest response is requested from a respondent with the sensitive characteristic, whereas the randomizing device instructions have to be followed by respondents of the nonsensitive class. Also, two-stage RR methods requiring two randomizing devices (e.g., Kim & Elam, 2005; Kim & Warde, 2005; Mangat, 1994) and optional randomized response methods (e.g., Chaudhuri & Mukerjee, 1985; Chaudhuri & Saha, 2005; Saha, 2007) were developed to reduce the sampling variance. Many authors explored RR techniques handling two sensitive characteristics at the same time (e.g., Christofides, 2005; Tian, Yu, Tang, & Geng, 2007). Methods not requiring direct answers to sensitive questions among others were addressed by Kuk (1990), Christofides (2003) and Tian et al. (2007).

5.4.2 Unrelated-Question Method (UQM)

The fact, that in Warner's design the sensitive question is paired with its logical opposite such that both questions are sensitive in nature, received a lot of criticism. It can provoke distrust among the interviewees and thus jeopardize the merits of RR. It is this feature that led to the development of the unrelated-question model (Horvitz et al., 1967; Greenberg et al., 1969), where a nonsensitive question is introduced.

Greenberg et al. (1969) assumed that respondents will be more cooperative and truthful if the sensitive question is paired with a question unrelated to the underlying sensitive characteristic. Such a pair of questions can be:

Question A: Are you a member of "A"?

Question B*: Do you subscribe to "Newspaper X"?

The proportion in the population with the innocuous attribute π_y is not known beforehand. As a result, in this design the challenge is to estimate the proportion with the sensitive characteristic π but also the proportion π_y. Therefore, the unrelated-question model requires two independent samples of size n_1 and n_2 to estimate two unknown parameters π and π_y. Respondents of each sample are questioned using similar randomizing devices, but with different probabilities of the sensitive question selection, i.e. $p_1 \neq p_2$.

The data collection procedure is similar to that proposed by Warner. A participant in sample i is instructed to use the randomizing device i where the sensitive question is selected with probability p_i and the innocuous question with probability $1 - p_i$. Depending on the outcome of a randomizing device a participant is asked to give an undisclosed honest answer to the question selected.

In Figure 5.3 the probability tree corresponding to this RR design is presented. Each respondent belongs to the class with the sensitive attribute A or its compliment without the sensitive attribute $\bar{A}$, and has a positive or negative status on the innocuous attribute. The affirmative responses come from two distributions and can be observed according to three different situations. An affirmative response with probability πp_i will be recorded, when instructing a respondent with the sensitive characteristic to respond to the sensitive question Q_A. However, if the innocuous question Q_B* is selected, an affirmative response is obtained with probability $\pi(1 - p_i)\pi_y$ according to the status of the respondent on the non-sensitive characteristic. Note that with probability $(1-\pi)(1-p_i)\pi_y$ an affirmative response can also be observed when instructing a respondent belonging to $\bar{A}$ to respond to question Q_B*.

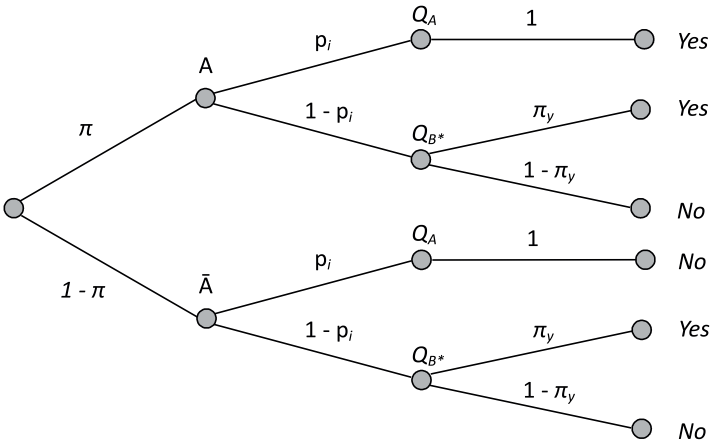

Figure 5.3: Probability tree of the unrelated-question model for a dichotomous item.

The proportion of observed affirmative responses $\lambda_1 = \frac{n_1^*}{n_1}$ and $\lambda_2 = \frac{n_2^*}{n_2}$, where n_1^* and n_2^* are the number of affirmative responses in the two samples, are equal

$$\lambda_i = p_i\pi + (1 - p_i)\pi_y, \tag{5.4}$$

where $i = 1, 2$. The true proportion of prevalence π can be estimated based on λ_1 and λ_2 as follows

$$\hat{\pi} = \frac{\lambda_1(1-p_2) - \lambda_2(1-p_1)}{p_1 - p_2} \tag{5.5}$$

with variance

$$var(\hat{\pi}) = \frac{1}{(p_1-p_2)^2}\left[\frac{\lambda_1(1-\lambda_1)(1-p_2)^2}{n_1} + \frac{\lambda_2(1-\lambda_2)(1-p_1)^2}{n_2}\right]. \tag{5.6}$$

For large samples this method has proven to perform better than Warner's method.

To circumvent the need of sample splitting, Greenberg et al. (1969) and Horvitz, Greenberg, and Abernathy (1976) proposed using an unrelated question with a known outcome distribution. A classical example of such an (unrelated) innocuous question is a question concerning the birth month of a respondent such that $\pi_y = 1/12$. In case of known π_y, Equation 5.4 reduces to a single equation and the expressions (5.5) and (5.6) become

$$\hat{\pi}|\pi_y = \frac{\lambda_1 - (1-p_1)\pi_y}{p_1}$$

and

$$var(\hat{\pi}|\pi_y) = \frac{\lambda_1(1-\lambda_1)}{n_1 p_1^2}, \tag{5.7}$$

respectively.

This method is more efficient than Warner's design because the variance of its maximum likelihood estimate is smaller that that of Warner's model for any $p_1 \in [0.33933, 1]$, as shown in Dowling and Shachtman (1975). However, application of this method is limited due to the scarcity of nonsensitive questions with known probability distributions. In addition, most demographic-type questions might raise well-placed suspicion among respondents. For example, in case of aforementioned birth month question certain groups, e.g. in- and outpatients of a clinic, students of an educational institution, incarcerated populations etc., might expect that such data are known prior to the survey.

Two other modifications to the unrelated-question model were proposed by Moors (1971) and Folsom et al. (1973). In order to achieve the efficiency of the model with known π_y, Moors proposed to use one of the samples to estimate the prevalence of nonsensitive characteristic by means of direct-questioning and use this parameter in the estimation of π in the other sample where the randomizing device is used (e.g. Soeken and Damrosch (1986)). In Folsom's design two nonsensitive questions are used with a sensitive one in both samples. The extension of Moors' idea is reflected in the fact that in the first sample one of the innocuous questions is asked directly while the other one is paired with the sensitive question. The roles of the nonsensitive questions are reversed in the second sample. The sensitive question is selected with probability p in both samples.

5.4.3 Forced Response Method (FRR)

Another well-known modification of the unrelated-question method is the forced response method (Boruch, 1971a). As mentioned by Fox and Tracy (1986), the nonsensitive question appears redundant if the nonsensitive response generation is built into the randomizing device such that the proportion of nonsensitive responses is also known a priori.

Tracy and Fox (1981) used an introductory RR example to measure prevalence of spouse abuse. It is not very likely that a question of such a sensitive nature would be honored with an honest answer when asked directly. Therefore, each man, in a sample of hundred men, was asked to toss a coin before responding and to raise his hand in case he had ever abused his wife or if the outcome of the coin toss revealed a head. Even though the individual coin toss outcome is not known to the researcher it is still possible to obtain reliable prevalence estimates based on the masked responses and probability distribution of coin tosses. For example, when fifty eight hands were raised one could argue that fifty men out of hundred tossed head and raised their hand for that reason, while eight remaining hand-raisers tossed tail but raised their hand due to the fact of abuse. Therefore, the prevalence of abuse can be estimated as eight out of fifty or sixteen percent. This implies that eight individuals in the head tossing group might have abused their spouses as well but tossed head and raised their hands without revealing their true status, i.e. did not have to answer the posed question. Due to the protective quality of the randomizing device individual hand-raisers cannot be associated with abusive behavior. That is, the true status of a participant cannot be determined. Nonetheless, an estimate of abuse prevalence on an aggregate (population) level can be obtained.

In the setting of a dichotomous forced RR method, the outcome of the randomizing device determines whether an honest or a predefined/forced response has to be provided. Figure 5.4 illustrates this response procedure. The true proportion of A, the class possessing the sensitive characteristic, is denoted by π. Each respondent belongs to one of the two mutually exclusive classes. All respondents in a sample receive the same dichotomous question. Before responding each participant performs a randomization of the response type, i.e. an honest response is chosen with probability p_1, and a forced response with probability $1 - p_1$. In addition, in case the forced response is administered, forced affirmative and forced negative responses are requested with probabilities p_2 and $1 - p_2$, respectively. Depending on the individual status and the type of the response selected by the randomizing device, respondent gives an affirmative or a negative response.

The probability of observing an affirmative response under the forced randomized response design is the sum of three probabilities from two classes: (1) the honest and the forced affirmative responses received from respondents of the sensitive class with probabilities πp_1 and $\pi(1-p_1)p_2$, respectively, and (2) the forced affirmative responses obtained from the complimentary nonsensitive class with probability $(1-\pi)(1-p_1)p_2$. The proportion of observed affirmative responses λ expressed in terms of randomizing device probabilities p_1 and p_2 and proportion of respondents possessing the sensitive attribute π takes the following form:

$$\lambda = p_1\pi + (1-p_1)p_2. \tag{5.8}$$

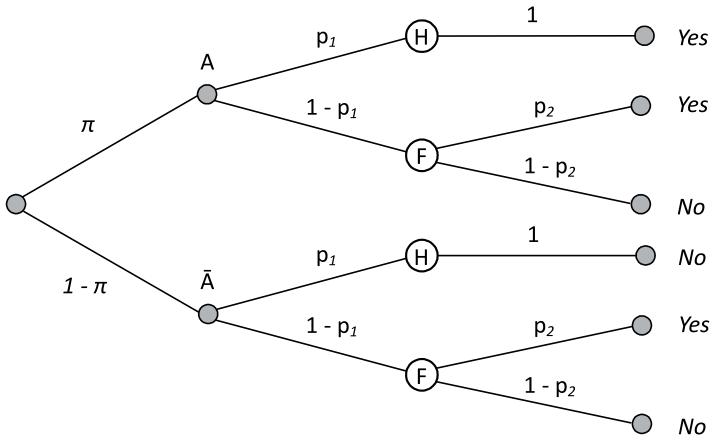

Figure 5.4: Probability tree of the forced RR model for a dichotomous item.

Then, the maximum likelihood estimate for π is

$$\hat{\pi} = \frac{\lambda - (1 - p_1)p_2}{p_1} \tag{5.9}$$

with

$$var(\hat{\pi}) = \frac{\lambda(1 - \lambda)}{np_1^2}. \tag{5.10}$$

This method pertains to dichotomously scored responses. However, many surveys use items with a multiple-point ordinal response format. An example of such a response format is a Likert scale with, for instance, four response categories: strongly disagree, disagree, agree, and strongly agree. For such items, Abul-Ela et al. (1967) reported about a trichotomous randomized response model, which was an extension of Warner's design. It required an additional sample and enabled proportion estimation of three related, mutually exclusive groups. One sample was shown to be enough for multinomial proportion estimation by Eriksson (1973), where the randomization was performed using a deck of cards.

The Forced RR method is easily extended to multiple response categories. Here, the probability of a forced response $p_2(c)$ is defined per category, $c = 1, \ldots, C$ and is built into the randomizing device. Note that at most $C - 1$ response categories are stigmatizing. For example, the category $c = 1$ represents the situation of no social stigma, and levels of social stigma attached to $c = 2, \ldots, C$ increase with the category index. The complete probability tree for a C-point ordinal item is shown in Figure 5.5 and is rather similar to that depicted in Figure 5.4. Each respondent belongs to one of C mutually exclusive groups A_c with probability π_c, where $\bigcup_{c=1}^{C} A_c = \Omega$ and $\sum_{c=1}^{C} \pi_c = 1$. When confronted with an item in a conventional survey setting, an individual belonging to the class A_1 provides an honest "Response 1" with probability π_1. However, in a forced randomized response setting an honest choice is observed with probability $p_1\pi_1$ because with probability $(1 - p_1)\pi_1$ a forced response is given. Thus, with probability $(1 - p_1)\pi_1 p_2(1)$

a forced "Response 1" is given by a respondent of group A_1. With probability $\sum_{c=2}^{C}(1-p_1)\pi_c p_2(1)$ a respondent not belonging to group A_1 is requested to give a forced category 1 response. The interpretation of the rest of the tree follows the same logic.

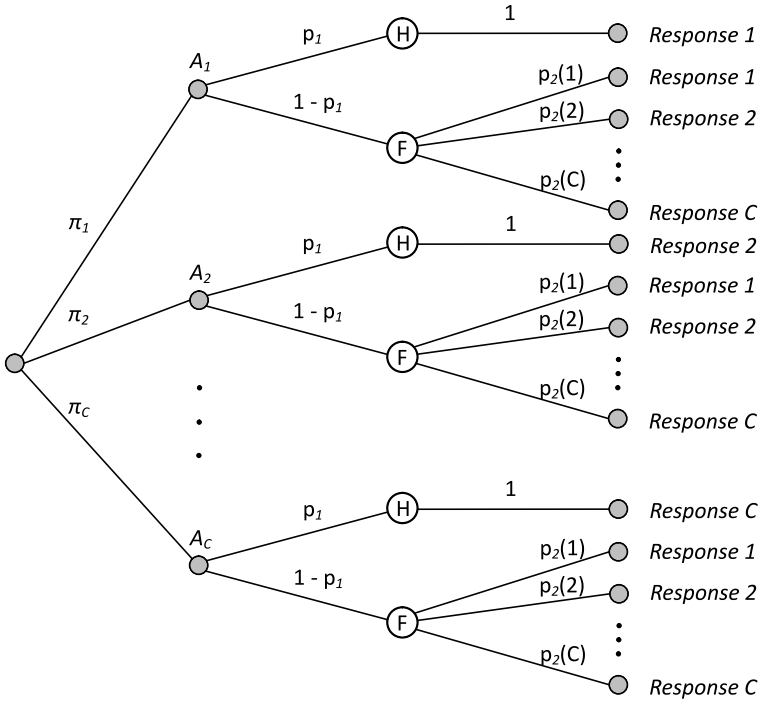

Figure 5.5: Probability tree of the forced RR model for an ordinal item.

For $c = 1, \ldots, C$, let n_c denote the observed counts of "Response c", in a random sample of n respondents. Then, the proportion $\lambda_c = \frac{n_c}{n}$ equals

$$\lambda_c = p_1\pi_c + (1-p_1)p_2(c), \tag{5.11}$$

with MLE for π_c

$$\hat{\pi}_c = \frac{\lambda_c - (1-p_1)p_2(c)}{p_1} \tag{5.12}$$

and

$$var(\hat{\pi}_c) = \frac{\lambda_c(1-\lambda_c)}{np_1^2}. \tag{5.13}$$

5.4.4 Smoke Study: "Do you smoke?"

The FRR method is illustrated with two empirical examples. First, a fully Bayesian approach is used to model univariate binary forced randomized responses. Second, univariate ordinal forced randomized response data will be analysed. Within full

Bayesian approach, prior distributions for all model parameters are specified and all parameters are estimated simultaneously using MCMC.

The dichotomous forced randomized response technique will be illustrated by its application to the first item of the Smoking scale questionnaire. Lung patients of Medical Spectrum Twente (MST) were questioned about their smoking behavior. They were interviewed directly or interviewed via the forced randomized response technique about their smoking behavior using a spinner where an honest response is required with probability $p_1 = 0.778$ and a forced positive answer with probability $p_2 = 0.5$. Table 5.1 presents the number of observations for the first item ("Do you smoke?") as well as the aggregate-level prevalence estimates among males and females in the randomized response group (i.e., those that were interviewed using RR).

The smoking prevalence estimates were obtained by MLE (see Equation 5.9) and using full Bayesian estimation. In a fully Bayesian approach, hyperpriors are defined for the prior parameters, and the model can be formulated as

$$\begin{aligned} Y_i \mid \pi_i &\sim \mathcal{BIN}(\Delta(\pi_i)), \\ \Delta(\pi_i) &\sim \mathcal{B}e(a + \beta_1 Female, b + \beta_2 Female), \\ a, b &\sim \mathcal{U}(1, 100), \\ \beta_1, \beta_2 &\sim \mathcal{N}(0, 1), \end{aligned}$$

where Y_i represents a binary response of patient i, $\Delta(\pi_i)$ the linearly transformed response rate of person i, which depends on the parameters of the forced randomized response design utilized,

$$\Delta(\pi_i) = p_1\pi_i + (1 - p_1)p_2,$$

and a and b are the prior parameters, and β_1 and β_2 the female effects. The expected population prevalence of smoking among males and females is

$$E(\pi_i; male) = \frac{a}{a + b}$$

and

$$E(\pi_i; female) = \frac{a + \beta_1}{a + \beta_1 + b + \beta_2},$$

respectively.

To estimate this model, a total of 150,000 MCMC iterations were performed and the first 50,000 iterations were discarded when the posterior means and standard deviations of the parameters were estimated. The MCMC estimates for males and females closely resembled the maximum-likelihood estimates. However, the Bayes estimates show slightly more shrinkage to a general population mean and were closer together. The estimated population proportion of female smokers was equal to the male smokers. The WinBUGS listing accompanying this example can be found in Appendix F.

Table 5.1: Prevalence of smoking among males and females based on item 1 of the smoking scale questionnaire.

Group		Response	N	ML mean (SD)	MCMC mean (SD)
Total					
		Agree	69	.305 (.044)	.312 (.042)
		Disagree	129	–	–
Gender					
	Female	Agree	37	.319 (.061)	.316 (.042)
		Disagree	66	–	–
	Male	Agree	32	.290 (.062)	.315 (.042)
		Disagree	63	–	–

5.4.5 Smoke Study: "How many cigarettes are you smoking per day?"

The analysis of an ordinal item under the forced randomized response model is illustrated with Item 11 of the Smoking scale questionnaire ("How many cigarettes are you smoking per day?" with response alternatives "None", "10 or less" and "More than 10"). Respondents in the RR group were interviewed via the forced randomized response technique using a spinner where an honest response is required with probability $p_1 = 0.611$ and a forced response alternatives are prescribed with probabilities $p_2(1) = p_2(3) = 0.25$ and $p_2(2) = 0.5$.

The smoking prevalence estimates are obtained using a full Bayesian model defined by

$$\begin{aligned} Y_i \mid \pi_{ic} &\sim \mathcal{C}ategorical(\Delta(\pi_{ic})), \\ \Delta(\pi_{ic}) &\sim \mathcal{D}(a_1, a_2, a_3), \\ a_1, a_2, a_3 &\sim \mathcal{U}(1, 100), \end{aligned}$$

where Y_i is an observed ordinal response of patient i, $\Delta(\pi_{ic})$ the linearly transformed response rate of person i, which depends on the parameters of the forced randomized response design utilized,

$$\Delta(\pi_{ic}) = p_1\pi_{ic} + (1 - p_1)p_2(c).$$

The prior expected population prevalence of cigarette smoking per category is

$$E(\pi_{ic}) = \frac{a_c}{\sum_{c=1}^{3} a_c}.$$

Table 5.2: Prevalence of smoking among males and females based on Item 11 of the smoking scale questionnaire.

Group	Response	N	ML mean (SD)	MCMC mean (SD)
Total				
	"None"	102	.683 (.058)	.667 (.057)
	"10 or less"	58	.161 (.053)	.170 (.054)
	"More than 10"	38	.155 (.046)	.163 (.046)

The model in Equation 5.14 was not directly defined in WinBUGS (Lunn et al., 2000). The model was restated in terms of series of Beta distributions, which was implemented in WinBUGS. See Section 2.5.2, for an extensive description for defining the Dirichlet-multinomial model in components of beta-binomials.

Table 5.2 presents the number of observations per category for this item, as well as the MLE and the Bayesian prevalence estimates. A total of 150,000 MCMC iterations were performed with a burn-in of 50,000 iterations. The MCMC estimates per scoring category closely resembled the maximum-likelihood estimates. The WinBUGS listing accompanying this example can be found in Appendix G.

5.5 Nonrandomized Response Techniques

As an alternative to the randomized response techniques, nonrandomized methods have been developed that do not require a randomizing device. The methods were developed in the late 70's (Takahasi & Sakasegawa, 1977) and have received increased attention in the past five years. Related developments are the hidden sensitivity method (Tang, Tian, Tang, & Liu, 2009; Tian et al., 2007), the triangular method (Tan, Tian, & Tang, 2009; Tian, Tang, Liu, Tan, & Tang, 2011; Yu, Tian, & Tang, 2008), the crosswise method (Jann, Jerke, & Krumpal, 2012; Yu et al., 2008), the method of single sample count (Petróczi et al., 2011), and the unmatched count technique (Coutts & Jann, 2011). The techniques are usually claimed to be free from the limitations of the randomized approach and to increase the relative efficiency and the degree of privacy protection. Four models that will be discussed in detail are the Takahasi's RR technique, the triangular and the crosswise methods, and the hidden sensitivity method.

5.5.1 Takahasi's RR Technique

In 1977, Takahasi and Sakasegawa proposed a technique that randomized responses without using a randomizing device. They argued that the developed technique is superior to traditional RR approaches due to its extended applicability. The technique can be applied not only to face-to-face interviews but also to mail surveys,

since a randomizing device is not required. The randomizing device is substituted by an auxiliary question. For example, a respondent can be asked to make an undisclosed choice between two (or more) "options". A possible question could be "Which do you prefer 'spring' or 'autumn'?" or "Is your favorite choice of color red, blue or yellow?" The choices of response categories must be mutually exclusive.

Let π be the proportion of respondents having sensitive attribute A, and C the number of response categories of the non-sensitive auxiliary question with outcomes B_c, where $c = 1, \ldots, C$. The technique requires C independent samples. Each respondent is asked to respond to the auxiliary question and proceed with the sensitive question according to instructions, which are summarized in Table 5.3.

Table 5.3: Takahasi model: Response instruction scheme.

	Sample 1		Sample 2		...	Sample $C-1$		Sample C	
	A	not A	A	not A	...	A	not A	A	not A
B_1	0	1	1	0	...	1	0	1	0
B_2	1	0	0	1	...	1	0	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
B_{C-1}	1	0	1	0	...	0	1	1	0
B_C	1	0	1	0	...	1	0	0	1

Table 5.3 reads as follows. The affirmative and negative responses are coded as 1 and 0, respectively. An untraceable mix of affirmative and negative responses is obtained, since participants in different samples receive different instructions. For example, instructions received by respondents in "Sample 1" are summarized in two columns. All respondents possessing the sensitive attribute A except the ones that have chosen B_1 are instructed to give an affirmative response (i.e. response of type 1), and otherwise a negative response (i.e. response of type 0). Participants with outcome B_1 are requested to respond 0 if they belong to the sensitive class and 1 otherwise. Generalizing to "Sample c" ($c = 2, \ldots, C$), respondents with outcome B_c ($c = 2, \ldots, C$) respond 0 in case of possessing the sensitive characteristic and 1 otherwise, while the others of "Sample c" respond vice versa.

Let $P(A, c)$ denote the proportion of persons in the population possessing attribute A and choose outcome B_c, and $\sum_{c=1}^{C} P(A, c) = \pi$. Similarly, $P(\bar{A}, c)$ is the proportion of persons not possessing attribute A that choose B_c, and $\sum_{c=1}^{C} P(\bar{A}, c) = 1 - \pi$. Let n_c denote the sample size and λ_c the proportion of affirmative responses. The probability of observing λ_c positive responses equals

$$\lambda_c = \pi + P(A, c) + P(\bar{A}, c).$$

The MLE estimate of π is

$$\hat{\pi} = \sum_{c=1}^{C} \lambda_c - 1 \tag{5.14}$$

with variance

$$var(\hat{\pi}) = \frac{\sum_{c=1}^{C} \lambda_c(1-\lambda_c)}{n_c}. \tag{5.15}$$

This method circumvents a problem of the traditional unrelated-question technique, where the outcome distribution of the unrelated question have to be determined or known beforehand (see Fox and Tracy (1986) for scarcity of questions with known probability distributions and problems with demographic-type questions). This follows directly from expressions (5.14) and (5.15) that do not contain parameters related to the auxiliary question.

The disadvantage of the method is that to make inferences about the prevalence of A more than one data sample is required. The instructions can be modified such that the resulting model will resemble the forced RR model with the same prevalence estimate and variance as stated by expressions (5.14) and (5.15), respectively. In that case, two samples are required. If the auxiliary question probabilities are known and the non-sensitive characteristic is independent of the sensitive characteristic, the method resembles Warner's method.

5.5.2 The Triangular and Crosswise RR Methods

Two other randomized response models that also do not require a (physical) randomizing device, are the triangular model (TM) and the crosswise method (CM) proposed by Yu et al. (2008). Two questions are asked, a sensitive question and an innocuous question with a known population prevalence, and a joint answer to both questions is required.

Let $U = 1$ denote the group of people that possess the sensitive characteristic, and $U = 0$ otherwise. Let W be an innocuous dichotomous random variable independent of U. It is assumed that W is selected such that proportion $p = P(W = 1)$ is known. The quantity in question is the proportion of the sensitive characteristic $\pi = P(U = 1)$.

The sensitive and innocuous response options are placed in a two by two contingency table where two quadrants represent response categories of the innocuous question with cells $\{U = 0, W = 0\}$ and $\{U = 0, W = 1\}$, while the other two quadrants represent response categories of the sensitive question with outcomes $\{U = 1, W = 0\}$ and $\{U = 1, W = 1\}$.

In the triangular model, a response is requested to both questions, which reveals whether he/she belongs to subgroup $\{U = 0, W = 0\}$ or subgroup $\{U = 1\} \bigcup \{U = 0, W = 1\}$. Both subgroups are considered to be non-sensitive since the second subgroup merely implies that a respondent possesses the nonsensitive characteristic $W = 1$, which stimulates truthful responses. The proportion of respondents scoring in the first subgroup equals

$$\lambda = (1 - \pi)(1 - p),$$

with MLE estimate of π

$$\hat{\pi} = 1 - \frac{\lambda}{1-p},$$

with variance

$$var(\hat{\pi}) = \frac{\lambda(1-\lambda)}{n(1-p)^2}.$$

Tian et al. (2011) derived a sample size formulae for this method based on a power analysis.

In the crosswise model, different sets of non-sensitive subgroups are defined, namely $\{U = 0, W = 0\} \bigcup \{U = 1, W = 1\}$ and $\{U = 0, W = 1\} \bigcup \{U = 1, W = 0\}$. The probability of respondent i scoring in subgroup one equals

$$\lambda = (1-\pi)(1-p) + \pi p, \tag{5.16}$$

where $p \neq 0.5$. The crosswise model is also referred to as the nonrandomized Warner model by Tian et al. (2011) (see Equation 5.1), and the resulting prevalence estimate and variance have the same form as those of the opposite-question method (see Equations 5.2 and 5.3).

Jann et al. (2012) implemented and evaluated this technique, and compared the performance to the direct questioning method. They applied the crosswise model to measure the prevalence of plagiarism in student papers. They noted that in this crosswise method a respondent never have to give a self-protective answer, which, for example, would be possible for the triangular method when falling in subgroup $\{U = 0, W = 0\}$.

An extensive comparison of the efficiency of these techniques with the traditional randomized response models can be found in Tan et al. (2009). The attractive features of the triangular and the crosswise models are (1) redundance of a randomizing device, (2) reproducibility of results (in case the survey was repeated respondents would score in the same subgroups), (3) the fact that respondents are not asked to give an answer to the sensitive question, and (4) their applicability to face-to-face as well as to mail and internet surveys. Nevertheless, the major disadvantage of the nonrandomized methods is that they require prior knowledge of the population prevalence for some innocuous question. This feature was the main drawback of the traditional unrelated-question method.

5.5.3 The Hidden Sensitivity RR Method

In the method of hidden sensitivity (HS) described by Tian et al. (2007), two sensitive questions (e.g. a question on drug abuse history and a question on AIDS history) with binary outcomes are handled simultaneously. This method makes no use of a randomizing device, and to stimulate respondents' cooperation a non-sensitive question is used to indirectly obtain respondents' answers to two sensitive questions. The innocuous question is selected such that the outcome variable X assumes four different values with known probability p_i $(i = 1, \ldots, 4)$ and is independent of U and W. Let $U = 1$ and $W = 1$ denote the sensitive characteristics of a respondent. Then, response combinations $\{U = 0, W = 1\}$, $\{U = 1, W = 0\}$, and $\{U = 1, W = 1\}$ all carry an incriminating loading.

A four by four contingency table can be constructed as follows. First, each respondent is asked to respond to the non-sensitive question and then the two sensitive questions. If a participant does not possess any of the two sensitive attributes, his answer falls in one of the cells $X = i$ $(i = 1, \ldots, 4)$ of a non-sensitive category $\{U = 0, W = 0\}$. If a respondent belongs to one of the sensitive groups, he is asked to tick a non-sensitive cell $\{X = i\} \bigcap \{U = 0, W = 0\}$. As a result all answers will be accumulated under the outcome $\{U = 0, W = 0\}$. Probabilities are defined as $p_1^* = P(U = 0, W = 0)$, $p_2^* = P(U = 0, W = 1)$, $p_3^* = P(U = 1, W = 0)$, and $p_4^* = P(U = 1, W = 1)$.

Let n_i denote the observed frequency of response $X = i$, where $n = \sum_{i=1}^{4} n_i$. The probability to observe n_i category i responses is given by

$$\lambda_i = \frac{n_i}{n} = \begin{cases} p_i p_1^* & \text{if } i = 1 \\ p_i p_1^* + p_i^* & \text{if } i = 2, 3, 4. \end{cases} .$$

The Expectation-Maximization algorithm can be used to obtain MLEs of the cell probabilities and the bootstrap approach to estimate standard errors.

The (dis-)advantages of this method are comparable to that of the triangular and crosswise methods. It has to be stressed that in this design a respondent only gives a self-protective answer when scoring in the non-sensitive subgroup $\{U = 0, W = 0\}$. The method has the same implementation problem as the triangular and crosswise method, namely the population proportions related to the innocuous question have to be known.

Tang et al. (2009) adopted the idea of assembling the observed frequencies by means of adding up the number of respondents scoring to non-sensitive response category 1 of both sensitive questions and and the number of respondents scoring to category i of the non-sensitive question. They applied the hidden sensitivity model to a multi-category response model for a single sensitive question. The method was illustrated using a real data set from a questionnaire on sexual activities in Korean adolescents.

Other recently developed methods in the field of nonrandomized methods, that are not making use of a randomizing device, are the single sample count method (Petróczi et al., 2011) and the unmatched count method (Coutts & Jann, 2011).

Despite the fact that nonrandomized response techniques are claimed to be more efficient, and more cost-effective than the traditional randomized response techniques, they have some shortcomings. Either prior knowledge of the population prevalence for an innocuous question is required or more than one sample is required for making accurate inferences. Furthermore, the use of a randomizing device is avoided but this does not remove the persistent dilemma of cooperation versus efficiency. The nonsensitive question controls in a way the cooperation and efficiency of the nonrandomized method, which makes it much more complicated to take active control in establishing a certain accuracy. Methods that use a randomizing device have the advantage that the properties of the randomizing device, which influence the respondent's cooperation and the efficiency, can be adjusted. This seems to be rather difficult when a nonsensitive question is used. Furthermore, it is not clear if the nonrandomized methods can be generalized to perform a scale analysis.

5.6 Randomized Response Methods and Multi-Item Measurements

Traditionally, common randomized response techniques are limited to single-item measures. However, in the social and behavioral sciences measures usually consists of multiple items. The fact that univariate randomized response techniques limit inferences to the population level has also hampered their application. Population level inferences for multi-item scale data were discussed by Himmelfarb (2008). Contrary to the univariate randomized response techniques, individual-level inferences can be made when observing multiple randomized responses leading to multivariate outcome data. This can be done by extending the randomized response procedure in such a way that each scale item is administered using the randomized response technique. In such a multivariate setting, the variability contributed by the randomized response procedure will be reduced. This leads to more accurate population parameter estimates than those obtained via single-item randomized response methods.

5.6.1 Multi-Item Randomized Response Models

Different approaches to the randomized response analysis of multivariate data are reported in the literature. The methods, that have been developed to analyse multivariate randomized response data, differ with respect to the sample size and the assumed sampling distribution. In 2005, Fox proposed a class of randomized IRT models within a Bayesian framework, while Böckenholt and van der Heijden (2007) discussed a comparable class of so-called item randomized-response models in a frequentist framework.

In their monograph on item response theory, Embretson and Reise (2000) indicate 500 respondents as the minimal sample size delivering stable parameter estimates. As was mentioned before, data collected using a randomized response procedures contain a substantial amount of noise, thus less information about the underlying construct will be available compared to the conventional direct-questioning method. To achieve the same level of precision, even larger samples are required under randomized response modifications of IRT models. For situations where sufficiently large data sets are not available, individual-level inferences are possible using a different methodology. That is, for this case models for relatively small data sets, as the beta-binomial model for dichotomous data (Fox, 2008) and the Dirichlet-multinomial model for categorical ordinal data (Avetisyan & Fox, 2012), have been developed.

The traditional FRR method for analysing randomized response data is extended for a multiple item scale analysis. The randomizing device characteristics governing the RR data collection are allowed to vary over the scale items. However, in the present work this option will not be considered. Let the randomizing device with honest response selection probability p_1 and forced affirmative response probability p_2 be specified for dichotomous item k, where $k = 1, \ldots, K$. Then, U_{ik} is the latent honest response of person i, $i = 1, \ldots, N$, to item k if asked directly, further to be called the true response of person i to item k. It is required that each participant performs an independent forced randomized response trial for

each item k before responding. Then, the probability of an observed affirmative randomized response Y_{ik} is expressed by

$$P(Y_{ik} = 1 \mid p_1, p_2) = p_1 P(U_{ik} = 1) + (1 - p_1)p_2. \tag{5.17}$$

For a polytomous randomized response, let $p_2(c)$ denote the probability of a forced response in category c for $c = 1, \ldots, C_k$, such that the number of response categories can vary over items and that at most $C_k - 1$ response categories are stigmatizing. The probability of an observed randomized response of individual i in category c of item k is given by

$$P(Y_{ic} = c \mid p_1, p_2) = p_1 P(U_{ik} = c) + (1 - p_1)p_2(c). \tag{5.18}$$

It is assumed, that if $p_1 > 1/2$, the randomized response data will contain sufficient information to make meaningful inferences.

5.6.2 The Beta-Binomial and Dirichlet-Multinomial Modeling Approach

For relatively small data sets, multivariate randomized response data can be analyzed using the beta-binomial or the Dirichlet-multinomial model. The analysis allows for the estimation of individual response probabilities given individual count data, binomial, or categorical responses.

Let N participants respond to K binary items. Separate observation u_{ik} of i^{th} participant to k^{th} item is Bernoulli distributed with probability π_{ik}. The binomial probability function can be used for describing the respondent's affirmative response count data if all responses are independent from one another and if the response rate is constant, i.e. π_i, for $k = 1, \ldots, K$. A constant response probability can be assumed for multi-item scales measuring behavior, interest or opinion, since the items measuring this constructs are often descriptive in nature. Originally proposed by Lord (1965), the beta-binomial model allows modelling of the variation in individual responses via a binomial distribution, and the response rate variation over respondents by means of conjugated beta prior distribution with parameters $\tilde{a}$ and $\tilde{b}$, that is,

$$\begin{aligned} U_{i.} \mid \pi_i &\sim \mathcal{BIN}(K, \pi_i), \\ \pi_i \mid \tilde{a}, \tilde{b} &\sim \mathcal{B}(\tilde{a}, \tilde{b}), \end{aligned}$$

where $U_{i.} = \sum_k U_{ik}$. If the multivariate dichotomous data are observed via forced randomized response design, the probability of observing a positive response from participant i to item k is related to the true response by the expression

$$\begin{aligned} P(Y_{ik} = 1 \mid \pi_i) &= p_1\pi_i + (1 - p_1)p_2 \\ &= \Delta(\pi_i). \end{aligned} \tag{5.19}$$

A Beta-binomial model accommodates the forced randomized response sampling mechanism by modeling linearly transformed response rates. The model is given by

$$\begin{aligned} Y_{i.} \mid \pi_i &\sim \mathcal{BIN}(K, \Delta(\pi_i)), \\ \Delta(\pi_i) &\sim \mathcal{B}(a, b), \end{aligned}$$

where the transformation parameters p_1 and p_2, the characteristics of the randomizing device, are known beforehand. Bayes estimate of the individual's response rate and its variance is obtained by using (5.19), and is described in detail in Section 2.3.

The Beta-binomial model for multivariate randomized response data can be generalized to multinomial data such that individual category-response rates can be estimated. This can be accomplished using the conjugate Dirichlet prior distribution.

With the Dirichlet-multinomial model, the variation in individual category-response count data via a multinomial distribution is explicitly modeled. The category-response rate variation over respondents are modeled through a conjugate Dirichlet prior distribution with parameters $\tilde{a}_1, \ldots, \tilde{a}_C$,

$$\begin{aligned} U_{i.1}, \ldots, U_{i.C} \mid \pi_{i1}, \ldots, \pi_{iC} &\sim Mult(K, \pi_{i1}, \ldots, \pi_{iC}), \\ \pi_{i1}, \ldots, \pi_{iC} &\sim \mathcal{D}(\tilde{a}_1, \ldots, \tilde{a}_C), \end{aligned}$$

where $U_{i.c} = \sum_k U_{ikc}$ for $c = 1, \ldots, C$. The individual observed forced randomized response count data per response category across K items for respondent i ($i = 1, \ldots, N$) are stored in a vector $\mathbf{y}_{i.} = (y_{i.1}, \ldots, y_{i.C})^t$ and are assumed to be multinomially distributed given the individual transformed category-response rates $\Delta(\pi_i) = (\Delta(\pi_{i1}), \ldots, \Delta(\pi_{iC}))$. The latter are obtained by linear transformation from the true individual category-response rates π_i,

$$\begin{aligned} P(Y_{ik} = c \mid \pi_{ic}) &= p_1 \pi_{ic} + (1 - p_1) p_2(c) \\ &= \Delta(\pi_{ic}), \end{aligned}$$

and are assumed to follow a Dirichlet distribution with parameters $a_1, \ldots, a_C$. The Dirichlet-multinomial model for the observed randomized count data per category is given by

$$\begin{aligned} Y_{i.1}, \ldots, Y_{i.C} \mid \pi_{i1}, \ldots, \pi_{iC} &\sim Mult(K, \Delta(\pi_{i1}), \ldots, \Delta(\pi_{iC})), \\ \Delta(\pi_{i1}), \ldots, \Delta(\pi_{iC}) &\sim \mathcal{D}(a_1, \ldots, a_C). \end{aligned}$$

Analytic expressions of the posterior mean and standard deviation of the true individual categorical-response rates are derived in Section 2.4. The expressions can be used for estimation given prior knowledge or empirical Bayes estimates of the population response rates. For full Bayes estimation see Section 2.5.2 and Appendix B.

5.6.3 The Randomized Item Response Theory Modeling Approach

Item response theory models relate item characteristics to individual differences in the latent characteristic being measured. Since the paramount idea of IRT modeling is that it scales the item difficulty and the person trait onto the same continuum regardless of the type of the item and response format, IRT models allow simultaneous estimation of mixed data, i.e. dichotomous and polytomous. IRT analysis allows for more complex designs, where different sets of items can

be used to measure persons on one common scale, it handles measurement error at the individual level. Randomized item response technique models provide an opportunity of measuring latent individual sensitive characteristics given observed randomized item responses.

In large-scale surveys on sensitive topics, multiple items aimed at measuring a certain latent construct can be administered in a randomized response manner. However, the resulting set of multiple correlated randomized item responses $\mathbf{Y}$ cannot be analyzed using standard IRT techniques. The techniques have to be modified such that the true item response $\mathbf{U}$, which is latent due to randomization, is modeled. Item response theory models have been extended to handle multi-item RR data by Böckenholt and van der Heijden (2007), Böckenholt et al. (2009), (Fox & Meijer, 2008), Fox (2005b), and Fox and Wyrick (2008), among others.

The two-parameter normal ogive (2PNO) model can be used to describe the probability of an affirmative response to a dichotomous item. The model is given by

$$\begin{aligned} \pi_{ik} &= P(U_{ik} = 1 \mid \theta_i, a_k, b_k) \\ &= \Phi(a_k(\theta_i - b_k)), \end{aligned}$$

where $\Phi(.)$ stands for the cumulative normal distribution function, θ_i is a person parameter, and a_k and b_k are the item discrimination and difficulty, respectively. The FRR model with a two-parameter normal ogive model for the true responses leads to

$$\begin{aligned} P(Y_{ik} = 1 \mid \theta_i, a_k, b_k) &= p_1 P(U_{ik} = 1 \mid \theta_i, a_k, b_k) + (1 - p_1)p_2 \\ &= p_1 \Phi(a_k(\theta_i - b_k)) + (1 - p_1)p_2. \end{aligned}$$

For polytomous item response data, where item responses can be characterized as ordered categorical, the graded response model of Samejima (1969) can be applied. The probability of a response falling in a category c given by respondent i is defined as

$$\begin{aligned} \pi_{ik}(c) &= P(U_{ik} = c \mid \theta_i, a_k, \boldsymbol{b}_k) \\ &= \Phi\left(a_k\left(\theta_i - b_{k,c-1}\right)\right) - \Phi(a_k(\theta_i - b_{k,c})), \end{aligned}$$

where $\boldsymbol{b}_k$ are the ordered vector of thresholds of item k for response categories $c = 1, \ldots, C$: $b_{k1} < \ldots < b_{kC}$. The combination of forced randomized response model with a graded response model for the true responses results in

$$\begin{aligned} P(Y_{ik} = c \mid \theta_i, a_k, \boldsymbol{b}_k) &= p_1 P(U_{ik} = c \mid \theta_i, a_k, \boldsymbol{b}_k) + (1 - p_1)p_2(c) \\ &= p_1 \left[\Phi\left(a_k\left(\theta_i - b_{k,c-1}\right)\right) - \Phi(a_k(\theta_i - b_{k,c}))\right] + (1 - p_1)p_2(c). \end{aligned}$$

In the setting where questionnaire items are measuring more that one sensitive behavior, randomized item response models are extended to multiple dimensions. The extended to multiple latent constructs model consists of two components. First, the multivariate RR data are related to individual's response probabilities via an RR model. Second, the multidimensional model of the latent true responses is assumed to model the actual response process. Suppose, the multidimensional

ability vector $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iQ})^t$ denotes the sensitive characteristics of person i. The probability of a true response in category c is given by

$$\begin{aligned}\pi_{ik}(c) &= P(U_{ik} = c \mid \boldsymbol{\theta}_i, \boldsymbol{A}_k, \boldsymbol{\kappa}_k) \\ &= \Phi\left(\boldsymbol{A}_k^t\boldsymbol{\theta}_i - \kappa_{k,c-1}\right) - \Phi(\boldsymbol{A}_k^t\boldsymbol{\theta}_i - \kappa_{k,c}) \\ &= \Phi\left(\sum_q A_{kq}\theta_{iq} - \kappa_{k,c-1}\right) - \Phi\left(\sum_q A_{kq}\theta_{iq} - \kappa_{k,c}\right),\end{aligned}$$

where $\boldsymbol{A}_k$ is the vector of item discrimination parameters or factor loadings of dimension Q, and $\kappa_{k,c}$ denotes the ordered vector of thresholds (also see Section 4.3.2).

The forced randomized response model with a multidimensional graded response model for the true responses leads to

$$\begin{aligned}P(Y_{ik} = c \mid \boldsymbol{\theta}_i, \boldsymbol{A}_k, \boldsymbol{\kappa}_k) &= p_1 P(U_{ik} = c \mid \boldsymbol{\theta}_i, \boldsymbol{A}_k, \boldsymbol{\kappa}_k) + (1 - p_1)p_2(c) \\ &= p_1\left[\Phi\left(\boldsymbol{A}_k^t\boldsymbol{\theta}_i - \kappa_{k,c-1}\right) - \Phi(\boldsymbol{A}_k^t\boldsymbol{\theta}_i - \kappa_{k,c})\right] + (1 - p_1)p_2(c) \\ &= p_1\left[\Phi\left(\sum_q A_{kq}\theta_{iq} - \kappa_{k,c-1}\right) - \Phi\left(\sum_q A_{kq}\theta_{iq} - \kappa_{k,c}\right)\right] \\ &+ (1 - p_1)p_2(c).\end{aligned}$$

5.6.4 Mixture Modeling for Compliance and Non-Compliance

Inferences based on RR data are most often more accurate than inferences from direct questioning data, if respondents comply with the RR instructions. However, many authors (e.g., Campbell, 1987; Moshagen, Musch, & Erdfelder, 2012; Soeken & MacRready, 1982) argue that randomized response techniques are susceptible to non-compliant behavior. Thus, despite the privacy protection offered by the RR questioning procedure some respondents will exhibit non-compliant behavior. It expresses in the consistent selection of the least stigmatizing response, which violates the integrity of the design.

In the history of the RR methodology, the issue of non-compliance with the RR instructions was addressed as early as in 1969 by Greenberg et al. They implied that Warner's as well as the unrelated-question procedures will be equally affected by the incorrect responses due to the genuine misunderstanding of the design. They claimed that missreporting was resulting exclusively due to self-representational concerns. In addition, Edgell, Duchan, and Himmelfarb (1992) and Lensvelt-Mulders and Boeije (2007) mention that in the forced response setting participants that do not possess the sensitive characteristic might find it difficult to comply with the forced incriminating response of the randomizing device.

Landsheer, van der Heijden, and van Gils (1999) argued that the understanding of and trust in the protection against exposure is not automatically established by the randomized response. They stressed the fact that respondents who fully understand the protective power of the randomized response were more willing to comply with the randomizing device instructions. The interviewer's ability to instruct as well as the respondent's ability to understand the instructions contribute to the level of trust established by the randomized response method.

Clark and Desharnais (1998) proposed to estimate the extent of non-compliance on two different samples by means of collecting RR data using different randomized response designs with an assumption that both randomized designs induce the same level of cheating. Böckenholt and van der Heijden (2007), and Cruyff, van den Hout, van der Heijden, and Böckenholt (2007) proposed a two-component latent class model, where one group follows the instructions, and therefore can be used in useful inference process, while the other group ignores the instructions by providing self-protective response, that does not have to be included in the useful data set. The randomized item response technique model is easily extended to accommodate the non-compliant class. Let G_{ik} denote the binary latent class variable adopting value 1 when person p responds to item i in a self-protective manner, and 0 otherwise. The RIRT is then given by

$$P(U_{ik} = 0) = P(G_{ik} = 0)P(U_{ik} = 0 \mid \theta_i, a_k, b_k) + P(G_{ik} = 1)P(U_{ik} = 0).$$

The resulting model is a mixture consisting of an RIRT model for the compliant class and the non-compliant class.

5.7 Discussion

Respondents possessing a stigmatized characteristic are rather reluctant to give self-incriminating answers. As a result, self-reports on sensitive behavior are susceptible to response and nonresponse bias. To overcome this problem survey techniques assuring confidentiality of given responses are necessary. In this respect, a randomized response technique is a very promising data collection tool when sensitive topic is inquired.

The family of randomized response techniques provides assurances of confidentiality by means of a chance mechanism introduction in the response process. An advantage of randomized response models for sensitive topics characterizes by that individual answers on stigmatized or disapproved behavior are uncoupled from individuals. Resulting data set partially reflects masking properties of the mechanism used. However, based on the known properties of this mechanism, meaningful inferences can be made. Randomized response techniques outperform direct questioning formats with respect to the accuracy of inferences.

Randomized response techniques can be used for estimation of population proportions on the basis of single-item and multi-item measures. Numerous traditional and more recent modifications of techniques operating on binary and categorical ordinal data in a single-item setting have been discussed in literature.

The present overview focused on three popular techniques, namely the Warner's opposite-question, the unrelated-question, and the forced response technique. Randomized response method of Warner is usually criticized due to the sensitivity of both response alternatives, which led to the development of the unrelated-question design. However, it requires an additional innocuous question with a known outcome probability distribution. Such questions are not always readily available and to determine a probability distribution parameter of such unrelated question an additional sample has to be used. The forced response modification of the latter was introduced to overcome the aforementioned requirement of extra parameter estimation. It incorporates the nonsensitive response distribution

into the device governing the chance mechanism. The forced response method is more efficient than traditional Warner's and unrelated-question methods. Texts on the efficiency of randomized response techniques can be found in Bhargava and Singh (2002), Dowling and Shachtman (1975), Lensvelt-Mulders, Hox, and van der Heijden (2005) and Liu and Chow (1976a), among others. Moreover, all abovementioned designs require a physical randomizing device, e.g. a die, a spinner, etc., which makes them not very cost effective.

Nonrandomized response techniques do not require a randomizing device, which makes them rather interesting when considering mail surveys etc. We discussed Takahasi's and the hidden sensitivity methods, as well as the triangular and crosswise designs that rely on an innocuous question for response randomization. Contrary to the traditional RR techniques, nonrandomized techniques are possessing the property of reproducibility. When a nonrandomized technique is used for repeated response collection, equal data sets are obtained. The major disadvantage of the nonrandomized response methods is that an innocuous characteristic with a known population prevalence is necessary, as it was in case of the traditional unrelated-question design.

Different approaches are used to model multivariate RR data depending on the size of the population, from which data were obtained. That is, for small samples individual categorical (binary or ordinal) count data can be used to make inferences on both population and individual levels. Models described are beta-binomial for binary RR data and Dirichlet-multinomial for ordinal RR data. Individual response rates are related to the observed randomized count data. The observed individual response rates are related to the true latent individual response rates by linear transformation possessing characteristics of the masking device used.

Randomized response techniques provide more accurate estimates on sensitive topics compared to conventional direct questioning. However, since respondents are not acquainted with (non-)randomized response questioning mode, the instruction stage requires considerable additional effort. Aims of the research have to be clear to the respondent as well as the procedure of responding. In addition, it is of crucial importance to make sure that a respondent comprehends the full power of randomized response technique in the uncertainty it introduces over each response. If a respondent does not trust the protection offered by the method, issues of noncompliance, where a respondent may tend to lie under randomized response as it was the case under direct questioning interviewing method (e.g. Abul-Ela et al., 1967; Böckenholt et al., 2009; Clark & Desharnais, 1998; Fox & Tracy, 1986), are starting to play a leading role in his response pattern. Non-compliance is usually characterized by choice of the least stigmatizing response alternative. This causes extra perturbation that has to be taken into account in the model to obtain valid statistical inferences.

Appendix A

Derivation of the Marginal Log-Likelihood Function

Here, the derivation of the marginal likelihood expressed by Equation 2.7 is given. The marginal distribution of the RR data given the prior parameters $\boldsymbol{\alpha}$ can be stated as

$$\begin{aligned} p(\mathbf{y} \mid \boldsymbol{\alpha}) &= \prod_i \int_{\Delta(\mathbf{p}_i)} p\left(\mathbf{y}_{i\cdot} \mid \Delta(\mathbf{p}_i)\right) p\left(\Delta(\mathbf{p}_i) \mid \boldsymbol{\alpha}\right) d(\Delta(\mathbf{p}_i)) \\ &= \prod_i \frac{K!}{\prod_c y_{i\cdot c}!} \frac{\Gamma(\alpha_0)}{\prod_c \Gamma(\alpha_c)} \int_{\Delta(\mathbf{p}_i)} \prod_c \Delta(p_{ic})^{\alpha_c + y_{i\cdot c} - 1} d(\Delta(\mathbf{p}_i)) \\ &= \prod_i \frac{K!}{\prod_c y_{i\cdot c}!} \frac{\Gamma(\alpha_0)}{\prod_c \Gamma(\alpha_c)} \frac{\prod_c \Gamma(\alpha_c + y_{i\cdot c})}{\Gamma(\alpha_0 + K)}. \end{aligned}$$

The gamma function Γ can be represented as a factorial function, where $\Gamma(n) = (n-1)!$. Therefore, the marginal distribution can be rewritten in terms of factorial multipliers.

$$p(\mathbf{y} \mid \boldsymbol{\alpha}) = \prod_i \frac{K!}{\prod_c y_{i\cdot c}!} \frac{(\alpha_0 - 1)!}{\prod_c (\alpha_c - 1)!} \frac{\prod_c (\alpha_c + y_{i\cdot c} - 1)!}{(\alpha_0 + K - 1)!}.$$

The factorial multipliers of the last fraction can be manipulated such that

$$\begin{aligned} (\alpha_c + y_{i\cdot c} - 1)! &= \left[\prod_{j=1}^{y_{i\cdot c}} ((\alpha_c - 1) + j) \right] (\alpha_c - 1)! \\ &= \left[\prod_{j=0}^{y_{i\cdot c} - 1} (\alpha_c + j) \right] (\alpha_c - 1)! \end{aligned}$$

and

$$\begin{aligned} (\alpha_0 + K - 1)! &= \left[\prod_{j=1}^{K} ((\alpha_0 - 1) + j) \right] (\alpha_0 - 1)! \\ &= \left[\prod_{j=0}^{K-1} (\alpha_0 + j) \right] (\alpha_0 - 1)!. \end{aligned}$$

The density $p(\mathbf{y} \mid \boldsymbol{\alpha})$ can be rewritten as

$$p(\mathbf{y} \mid \boldsymbol{\alpha}) = \prod_i \frac{K!}{\prod_c y_{i.c}!} \frac{\left[\prod_{j=0}^{y_{i.1}-1} (\alpha_1 + j) \right] \ldots \left[\prod_{j=0}^{y_{i.C}-1} (\alpha_C + j) \right]}{\prod_{j=0}^{K-1} (\alpha_0 + j)}$$

and the the logarithm of the density $p(\mathbf{y} \mid \boldsymbol{\alpha})$ can be stated as

$$l(\boldsymbol{\alpha} \mid \mathbf{y}) \propto \sum_{i=1}^{N} \left[\sum_{j=0}^{y_{i.1}-1} log\,(\alpha_1 + j) + \ldots + \sum_{j=0}^{y_{i.C}-1} log\,(\alpha_C + j) - \right.$$
$$\left. - \sum_{j=0}^{K-1} log\,(\alpha_0 + j) \right],$$

leaving out the first term, which is a constant.

Appendix B

WinBUGS Code: Multinomial-Dirichlet Model Specification

The code of the Dirichlet-multinomial model for RR data, expressed in a form of series of beta-binomial distributions, is given for N persons, K items, and five response categories. The randomizing device has parameters ϕ_1 and ϕ_2.

```
model
{
    for (i in 1:N){
        y[i,1:5] ~ dmulti(q[i,],K)
        q[i,1] ~ dbeta(alpha[1],betatot1)
        q2_star[i] ~ dbeta(alpha[2],betatot2)
        q[i,2] <- q2_star[i] * (1-q[i,1])
        q3_star[i] ~ dbeta(alpha[3],betatot3)
        q[i,3] <- q3_star[i] * (1-q[i,1]-q[i,2])
        q4_star[i] ~ dbeta(alpha[4],alpha[5])
        q[i,4] <- q4_star[i] * (1-q[i,1]-q[i,2]-q[i,3])
        q[i,5] <- 1-q[i,1]-q[i,2]-q[i,3]-q[i,4]
    }

    alpha[1] ~ dunif(0,10)
    alpha[2] ~ dunif(0,10)
    alpha[3] ~ dunif(0,10)
    alpha[4] ~ dunif(0,10)
    alpha[5] ~ dunif(0.5,10)

    alpha0 <- sum(alpha[1:5])

    betatot1 <- sum(alpha[2:5])
    betatot2 <- sum(alpha[3:5])
```

```
    betatot3 <- sum(alpha[4:5])

    for (i in 1:N){
        for (c in 1:5){
            p[i,c] <- (q[i,c]-(1-phi1)*phi2) / phi1
        }
    }
}
```

Appendix C

CAPS-AEQ Questionnaire

The questionnaire comprises seventeen polytomous items.
CAPS: Socio-emotional and community problems
How often (almost always (5), often (4), sometimes (3), seldom (2), almost never (1)) have you had any of the following problems over the past years as a result of drinking too much alcohol?

1. Feeling sad, blue or depressed
2. Nervousness or irritability
3. Hurt another person emotionally
4. Family problems related to your drinking
5. Spent too much money on drugs
6. Badly affected friendship or relationship
7. Hurt another person physically
8. Caused other to criticize your behavior
9. Nausea or vomiting
10. Drove under the influence
11. Spent too much money on alcohol
12. Feeling tired or hung over
13. Illegal activities associated with drug use

AEQ: Sexual enhancement

14 I often feel sexier after I've had a couple of drinks
15 I'm a better lover after a few drinks
16 I enjoy having sex more if I've had some alcohol
17 After a few drinks, I am more sexually responsive

Appendix D

Smoking Scale Questionnaire

The questionnaire comprises nine dichotomous items followed by three polytomous items. In the RR condition, for each item a spinner was spun, and wherever the arrow landed determined whether the item was to be answered honestly or dictated the answer choice to be recorded by a participant. For participants with a condition called color blindness (of any type) a modified digitized spinner disc was available.

1. Do you smoke?
2. Do you feel like smoking when having a cup of coffee or a glass of beer?
3. Do you feel like smoking when somebody else is smoking next to you?
4. Does it disturb you when somebody else is smoking in the same room? (Score reversed)
5. Do you think that smoking is an unhealthy habit? (Score reversed)
6. Do you think that smoking is disgusting? (Score reversed)
7. Do you enjoy the smell of somebody else's cigarette?
8. Do you support the governmental policy concerning the indoor smoking ban? (Score reversed)
9. Are your guests allowed to smoke at your house?
10. How many years have you been smoking/had been smoking? (response categories: never, 15 or less, more than 15)
11. How may cigarettes are you smoking per day? (none, 10 or less, more than 10)
12. How many days per week are you smoking? (none, 4 or less, more than 4)

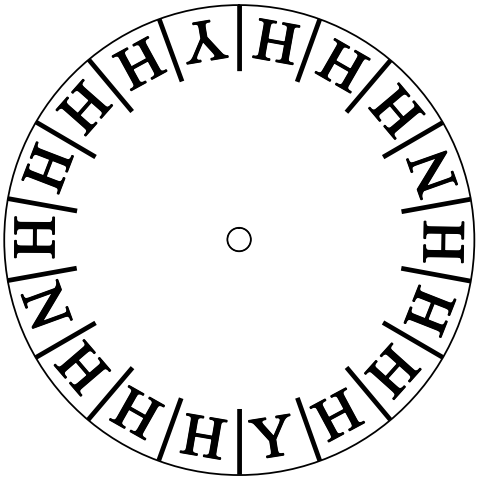

Figure D.1: Spinner disc used for data collection on dichotomous item under the forced RR model.

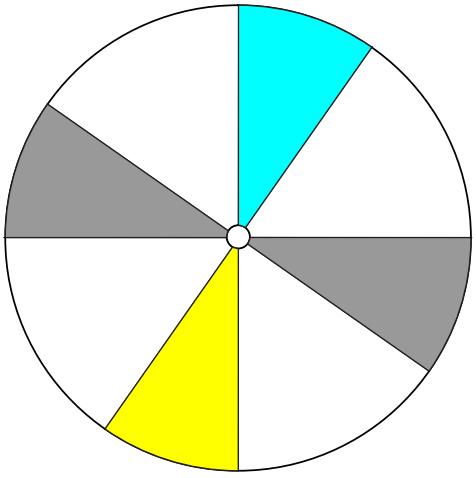

Figure D.2: Spinner disc used for data collection on ordinal item with four categories under the forced RR model.

Appendix E

WinBUGS Code: Mixture Randomized Item Response Model

The code of the mixture randomized item response model for RR and DQ data is given for N persons and K dichotomous and Kp polytomous items. The randomizing device parameters are p1 and p2 for dichotomous response data, and pp1 pp2(c), for polytomous response data. The indicator RR[i] equals zero (one) when participant i belongs to the DQ group (RR group).

```
model{
    for(i in 1:N){              #subjects
        for(k in 1:K){          #dichotomous items
            Q[i,k] <- phi(theta[i] - b[k])
            p[i,k,1] <- (1-RR[i])*(1-Q[i,k])+ RR[i]*(pi1[i]
                        + (1-pi1[i])*(1-(p1*Q[i,k]+(1-p1)*
                        p2)))
            p[i,k,2] <- (1-RR[i])*Q[i,k]+RR[i]*(1-pi1[i])*
                        (p1*Q[i,k]+(1-p1)*p2)
            Y[i,k] ~ dcat(p[i,k,1:2])
        }
        for(kk in 1:Kp){        #polytomous items
            for (c in 1:2){ #response categories
                Qp[i,kk,c] <- phi(bp[kk,c]-theta[i])
            }
            pp[i,kk,1] <- (1-RR[i])*Qp[i,kk,1]+RR[i]*
                          (pi2[i]+(1-pi2[i])*
                          (pp1*Qp[i,kk,1]+(1-pp1)*pp2[1]))
            pp[i,kk,2] <- (1-RR[i])*(Qp[i,kk,2]-Qp[i,kk,1])
                          + RR[i]*(1-pi2[i])*
                          (pp1*(Qp[i,kk,2]-Qp[i,kk,1])+
                          (1-pp1)*pp2[2])
```

```
            pp[i,kk,3] <- (1-RR[i])*(1-Qp[i,kk,2])+
                          RR[i]*(1-pi2[i])*(pp1*(1-
                          Qp[i,kk,2])+(1-pp1)*pp2[3])
            Yp[i,kk] ~ dcat(pp[i,kk,1:3])
        }
        pi1[i] ~ dbern(pi01) #mixture model dichotomous
                             data
        pi2[i] ~ dbern(pi02) #mixture model polytomous
                             data
        theta[i] ~ dnorm(mutheta[i],sigmathetaP)
        mutheta[i] <- beta[1] + beta[2]*RR[i]
    }
    #prior distributions
    for(k in 1:K){
        b[k] ~ dnorm(mub,sigmabP)
    }
    for(kk in 1:Kp){
        bp[kk,1] <- mu[kk]
        bp[kk,2] <- bp[kk,1] + exp(mu2[kk])
        mu[kk] ~ dnorm(-1,0.1)
        mu2[kk] ~ dnorm(0,1)
    }
    pi01 ~ dbeta(1,1)
    pi02 ~ dbeta(1,1)
    beta[1] ~ dnorm(0,100)  #identification latent scale
    beta[2] ~ dnorm(0,1.0E-1)
    sigmathetaP ~ dgamma(1,1)
    sigmatheta <- 1/sigmathetaP
    mub ~ dnorm(0,1.0E-2)
    sigmabP ~ dgamma(1,1)
    sigmab <- 1/sigmabP
}
```

Appendix F

WinBUGS Code: Dichotomous FRR with Gender Effect

Univariate forced randomized response WinBUGS code listing: Estimation of prevalence of smoking among pulmonary patients based on responses to dichotomous Item 1: Do you smoke?

```
model
{
    for (i in 1 : N) {
        PI[i,gender[i]] ~ dbeta(a[gender[i],b[gender[i]])
        Q[i,gender[i]] <- p1*PI[i,gender[i]] + (1-p1)*p2
        Y[i] ~ dbern(Q[i,gender[i]])
    }
    a[1] ~ dunif(1,100)
    b[1] ~ dunif(1,100)
    a[2] <- a[1] + dummy[1]
    b[2] <- b[1] + dummy[2]
    dummy[1] ~ dnorm (0,1)
    dummy[2] ~ dnorm (0,1)
    Pm <- a[1]/(a[1]+b[1])
    Pf <- a[2]/(a[2]+b[2])
}
```

Appendix G

WinBUGS Code: Polytomous FRR

Univariate forced randomized response WinBUGS code listing: Estimation of prevalence of smoking among pulmonary patients based on responses to polytomous Item 11: How many cigarettes are you smoking per day? (none, 10 or less, more than 10)

```
model
{
    for (i in 1:N) {
        Q[i,1] ~ dbeta(a[1],btot1)
        Q2_star[i] ~ dbeta(a[2],a[3])
        Q[i,2] <- Q2_star[i] * (1-Q[i,1])
        Q[i,3] <- 1 - (Q[i,1] + Q[i,2])

        Y[i] ~dcat(Q[i,1:3])
    }
    for (j in 1:3) {
        a[j] ~ dunif(0,10)
    }
    btot1 <- sum(alpha[2:3])
    atot <- sum(a[1:3])
    for (j in 1:3) {
        P[j] <- (a[j]/atot-(1-p1)*p2[j])/p1
    }
}
```

References

Abul-Ela, A. A., Greenberg, B. G., & Horvitz, D. G. (1967). A multi-proportions randomized response model. *Journal of the American Statistical Association*, *62*, 990–1008.

Akers, R. L., Massey, J., Clarke, W., & Lauer, R. M. (1983). Are self-reports of adolescent deviance valid? Biochemical measures, randomized response and the bogus pipeline in smoking behavior. *Social forces*, *62*, 234–251.

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using gibbs sampling. *Journal of Educational Statistics*, *17*, 251–269.

Attebring, M., Herlitz, J., Berndt, A.-K., Karlsson, T., & Hjalmarson, A. (2001). Are patients truthful about their smoking habits? A validation of self-report about smoking cessation with biochemical markers of smoking activity amongst patients with ischaemic heart disease. *Journal of Internal Medicine*, *249*, 145–151.

Avetisyan, M., & Fox, J.-P. (2012). The dirichlet-multinomial model for multivariate randomized response data and small samples. *Psicológica*, *33*, 362–390.

Bargh, J. A., Gollwitzer, P. M., Lee–Chai, A., Barndollar, K., & Trtschel, R. (2001). The automated will: Nonconscious activation and pursuit of behavioral goals. *Journal of Personality and Social Psychology*, *81*, 1014–1027.

Béguin, A. A., & Glas, C. A. W. (2001). Mcmc estimation of multidimensional irt models. *Psychometrika*, *66*, 541–562.

Bhargava, M., & Singh, R. (2002). On the efficiency comparison of certain randomized response strategies. *Metrika*, *55*, 191–197.

Böckenholt, U., Barlas, S., & van der Heijden, P. G. M. (2009). Do randomized-response designs eliminate response biases? an empirical study of noncompliance behavior. *Journal of Applied Econometrics*, *24*, 377–392.

Böckenholt, U., & van der Heijden, P. G. M. (2007). Item randomized-response models for measuring non-compliance: Risk-return perceptions, social influences, and self-protective responses. *Psychometrika*, *72*, 245–262.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompen-

satory multidimensional item response models using markov chain monte carlo. *Applied Psychological Measurement*, *27*, 395–414.

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

Boruch, R. F. (1971b). Maintaining confidentiality on data in educational research: A systemic analysis. *American Psychologist*, *26*, 413–430.

Boruch, R. F., & Cecil, J. S. (1979). *Assuring the confidentiality of social research data*. Philadelphia: University of Pennsylvania Press.

Bradburn, N. M., & Sudman, S. (1979). *Improving interview method and questionnaire design: Response effects to threatening questions in survey research*. San Francisco, CA: Jossey-Bass.

Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking questions: The definitive guide to questionnaire design for market research, political polls, and social and health questionnaires*. San Francisco, CA: Jossey-Bass.

Brier, S. S. (1980). Analysis of contingency tables under cluster sampling. *Biometrika*, *67*, 591–596.

Broemeling, L. D. (2007). *Bayesian biostatistics and diagnostic medicine*. Boca Raton, FL: Chapman and Hall.

Brooks, S., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434–455.

Brown, S. A., Christiansen, B. A., & Goldman, M. S. (1987). The alcohol expectancy questionnaire: An instrument for the assessment of adolescent and adult alcohol expectancies. *Journal of Studies on Alcohol*, *48*, 483–491.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, *75*, 33–57.

Campbell, A. A. (1987). Randomized response technique. *JSTOR Science*, *236*, 1049.

Cannell, C. F., Miller, P. V., & Oksenberg, L. (1981). Research on interviewing techniques. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 389–437). San Francisco, CA: Jossey-Bass.

Chambers, R. (2010). mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software*, *48*.

Chartrand, T. L., & Bargh, J. A. (1996). Automatic activation of impression formation and memorization goals: Nonconscious goal priming reproduces effects of explicit task instructions. *Journal of Personality and Social Psychology*, *71*, 464–478.

Chaudhuri, A., & Mukerjee, R. (1985). Optionally randomized response techniques. *Calcutta Statistical Association Bulletin*, *34*, 225–229.

Chaudhuri, A., & Mukerjee, R. (1988). *Randomized response: Theory and techniques*. New York: Marcel Dekker.

Chaudhuri, A., & Saha, A. (2005). Optional versus compulsory randomized response techniques in complex surveys. *Journal of Statistical Planning and Inference*, *135*, 516–527.

Christofides, T. C. (2003). Honest answers to embarrassing questions: Detecting cheating in the randomized response model. *Metrika*, *57*, 195–200.

Christofides, T. C. (2005). Randomized response technique for two sensitive characteristics at the same time. *Metrika*, *62*, 53–63.

Clark, S. J., & Desharnais, R. A. (1998). Honest answers to embarrassing questions: Detecting cheating in the randomized response model. *Psychological Methods*, *3*, 160–168.

Corstange, D. (2009). Sensitive questions, truthful answers? modeling the list experiment with listit. *Political Analysis*, *17*, 45–63.

Coutts, E., & Jann, B. (2011). Sensitive questions in online surveys: Experimental results for the randomized response technique (rrt) and the unmatched count technique (uct). *Sociological Methods & Research*, *40*, 169–193.

Cross, P., Edwards-Jones, G., Omed, H., & Williams, A. P. (2010). Use of a randomized response technique to obtain sensitive information on animal disease prevalence. *Preventive Veterinary Medicine*, *96*, 252–262.

Cruyff, M. J. L. F., van den Hout, A., van der Heijden, P. G. M., & Böckenholt, U. (2007). Log-linear randomized-response models taking self-protecting response behavior into account. *Sociological Methods & Research*, *36*, 266–282.

Daly, R. J., & Blann, A. D. (1996). Self-reported smoking in vascular disease: the need for biochemical confirmation. *British Journal of Biomedical Science*, *53*, 204–208.

Danaher, P. J. (1988). Parameter estimation for the dirichlet-multinomial distribution using supplementary beta-binomial data. *Communications in Statistics - Theory and Methods*, *17*, 1777–1788.

De Jong, M., Pieters, R., & Fox, J.-P. (2010). Reducing social desirability bias via item randomized response: An application to measure underreported desires. *Journal of Marketing Research*, *47*, 14–27.

De Schrijver, A. (2012). Sample survey on sensitive topics: Investigating respondents understanding and trust in alternative versions of the randomized response technique. *Journal of Research Practice*, *8, issue 1, M1*.

Dishon, M., & Weiss, G. H. (1980). Small sample comparison of estimation methods for the beta distribution. *Journal of Statistical Computation and Simulation*, *11*, 1–11.

Dowling, T. A., & Shachtman, R. H. (1975). On the relative efficiency of randomized response models. *Journal of the American Statistical Association*, *70*, 84–87.

Edgell, S. E., Duchan, K. L., & Himmelfarb, S. (1992). An empirical test of the unrelated question randomized response technique. *Bulletin of the Psychonomic Society*, *30*, 153–156.

Edgell, S. E., Himmelfarb, S., & Duchan, K. L. (1982). The validity of forced responses in a randomized response model. *Sociological Methods Research*, *11*, 89–100.

Edwards, M. C. (2010). A markov chain monte carlo approach to confirmatory item factor analysis. *Psychometrika*, *75*, 474–497.

Elffers, H., Robben, H. S. J., & Hessing, D. J. (1992). On measuring tax evasion. *Journal of Economic Psychology*, *13*, 545–567.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: , EUA : Lawrence Erlbaum Associates.

Eriksson, S. A. (1973). On measuring tax evasion. *International Statistical Review / Revue Internationale de Statistique*, *41*, 101–113.

Folsom, R. E., Greenberg, B. G., Horvitz, D. G., & Abernathy, J. R. (1973). The two alternate questions randomized response model for human surveys. *Journal of the American Statistical Association*, *68*, 525–530.

Formann, K. A., & Kohlmann, T. (1996). Latent class analysis in medical research. *Statistical Methods in Medical Research*, *5*, 179–211.

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

Fox, J.-P. (2005a). Multilevel irt using dichotomous and polytomous items. *British Journal of Mathematical and Statistical Psychology*, *58*, 145–172.

Fox, J.-P. (2005b). Randomized item response theory models. *Journal of Educational and Behavioral Statistics*, *30*, 1–24.

Fox, J.-P. (2008). Beta-binomial anova for multivariate randomized response data. *British Journal of Mathematical and Statistical Psychology*, *61*, 453-470.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York: Springer.

Fox, J.-P., & Meijer, R. R. (2008). Using item response theory to obtain individual

information from randomized response data: An application using cheating data. *Journal of Applied Psychological Measurement*, *32*, 595–610.

Fox, J.-P., & Wyrick, C. (2008). A mixed effects randomized item response model. *Journal of Educational and Behavioral Statistics*, *33*, 389–415.

Franklin, L. A. (1989). A comparison of estimators for randomized response sampling with continuous distributions from a dichotomous population. *Communications in Statistics - Theory and Methods*, *18*, 489–505.

Garrett, E. S., Eaton, W. W., & Zeger, S. (2002). Methods for evaluating the performance of diagnostic tests in the absence of a gold standard: A latent class model approach. *Statistics in Medicine*, *21*, 1289–1307.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., & Smith, A. F. M. (1990). Illustration of bayesian inference in normal data models using gibbs sampling. *Journal of the American Statistical Association*, *85*, 972–985.

Gelfand, A. E., & Smith, A. F. M. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.

Gilks, W., Richardson, S., & Spiegelhalter, D. (Eds.). (1996). *Markov chain monte carlo in practice*. London: Chapman & Hall.

Gingerich, D. W. (2010). Understanding off-the-books politics: Conducting inference on the determinants of sensitive behavior with randomized response surveys. *Political Analysis*, *18*, 349–380.

Goodhardtn, G. J., Ehrenberg, A. S. C., & Chatfield, C. (1984). The dirichlet: A comprehensive model of buying behaviour. *Journal of the Royal Statistical Society. Series A (General)*, *147*, 621–655.

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

Greenberg, B. G., Kuebler, R. R., Abernathy, J. R., & Horvitz, D. G. (1971). Application of the randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association*, *66*, 243–250.

Hill, P., Haley, N. J., & Wynder, E. L. (1983). Cigarette smoking: Carboxyhemoglobin, plasma nicotine, cotinine and thiocyanate vs self-reported smoking data and cardiovascular disease. *Journal of Chronic Diseases*, *6*, 439–449.

Himmelfarb, S. (2008). The multi-item randomized response technique. *Sociological Methods and Research*, *36*, 495–514.

Horvitz, D. G., Greenberg, B. G., & Abernathy, J. R. (1976). Randomized response: A data-gathering device for sensitive questions. *International Statistical Review*, *44*, 181–196.

Horvitz, D. G., Shah, B. U., & Simmons, W. R. (1967). The unrelated question randomized response model. *Proceedings of the Social Statistics Section, Americaan Statistical Association*, 65–72.

Jackman, S. (2001). Multidimensional analysis of roll call data via bayesian simulation: Identification, estimation, inference, and model checking. *Political Analysis*, *9*, 227–241.

Jann, B., Jerke, J., & Krumpal, I. (2012). Asking sensitive questions using the crosswise model an experimental survey measuring plagiarism. *Public Opinion Quarterly*, *76*, 32–49.

Jarvis, M. J., Tunstall-Pedoe, H., Feyerabend, C., Vesey, C., & Saloojee, Y. (1987). Comparison of tests used to distinguish smokers from nonsmokers. *American Journal of Public Health*, *11*, 1435–1438.

Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis.* Upper Saddle River, NJ: Prentice-Hall.

Jones, E., & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin*, *76*, 349-364.

Junker, B. W. (2001). On the interplay between nonparametric and parametric IRT, with some thoughts about the future. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 247–276). New York: Springer-Verlag.

Kim, J. M., & Elam, M. E. (2005). A two-stage stratified warners randomized response model using optimal allocation. *Metrika*, *61*, 1-7.

Kim, J. M., & Warde, W. D. (2005). A mixed randomized response model. *Journal of Statistical Planning and Inference*, *133*, 211-221.

Kuk, A. Y. C. (1990). Asking sensitive questions indirectly. *Biometrika*, *77*, 436–438.

Lamb, C. W., & Stem, D. E., Jr. (1978). An empirical validation of the randomized response technique. *Journal of Marketing Research*, *15*, 616–621.

Landsheer, J. A., van der Heijden, P., & van Gils, G. (1999). Trust and understanding, two psychological aspects of randomized response. *Quality & Quantity*, *33*, 1–12.

Lensvelt-Mulders, G. J. L. M., & Boeije, H. R. (2007). Evaluating compliance with a computer assisted randomized response technique: A qualitative study into the origins of lying and cheating. *Computers in Human Behavior*, *23*, 591-608.

Lensvelt-Mulders, G. J. L. M., Hox, J. J., & van der Heijden, P. G. M. (2005). How to improve the efficiency of randomized response designs. *Quality and Quantity*, *39*, 253–265.

Lensvelt-Mulders, G. J. L. M., Hox, J. J., van der Heijden, P. G. M., & Maas, C. (2005). Meta-analysis of randomized response research: 35 years of validation. *Sociological Methods & Research*, *33*, 319–348.

Liu, P. T., & Chow, L. P. (1976a). The efficiency of the multiple trial randomized response technique. *Biometrics*, *32*, 607–618.

Liu, P. T., & Chow, L. P. (1976b). A new discrete quantitative randomized response model. *Journal of the American Statistical Association*, *71*, 72–73.

Liu, P. T., Chow, L. P., & Mosley, W. H. (1975). Use of the randomized response technique with a new randomizing device. *Journal of the American Statistical Association*, *70*, 329–332.

Lopes, H. F., & West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, *14*, 41-67.

Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika*, *30*, 239–270.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Oxford, England: Addison-Wesley.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

Mangat, N. S. (1994). An improved randomized response strategy. *Journal of the Royal Statistical Society. Series B (Methodological)*, *56*, 93–95.

Mangat, N. S., & Singh, R. (1990). An alternative randomized response procedure. *Biometrika*, *77*, 439–442.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*, 187–212.

Middleton, E. T., & Morice, A. H. (2000). Breath carbon monoxide as an indication of smoking habit. *Chest*, *117*, 758–763.

Monninkhof, E. M., van der Valk, P. D., van der Palen, J., Mulder, H., Pieterse, M., van Herwaarden, C. L., & Zielhuisd, G. (2004). The effect of a minimal contact smoking cessation programme in out-patients with chronic obstructive pulmonary disease: a prepost-test study. *Patient Education and Counseling*, *52*, 231–236.

Moors, J. J. A. (1971). Optimization of the unrelated question randomized re-

sponse model. *Journal of the American Statistical Association*, *335*, 627–629.

Moshagen, M., Musch, J., & Erdfelder, E. (2012). A stochastic lie detector. *Behavior Research Methods*, *44*, 222–231.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, *49*, 65–82.

O'Hare, T. (1997). Measuring problem drinking in first time offenders: Development and validation of the college alcohol problem scale (caps). *Journal of Substance Abuse Treatment*, *14*, 383-387.

Ostapczuk, M., Musch, J., & Moshagen, M. (2011). Improving self-report measures of medication non-adherence using a cheating detection extension of the randomized-response-technique. *Statistical Methods in Medical Research*, *20*, 489–503.

Patz, R. J., & Junker, B. W. (1999a). Applications and extensions of mcmc in irt: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*, 342–366.

Patz, R. J., & Junker, B. W. (1999b). A straightforward approach to markov chain monte carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146–178.

Paul, S. R., Balasooriya, U., & Banerjee, T. (2005). Fisher information matrix of the dirichlet-multinomial distribution. *Biometrical Journal*, *47*, 230-236.

Petróczi, A., Nepusz, T., Cross, P., Taft, H., Shah, S., Deshmukh, N., ... Naughton, D. P. (2011). New non-randomised model to assess the prevalence of discriminating behaviour: a pilot study on mephedrone. *Statistical Methods in Medical Research*, 1–18.

Rasinski, K. A., Visser, P. S., Zagatsky, M., & Rickett, E. M. (2005). Using implicit goal priming to improve the quality of self-report data. *Journal of Experimental Social Psychology*, *41*, 321-327.

Reckase, M. D. (2009). *Multidimensional item response theory.* New York: Springer.

Reinmuth, J. E., & Geurts, D. G. (1975). The collection of sensitive information using a two-stage, randomized response model. *Journal of Marketing Research*, *12*, 402–407.

Rittenhouse, B. E. (1996a). A novel compliance assessment technique: The randomized response interview. *International Journal of Technology Assessment in Health Care*, *12*, 498–510.

Rittenhouse, B. E. (1996b). Respondent-specific information from the randomized

response interview: Compliance assessment. *Journal of Clinical Epidemiology*, *49*, 545-549.

Saha, A. (2007). Optional randomized response in stratified unequal probability samplinga simulation based numerical study with kuks method. *Test*, *16*, 346-354.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34(4, Pt. 2)*.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hill.

Schafer, J. L., & Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, *11*, 437–457.

Scheers, N. J. (1992). A review of randomized response techniques. *Measurement and Evaluation in Counseling and Development*, *25*, 27–41.

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

Sheng, Y. (2010). Bayesian estimation of mirt models with general and specific latent traits in matlab. *Journal of Statistical Software*, *34*, 1–27.

Sheng, Y., & Wikle, C. K. (2007). Comparing multiunidimensional and unidimensional item response theory models. *Educational and Psychological Measurement*, *67*, 899–919.

Shi, J.-Q., & Lee, S.-Y. (1998). Bayesian sampling-based approach for factor analysis models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology*, *51*, 233–252.

Soeken, K. L. (1987). Randomized response research in health research. *Evaluation & the Health Professions*, *10*, 58-66.

Soeken, K. L., & Damrosch, S. P. (1986). Randomized response technique: Applications to research on rape. *Psychology of Women Quarterly*, *10*, 119–126.

Soeken, K. L., & MacRready, G. B. (1982). Respondents' perceived protection when using randomized response. *Psychological Bulletin*, *92*, 487–489.

Song, X.-Y., & Lee, S.-Y. (2001). Bayesian estimation and test for factor analysis model with continuous and polytomous data in several populations. *British Journal of Mathematical and Statistical Psychology*, *54*, 237–259.

Stem, D. E., Jr., & Steinhorst, R. K. (1984). Telephone interview and mail questionnaire applications of the randomized response model. *ournal of the American Statistical Association*, *79*, 555–564.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology.* San Francisco, CA: Jossey-Bass.

Takahasi, K., & Sakasegawa, H. (1977). A randomized response technique without making use of any randomizing device. *Annals of the Institute of Statistical Mathematics*, *29*, 1–8.

Tan, M. T., Tian, G. L., & Tang, M. L. (2009). Sample survey with sensitive questions: A nonrandomized response approach. *American Statistical Association*, *63*, 9–16.

Tang, M. L., Tian, G. L., Tang, N. S., & Liu, Z. (2009). A new non-randomized multi-category response model for surveys with a single sensitive question: Design and analysis. *Journal of the Korean Statistical Society*, *38*, 339-349.

Tian, G. L., Tang, M. L., Liu, Z., Tan, M., & Tang, N. S. (2011). Sample size determination for the non-randomised triangular model for sensitive questions in a survey. *Statistical Methods in Medical Research*, *20*, 159-173.

Tian, G. L., Yu, J. W., Tang, M. L., & Geng, Z. (2007). A new non-randomized model for analysing sensitive questions with binary outcomes. *Statistics in Medicine*, *26*, 4238-4252.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge, England: Cambridge University Press.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*, 859-883.

Tracy, P. E., & Fox, J. A. (1981). The validity of randomized response for sensitive measurements. *American Sociological Review*, *46*, 187–199.

Umesh, U. N., & Peterson, R. A. (1991). A critical-evaluation of the randomized-response method - applications, validation, and research agenda. *Sociological Methods & Research*, *20*, 104–138.

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

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory.* New York: Springer.

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

Werner, M. J., Walker, M. D., & Greene, J. W. (1995). Relation of alcohol expectancies to changes in problem drinking among college students. *Archives of Pediatrics & Adolescent Medicine*, *149*, 733–739.

Wetter, D. W., Kenford, S. L., Welsch, S. K., Smith, S. S., Fouladi, R. T., Fiore, M. C., & Baker, T. B. (2004). Prevalence and predictors of transitions in smoking behavior among college students. *Health Psycholology*, *23*, 168–177.

Williams, B. L., Suen, H. K., & Baffi, C. R. (1993). A controlled randomized response technique. *Evaluation & the Health Professions*, *16*, 225–245.

Wilson, J. R., & Chen, G. S. C. (2007). Dirichlet-multinomial model with varying response rates over time. *Journal of Data Science*, *5*, 413–423.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*, 5879.

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, *31*, 83–105.

Yu, J. W., Tian, G. L., & Tang, M. L. (2008). Two new models for survey sampling with sensitive characteristic: Design and analysis. *Metrika*, *67*, 251–263.

Samenvatting

Binnen de gedrags-, gezondheids- en sociale wetenschappen berusten de meeste metingen op zelfrapportage, waarbij de respondent zelf optreed als waarnemer van zijn eigen mening. De meest geaccepteerde manier van data verzameling bij zelfrapportage is het afnemen van een enqute met directe vragen. De respondent wordt gevraagd één of meerdere items te beantwoorden. Voor het waarnemen van een mening, houding of gedrag bij een persoon gaat men er vanuit dat deze methode met directe vragen aan het benodigde betrouwbaarheidsniveau voldoet. De nauwkeurigheid van de verkregen data bepaalt de validiteit van de getrokken conclusies. Als een onderzoek echter over een gevoelig onderwerp gaat, dan is de betrouwbaarheid van zelfrapportage niet zo vanzelfsprekend. Voorbeelden van gevoelige onderwerpen zijn verslavingen (drugs, alcohol en tabak), seksualiteit, welzijn, fraude en belasting ontduiking. Bij een vraag over zo'n gevoelig onderwerp zijn respondenten geneigd een sociaal geaccepteerd antwoord te geven. Deze opzettelijke, systematische en vaak onmeetbare onjuiste weergave van de daadwerkelijke mening, houding of gedraging resulteert in systematisch onjuiste conclusies.

Er is onderzoek gedaan naar betrouwbare methoden voor valide data-verzameling over gevoelige onderwerpen. Het zelf afnemen van de test en het gebruik van computers, alsmede zorgvuldige formulering van vragen en aanpassingen aan de manier waarop de respondent kan antwoorden, wordt van verwacht dat respondenten eerlijker antwoorden. Deze methoden blijken echter tekort te schieten bij vragenlijsten over gevoelige onderwerpen.

Een alternatieve benadering heeft veel vooruitgang geboekt door het verzamelen van data met indirecte methoden die gebruik maken van randomisatie of opzettelijke onzekerheid. De respondent wordt beschermd door anonimiteit te garanderen via een randomisatie procedure. Voorbeelden van randomisatie processen zijn het opgooien van een munt, het gooien van een dobbelsteen, het draaien aan een wiel etc. De uitkomst van dit proces bepaalt de instructie voor respondent. Omdat de onderzoeker geen controle heeft over het randomisatie proces kunnen de antwoorden achteraf niet getraceerd worden. Hierdoor wordt de bescherming van de respondent verhoogd. Bij de geforceerde gerandomiseerde respons methode die wordt gebruikt in dit proefschrift berust het antwoord van de respondent op de uitkomst van een randomisatie proces (eerlijk of vooraf bepaald antwoord). De gerandomiseerde antwoorden en de bekende verdeling van het randomisatie proces leveren voldoende informatie op voor schattingen met betrekking tot het gevoelige onderwerp.

Alhoewel oorspronkelijk ontworpen voor dataverzameling omtrent een vraag, zijn de gerandomiseerde respons technieken (RR) succesvol toegepast op schalen met meerdere items. In dit proefschrift worden Bayesiaanse modellen voor gevoelige multivariate metingen, waarbij observaties worden verzameld met gerandomiseerde respons technieken, uitgebreid op meerdere manieren, waarbij rekening wordt gehouden met de grootte van de steekproef.

In hoofdstuk 2 wordt een Dirichlet-multinomiaal model voorgesteld voor ordinale gerandomiseerde data en kleine steekproeven. Het hoofdstuk laat zien dat dit een generalisatie is van het beta-binomiale model voor binaire data. Individuele niet-geobserveerde categorische responspercentages worden geschat door middel van een lineaire transformatie van categorische responspercentages gebaseerd op geobserveerde RR data. De empirisch Bayesiaanse schattingen worden vergeleken met de volledig Bayesiaanse schattingen. De volledig Bayesiaanse procedure is gemplementeerd in WinBUGS door de Dirichlet als een serie van beta-binomiale distributies weer te geven. Het model is uitgebreid tot een gerestricteerd Dirichlet-multinomiaal model op zo'n manier dat de homogeniteit van de categorische respons percentages over individuen, of groepen individuen, expliciet kunnen worden getest. In het tweede gedeelte van dit hoofdstuk wordt een simulatiestudie gepresenteerd waarin de terug schatting van de gesimuleerde parameters voor de volledig Bayesiaanse methode, als ook de gevoeligheid van de parameter schattingen voor verschillende condities wordt onderzocht. De invloed van prior instellingen wordt getoetst. De College Alcohol Problem Scale (CAPS) wordt gebruikt om de prestaties van het volledig Bayesiaanse model te illustreren, waarbij groepsspecifieke populatie proporties worden onderzocht.

In hoofdstuk 3 wordt een validatie studie van de gerandomiseerde respons techniek gepresenteerd, waarbij een vragenlijst met meerdere items wordt afgenomen om rookgedrag te meten. Daarnaast wordt een klinische adem test gedaan om te meten of een patiënt rookt. Voor het meten van individueel rookgedrag wordt een gerandomiseerd item respons model gebruikt voor binaire en ordinale data. Data zijn verzameld met zowel een gerandomiseerde antwoorden als met een directe vraag techniek. Voor elk van beide condities wordt de uitkomst van de ademtest vergeleken met de schatting van latent rookgedrag op basis van de vragenlijst. De data in de conditie met gerandomiseerde antwoorden zijn accurater dan de data in de directe vraag conditie, wanneer de uitkomsten worden vergeleken met de resultaten van de ademtest. Verder wordt een Bayesiaanse methode om de accuraatheid van de test vast te stellen geïntroduceerd. Deze methode maakt het mogelijk om classificatie kansen zoals de sensitiviteit en specificiteit te berekenen. Deze waarden worden gebruikt om de diagnostische accuraatheid van de test vast te stellen. De gerandomiseerde respons techniek wordt verder gevalideerd door middel van de positieve en negatieve voorspellende waarden van de test.

In hoofdstuk 4 wordt een multidimensionaal gerandomiseerd item respons theorie model geïntroduceerd om meerdere onderliggende factoren te meten bij een vragenlijst met meerdere items. Dit model bestaat uit drie stadia. Eerst worden de gerandomiseerde respons data gerelateerd aan individuele respons kansen. Ten tweede wordt het antwoordproces beschreven door een multidimensionaal IRT model. Ten derde worden latente sensitieve karakteristieken beschouwd als uitkomsten van een multivariaat regressie model. Een MCMC algoritme met een

dubbele data augmentatie stap is ontwikkeld om gelijktijdig alle model parameters te kunnen schatten. Na een simulatie studie om het terugvinden van de geschatte parameters te beoordelen, wordt de data van de College Alcohol Problem Scale geanalyseerd.

In hoofdstuk 5 wordt een overzicht van RR technieken gepresenteerd, waarbij onderscheid wordt gemaakt tussen traditionele en recent ontwikkelde technieken. Traditionele methoden worden besproken waarbij de redenering achter uitbreidingen wordt gegeven. Verschillende typen dataverzamelingsstrategien met gerandomiseerde antwoorden worden in detail besproken. Verschillende methoden om parameters te schatten worden gepresenteerd. Meer recente, niet gerandomiseerde respons technieken worden samengevat en vergeleken met standaard procedures. De kwestie van inferentieniveau gegeven gerandomiseerde respons data voor verschillende vragen types wordt besproken: inferenties op individueel niveau wanneer meerdere observaties beschikbaar zijn en inferenties op populatie niveau wanneer sprake is van metingen met slechts een enkel item. Een paar gerandomiseerde respons technieken worden gellustreerd met voorbeelden.

De parameters voor complexe Bayesiaanse modellen in dit proefschrift worden geschat met Markov Chain Monte Carlo (MCMC) methoden. De Bayesiaanse benadering heeft een aantal voordelen. Alle onbekende parameters worden gedefinieerd als random parameters. Elke parameter krijgt een prior distributie, die het mogelijk maakt om van tevoren beschikbare informatie te kwantificeren. Nadat de data geobserveerd zijn, wordt deze informatie aangepast op zo'n manier dat de data en de prior informatie leiden tot posterior informatie. Wanner nieuwe data beschikbaar zijn, kan de posterior informatie uit de vorige analyse gebruikt worden als prior informatie in een volgende analyse.