Generalized Linear Mixed Models for Randomized Responses

Jean-Paul Fox *[1], Duco Veen[2] and Konrad Klotzke[3]

[1]University of Twente

[2]University of Utrecht

[3]University of Twente

May 30, 2018

*Correspondence should be sent to Jean-Paul Fox, University of Twente, Department of research methodology, measurement and data analysis, P.O. Box 217, 7500 AE Enschede, The Netherlands

Abstract

Response bias (nonresponse and social desirability bias) is one of the main concerns when asking sensitive questions about behavior and attitudes. Self-reports on sensitive issues as in health research (e.g., drug and alcohol abuse), and social and behavioural sciences (e.g., attitudes against refugees, academic cheating) can be expected to be subject to considerable misreporting. To diminish misreporting on self-reports indirect-questioning techniques have been proposed such as the randomized response techniques. The randomized response techniques avoid a direct link between individual's response and the sensitive question thereby protecting the individual's privacy. Next to the development of the innovative data-collection methods, methodological advances have been made to enable a multivariate analysis to relate responses to sensitive questions to other variables. It is shown that the developments can be represented by a general response-probability model (including all common designs) by extending it to a generalized linear (GLM) or a generalized linear mixed effects model (GLMM). The general methodology is based on modifying common link functions to relate a linear predictor to the randomized response. This approach makes it possible to use existing software for GLMs and GLMMs to model randomized response data. The R-package GLMMRR makes the advanced methodology available to applied researchers. The extended models and software will seriously improve the application of the randomized response methodology. Three empirical examples are given to illustrate the methods.

Keywords: generalized linear model, mixed effect model, randomized response, social desirability, survey designs

1 Introduction

In survey studies about sensitive topics, it is well known that people are tended to respond in a socially desirable way or even refuse to cooperate. The sensitivity of the survey questions can induce misreporting. Guaranteeing confidentiality and anonymity of the responses is often not sufficient to obtain full cooperation of the survey respondents. When asking questions about fraud, corruption, or sexual behavior, direct questioning will lead to under- or overreporting and nonresponse, which makes it impossible to estimate true population prevalence of sensitive behavior. Innovative data collection techniques have been developed to avoid response bias and to stimulate cooperation of respondents to answer truthfully. Originally developed by Warner (1965), randomized response (RR) techniques protect respondent's privacy by introducing a randomization device to mask individual responses. In Warner's model, the data collection procedure is based on two mutually exclusive questions and a randomization device (e.g., dice, coin) determines which question is being answered. The respondent answers the selected question and only reveals the response and not the question to the interviewer and, subsequently, not the true status on the attribute of interest.

This indirect-questioning technique encourages respondents to cooperate and reduce socially desirable response behavior. The method has been supported by substantial empirical evidence. Several studies showed higher prevalence estimates of a sensitive behavior, for instance in academic fraud (Fox, 2005; Fox & Meijer, 2008; Jann, Jerke, & Krumpal, 2012), cheating (Moshagen, Hilbig, Erdfelder, & Moritz, 2014), excessive alcohol consumption (Fox & Wyrick 2008; De Jong, Fox, & Steenkamp, 2015), desires for products and services in the adult entertainment (De Jong, Pieters, & Fox, 2010), violating regulations for social benefit (van den Hout, Böckenholt , & van der Heijden, 2010), smoking behavior among lung patients (Fox, Avetisyan, van der Palen, 2013), and self-presentational behavior in job interviews (Jansen, König, Stadelmann, & Kleinmann, 2012). Lensvelt-Mulders et al. (2005) give an overview of various applications of the randomized response technique.

Beside the increasing number of applications of RR studies, many methodological innovations have been made to improve the statistical analysis. Multivariate methods have been developed to analyse prevalence of the sensitive behavior in relation to other variables, and to analyse multiple (scale) items (e.g., Böckenholt and van der Heijden, 2007; De Jong et al., 2015; Fox, 2005). Furthermore, the RR methods have been improved to make individual inferences, besides the computation of a population prevalence. Böckenholt and van der Heijden (2007), Fox (2005), Fox et al. (2013), and De Jong et al. (2010, 2015) developed RR models including person parameters, which make it possible to measure person-specific sensitive behavior, given multiple RR observations measuring different characteristics of the sensitive behavior. Böckenholt and van der Heijden (2007), Fox, Klein Entink, and Avetisyan (2014) and Fox (2016) considered measurement models for multiple latent constructs measured by randomized item-responses. Thirteen items of the college alcohol problem scale and four items of the alcohol expectancy questionnaire were used to measure multiple sensitive constructs; sexual enhancement expectancy, socio-emotional and community problems. The RR method was used since participants were expected to conceal undesirable behavior related to drinking alcohol. Cruyff, Böckenholt, and van der Heijden (2016) considered a multidimensional RR design analysis to capture all dimensions of a sensitive construct.

Despite the advantages of the RR method and the advanced methodology, the RR technique is not commonly applied to measure sensitive constructs. A few reasons can be given. The increased complexity of the statistical analysis of RR data, in comparison to directly-questioned data, might diminish the use of RR techniques. From the literature, it appears as if the different RR techniques require different statistical tools, and this might avoid a wide use of RR methods in applied research (Jann, Jerke, Krumpal, 2012). For instance, Tourangeau and Yan (2007) and Jansen et al. (2012) incorrectly remarked that, when using RR methods, it is not possible to examine relationships between a sensitive construct and other participant's characteristics. Scheers and Dayton (1988) showed how to perform a logistic regression on RR data, and Böckenholt and van der Heijden (2007), De Jong et al. (2010, 2015), and Fox (2005, 2016) have shown how to measure sensitive constructs using item response theory (IRT) and relate constructs to background variables. However, the problem might be that (advanced) RR methods have not been implemented in easy-to-use (open-access) software. There are a few R-packages for RR data (e.g., Blair, Zhou, & Imai, 2015b; Heck & Moshagen, 2014; Jann 2011) but they are limited to logistic and linear regression models for RR data and do not cover all methodological advances (e.g., cross-classified random effects, mixed models, complex hierarchical sampling designs, multidimensional models, models for non-compliance behavior).

Furthermore, developments have been made using the Bayesian modeling framework and Markov chain Monte Carlo (MCMC) methods to estimate model parameters. As far as we know, current (open-source) software are based on likelihood estimation methods, and do not cover all these developments.

Höglinger and Jann (2016) argued that the respondents may not understand the procedure, which will undermine the provided protection mechanism of the RR method (e.g., Böckenholt, Barlas, van der Heijden, 2009). As remarked by Jann et al, (2012), the RR method requires a randomizing device, which supports the protection mechanism. The randomizing device should not be too complex or abstract leading respondents to distrust the procedure. Recently, a crosswise RR-questioning technique has been developed (Yu, Tian, & Tang, 2008) which avoids the use of a randomization device. Jann et al. (2012) provided formulas for applications using regression techniques for the crosswise data-collection method. However, a general modeling framework with methodological tools is required for data collected by any RR design, including the crosswise method, to fully support RR methods in applied research.

A more general modeling approach was developed by van den Hout, van der Heijden, and Gilchrist (2007), and Blair, Zhou, & Imai, (2015a), who discussed a general probability model for four RR designs. Van den Hout et al. (2007) showed that a logistic regression analysis of the RR data for the four RR designs is based on the same single likelihood function. It will be shown that this methodology can be generalized in different ways. First, it is shown that for the general response-probability model recently developed RR designs, as the crosswise and triangular method (Yu et al., 2008), are also included. Subsequently, a general maximum likelihood estimator for prevalence is given, which is applicable to all the RR designs. Second, a generalized linear regression (GLM) model can be defined for all RR designs. This only requires a linear transformation of the common link functions (e.g., logistic, probit), which relate a linear predictor to the randomized responses. These GLM models are based on the same likelihood function such that standard GLM software can be used. Third, in the same way, generalized linear mixed models (GLMM: McCulloch & Searle, 2001; Tutz, 2012) can be defined for all RR designs, which also only require a linear transformation of the link functions. Furthermore, the GLMMs for RR data are also based on a common likelihood function and standard GLMM software can be used. Fourth, for all RR designs, both the GLM and the GLMM can be extended to relate responses to sensitive questions to multiple linear predictors. For instance, by introducing a composite link function (Thompson & Baker, 1981), a linear predictor for the regular responses together with a linear predictor for the self-protective responses can be included in the model. Fifth, general goodness-of-fit tests, diagnostic checks, and residual analysis for RR data can be used to evaluate model fit. The procedures follow directly from standard procedures for the GLM and GLMM (e.g., Tutz, 2012).

All the developed methodological tools have been implemented in the R-package GLMMRR (Fox, Klotzke, & Veen, 2016) for the statistical software program R (R Core Team, 2014). The GLM and the GLMM package (lme4; Bates, Mächler, Bolker, & Walker, 2015) were extended with additional link functions to accommodate the various RR designs in combination with four possible cumulative distributions (i.e., logistic, cumulative normal, Gumbel, Cauchy). The functionality of both packages is remained, which has led to an R-package with extensive functionality for RR modeling.

The different RR designs are discussed. Then, different link functions are given to extent the GLM and GLMM for RR data. A discussion is given about composite link functions to relate multiple linear predictors to randomized response observations. Subsequently, the maximum likelihood estimation methodology is discussed, and a simulation study is presented about the parameter recovery properties of the estimation methods. Next, goodness-of-fit statistics are discussed. Finally, it is shown how to apply the methodology using three empirical examples.

2 Randomized Response Survey Designs

The RR designs describe different data collection techniques with the same purpose of scrambling an individual's response using some sort of a randomization method. Although the designs differ with respect to the random process to scramble individual responses, a general response-probability model can be given that describes the response process for each design. In order to explain the RR methodology, the different RR designs are discussed and represented in a general form.

Following the notation of van den Hout et al. (2007) and Böckenholt and van der Heijden (2007), the probability model for a positive response can be written as a linear equation of the answer to the sensitive question. Let $\tilde{Y}_i$ denote the honest response (also referred to as the true response) of 1 of subject i to the sensitive question, before it is randomized due to a randomizing device. The probability of the event that the RR variable Y_i equals 1 is linearly related to the probability of an honest response of 1. This model is represented by

$$P\left(Y_i = 1\right) \;=\; c + dP\left(\tilde{Y}_i = 1\right), \tag{1}$$

where parameters c and d describe the random response process defined by the RR design. Before showing that the various RR designs can be represented in the general form of Equation (1), Table 1 gives an overview with the parameters c and d for each RR design. Note that the parameters c and d are known, and a linear probabilistic relationship between observed responses and the sensitive question is defined in Equation (1), which makes it possible to construct an estimator for the prevalence $\pi = P(\tilde{Y}_i = 1)$.

Insert Table 1 about here

Warner Design

In the design of Warner (1965), a respondent can be asked to belong to group A (i.e., you have the sensitive characteristic) or asked to belong to group B (i.e., you do not have the sensitive characteristic). A randomizing device, for example, a spinner, is used to facilitate a random process, where the spinner points to either an A or B, to select at random a question. Let Y_i denote the dichotomous randomized response of subject i, and the event $Y_i = 1$ ($Y_i = 0$) corresponds to belonging to group A (group B). Furthermore, let p_1 denote the probability that the question is selected whether the respondent belongs to group A. According to Warner's design, the probability of a positive response can be stated as

$$\begin{aligned} P\left(Y_i = 1\right) &= p_1\pi + \left(1 - p_1\right)\left(1 - \pi\right) \\ &= \left(1 - p_1\right) + \left(2p_1 - 1\right)\pi, \end{aligned} \tag{2}$$

where π denotes the proportion of persons in the population belonging to group A. It follows that Warner's response model can be represented in the form of the general response-probability model with $c = 1 - p_1$ and $d = 2p_1 - 1$.

Unrelated Question Design

Greenberg, Abul-Ela, Simmons and Horvitz (1969) proposed an alternative to Warner's design. In their design, the sensitive question (i.e., you have the sensitive characteristic) is selected at random with probability p_1, or an unrelated non-sensitive question is selected with $1 - p_1$. In contrast to Warner's model, the questions are not each other's antonyms and the prevalence of the nonsensitive question is independent of the sensitive question. Let π denote the prevalence of the sensitive question, and p_2 the prevalence to the unrelated question. Then, the probability of a positive response can be stated as a general response-probability model

$$P\left(Y_i = 1\right) = \left(1 - p_1\right)p_2 + p_1\pi, \tag{3}$$

with $c = \left(1 - p_1\right)p_2$ and $d = p_1$.

Forced Response Design

The unrelated-question design becomes less efficient, when the prevalence to the non-sensitive question needs to be estimated from the data. Therefore, Boruch (1971) used the randomizing device to select at random the sensitive question, with probability $1-p_1$, or when the sensitive question was not selected, to select a predetermined positive response with probability p_2. Therefore, a positive or affirmative response can then either be an honest response to the sensitive question or an obliged response. According to the forced response design, the response-probability model is also given by Equation (3), where p_2 denotes the proportion of respondents who were prompted to give a positive answer irrespective of the sensitive question.

Kuk's Design

The method that Kuk (1990) developed is based on a randomizing device, which generates two binary outcomes (e.g., two stacks of cards). The idea is that a respondent is given two stacks of cards with different proportions of red cards, where red represents a positive response. A positive response is given by reporting the color of the drawn card. At random a card is drawn from each stack, where a card is drawn from the stack with a higher (lower) proportion of reds to express a positive (negative) response. Additional details about the implementation of Kuk's design can be found in van der Heijden, van Gils, Bouts, and Hox (2000). Let p_1 and p_2 denote the proportion of red cards of the stack with the higher and lower proportion of reds, respectively. Then, the response probability model is given by

$$\begin{aligned} P\left(Y_i = 1\right) &= p_2\left(1 - \pi\right) + p_1\pi \\ &= p_2 + \left(p_1 - p_2\right)\pi \end{aligned} \tag{4}$$

where $c = p_2$ and $d = p_1 - p_2$.

Crosswise Design

A variant of Warner's design developed by Yu et al. (2008) is the crosswise model. The sensitive question is paired with an unrelated nonsensitive question of which the prevalence is known. Both questions, the sensitive and nonsensitive question, are answered simultaneously with one positive or one negative response. A positive response is given when the answer is either positive to both questions or negative to both questions, and a negative response when the answer is positive to one of the two questions. The corresponding general response-probability model is equal to Warner's model, Equation (2). However, the crosswise design has several advantages over Warner's design. The randomization procedure does not require a device, since this element is integrated in the question. The instructions are easier to understand and it is not required to answer directly the sensitive question. Jann et al. (2012) showed improved prevalence estimates under the crosswise model. Höglinger and Jann (2016) showed that the crosswise model yielded higher prevalence estimates partly due to false-positives, where non-cheaters were incorrectly classified as cheaters.

Triangular Design

Yu et al. (2008) also proposed the triangular model, which is based on the same principle as the crosswise method. A nonsensitive question with a known prevalence, p_1, is paired with a sensitive question of interest. Both questions are answered with one response, positive or negative. If the answers to both questions are negative a negative response is given. If at least one answer to one of the questions is positive a positive response is given. The response-probability model for the triangular design is given by

$$\begin{aligned} P\left(Y_i = 1\right) &= p_1\pi + \left(1 - p_1\right)\pi + p_1\left(1 - \pi\right) \\ &= p_1 + (1 - p_1)\pi \end{aligned}$$

where p_1 denotes the prevalence of the nonsensitive characteristic. The triangular method also do not require a randomizing device, and it is expected to be more efficient than the crosswise design, which corresponds to Warner's design.

3 Estimating Population Prevalence

The general response-probability model in Equation (1) will be used to derive a common maximum likelihood estimator for prevalence. Let π denote an (unknown) prevalence rate in a population and n persons are randomly selected and interviewed using an RR technique with parameters c and d. The general RR model states that responses Y_i $(i = 1, \ldots, n)$ are Bernoulli distributed with success rate $c + d\pi$; that is, $Y_i \sim B\left(c + d\pi\right)$. Subsequently, in Appendix A it is shown that the maximum likelihood estimator of π is given by

$$\hat{\pi}_{mle} = \frac{1}{d}\left(\overline{y} - c\right), \tag{5}$$

where $\overline{y}$ represents the sample mean. Furthermore, it is shown that this unbiased estimator is efficient, since the variance of the estimator is equal to the Cramér-Rao lower bound. The variance of the general maximum likelihood estimator $\hat{\pi}_{mle}$ is given by

$$var\left(\hat{\pi}_{mle}\right) = \frac{1}{d^2 n}\left(\left(c + d\pi\right)\left(1 - \left(c + d\pi\right)\right)\right). \tag{6}$$

As a result, for each RR design, the maximum likelihood estimator and variance are given in Table 1. The asymptotic distribution of $\hat{\pi}_{mle}$ can be used to define the Wald statistic. It follows that,

$$\hat{\pi}_{mle} \sim N\left(\pi, \sigma^2_{\hat{\pi}}\right), \tag{7}$$

where $\sigma^2_{\hat{\pi}} = var\left(\hat{\pi}_{mle}\right)$, as defined in Equation (6). In Table 1, for each RR design the expressions for the ML estimator and variance are given.

4 Generalized Linear RR Models

The generalized linear model is characterized by independent observations distributed according to an exponential family distribution. Explanatory variables are available on each observation, which are used to describe a linear predictor (i.e., systematic linear component). Then, the generalized linear model (e.g., McCullagh & Nelder, 1989; Scheers & Dayton, 1988) is used to model the binary responses given the linear predictor. An important element is the link function to relate the expectation of an observation to a linear predictor.

Let $\mathbf{x}_i$ $(i = 1, \ldots, n)$ denote the covariates for subject i, and $\eta_i = \mathbf{x}_i^t \boldsymbol{\beta}$ the linear predictor. The linear predictor is linked to the conditional expectation of a single observation via a link function $g(.)$. Using the general response-probability model in Equation (1), the inverse of the link function, $g^{-1}(.)$, is used to relate the expected randomized response to the linear predictor of the sensitive question. This relationship is given by

$$\begin{aligned} E\left(y_i \mid \mathbf{x}_i\right) = \mu_i &= c + d\, P\left(\tilde{Y} = 1 \mid \mathbf{x}_i\right) \\ &= c + d\, g^{-1}\left(\eta_i\right) \\ &= c + d\, g^{-1}\left(\mathbf{x}_i^t \boldsymbol{\beta}\right) \end{aligned} \tag{8}$$

where $g^{-1}(.)$ represents the response function or (cumulative) distribution function to model an affirmative honest response probability. For binary data, the expected response corresponds to the probability of success. Therefore, the general link function $g^{-1}(.)$ will also be referred to as $\pi(.)$, representing the success probability. This success probability can depend on explanatory variables, and the $\pi(\mathbf{x}_i^t \boldsymbol{\beta})$ represents the success probability given the linear predictor.

When assuming a logistic distribution function, $\pi(.)$, for the honest response $\tilde{y}$, the inverse link function is given by

$$\begin{aligned}\mu_i &= c + d\pi\left(\eta_i\right) \\ &= c + d\,\frac{\exp\left(\eta_i\right)}{1+\exp\left(\eta_i\right)},\end{aligned}$$

and, subsequently, the modified link function is given by

$$g\left(\mu_i\right) = \log\left(\frac{\mu_i - c}{c + d - \mu_i}\right) = \eta_i.$$

It follows that this modified logit link function generalizes the (common) link function for the standard logistic regression model, which is represented with $c = 0$ and $d = 1$. Van den Hout et al. (2007) also considered this modification of the logistic regression link function.

A more general procedure can be given, since different response distributions, $\pi(.)$, for the honest response can be considered. For each response distribution, it is only required to specify the appropriate link function. For example, when considering a probit model for the honest responses, the honest success probability is defined by the cumulative normal distribution function, where the probit link function (i.e., the inverse of the cumulative normal distribution function) is used to define the link with the linear predictor. In Table 2, the logit, probit, complementary log-log, and cauchit link functions are given for binary RR data. The represented link functions are modifications of the common link functions for binary (response) data. It follows that for each response function, the expected response, μ_i, can be expressed according to Equation (8), where the $g^{-1}(.)$ represents the cumulative distribution function. An advantage of this general GLM modeling approach for binary RR data is that standard GLM software can be used, since only a linear transformation of the common link functions is required.

The cauchit link function, which is based on the standard Cauchy quantile function, is not often used. However, the corresponding cauchit model can be more appropriate when the observations contain some surprising values. For example, when observations for which the linear predictor is large in absolute value, indicating an accurate prediction of the outcome, and yet the linear predictor is incorrect. Such inconsistencies in the response observations are more easily captured by the cauchit model, than the probit or logit, since the Cauchy distribution has heavier tails compared to the normal and logistic distribution.

Insert Table 2 about here

GLMs With Composite Link Functions

The composite link function was introduced by Thompson and Baker (1981). They defined a link function that comprises different linear predictors instead of one linear predictor for each observation. The simple composite link function has known weights for each link function. When the weights are multiplied with unknown parameters, the composite link function is referred to as a bilinear composite link function, which extends the simple composite link function. The bilinear composite link function can be stated as

$$\mu_i \quad = \quad \sum_q \boldsymbol{\alpha}^t \mathbf{w}_{qi} g_q^{-1}\left(\eta_{qi}\right),$$

where $\boldsymbol{\alpha}$ are unknown parameters, $\boldsymbol{w}$ are known weights, and g_q^{-1} the inverse link function for linear predictor η_{qi}. For binary data, each inverse-link function is a cumulative distribution function and the notation $\pi_q = g_q^{-1}$ is used.

Rabe-Hesketh and Skrondal (2007) considered composite link functions for different latent variable models, and showed that Warner's logistic regression model can be represented as a generalized linear model with a simple composite link function. This result can be generalized, since it implies directly to the RR designs represented by Equation (1). For each survey design stated in Table 2, the generalized linear RR model defined in Equation (8), and in Table 2, has a simple composite link function. For example, when considering the unrelated question design, and using a linear component for the related and unrelated question, the expression for the conditional expected value is given by,

$$\begin{aligned} \mu_i &= p_1 \pi_1 \left(\eta_{1i}\right) + (1 - p_1) \pi_2 \left(\eta_{2i}\right) \\ &= \sum_{q=1}^{2} w_q \pi_q^{-1} \left(\eta_{qi}\right), \end{aligned}$$

where weights w_1 and w_2 determine the random selection of the related and unrelated question. The inverse-link functions $\pi_1(.)$ and $\pi_2(.)$ relate each linear term to the expected value, and each function can be a logit, probit, complementary log-log or cauchit function. Different link functions can be specified. This also implies that a different response distribution can be defined for honest responses to the related question compared to honest responses to the unrelated question.

However, in RR modeling, interest is mainly focused on the linear predictor in relation to responses to the sensitive question. This makes the composite link function not particularly relevant for generalized linear RR modeling. However, in a more complex some situation, some respondents may not follow the RR instructions and always respond, for example, in the least stigmatizing way (Böckenholt & van der Heijden, 2007). Often the the least stigmatizing answer is "No", and therefore this response behavior is also referred to as self-protective no-saying. Van den Hout, Gilchrist, and van der Heijden (2010), considered the log-linear RR model with a composite link function to include self-protective response behavior. In the generalized linear modeling approach with a bilinear composite link function, an additional component for the self-protective response behavior can be included. Consider the general response-probability model in Equation (1), a second linear predictor for the self-protective responses, and a mixture parameter α_i. Then, the expected response can be written using a bilinear composite function

$$\mu_i = \alpha_i \left(c + d\pi_1\left(\eta_{1i}\right)\right) + (1 - \alpha_i)\pi_2\left(\eta_{2i}\right), \tag{9}$$

where the mixture parameter can be linked to a third linear predictor with $\alpha_i = \pi_3\left(\eta_{3i}\right)$.

5 Generalized Linear Mixed RR Models

Generalized linear mixed models include both fixed and random effects in the linear predictor. Consider response variable y_{ij}, where i refers to the subject ($i = 1, \ldots, n_j$) and j refers to the cluster ($j = 1, \ldots, J$). The clustering of responses leads to an additional correlation between the responses. For instance, the cluster can refer to a single subject who was repeatedly measured, or to a class of students from which single observations were obtained. Random effect variables, denoted as $\mathbf{b}_j$, are used to model the dependencies between clustered responses. Besides the choice of link function and the specification of explanatory variables, the effect of each predictor variable can be specified to be fixed or random.

Let the predictor variables $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ denote the explanatory variables associated with the response variable y_{ij}. The explanatory variables $\mathbf{x}_{ij}$ are used to define population-specific effects and variables $\mathbf{z}_{ij}$ to define cluster-specific effects. The GLM for RR responses defined in Equation (8) can be extended by including the random effects. This is done by modeling the mean response conditionally on the random effects $\mathbf{b}_j$. The expected response for subject i in cluster j is extended with random effect $\mathbf{b}_j$, and is given by

$$\begin{aligned} E\left(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_j\right) &= c + d\pi(\eta_{ij}) \\ &= c + d\pi\left(\mathbf{x}_{ij}^t\boldsymbol{\beta} + \mathbf{z}_{ij}^t\mathbf{b}_j\right), \end{aligned} \tag{10}$$

where $\pi(.)$ is again a cumulative distribution function to model an affirmative honest response. The response function (i.e. inverse-link function), which defines the link between the predictor and the expected response is again a linear transformation of the $\pi(.)$. In this generalized linear mixed effects model, the $\boldsymbol{\beta}$ parameters are considered to be the fixed effects model parameters and the $\boldsymbol{b}_j$ the random effects. Although not strictly necessary, the random effects $\mathbf{b}_j$ are assumed to be multivariate normally distributed with mean zero and covariance matrix $\mathbf{Q}_b$. The random effects define a common correlation between responses from the same cluster.

6 ML Estimation: Parameter Recovery Study

The general method to estimate the model parameters given binary data is maximum likelihood. The basic procedure is to define the log-likelihood of the model parameters given the binary response data. The maximum likelihood estimates will maximize the log-likelihood, and are obtained by solving the maximum likelihood equations. Fisher scoring algorithm can be used to find the maximum likelihood estimates as the roots of the maximum likelihood equations. The details of the estimation method for the GLM, GLMM, and composite link functions are given in Appendix B.

The performance of the ML estimation method was evaluated for GLMMs with four different link functions. Data were generated according to a forced RR design with $p_1 = .778$ and $p_2 = .50$. The linear predictor consisted of a random intercept b_i, for $i = 1, \ldots, N$ ($N = 200$ and $N = 500$) assumed to be normally distributed with variance $\sigma^2 = 0.50$. For each i, $J = 10$ and $J = 20$ repeated observations were generated. A predictor effect was set at $\beta = 1$, and the predictor values were sampled from a normal distribution with a mean of 0 and standard deviation of 0.50. Fixed effects λ_j were set from -.50 to .50 with a step size of J^{-1}. The corresponding success probability is given by

$$\begin{aligned} P\left(Y_{ij}=1 \mid b_i, \lambda_j, X_{ij}\right) &= .111 + .778\,\pi\left(\eta_{ij}\right) \\ \eta_{ij} &= \lambda_j + \beta X_{ij} + b_i \end{aligned}$$

where $b_i \sim N\left(0, \sigma^2\right)$. For each condition, 50 data sets were generated for four different cumulative distributions (i.e., logistic, cumulative normal, Gumbel, Cauchy) using the corresponding link functions, as described in Table 2.

In Table 3, the averaged parameter estimates of the predictor effect, β, and random effect variance, σ^2, are given across the 50 data replications. For each condition, it can be seen that the 95% confidence intervals include the true parameter values. The size of the confidence intervals become smaller, when increasing the sample size. The variance estimates were most often slightly below the true value. The sample size was of moderate size, and the randomized response mechanism further reduced the effective sample size. Therefore, it is likely that the random effect variance was not always realized in the sample data. Furthermore, for the cauchit link function, the predictor effects were slightly underestimated, when increasing the sample size. The wider tails of the Cauchy distribution restricts the more extreme linear predictor values less stringent than the logistic, normal, and Gumbel distribution. Extreme linear predictor values are less likely related to extreme success probabilities under the Cauchy distribution than under the other cumulative distributions. Therefore, the Cauchy distribution induced less structure on the response probabilities given the linear predictor, leading to lower estimated fixed and random effects.

In another simulation study it is shown that the Cauchy distribution is particularly useful for skewed distributed responses and that the GLM with the cauchit link is preferred when the RR data are Cauchy distributed (see the Web Appendix). In general it was concluded that the maximum likelihood estimation method was capable of recovering the true parameter values.

Insert Table 3 about here

7 Goodness-of-fit Statistics

The general goodness-of-fit tests, diagnostic checks, and residual analysis for binary response data can be used to evaluate the fit of generalized linear RR models. The discrepancy between the fitted values under the generalized linear RR model and the observations y_i $(i = 1, \ldots, n)$ can be expressed by the deviance. The deviance of the saturated model is zero. Therefore, for the generalized linear RR model the deviance can be expressed as

$$\begin{aligned} D\left(\mathbf{y}, \hat{\boldsymbol{\pi}}\right) &= -2\sum_{i=1}^{n} y_i \log\left(c + d\hat{\pi}_i\right) + \left(1 - y_i\right)\log\left(1 - \left(c + d\hat{\pi}_i\right)\right) \\ &= -2\log L\left(\hat{\boldsymbol{\beta}}, \mathbf{y}\right), \end{aligned} \qquad (11)$$

where $\hat{\pi}_i = \pi\left(\mathbf{x}_i^t\hat{\boldsymbol{\beta}}\right)$ and the log-likelihood is defined in Equation (15). Following Table 2, a link function is defined for each cumulative distribution function $\pi(.)$.

The deviance function can be used as a model-fit statistic, when the number of model-specific success probabilities is less than the number of observations. Therefore, consider $i = 1, \ldots, I$ different repeated trials of binary observations. Let n_i denote the number of observations at measurement point $\mathbf{x}_i$. For each set of repeated trials, $\mathbf{y}_i$, a unique success probability $(c + d\pi_i)$ applies such that the observed trials are binomially distributed. The deviance function for the binomially distributed sets of observations is given by

$$D\left(\overline{\mathbf{y}}, \hat{\boldsymbol{\pi}}\right) \quad = \quad 2\sum_{i=1}^{I} n_i \left(\overline{y}_i \log\left(\frac{\overline{y}_i}{(c + d\hat{\pi}_i)}\right) + (1 - \overline{y}_i) \log\left(\frac{1 - \overline{y}_i}{1 - (c + d\hat{\pi}_i)}\right)\right), \tag{12}$$

where $\overline{\mathbf{y}} = (\overline{y}_1, \ldots, \overline{y}_I)$ are the average number of successes for the sets of repeated trials. This deviance function evaluates the fit of the average success probabilities for each set of repeated trials. This deviance function is also referred to as the deviance for grouped binary RR data. Categorical predictor variables, $\mathbf{x}_i$, give support to a clustered interpretation of the binary observations. Conditional on the combined level of the categorical predictors, the observations are assumed to be binomially distributed. All categories of the categorical predictors should contain enough observations.

For grouped observations, the deviance is considered to be a goodness-of-fit statistic. The deviance is asymptotically χ^2-distributed with I minus the number of categorical predictors as the degrees of freedom, for a fixed number of categories of the predictors and $n_i \rightarrow \infty$ for $i = 1, \ldots, I$ (Tutz, 2012, pp. 89-93). The deviance for single responses, Equation (11), is not asymptotically χ^2-distributed, since the degrees of freedom increase with the sample size. The deviance for single observations can still be used in a residual analysis.

When considering grouped observations, the Pearson statistic can also be used as a goodness-of-fit test, which is given by

$$X_P^2 = \sum_{i=1}^{I} n_i \frac{\left(\overline{y}_i - (c + d\hat{\pi}_i)\right)^2}{(c + d\hat{\pi}_i)\left(1 - (c + d\hat{\pi}_i)\right)}. \tag{13}$$

The Pearson statistic is also asymptotically χ^2-distributed, for fixed I and $n_i \to \infty$, and has the same asymptotic distribution as the deviance in Equation (12).

When dealing with strict continuous predictors, both tests cannot be used. Each level of the predictors will have one observation and the tests are not χ^2-distributed. For semi-continuous predictors it might still be possible to find categories with sufficient observations. Hosmer and Lemeshow (1980) proposed a statistic where the fitted values are categorized in I equally sized groups. The Hosmer-Lemeshow statistic, X_{HL}^2, is assumed to be χ^2-distributed with $I-2$ degrees of freedom. The grouping is based on the fitted values under the model. This leads to a reduced power of the statistic and, as for all global tests, a misspecified linear term will often not be recognized.

The goodness-of-fit fit statistics provide insight about the global fit of the model. A residual analysis can give more detailed information about the a possible misfit of the model. For binary RR data, the Pearson residual can be defined for the generalized linear RR model, which has the form

$$r_p\left(\overline{y}_i, \hat{\pi}_i\right) = \frac{\overline{y}_i - (c + d\hat{\pi}_i)}{\sqrt{(c + d\hat{\pi}_i)\left(1 - (c + d\hat{\pi}_i)\right)/n_i}}.$$

The Pearson residuals can be used to investigate large differences between RR observations and the expected values under the model, and a plot of the residuals can indicate outliers.

For the generalized linear mixed RR models, different residuals can be considered. The difference between the observation and the conditional expected value (given the random effect estimate) leads to a conditional residual, which is given by

$$\begin{aligned} r_c\left(y_{ij}, \hat{\pi}_{ij}\right) &= y_{ij} - E\left(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_j\right) \\ &= y_{ij} - \left(c + d\pi\left(\mathbf{x}_{ij}^t \hat{\boldsymbol{\beta}} + \mathbf{z}_{ij}^t \hat{\mathbf{b}}_j\right)\right). \end{aligned}$$

When the marginal expected value is computed, by integrating out the random effects, a marginal residual can be computed. This is the difference between the observation and the marginal mean, and is given by

$$\begin{aligned} r_m\left(y_{ij}, \hat{\pi}_{ij}\right) &= y_{ij} - E\left(y_{ij} \mid \mathbf{x}_{ij}\right) \\ &= y_{ij} - \left(c + d\pi\left(\mathbf{x}_{ij}^t \hat{\boldsymbol{\beta}}\right)\right). \end{aligned}$$

The residuals can be standardized by dividing them by their scaled standard deviation, which leads again to a Pearson residual.

The Akaike information criterion (AIC) is used to compare non-nested models. A general description can be given since the AIC consists of a deviance term, which represents the discrepancy between the data and the fitted values, and a penalty term. The deviance term is represented by the log-likelihood evaluated at the parameter estimates. The AIC is given by,

$$AIC = -2 \log L\left(\hat{\boldsymbol{\beta}}, \mathbf{y}\right) + 2k$$

where k represents the number of model parameters. For the generalized linear mixed RR model the number of model parameters, k, is estimated, since it depends on the estimated random effect variance.

8 Real Data Study

8.1 Measuring Plagiarism using the Crosswise Design

Jann et al. (2012) discussed an RR survey study concerning plagiarism among 474 German and Swiss students conducted between June and July 2009. They used a randomized experimental design: students in the study were assigned to an interview using the direct-questioning technique ($n_1 = 116$) or the crosswise (CW) technique ($n_2 = 358$). Students from the university of Leipzig, the ETH Zurich, and the LMU Munich were assigned to one of two experimental conditions.

It was assumed that students would underreport about their involvement in plagiarism, since it is socially undesirable behavior. Therefore, the CW technique was expected to lead to higher prevalence estimates than the direct-questioning (DQ) technique, when it was successful in eliciting more honest answers. Levels of plagiarism were measured using two questions and the same questions were used in both experimental conditions. The two questions about plagiarism were followed by several personal questions and questions about writing student papers. The (translated) questions related to plagiarism were, as reported in Jann et al., (2012),

1. When writing an assignment(e.g., seminar paper, term paper, thesis), have you ever intentionally adopted a passage from someone elses work without citing

the original? (partial plagiarism)

2. Did you ever have someone else write a large part of an assignment for you or hand in someone elses work (e.g., from www.hausarbeiten.de) as your own? (severe plagiarism)

In the CW condition, each sensitive question about plagiarism was paired with a non-sensitive question, and a joint answer was requested to both questions. The non-sensitive questions were:

1. Is your mothers birthday in January, February, or March? (paired with the partial plagiarism question)
2. Is your fathers birthday in October, November, or December? (paired with the severe plagiarism question)

The non-sensitive items were assumed to be independent of the sensitive items, and the probability of a positive response to each nonsensitive question was set to .25 assuming a uniform distribution of birthdays in the sample. According to Table 1, the parameters c and d were equal to 0 and 1 for the DQ responses and $.75\,(= 1 - .25)$ and $-.50\,(= 2 \cdot .25 - 1)$ for the CW responses. In Jann et al., (2012), the partial and severe prevalence estimates were computed for students who were questioned directly and students who were questioned using the CW method. The reported partial plagiarism prevalence estimates were 7.3% and 22.3%, and the severe plagiarism prevalence estimates were 1% and 1.6%, for the DQ group and CW group, respectively. These estimates follow directly from the maximum likelihood estimator defined in Equation (5), where the variance can be computed from Equation (6).

The partial plagiarism prevalence estimates are significantly different. This result shows that the CW method avoids the self-protective behavior of students, when responding to the sensitive plagiarism question.

A multivariate analysis of the response data was carried out to investigate effects of the questioning technique and background variables (e.g., student characteristics) on (severe and partial) plagiarism. The response data were stacked to analyse the effects of background variables on the probability of partial and severe plagiarism. The stacked response data provided support to a joint analysis of all response data, while accounting for the different link functions. Therefore, the GLM for RR data was used with appropriate values for the c and d parameters. Subsequently, different GLMs were fitted by considering different cumulative distribution functions to model the responses given a linear predictor, as described in Table 2.

In Table 4, the ML estimates and standard deviations are given of the RR indicator variable (CW was coded as 1 and DQ as 0), of the variable student age (ranging from 18 to 41), and of the question indicator variable (partial plagiarism was coded as 1 and severe plagiarism as 0). It follows that the three variables have significant effects using the logistic, probit, and the complementary log-log link function. From the AIC estimates follows that the probit link function leads to the best goodness-of-fit, but differences are very small. When differences are small, the logit link function can be preferred, since it leads to easily interpretable estimates in terms of log-odds or odds.

Given the estimated effects from Table (4), the estimated odds ratio is given by,

$$\frac{P\left(Y_i=1\right)}{P\left(Y_i=0\right)}=\exp\left(.63-.26\text{Age}_i+1.34\text{RR}_i+2.40\text{Partial}_i\right),$$

where Y_i represents the true response of student i. The RR effect of 1.34 represents the effect on the log-odds ratio when comparing indirectly questioning (CW) with directly questioning (DQ). It follows that the odds ratio increased with a factor of $3.82 = exp(1.34)$ for students who were questioned via the CW method, compared to students who were questioned directly. Students questioned with the CW method were more likely to admit to plagiarism (partial and/or severe). The effect of variable age shows that there is a negative relationship between student age and plagiarism, where older students are less likely to be involved in plagiarism. Finally, when comparing the odds ratio of the partial plagiarism question with the odds ratio of the severe plagiarism question, it follows that the log odds ratio increased with a factor of $exp(2.40) = 11.02$ when it concerned the partial plagiarism question. Only a few students admitted to severe plagiarism.

The generalized linear multivariate modeling of the plagiarism data was not optimal due to the small prevalence estimate of severe plagiarism. However, the analysis did show a combined effect of the CW method, given the higher prevalence estimates based on the RR data compared to the DQ data. The general goodness-of-fit statistics, (e.g., Deviance, Pearson, and Hosmer-Lemeshow) did not indicate a misfit of the model. For the logistic distribution, the test statistic values were 51.1 (p=.88), 54.5 (p=.80), and 6.2 (p=.62), for the Deviance, Pearson, and Hosmer-Lemeshow statistic, respectively, where the corresponding p-values are given within brackets. For the Cauchy distribution, the statistic values were slightly higher but the p-values were still less than .95, and did not indicate a misfit of the model.

Insert Table 4 about here

8.2 A Smoking Behaviour Study Among Lung Patients

In the smoking behaviour study of Fox et al. (2013), a multi-item questionnaire was used to measure smoking behavior of lung patients over 16 years of age who were asked to voluntarily participate in the survey. We focused on three items of the questionnaire for which count data were observed. The three items are about regular smoking and the amount of smoking; item 1 “”How many years have you been smoking/had been smoking?”, item 2, “How many cigarettes are you smoking per day?”, and item 3, “”How many days per week are you smoking?”. The object was to identify regular smokers from non-regular smokers, and to examine whether there were differences between patients responding using the RRT and those responding directly. In total, 305 patients were assessed, and 198 of them completed the test using the forced RR technique. Using a randomizing device, an honest response was requested with probability $p_1 = .611$, and a forced response to one of four response categories was given with probability $p_2 = .25$, where the response categories were ’never/none’ and three others representing different frequency intervals. Demographic characteristics such as age (in years) and educational level of the patient were collected as background characteristics and information on medical condition. Patients were randomly assigned to the randomized response condition or to the direct-questioning technique.

Let Z_{ik} denote the (true) response of patient i to item k, which represents the number of occasions of the questioned event. The response variable Z_{ik} is assumed to be Poisson distributed with parameter λ_i. Let Y_{ik} denote patient's i randomized response of "never/none" ($Y_{ik} = 0$) or a non-zero frequency ($Y_{ik} = 1$). The complementary log-log link function can be used to model the probability of a randomized response of a non-zero observation, $P(Y_{ik} = 1)$, given the linear predictor $\mathbf{x}_i\boldsymbol{\beta}$. It follows that,

$$\begin{aligned}
P\left(Y_{ik} = 1 \mid \boldsymbol{\beta}\right) &= p_{i1} P\left(Z_{ik} > 0 \mid \lambda_i\right) + (1 - p_{i1}) p_{i2} \\
&= p_{i1}\left(1 - P\left(Z_{ik} = 0 \mid \lambda_i\right)\right) + (1 - p_{i1}) p_{i2} \\
&= p_{i1}\left(1 - \exp\left(-\lambda_i\right)\right) + (1 - p_{i1}) p_{i2} \\
&= p_{i1}\left(1 - \exp\left(-\exp\left(\mathbf{x}_i\boldsymbol{\beta}\right)\right)\right) + (1 - p_{i1}) p_{i2}, \\
&= c_i + d_i\left(1 - \exp\left(-\exp\left(\mathbf{x}_i\boldsymbol{\beta}\right)\right)\right),
\end{aligned}$$

with $c_i = (1 - p_{i1})p_{i2}$ and $d_i = p_{i1}$ and where the linear predictor is linked to the expected number of events λ_i through the function, $\log\left(\lambda_i\right) = \mathbf{x}_i\boldsymbol{\beta}$. It follows that the success probability of a randomized non-zero count response is modeled as a function of the linear predictor $\mathbf{x}_i\boldsymbol{\beta}$ through the complementary log-log link function. For patients responding directly $p_{i1} = 1$ and $p_{i2} = 0$, and for those using the RRT $p_{i1} = .611$ and $p_{i2} = .25$

GLMs for RR data with a complementary log-log link were fitted to examine differences in smoking behavior across patients. In Table 5, the parameter estimates are given of two fitted GLMs referred to as Model 1 and Model 2. The RR variable indicates whether the response was given using the forced response technique (RR equals 1) or without the forced response technique (RR equals 0). The significant estimate of .79 of the RR variable in Model 1 shows that patients in the RR condition scored more often a non-zero frequency than those responding directly. For instance, for item 2 the probability that a patient is smoking daily is around .15 ($1 - \exp(-\exp(-1.80)) \approx .15$) for those who responded directly, and around .31 ($1 - \exp(-\exp(-1.80 + 0.79))$) for those who responded using the RRT.

For Model 2, the type of disease and age are added as predictors. The estimated intercept is around 2.33, and it shows that the COPD patients (in the RR and DQ condition) have or had smoked for one or more years (item 1). Patients with asthma or bronchitis scored much lower and were less likely to smoke, although those questioned with RR scored significantly higher. The effect of age is negative, and shows that youngsters were more likely to smoke.

Insert Table 5 about here

8.3 A Student Cheating Study

Fox (2005) and Fox and Meijer (2008) discussed a study about cheating of Dutch University students of different education programs. In total 349 students were questioned about types of cheating, where 36 items were used to measure an academic dishonesty (cheating) score. Students were questioned from 7 different programs; computer science (CS), educational science and technology (EST), philosophy of science (PS), mechanical engineering (ME), public administration and technology (PAT), science and technology (ST), and applied communication sciences (ACS).

Students were questioned using the forced randomized response method and for each of 36 items they were asked whether they agreed or disagreed. A dice was used before a question could be answered and students answered yes when the sum of the outcomes equaled 2, 3, or 4; answered no when the sum equaled 11 or 12; or answered the sensitive question truthfully. As a result, the parameters of the used forced response technique were $p_1 = 3/4$ and $p_2 = 2/3$, such that $c = 1/6$ and $d = 3/4$.

The 36 items covered types, frequencies, and reasons of cheating, which were used to estimate an overall cheating behavior score via a Bayesian IRT model for randomized response data. A total of 13 items about types of cheating were selected from the 36 cheating items of Fox (2005) and Fox and Meijer (2008). In Table 6, these items are given, where the item numbers refer to the numbering used by Fox and Meijer (2008). Using a maximum likelihood estimation method, various GLMM models were fitted to investigate differences in the prevalence of types of cheating across items, students, and education programs. A cross-classified random item effects model was fitted, where random effects were defined for the clustering of randomized responses by items and students. Furthermore, different link functions were used to identify an optimal model. Ten students who answered zeroes and one student who answered ones to all items were excluded from the analysis, since their behavior scores could not be computed. As a result, a total of 4,394 observations were analysed of 338 students responding to 13 items.

Insert Table 6 about here

In Table 7, the parameter estimates of the different GLMMs for RR data are given. A random item effects model, referred to as Model0, was fitted to explore the variability in the prevalence of cheating across items and persons. It can be seen that the variability in student behavior is around .84, which is much higher than the variability in types of cheating which is around .07. The prevalence of cheating differs across students but not much across types of cheating. Under the header "Model1", the results are given of the GLMM with variable Male and educational program to explain differences in student cheating behavior. The estimated negative effect of Male indicates that males are less likely to cheat than females, but the effect is not significant. The Wald test score equalled -0.41, which led to a p-value of 0.68. The prevalence estimates differ across educational programs, where CS students scored much lower than students from other programs. Female students of ACS reported the highest mean score of cheating behavior. The student predictor variables reduced the student-level variance with 9.5%, and as expected did not reduce the variance in prevalence across items. In Model2, the variability in prevalence across items is modeled using fixed item effects, where the effect of item 3 is fixed to zero for reasons of identification. It can be seen that the prevalence estimates of types of cheating is highest for item 10 ("Used crib notes or cheat sheets") and lowest for item 33 ("Minimized effort in a joint assignment"). Students who are likely to cheat most often use a crib note. The standard errors of the item effects are relatively high, and according to the Wald test only items 4, 10, 17, and 31 have effects which are significantly different from zero. When comparing Model2 to Model1, it follows that the deviance is reduced by introducing fixed item effects, and the AIC shows a slight preference for Model2. Note that with Model2, prevalence estimates for the sensitive questions are computed given hierarchically structured data, since responses are clustered in students, who are clustered in educational programs.

The discussed models had a logit-link function, but it is possible that another link improves the model fit. The goodness-of-fit of models with different link functions can be compared using the AIC and Deviance. Model2 was fitted using each of the four link functions, where the probit link shows the lowest AIC and Deviance value. In Table 7, the estimated effects of this GLMM (probit) model is given under the header "Model3 (probit)". The variance of the errors under the cumulative normal distribution (probit model) is equal to one, where it is $\pi^2/3$ under the logistic distribution function. Therefore, the estimated fixed and random effects of Model3 are smaller than those of Model2 but they lead to similar conclusions. The fit of the random item effects model and fixed item effects model is almost equal when using the probit link function.

To interpret the outcomes of Model3, the response probability of cheating of student i on item k can be expressed as

$$\begin{aligned} P\left(Y_{ik} = 1 \mid \eta_{ik}\right) &= c + d\Phi\left(\eta_{ik}\right) \\ \eta_{ik} &= \text{Item}_k + \text{Male}_i + \text{Program}_i + \text{Student}_i \end{aligned}$$

where η_{ik} is the linear predictor and Student_i the cheating behavior score. This expression can be used to compute prevalence estimates of cheating for different types of items and groups of students. For example, it follows that the prevalence estimates of CS students using a crib note is around 32%, which is around 51% for

ACS students. Although, only RR data was observed, student-specific predictions can be made using the random effects estimates for the student cheating behavior. Finally, note that the prevalence estimates are based on an estimate of the linear predictor, η_{ik}, and the predictions cannot be related to observed RR data. The defined link functions make it possible to compute predictions under the model, while accounting for the fact that RR data are observed.

Insert Table 7 about here

9 Discussion

When investigating sensitive attributes direct-questioning techniques can induce non-cooperation of respondents, which will lead to response bias. More truthful answers can be elicited through indirect-questioning techniques. Several RR validity studies have shown improved prevalence estimates (e.g., Fox et al., 2013; Hoffmann, Diedenhofen, Verschuere, & Musch, 2015; Höglinger & Jann, 2016; Moshagen et al., 2014; van der Heijden et al., 2000). Fox et al. (2013) investigated the validation of the RR technique using a treatment-control design, where lung patient's smoking behavior was assessed using a multi-item measure through either direct or indirect-questioning. A breath test was also administered using a carbon monoxide (CO) monitor to determine patient's smoking status, which served as a gold standard. Patients, detected as smokers, scored significantly higher, when questioned via the forced RR technique, compared to detected smokers who were directly questioned. The study provides insight in the sensitivity and specificity of the self-report under direct and indirect questioning. Note that higher prevalence estimates might not always validate the RR technique, since this requires insight in the over and underreporting behavior of respondents (e.g., de Jong et al., 2015; Höglinger & Jann, 2016) or the true status of the respondent with respect to the sensitive attribute (Fox et al., 2013).

Despite recent work on validating RR methods, there is not much applied substantive research investigating sensitive attributes that might be distorted by response bias, when using direct-questioning techniques. To provide support to substantive applications using RR techniques, a general modeling framework is proposed, which encompasses the different RR designs and provides support to advanced GLM and GLMM methods for RR data. Applied researchers familiar with GLM and GLMMs can easily include RR data in their analysis using the developed R-package GLMMRR. By extending the common class of link functions to model binary data, the general class of RR models are integrated into the GLMM framework, while maintaining the general features included in the GLM and GLMM software (e.g., lme4 Bates et al., 2015). The modified link parameters are the RR design characteristics, and by specifying the design parameters for each response, it is possible to jointly model responses collected via the different RR designs, including direct-questioning. It is our objective to stimulate applied and methodological RR research by offering the open-source software, GLMMRR: Generalized Linear Mixed Modeling of RR data (Fox et al., 2016).

To account for noncompliance behavior, where respondents do not follow the design instructions, more advanced models are required. By introducing a composite link function, different linear predictors can be linked to each response. This makes it possible to define a coherent modeling framework for RR data. However, when using (modified) composite link functions in combination with GLMMs, more complex estimation methods are required to estimate all model parameters. MCMC methods can be used to estimate more complex models, which can account for noncompliance behavior (e.g., Böckenholt & van der Heijden, 2007; De Jong et al., 2010; Fox et al., 2013). However, more research is needed to develop MCMC methods that can handle in a flexible way GLMMs with composite link functions.

The GLMMs are computationally intensive methods and are usually applied to relatively large sample sizes (Bates et al., 2015). Furthermore, the RR designs require larger samples to achieve the same level of accuracy as direct-questioning. The prevalence of sensitive behaviors are often relatively low, and more data are required to obtain accurate estimates. Our R-package can handle large sample sizes ($10^4 - 10^6$ observations) to support real applications. Furthermore, the RR design characteristics can be adjusted across questions, such that less noise might be added to responses of the more sensitive questions with probably lower prevalence estimates. The efficiency of the RR designs can be adjusted to improve the control of the bias-variance trade-off of the indirect-questioning technique (Jann et al., 2012; Rosenfeld, Imai, & Shapiro, 2015). By reducing the amount of noise for the most sensitive questions, accurate prevalence estimates might still be obtained without requiring a substantially larger sample size compared to direct questioning.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1-48. doi = 10.18637/jss.v067.i01.

Blair, G., Imai, K., & Zhou, Y-Y. (2015). Design and analysis of the randomized response technique. *Journal of the American Statistical Association, 110*, 1304-1319. doi=10.1080/01621459.2015.1050028.

Blair, G., Zhou,Y.-Y., & Imai, K. (2015b). *rr: Statistical Methods for the Randomized Response,* Comprehensive R Archive Network (CRAN). Available at http://CRAN.R-project.org/package=rr.

Böckenholt, U., Barlas, S., & van der Heijden, P. G. M. (2009). Do randomized-response designs eliminate response biases? An empirical study of non-compliance behavior. *Journal of Applied Econometrics*, 24, 377-392. doi = 10.1002/jae.1052.

Böckenholt, U. & van der Heijden, P. G. M. (2007). Item randomized–response models for measuring noncompliance: Riskreturn perceptions, social influences, and self-protective responses. *Psychometrika, 72*, 24562. doi = 10.1007/s11336-005-1495-y .

Boruch, R. F. (1971). Maintaining confidentiality of data in educational research: A systematic analysis. *American Psychologist, 26*, 413-430.

Cruyff, M. J. L. F., Böckenholt, U., & van der Heijden, P. G. M. (2016). The multidimensional randomized response design: Estimating different aspects of the same sensitive behavior. *Behavior Research Methods*, doi; 10.3758/s13428-015-0583-2.

De Jong, M. G., Fox, J.-P., & Steenkamp, J. B. E. M. (2015). Quantifying under- and overreporting in surveys through a dual-questioning-technique design. *Journal of Marketing Research*, 52, 737-753. doi: 10.1509/jmr.12.0336.

De Jong, M. G., Pieters, F. G. M., & Fox, J.-P. (2010). Reducing social desirability bias through item randomized response: An application to measure underreported desires. *Journal of Marketing Research, 47*, 14-27. doi=10.1509/jmkr.47.1.14 .

Fox, J.-P. (2005). Randomized item response theory models. *Journal of Educational and Behavioral Statistics, 30*, 1-24. doi = 10.3102/10769986030002189 .

Fox, J.-P. (2016). Bayesian randomized item response theory models for sensitive measurements. In W.J. van der Linden (Ed.), Handbook of item response theory: Vol. 1. Models. Boca Raton, FL: Chapman & Hall/CRC.

Fox, J.-P., Avetisyan, M., & van der Palen, J. (2013). Mixture randomized item-response modeling: A smoking behavior validation study. *Statistics in Medicine, 32*, 48214837. doi=10.1002/sim.5859.

Fox, J.-P., Klein Entink, R. K., & Avetisyan, M. (2014). Compensatory and non-compensatory multidimensional randomized item response models. *British Journal of Mathematical and Statistical Psychology, 67*, 133-152. doi= 10.1111/bmsp.12012.

Fox, J.-P., Klotzke, K., & Veen, D. (2016). *GLMMRR: Generalized Linear Mixed Modeling of RR data.* Comprehensive R Archive Network (CRAN). Available at https://cran.r-project.org/web/packages/GLMMRR.

Fox, J.-P., & Meijer, R. R. (2008). Using IRT to obtain individual information from randomized response data: An application using cheating data. *Applied Psychological Measurement, 32*, 595-610. doi=10.1177/0146621607312277.

Fox, J.-P., & Wyrick, C. (2008). A mixed effects randomized item response model. *Journal of Educational and Behavioral Statistics, 33*, 389-415. doi=10.3102/1076998607306451.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian Data Analysis, Second Edition*, London: Chapman & Hall.

Green, P. J. (1984). Iteratively reweigthed least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society, series B, 46*, 149-192.

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

Heck, D. W., and Moshagen, M. (2014), RRreg: Correlation and Regression Analyses for Randomized Response Data, Comprehensive R Archive Network (CRAN). Available at http://cran.r-project.org/package=RRreg.

Hoffmann, A., Diedenhofen, B., Verschuere, B., & Musch, J. (2015). A strong validation of the crosswise model using experimentally-induced cheating behavior. *Journal of Experimental Psychology*, 62, 403-414. doi= 10.1027/1618-3169/a000304.

Höglinger, M. & Jann, B. (2016). More is not always better: An experimental individual-level validation of the randomized response technique and the crosswise model. (Working paper No. 18). Retrieved from University of Bern Social Sciences: http://ideas.repec.org/p/bss/wpaper/18.html.

Hosmer, D. H., & Lemeshow, S. (1980). Goodness-of-fit tests for the multiple logistic regression model. *Communications in Statistics - Theory & Methods, 9*, 1043-1069.

Jann, B. (2011). *RRLOGIT: Stata Module to Estimate Logistic Regression for Randomized Response Data,* available at https://ideas.repec.org/c/boc/bocode/s456203.html.

Jann, B., Jerke, J. & Krumpal, I. (2012). Asking sensitive questions using the crosswise model: Questions using the crosswise model an experimental survey measuring plagiarism. *Public Opinion Quarterly, 76*, 3249. doi= 10.1093/poq/nfr036

Jansen, A., König, C. J., Stadelmann, E. H., & Kleinmann, M. (2012). Applicants self-presentational behavior: What do recruiters expect and what do they get? *Journal of Experimental Psychology, 11*, 77-85. doi= 10.1027/1866-5888/a000046

Kuk, A. Y. C. (1990). Asking sensitive questions indirectly. *Biometrika, 77*, 436-438. doi=10.1093/biomet/77.2.436.

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*, 319348. doi= 10.1177/0049124104268664.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear model* (2nd ed.). London: Chapman & Hall.

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized linear, and mixed models* (2nd ed.). New York: Wiley.

Moshagen, M., Hilbig, B. E., Erdfelder, E., & Moritz, A. (2014). An experimental validation method for questioning techniques that assess sensitive issues. *Experimental Psychology, 61*, 48-54. doi=10.1027/1618-3169/a000226.

R Core Team. (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from `http://www.R-project.org/`

Rabe-Hesketh, S., & Skrondal, A. (2007). Multilevel and latent variable modeling with composite links and exploded likelihoods. *Psychometrika, 72,* 123-140. doi=10.1007/s11336-006-1453-8.

Rosenfeld, B., Imai, K., & Shapiro, J. (2015). An empirical validation study of popular survey methodologies for sensitive questions. *American Journal of Political Science.* doi: 10.1111/ajps.12205.

Scheers, N. J., & Dayton, C. (1988). Covariate randomized response model. *Journal of the American Statistical Association, 83*, 969-974. doi=10.1080/01621459.1988.10478686

.

Thompson, R., & Baker, R. J. (1981). Composite link functions in generalized linear models. *Journal of the Royal Statistical Society, Series C, 30*, 125-131.

Tourangeau, R. & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*, 859-883. doi=10.1037/0033-2909.133.5.859.

Tutz, G. (2012). *Regression for Categorical Data.* Cambridge: Cambridge University Press.

van den Hout, A., Böckenholt, U. , & van der Heijden, P. G. M. (2010). Estimating the prevalence of sensitive behaviour and cheating with a dual design for direct questioning and randomized response. *Journal of the Royal Statistical Society, Series C*, 59, 723736. doi=10.1111/j.1467-9876.2010.00720.x.

van den Hout, A., van der Heijden, P. G. M. & Gilchrist, R. (2007). The logistic regression model with response variables subject tot randomized response. *Computational Statistics & Data Analysis, 51*, 6060-6069. doi=10.1016/j.csda.2006.12.002.

van den Hout, A., Gilchrist, R., & van der Heijden, P. G. M. (2010). The randomized response log linear model as a composite link model. *Statistical Modelling, 10*, 5767. doi=10.1177/1471082X0801000104.

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*, 505537.

doi=10.1177/0049124100028004005.

Warner S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association, 60*, 63-69. doi=10.1080/01621459.1965.10480775.

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. doi = 10.1007/s00184-007-0131-x.

Appendix A: ML estimation of prevalence

Assume $y_1, \ldots, y_n$ RR observations according to a survey design with parameters c and d, and let $\lambda = c + d\pi$. The likelihood can be written in the form

$$\begin{aligned} L(\pi; \mathbf{y}) &= \prod_{i=1}^{n} \lambda^{y_i} (1-\lambda)^{1-y_i} \\ &= \lambda^{\sum_i y_i} (1-\lambda)^{n-\sum_i y_i}. \end{aligned}$$

The maximum likelihood estimate of π maximizes the log-likelihood function. This estimate solves

$$\max_{\pi} \log L(\pi; \mathbf{y}) = \max_{\pi} \left[\sum_i y_i \log \lambda + \left(n - \sum_i y_i \right) \log (1-\lambda) \right].$$

The score function for the Bernoulli log-likelihood given RR data equals

$$S(\pi \mid \mathbf{y}) = \frac{\partial \log L(\pi; \mathbf{y})}{\partial \pi} = \frac{d \sum_i y_i}{c + d\pi} - \frac{d\left(n - \sum_i y_i\right)}{1 - (c + d\pi)}.$$

The maximum likelihood estimate satisfies $S(\hat{\pi}_{mle} \mid \mathbf{y}) = 0$, and it follows that

$$\hat{\pi}_{mle} = \frac{\sum_i y_i / n - c}{d} = \frac{1}{d} (\overline{y} - c).$$

As a result, given the survey design characteristics c and d, the maximum likelihood estimate is the sample average subtracted with c and divided by d. In Table 1, the maximum likelihood estimators are given for each survey design.

It can be shown that the unbiased maximum likelihood estimator of π is efficient.

Therefore, the variance of the estimator should be equal to the Cramér-Rao lower bound. The variance of the estimator is given by

$$\begin{aligned} var\left(\hat{\pi}_{mle}\right) &= var\left(\frac{\sum_{i=1}^{n} y_i/n - c}{d}\right) \\ &= \frac{\sum_{i=1}^{n} var(y_i)}{d^2 n^2} \\ &= \frac{1}{d^2 n}\left((c + d\pi)(1 - (c + d\pi))\right). \end{aligned} \tag{14}$$

The Cramér-Rao lower-bound of the variance of the maximum likelihood is estimated by the inverse of the information matrix. Since the Bernoulli density function is regular, and $\hat{\pi}_{mle}$ is an unbiased estimator of π, it follows that

$$var\left(\hat{\pi}_{mle}\right) \geq I\left(\pi \mid \mathbf{y}\right)^{-1},$$

where the sample information matrix is denoted by $I\left(\pi \mid \mathbf{y}\right) = -E\left(H\left(\pi \mid \mathbf{y}\right)\right)$ and $H(.)$, the Hessian, is the second derivative of the log-likelihood.

Using the chain rule, the Hessian of a single observation y_i can be expressed as,

$$\begin{aligned} H\left(\pi \mid y_i\right) &= \frac{\partial S\left(\pi \mid y_i\right)}{\partial \pi} \\ &= \frac{\partial}{\partial \pi}\left[\frac{dy_i}{\lambda} - \frac{d(1 - y_i)}{1 - \lambda}\right] \\ &= -\left[\frac{d^2 + d^2 S\left(\pi \mid y_i\right) - 2d\lambda S\left(\pi \mid y_i\right)}{\lambda(1 - \lambda)}\right]. \end{aligned}$$

The negative expectation of the Hessian is then

$$-E\left(H\left(\pi \mid y_i\right)\right) = \frac{d^2}{\lambda(1 - \lambda)},$$

since $E\left(S\left(\pi \mid y_i\right)\right) = 0$. The n observations are independently sampled. Subsequently, the sample information matrix is given by

$$\begin{aligned} I\left(\pi \mid \mathbf{y}\right) &= \frac{d^2 n}{\lambda(1-\lambda)} \\ &= \frac{d^2 n}{(c+d\pi)(1-(c+d\pi))}, \end{aligned}$$

which provides the Cramer-Rao lower bound of the variance of $\hat{\pi}_{mle}$,

$$var\left(\hat{\pi}_{mle}\right) \geq \frac{(c+d\pi)(1-(c+d\pi))}{d^2 n}.$$

It follows that this lower bound is equal to the variance in Equation (14).

Appendix B: ML Parameter Estimation

GLM Estimation Methodology

It can be shown that the maximum likelihood equations for the parameters of the GLM model for RR data, as defined in Equation (8), are equal to the general form of the GLM maximum likelihood equations. The only difference is that the success probability includes the RR design parameters c and d. The true probability of success, modeled as function of the linear predictors $\pi(\mathbf{x}^t\boldsymbol{\beta})$, is linearly transformed according to the used RR design. It will follow that this linear transformation of the true probability of success does not modify the structure of the maximum likelihood equations. The only difference is that components in the equations also contain the RR design parameters. So, the modified link functions, defined in Table 2, do not influence the estimation equations, since they only define the link between the expected response and the linear predictor.

The log-likelihood function of the generalized linear RR model can be stated as

$$\log L\left(\boldsymbol{\beta};\mathbf{y}\right) \;=\; \sum_{i=1}^{n} y_i \log\big(c + d\pi(\mathbf{x}_i^t\boldsymbol{\beta})\big) + (1-y_i)\log\Big(1 - \big(c + d\pi(\mathbf{x}_i^t\boldsymbol{\beta})\big)\Big)\,. \tag{15}$$

This log-likelihood matches the general form of the log-likelihood of the GLM model for binary data, except for the modification of the success probability. Therefore, the score function and information matrix for the generalized linear RR model will take the same form as those of the GLM. It follows that,

$$\begin{aligned}
S\left(\boldsymbol{\beta};\mathbf{y}\right) &= \frac{\partial \log L\left(\boldsymbol{\beta};\mathbf{y}\right)}{\partial \boldsymbol{\beta}} \\
&= \sum_{i=1}^{n} \frac{\mathbf{x}_i y_i}{c + d\pi\left(\mathbf{x}_i^t\boldsymbol{\beta}\right)} \frac{d\,\partial\pi\left(\eta_i\right)}{\partial\eta_i} - \frac{x_i(1-y_i)}{1-\left(c + d\pi\left(\mathbf{x}_i^t\boldsymbol{\beta}\right)\right)} \frac{d\,\partial\pi\left(\eta_i\right)}{\partial\eta_i} \\
&= \sum_{i=1}^{n} \mathbf{x}_i \frac{d\,\partial\pi\left(\eta_i\right)}{\partial\eta_i} \left[\frac{y_i - \left(c + d\pi\left(\mathbf{x}_i^t\boldsymbol{\beta}\right)\right)}{\left(c + d\pi\left(\mathbf{x}_i^t\boldsymbol{\beta}\right)\right)\left(1-\left(c + d\pi\left(\mathbf{x}_i^t\boldsymbol{\beta}\right)\right)\right)}\right] \\
&= \mathbf{X}^t\mathbf{D}\boldsymbol{\Sigma}^{-1}\left(\mathbf{y} - \boldsymbol{\mu}\right),
\end{aligned} \tag{16}$$

where the design matrix is given by $\boldsymbol{X}^t = (x_1, \dots, x_n)$, the first-order derivatives is given by $\mathbf{D} = \mathrm{Diag}\left(d\,\partial\pi(\eta_1)/\partial\eta, \dots, d\,\partial\pi(\eta_n)/\partial\eta\right)$, and the covariance matrix is given by $\boldsymbol{\Sigma} = \mathrm{Diag}\left(\sigma_1, \dots, \sigma_n\right)$, where $\sigma_i = (c + d\pi(\eta_i))(1 - (c + d\pi(\eta_i)))$. The score function in Equation (16) has the same structure as the general score function for binary response data under the generalized linear model (e.g., Tutz, 2012, p. 63-66).

The asymptotic variance is defined by the expected information matrix, which is given by,

$$\begin{aligned} I\left(\boldsymbol{\beta} \mid \mathbf{y}\right) &= E\left(\sum_{i=1}^{n} S\left(\boldsymbol{\beta}; y_i\right) S\left(\boldsymbol{\beta}; y_i\right)^t\right) \\ &= \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^t \left(\frac{d\,\partial\pi\left(\eta_i\right)}{\partial\eta_i}\right)^2 / \sigma_i \\ &= \mathbf{X}^t\left(\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}^t\right)\mathbf{X}^t, \end{aligned} \tag{17}$$

which resembles the common form of the information matrix for the generalized linear model (e.g., Tutz, 2012, p. 63-66). The matrix of first-order derivative is slightly different, since the elements are multiplied with RR design parameter d. Furthermore, the variance term is different, since it is based on the success probabilities defined in Equation (8). It follows that the asymptotic covariance of the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ is equal to

$$var\left(\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{X}^t\left(\hat{\mathbf{D}}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mathbf{D}}^t\right)\mathbf{X}\right)^{-1}, \tag{18}$$

where the matrix $\hat{\mathbf{D}}$ is the matrix of first-order derivatives evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ and the matrix $\hat{\boldsymbol{\Sigma}}$ is the covariance matrix evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$.

GLMM Estimation Methodology

The log-likelihood of the generalized linear mixed effect response model is equal to the GLM log-likelihood function defined in Equation (15), except that the linear term is extended with the random effect parameters as defined in Equation (10). The estimation methodology of the GLM can be used to estimate the fixed effects

$\boldsymbol{\beta}$. This follows from the fact that the random effect distribution does not include parameter $\boldsymbol{\beta}$. Consider the score function defined in Equation (16). This score function is defined conditionally on the random effects $\mathbf{b}$. Following the procedure of McCulloch et al, (2008, p. 195), the score function for the fixed effects under the generalized linear mixed effect model can be derived from the score function under the generalized linear model. It follows that,

$$\begin{aligned}
\frac{\partial \log L\left(\boldsymbol{\beta}; \mathbf{y}\right)}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}} \int p\left(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{b}\right) p\left(\boldsymbol{\beta} \mid \mathbf{Q}_b\right) d\mathbf{b} / p\left(\mathbf{y}\right) \\
&= \int S\left(\boldsymbol{\beta}; \mathbf{y}, \mathbf{b}\right) p\left(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{Q}_b\right) d\mathbf{b} \\
&= \int \left[\mathbf{X}^t \mathbf{D} \boldsymbol{\Sigma}^{-1} \left(\mathbf{y} - \boldsymbol{\mu}\right)\right] p\left(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{Q}_b\right) d\mathbf{b} \\
&= \mathbf{X}^t E\left(\mathbf{D}\boldsymbol{\Sigma}^{-1} \mid \mathbf{y}\right) \mathbf{y} - \mathbf{X}^t E\left(\mathbf{D}\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \mid \mathbf{y}\right), \qquad (19)
\end{aligned}$$

where the matrices $\mathbf{D}$ and $\boldsymbol{\Sigma}$ are defined below Equation (16), with $\sigma_{ij} = c + d\pi\left(\mathbf{x}_{ij}^t \boldsymbol{\beta} + \mathbf{z}_{ij}^t \mathbf{b}_j\right)$. For the generalized linear mixed effects model, the conditional expected value given the responses $\mathbf{y}$ of the matrices $\mathbf{D}$ and $\boldsymbol{\Sigma}$ are required to estimate the fixed effects. For low-dimensional random effects, Gauss-Hermite approximation can be used to compute the conditional expected values (see, e.g., Tutz, 2012, p. 407-409). In a similar way, the asymptotic variance can be computed using the expression in Equation (17), and by taking the conditional expectation over the matrices $\mathbf{D}$ and $\boldsymbol{\Sigma}$.

Restricted maximum likelihood estimates of the variance components can be obtained through a profile likelihood in which the fixed effects parameters are replaced by their maximum likelihood estimates (e.g., Breslow & Clayton, 1993). The random effects parameters can also be estimated similar to the estimation of the fixed effects (see, McCulloch et al., 2008, p. 196). Bates et al. (2015) give a more general overview of parameter estimation, which is also implemented in the R-package lme4.

It follows that the likelihood equations for the fixed effect parameters of the GLMM for RR data, as defined in Equation (19), have the same structure as the general GLMM likelihood equations for binary data. The modified link functions can be used to relate the expected response to the linear predictor, as defined in Table 2. This also depends on the cumulative distribution function, $\pi(.)$, to model the success probability.

Estimation with Composite Link Functions

Thompson and Baker (1981) and Green (1984) showed that an iteratively re-weighted least squares estimation procedure for the GLM model can be extended to account for composite link functions. They showed that the estimation equations can be adjusted to account for composite link functions.

Table 1: Parametrization of the different randomized response designs.

Survey Design	Par. c	Par. d	$\hat{\pi}_{mle}$	$Var(\pi_{mle})$
Warner	$1-p_1$	$2p_1-1$	$\frac{(\overline{y}-q_1)}{q_2}$	$\frac{(q_1+q_2\pi)(1-(q_1+q_2\pi))}{q_2^2n}$
Forced Response	$(1-p_1)p_2$	p_1	$\frac{(\overline{y}-(q_1p_2))}{p_1}$	$\frac{((q_1p_2)+p_1\pi)(1-((q_1p_2)+p_1\pi))}{p_1^2n}$
Unrelated Question	$(1-p_1)p_2$	p_1	$\frac{(\overline{y}-(q_1p_2))}{p_1}$	$\frac{((q_1p_2)+p_1\pi)(1-((q_1p_2)+p_1\pi))}{p_1^2n}$
Kuk	p_2	p_1-p_2	$\frac{(\overline{y}-p_2)}{p_1-p_2}$	$\frac{(p_2+(p_1-p_2)\pi)(1-(p_2+(p_1-p_2)\pi))}{(p_1-p_2)^2n}$
Crosswise	$1-p_1$	$2p_1-1$	$\frac{(\overline{y}-q_1)}{q_2}$	$\frac{(q_1+q_2\pi)(1-(q_1+q_2\pi))}{q_2^2n}$
Triangular	p_1	$1-p_1$	$\frac{(\overline{y}-p_1)}{q_1}$	$\frac{(p_1+q_1\pi)(1-(p_1+q_1\pi))}{q_1^2n}$

Note: $q_1 = (1-p_1), q_2 = (2p_1-1)$

Table 2: Parameterization of the different link functions for Bernoulli distributed binary RR data.

	Function	Logit	Probit	Compl. log-log	Cauchit
Distribution	$\pi\left(\mathbf{x}_i^t\boldsymbol{\beta}\right)$	Logistic	Cum. Normal	Gumbel	Cauchy
Expected value	μ_i	$c+\frac{d\,e^{\eta_i}}{1+e^{\eta_i}}$	$c+d\Phi(\eta_i)$	$c+d\left(1-\exp\left(1-\exp\left(\eta_i\right)\right)\right)$	$c+d\left(\arctan\left(\eta\right)/\pi+\frac{1}{2}\right)$
Link	$g\left(\mu_i\right)$	$\log\left(\frac{\mu_i-c}{c+d-\mu_i}\right)$	$\Phi^{-1}\left(\frac{\mu_i-c}{d}\right)$	$\log\left(-\log\left(\frac{c+d-\mu_i}{d}\right)\right)$	$\tan\left(\pi\left(\frac{1}{d}(\mu_i-c)\right)\right)$
Inverse link	$g^{-1}\left(\eta_i\right)$	$c+\frac{d\,e^{\eta_i}}{1+e^{\eta_i}}$	$c+d\Phi(\eta_i)$	$c+d\left(1-\exp\left(1-\exp\left(\eta_i\right)\right)\right)$	$c+d\left(\arctan\left(\eta\right)/\pi+\frac{1}{2}\right)$
Derivative	$\frac{dg^{-1}(\eta_i)}{d\eta_i}$	$\frac{d\,e^{\eta_i}}{\left(1+e^{\eta_i}\right)^2}$	$d\phi(\eta_i)$	$d\exp\left(\eta_i\right)\exp\left(-\exp\left(\eta_i\right)\right)$	$d/\left(\pi(1+\eta^2)\right)$

Note: $\eta_i=\boldsymbol{\beta}^t\mathbf{x}_i$, $g(\mu_i)=\eta_i$, $\arctan(.)=\tan^{-1}(.)$

Table 3: Parameter recovery (across 50 data replications) of the GLMM for the four different link functions. The RR data were generated according to the forced response method with $p_1 = .778$ and $p_2 = .50$.

		Logit		Probit		Cloglog		Cauchit	
N	J	Est.	95% CI	Est.	95% CI	Est.	95% CI	Est.	95% CI
Fixed Effect β									
200	10	0.97	[0.72, 1.20]	1.01	[0.75, 1.30]	0.96	[0.75, 1.20]	0.97	[0.64, 1.30]
	20	0.99	[0.78, 1.20]	0.98	[0.78, 1.20]	0.98	[0.75, 1.20]	1.00	[0.73, 1.30]
500	10	0.95	[0.75, 1.20]	0.97	[0.80, 1.10]	0.95	[0.79, 1.10]	0.93	[0.77, 1.10]
	20	0.97	[0.83, 1.10]	0.98	[0.84, 1.10]	0.97	[0.87, 1.10]	0.93	[0.78, 1.10]
Random Effect σ^2									
200	10	0.46	[0.22, 0.70]	0.48	[0.36, 0.60]	0.46	[0.33, 0.60]	0.49	[0.27, 0.71]
	20	0.49	[0.40, 0.58]	0.49	[0.40, 0.58]	0.48	[0.38, 0.58]	0.49	[0.34, 0.63]
500	10	0.47	[0.34, 0.60]	0.48	[0.39, 0.57]	0.48	[0.38, 0.57]	0.44	[0.30, 0.58]
	20	0.48	[0.41, 0.56]	0.48	[0.42, 0.54]	0.48	[0.42, 0.53]	0.47	[0.40, 0.55]

Table 4: Plagiarism study: Estimated effects of student age and type of questioning on partial and severe plagiarism using the GLM with four different link functions.

	Logit		Probit		Cauchit		Compl. log-log	
	Est.	S.E	Est.	S.E	Est.	S.E	Est.	S.E
Intercept	0.63	2.94	0.49	1.46	-21.53	28.29	0.16	2.73
Age	-0.26*	0.13	-0.14*	0.07	-0.34	0.24	-0.23.	0.12
RR	1.34*	0.52	0.69*	0.27	3.18.	1.79	1.24*	0.48
Partial Plag.	2.40*	1.09	1.11*	0.43	24.53	27.79	2.29*	1.07
AIC	813.32		813.25		815.18		813.39	

Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Table 5: Smoking Behavior Study: Influence of the questioning technique

	Model 1		Model 2	
	Est.	S.E	Est.	S.E
Intercept	.61*	.12	2.33*	.57
Item 2	-1.80*	.20	-2.04*	.22
Item 3	-1.79*	.20	-2.04*	.21
RR	.79*	.17	1.50*	.74
Asthma			-1.70*	.27
Lung Cancer			.24	.37
Bronchitis			-2.01*	.45
Lung Emphysema			-.30	.45
Other			-.90*	.23
Age			-.02*	.01
AIC	1084.3		1022.0	

Table 6: Student Cheating Study: Thirteen items about different types of cheating.

Number	Item
During an exam or test:	
3	Tried to confer with other students
4	Allowed others to copy your work
5	Others allowed you to copy their work
10	Used crib notes or cheat sheets
11	Used unauthorized material such as books or notes
12	Looked at another students test paper with his or her knowledge
17	Added information to authorized material
20	Took along an exam illegally
29	Invented data (i.e., entered nonexistent results into the database)
30	Altered data to obtain significant results
31	Lied to postpone a deadline
33	Minimized effort in a joint assignment
34	Submitted course work from others without their knowledge

Table 7: Student Cheating Study: Effects of educational program and gender on cheating behavior, while accounting for item and student differences.

	Model0 (Logit)		Model1 (Logit)		Model2 (Logit)		Model3 (Probit)	
	Est.	S.E	Est.	S.E.	Est.	S.E.	Est.	S.E.
Fixed Effects								
Intercept	-1.00	0.10						
Item 4					0.67	0.28	0.40	0.17
Item 5					0.21	0.29	0.12	0.17
Item 10					1.08	0.28	0.65	0.16
Item 11					0.24	0.30	0.14	0.17
Item 12					0.27	0.29	0.16	0.17
Item 17					0.98	0.28	0.59	0.16
Item 20					0.33	0.29	0.20	0.17
Item 29					0.29	0.30	0.18	0.17
Item 30					0.45	0.30	0.27	0.17
Item 31					0.92	0.28	0.55	0.16
Item 33					0.11	0.30	0.07	0.18
Item 34					0.25	0.30	0.15	0.17
Male			-0.07	0.17	-0.08	0.18	-0.04	0.11
Program								
CS			-1.46	0.26	-1.99	0.34	-1.19	0.20
PAT			-0.61	0.23	-1.10	0.31	-0.66	0.18
ACS			-0.59	0.19	-1.07	0.29	-0.65	0.17
ST			-0.97	0.24	-1.46	0.33	-0.88	0.19
EST			-0.93	0.17	-1.43	0.28	-0.86	0.16
ME			-1.19	0.23	-1.70	0.32	-1.02	0.19
PS			-0.93	0.27	-1.44	0.35	-0.86	0.20
Random Effects								
	Var.	SD.	Var.	SD.	Var.	SD.	Var.	SD.
Student	0.84	0.92	0.76	0.87	0.82	0.91	0.29	0.54
Item	0.07	0.26	0.07	0.26				
Information Criteria								
AIC	5771.7		5768.2		5759.8		5759.2	
Deviance	5765.7		5748.2		5717.8		5717.2	

Note: Adaptive Gauss-Hermite quadrature was used for estimation using 1 to 24 quadrature points for a single integral.