DEPARTMENT OF ENGINEERING MANAGEMENT

An argument for preferring Firth bias-adjusted estimates in aggregate and individual-level discrete choice modeling

Roselinde Kessels, Bradley Jones & Peter Goos

UNIVERSITY OF ANTWERP
Faculty of Applied Economics

Stadscampus
Prinsstraat 13, B.226
BE-2000 Antwerpen
Tel. +32 (0)3 265 40 32
Fax +32 (0)3 265 47 99
www.ua.ac.be/tew

FACULTY OF APPLIED ECONOMICS

DEPARTMENT OF ENGINEERING MANAGEMENT

An argument for preferring Firth bias-adjusted estimates in aggregate and individual-level discrete choice modeling

Roselinde Kessels, Bradley Jones & Peter Goos

RESEARCH PAPER 2013-013
AUGUST 2013

University of Antwerp, City Campus, Prinsstraat 13, B-2000 Antwerp, Belgium
Research Administration – room B.226
phone: (32) 3 265 40 32
fax: (32) 3 265 47 99
e-mail: joeri.nys@ua.ac.be

D/2013/1169/013

An argument for preferring Firth bias-adjusted estimates in aggregate and individual-level discrete choice modeling

Roselinde Kessels*†

Universiteit Antwerpen

Bradley Jones‡

SAS Institute

Universiteit Antwerpen

Peter Goos§

Universiteit Antwerpen

Erasmus Universiteit Rotterdam

*Roselinde Kessels is a postdoctoral researcher in econometrics at Universiteit Antwerpen, Faculty of Applied Economics and StatUa Center for Statistics, Prinsstraat 13, 2000 Antwerpen, Belgium. Tel.: +32 (0)3 265 40 95. E-mail: roselinde.kessels@uantwerpen.be

†Corresponding author.

‡Bradley Jones is a principal research fellow at SAS Institute Inc., SAS Campus Drive, Cary, NC 27513, USA. He is also a guest professor at Universiteit Antwerpen. Tel.: +1 919 531 4161. E-mail: bradley.jones@jmp.com

§Peter Goos is a full professor in statistics at Universiteit Antwerpen, Faculty of Applied Economics and StatUa Center for Statistics, Prinsstraat 13, 2000 Antwerpen, Belgium. He is also a professor in statistics at Erasmus Universiteit Rotterdam, Erasmus School of Economics, Postbus 1738, 3000 DR Rotterdam, The Netherlands. Tel.: +32 (0)3 220 40 59. E-mail: peter.goos@uantwerpen.be

An argument for preferring Firth bias-adjusted estimates in aggregate and individual-level discrete choice modeling

Abstract

Using maximum likelihood estimation for discrete choice modeling of small datasets causes two problems. The first problem is that the data often exhibit separation, in which case the maximum likelihood estimates do not exist. Also, provided they exist, the maximum likelihood estimates are biased. In this paper, we show how to adapt Firth's bias-adjustment method for use in discrete choice modeling. This approach removes the first-order bias of the estimates, and it also deals with the separation issue. An additional advantage of the bias adjustment is that it is usually accompanied by a reduction in the variance. Using a large-scale simulation study, we identify the situations where Firth's bias-adjustment method is most useful in avoiding the problem of separation as well as removing the bias and reducing the variance. As a special case, we apply the bias-adjustment approach to discrete choice data from individuals, making it possible to construct an empirical distribution of the respondents' preferences without imposing any *a priori* population distribution. For both research purposes, we base our findings on data from a stated choice study on various forms of employee compensation.

Keywords: discrete choice modeling, Firth's bias adjustment, penalized maximum likelihood, individual-level estimates, data separation

1 Introduction

Discrete choice models relate respondents' choices of one of two or more alternatives or profiles to the attributes of the respondents and the attributes of the alternatives. Data for discrete choice models are either collected via discrete choice experiments (DCEs), where respondents state their choices in hypothetical situations, or via observational studies, where respondents have actually made real-life choices. The former type of data are called stated choice data or conjoint choice data, the latter type revealed choice data. Stated choice data have been used to predict preferences for prospective goods in marketing, innovative health programs in health economics, new transportation systems in transport planning, and various other applications in social and economics research often involving new developments. Revealed choice data have been used to study actual choices of, for example, which car to buy, where to go to college and which mode of transport (car, bus, rail) to use for commuting to work.

A commonly used procedure to fit discrete choice models is maximum likelihood (ML) estimation. The resulting ML estimators possess a number of asymptotic properties, including efficiency and unbiasedness. However, for many applications, the sample data collected are small, especially those from DCEs, so one can no longer rely on the asymptotic properties. When fitting discrete choice models based on small datasets using traditional ML estimation techniques, two different issues cause concern. First, small datasets often exhibit separation. In that case, the ML estimates do not exist. Second, in case the ML estimates do exist, they are biased. In this paper, we show how to overcome these two problems using the bias-adjustment method of Firth (1993, 1995). A major advantage of the method is that it allows fitting a discrete choice model to individual-level data, and, subsequently, exploring the heterogeneity in the respondents' preferences.

Individual-level choice data often exhibit separation. In general, separation occurs in discrete choice data if the responses can be perfectly classified by a linear combination of the attributes of the alternatives (see, for studies on logistic regression, Albert and Anderson (1984), Santner and Duffy (1986), Lesaffre and Albert (1989) and Allison (2008)). When data separation exists, the ML estimator does not. In computer implementations of maximum likelihood estimation that do not recognize separation, the likelihood converges while at least one parameter gets large without bound. The actual parameter estimate reported is then a function of the convergence criterion for the likelihood rather than having any practical meaning. Albert and Anderson (1984) introduced the distinction between complete and quasi-complete separation. Complete separation occurs when a combination of the attributes classifies responses without error according to a strict inequality. Quasi-complete separation occurs when a combination of the attributes classifies responses without error up to a non-strict inequality.

Complete and quasi-complete separation are known to occur more often in small samples than in large samples. This has been recognized in biostatistics (Heinze, 2006), econometrics (Beggs et al., 1981), marketing (Chapman, 1984) and industry (Goos and Gilmour, 2012). A more recent insight is that separation is also more likely to occur for specific designs for data collection. Woods and van de Ven (2011) state that separation may also occur in larger experiments, for example, when an experimental design is used in which the success probability of every experimental observation is near 0 or 1. Kessels et al. (2011b,c) describe an example of an orthogonal design for a DCE involving eight choice sets of two alternatives. Such design is also called a utility-neutral optimal design because it relies on the assumption that people are ambivalent about any of the attribute levels, and thus also about any of the alternatives. The utility-neutral design of Kessels et al. (2011b,c) is poor since it leads to separation 20% of the time when there are 100 respondents and 4% of the time when there are 200 respondents. The authors also explain that Bayesian optimal designs for DCEs usually do not lead to data separation. That is because the Bayesian design methodology is state-of-the-art for constructing efficient DCEs as it incorporates the available information about people's preferences for various attributes in the choice design. A key feature of many DCEs is that they involve a small set of levels for the attributes, which makes them more vulnerable to data separation than studies with explanatory variables that take many different levels.

Various approaches have been proposed to overcome data separation in clinical trials (see, for instance, Clogg et al. (1991), Cardell (1993), Firth (1993, 1995), Bull et al. (2002), Heinze and Schemper (2002), Heinze (2006)). In all these approaches, this is accomplished by estimating the parameters using penalized maximum likelihood, which shrinks the estimates toward zero. In this paper, we propose using the approach of Firth (1993, 1995) to deal with the separation issue. A major advantage of that approach is that it has been designed to remove the first-order bias of ML estimates. Hence, Firth's method essentially kills two birds with one stone, overcoming the two weaknesses of ML estimation at the same time.

The strong theoretical underpinnings of Firth's method as a bias reducing technique give it a major advantage over other penalized likelihood approaches, as the bias in the ML estimates is often far from negligible. For instance, King and Ryan (2002) show that, for particular underlying parameter values and sample sizes, the absence of complete or quasi-complete separation does not guarantee a sufficiently small bias so as to result in satisfactory estimates.

Neither solutions to the separation problem nor bias reducing techniques have generated much attention in the literature on the analysis of stated or revealed choice data, due to the fact that data from individual respondents are usually pooled or aggregated. The data pooling obviously leads to large datasets and generally prevents the occurrence of data separation. The importance of modeling respondent heterogeneity, however, has lead to an increased interest in estimates of individual-level preference parameters and a large body of literature on the use of mixed logit models (Train, 2003), hierarchical Bayes models (Lenk et al., 1996), latent class models (Andrews et al., 2002), and convex optimization techniques (Evgeniou et al., 2007) to obtain individual-level preference estimates.

We advocate using Firth's method to obtain individual-level preference estimates, by fitting a model to each individual's responses separately when the number of choice sets evaluated by each respondent permits. In doing so, we can construct an empirical distribution of the respondents' preferences. Unlike panel mixed logit models, latent class models, hierarchical Bayes approaches or Evgeniou et al.'s (2007) convex optimization technique, this approach does not require imposing an *a priori* preference heterogeneity distribution for each of the model parameters. This is important since it is not at all clear what an appropriate *a priori* preference heterogeneity distribution would be in most practical applications. It is common to use the normal distribution or a lognormal distribution for the random preference parameters in a panel mixed logit model, but this is often merely a choice out of convenience. Similarly, Evgeniou et al. (2007) focus on unimodal representations of heterogeneity, but they indicate that extending their approach to other types of heterogeneity would be useful. Also the fact that Lenk and Orme (2009) demonstrate that the typical uninformative priors specified for hierarchical Bayes estimation have undesirable effects on the posterior in cases where the data is sparse pleads in favor of fitting a model to each individual's data, without imposing a population distribution.

The idea to model individuals' preferences directly, and not indirectly based on assumptions about preference distributions across samples of people is, however, not new. Mainly Louviere et al. (2008) proposed new ways of collecting additional information to expand the amount of available choice information for modeling the choices of individual decision makers. They call the approach that obtains individual-level parameters for the empirical distribution of sample preferences a "bottom-up" approach. They call the opposite approach that makes use of prior distributional assumptions a "top-down" approach. The authors state that, in theory, if one specifies correct preference distributions, and the number of choices per person is sufficiently large, top-down and bottom-up approaches should give the same results. In contrast, if assumptions about preference distributions are incorrect, the inferences from top-down models will be biased and incorrect. Furthermore, bottom-up approaches are more researcher-friendly since they are computationally simpler than the current top-down practices.

The discrete choice model that Louviere et al. (2008) used for individual-level preference estimation is the multinomial logit model. We use the same model for the application of Firth's bias-adjustment method. In Section 2, we review the multinomial logit model and explain the ML and Firth's estimation techniques for this model. In Section 3, we demonstrate the usefulness of Firth's method for individual-level preference estimation using an application that is concerned with employee compensation. To obtain a comprehensive overview of the situations in which Firth's method proves most effective, we describe the results of a detailed simulation study in Section 4. Finally, in Section 5, we summarize and discuss the results and point out some avenues for further research.

2 Model estimation

In this section, we sketch the multinomial logit model for analyzing stated or revealed choice data, and discuss the maximum likelihood estimation approach. Next, we explain how to adapt the bias-adjustment method of Firth (1993, 1995) for estimating the multinomial logit model and conclude with some inferential issues.

2.1 Multinomial logit model

The multinomial logit (MNL) model (McFadden, 1974) employs random utility theory which describes the utility that a respondent attaches to profile j ($j = 1, \ldots, J$) in choice set s ($s = 1, \ldots, S$) as the sum of a systematic and a stochastic component:

$$U_{js} = \mathbf{x}'_{js}\boldsymbol{\beta} + \varepsilon_{js}. \tag{1}$$

In the systematic component $\mathbf{x}'_{js}\boldsymbol{\beta}$, $\mathbf{x}_{js}$ is a $k \times 1$ vector describing the levels of the attributes of profile j in choice set s. The vector $\boldsymbol{\beta}$ is a $k \times 1$ vector of parameter values representing the effects of the attribute levels on the utility. The stochastic component ε_{js} is the error term, which is assumed to be independently and identically standard Gumbel distributed. Depending on the situation, the attributes may be continuous or categorical variables. For the sake of simplicity, we assume that the utility model involves main effects only, which are also called part-worths. When using aggregate data, the part-worth vector $\boldsymbol{\beta}$ is the same for every respondent.

Under the standard Gumbel distributional assumption, the MNL probability that a respondent chooses profile j in choice set s is

$$p_{js}(\mathbf{X}_s, \boldsymbol{\beta}) = \frac{\exp\left(\mathbf{x}'_{js}\boldsymbol{\beta}\right)}{\sum_{t=1}^{J} \exp\left(\mathbf{x}'_{ts}\boldsymbol{\beta}\right)}, \tag{2}$$

where $\mathbf{X}_s = [\mathbf{x}_{1s}, \ldots, \mathbf{x}_{Js}]'$ is the design matrix for choice set s. The stacked $\mathbf{X}_s$ matrices provide the design matrix $\mathbf{X}$ for the choice study. This mathematical notation assumes that only one survey is used for all respondents, but generalization to situations where a different survey is used for each respondent is trivial.

2.2 Maximum likelihood estimation

A standard estimation technique for the MNL model is maximum likelihood (ML) estimation. If we denote the choices from R respondents by a binary response variable, y_{jsr}, which takes the value one if respondent r, $r = 1, \ldots, R$, chooses profile j in choice set s and zero otherwise, then we obtain the ML estimator for the parameter vector $\boldsymbol{\beta}$ by maximizing the likelihood function

$$L(\boldsymbol{\beta}) = \prod_{r=1}^{R} \prod_{s=1}^{S} \prod_{j=1}^{J} (p_{js})^{y_{jsr}}, \tag{3}$$

or, alternatively, by maximizing the log-likelihood function

$$LL(\boldsymbol{\beta}) = \sum_{r=1}^{R}\sum_{s=1}^{S}\sum_{j=1}^{J} y_{jsr}\,\ln(p_{js}) \tag{4}$$

with respect to $\boldsymbol{\beta}$. We call the resulting estimator of the parameter vector $\hat{\boldsymbol{\beta}}_{\text{ML}}$. The ML estimates are usually found by equating the score function or the gradient of the log-likelihood function to zero and solving the resulting system of nonlinear equations. In the first part of Appendix A, we show that the score function is

$$\frac{\partial LL(\boldsymbol{\beta})}{\partial \beta_i} = \sum_{r=1}^{R}\sum_{s=1}^{S}\left(x_{msri} - \mathbf{p}_s'\mathbf{x}_{si}^*\right), \quad i = 1, \ldots, k, \tag{5}$$

where x_{msri} is the ith entry of vector $\mathbf{x}_{msr}$ denoting the profile that respondent r chooses from choice set s and $\mathbf{x}_{si}^*$ is the ith column vector of choice set matrix $\mathbf{X}_s$.

The ML estimator is known to be biased in finite data samples. Therefore, the expectation of the ML estimator $\hat{\boldsymbol{\beta}}_{\text{ML}}$ for the parameter vector $\boldsymbol{\beta}$ can be generally expressed as

$$\text{E}\left(\hat{\boldsymbol{\beta}}_{\text{ML}}\right) = \boldsymbol{\beta} + b\left(\hat{\boldsymbol{\beta}}_{\text{ML}}\right), \tag{6}$$

$$= \boldsymbol{\beta} + \frac{b_1(\hat{\boldsymbol{\beta}}_{\text{ML}})}{N} + \frac{b_2(\hat{\boldsymbol{\beta}}_{\text{ML}})}{N^2} + \frac{b_3(\hat{\boldsymbol{\beta}}_{\text{ML}})}{N^3} + \ldots, \tag{7}$$

where $b(\hat{\boldsymbol{\beta}}_{\text{ML}})$ is the bias, $b_1, b_2, b_3, \ldots$ are $O(1)$ functions of $\hat{\boldsymbol{\beta}}_{\text{ML}}$, which can be obtained explicitly once the model has been specified, and $N = RS(J-1)$ is the degrees of freedom (DF) for all R respondents. The first-order bias due to the term $b_1(\hat{\boldsymbol{\beta}}_{\text{ML}})/N$ is negligible for large samples, but can be severe with small or sparse datasets. Therefore, several techniques have been proposed in the literature to correct the first-order bias after obtaining the ML estimates (see, for instance, Quenouille (1949, 1956)). However, this type of after-the-fact bias reduction is possible only if the ML estimates exist. Hence, it fails in the presence of data separation.

2.3 Firth's method

The weakness of after-the-fact bias reduction techniques inspired Firth (1993, 1995) to propose a general method for removing the first-order term, $b_1(\hat{\boldsymbol{\beta}}_{\text{ML}})/N$, from the expression for the bias in Equation (7) in a way that does not rely on the existence of the ML estimator $\hat{\boldsymbol{\beta}}_{\text{ML}}$. This is achieved by modifying the score function, or, equivalently, by penalizing the likelihood function using the Jeffreys prior. The Jeffreys prior is a non-informative prior distribution (Jeffreys, 1946) which is proportional to the square root of the determinant of the Fisher information matrix of the model under study. For the MNL model, the Fisher information matrix is

$$\mathbf{M}\left(\boldsymbol{\beta}\right) = R\,\sum_{s=1}^{S}\mathbf{X}_s'\left(\mathbf{P}_s - \mathbf{p}_s\mathbf{p}_s'\right)\mathbf{X}_s, \tag{8}$$

where $\mathbf{p}_s = [p_{1s}, \ldots, p_{Js}]'$ and $\mathbf{P}_s = \text{diag}[p_{1s}, \ldots, p_{Js}]$.

Firth's penalized likelihood function is therefore

$$L_{\text{FIRTH}}(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) \sqrt{|\mathbf{M}(\boldsymbol{\beta})|}, \tag{9}$$

where the ordinary likelihood function $L(\boldsymbol{\beta})$ for the MNL model is given by Equation (3). Subsequently, Firth's penalized log-likelihood function becomes

$$LL_{\text{FIRTH}}(\boldsymbol{\beta}) = LL(\boldsymbol{\beta}) + \frac{1}{2} \ln |\mathbf{M}(\boldsymbol{\beta})|, \tag{10}$$

where the ordinary log-likelihood function $LL(\boldsymbol{\beta})$ for the MNL model is given by Equation (4).

Maximizing the penalized log-likelihood function requires equating the following modified score function to zero:

$$\frac{\partial LL_{\text{FIRTH}}(\boldsymbol{\beta})}{\partial \beta_i} = \frac{\partial LL(\boldsymbol{\beta})}{\partial \beta_i} + \frac{1}{2} \frac{\partial \ln |\mathbf{M}(\boldsymbol{\beta})|}{\partial \beta_i}, \quad i = 1, \ldots, k. \tag{11}$$

This equation consists of two parts: the original ML score function and the first-order derivative of the logarithm of the penalty function with respect to β_i. In Appendix A, we derive Firth's modified score function for the MNL model showing that the former part corresponds to Equation (5) and the latter part to

$$\frac{1}{2} \frac{\partial \ln |\mathbf{M}(\boldsymbol{\beta})|}{\partial \beta_i} = R \sum_{s=1}^{S} \left[\frac{1}{2} \operatorname{tr} \left(\left(\mathbf{X}_s \mathbf{M}^{-1} \mathbf{X}_s' \right) \frac{\partial \mathbf{P}_s}{\partial \beta_i} \right) - \mathbf{p}_s' \left(\mathbf{X}_s \mathbf{M}^{-1} \mathbf{X}_s' \right) \frac{\partial \mathbf{p}_s}{\partial \beta_i} \right]. \tag{12}$$

We denote the Firth bias-adjusted estimator of the parameter vector resulting from the modified score function by $\hat{\boldsymbol{\beta}}_{\text{FIRTH}}$.

2.4 Inferential issues

Once a model has been estimated, it is usually desirable to make inferences about its parameters. It is often informative to calculate the standard errors of the Firth bias-adjusted estimates. We estimate the standard errors by the square roots of the diagonal elements of the asymptotic variance-covariance matrix of the Firth estimates given by

$$\begin{aligned} \mathbf{Var}_{\text{FIRTH}} \left(\hat{\boldsymbol{\beta}}_{\text{FIRTH}} \right) &= \mathbf{M}_{\text{FIRTH}}^{-1} \left(\hat{\boldsymbol{\beta}}_{\text{FIRTH}} \right), & (13) \\ &= \left(- \frac{\partial^2 LL_{\text{FIRTH}} \left(\hat{\boldsymbol{\beta}}_{\text{FIRTH}} \right)}{\partial \hat{\boldsymbol{\beta}}_{\text{FIRTH}} \partial \hat{\boldsymbol{\beta}}_{\text{FIRTH}}'} \right)^{-1}. & (14) \end{aligned}$$

To determine which effects are statistically significant, it is standard to perform likelihood ratio (LR) tests. In such tests, we evaluate the difference in goodness of fit between nested models with Firth estimates. More specifically, we compare an unrestricted model, with estimate $\hat{\boldsymbol{\beta}}^{U}_{\text{FIRTH}}$, to a restricted model, with estimate $\hat{\boldsymbol{\beta}}^{R}_{\text{FIRTH}}$. To perform a LR test, one option would be to compare the penalized log-likelihood function values of the two models computed using Equation (10). However, this option is not feasible because the information matrices of the two models have different dimensions so that their determinants cannot be compared. We therefore suggest using the original log-likelihood function values of the two models computed using Equation (4) to perform a LR test. In other words, we suggest computing the test statistic

$$-2\left[LL\left(\hat{\boldsymbol{\beta}}^{R}_{\text{FIRTH}}\right) - LL\left(\hat{\boldsymbol{\beta}}^{U}_{\text{FIRTH}}\right)\right], \tag{15}$$

and comparing it to the χ^2_ν reference distribution, where ν denotes the number of restrictions imposed on the parameters in the restricted model. We can motivate this approach by the following theorem. If the ML estimates $\hat{\boldsymbol{\beta}}^{R}_{\text{ML}}$ and $\hat{\boldsymbol{\beta}}^{U}_{\text{ML}}$ exist, Equation (15) is asymptotically equivalent to the LR test statistic using the original log-likelihood function values of the restricted and unrestricted models with ML estimates,

$$-2\left[LL\left(\hat{\boldsymbol{\beta}}^{R}_{\text{ML}}\right) - LL\left(\hat{\boldsymbol{\beta}}^{U}_{\text{ML}}\right)\right]. \tag{16}$$

We provide a proof of this theorem in Appendix B.

Note that all derivations and results described in this section have been incorporated in the Choice Modeling platform of the statistical software package JMP (SAS Institute, Cary, NC, USA). More specifically, JMP provides Firth's method for estimating the MNL model, computes the variance-covariance matrix of the Firth estimates according to Equation (14) and performs LR tests according to Equation (15).

3 An application in employee compensation

We present an application using stated choice data related to employee compensation to illustrate the usefulness of Firth's bias-adjustment method for estimating the MNL model, in particular for estimating individual-level MNL models. We first describe the attributes and attribute levels of interest in the study, and the methodology used to design the study. We then proceed with the estimation of different MNL models using the traditional ML method and Firth's method, and conclude with a discussion on MNL model comparison.

3.1 Attributes and levels

We commissioned a study to investigate preferences for various forms of compensation of employees. The study was done using a choice experiment involving four three-level attributes, salary increase, bonus, extra vacation and flexible working time. In Table 1, we show the levels used for each of the attributes. A compensation scheme or profile is then a combination of attribute levels, one level for each attribute. We assume all attributes are categorical, which enables us to capture possible nonlinear effects of the attributes on the perceived utility of a compensation scheme. We modeled the attribute levels using effects-type coding which constrains the model parameters associated with each of the attributes to sum to zero. This means that the three levels of each attribute are coded as [1 0], [0 1] and $[-1\ \ -1]$, respectively.

<Insert Table 1 about here>

3.2 Design of the study

To design the questionnaire, we generated 24 choice sets of three profiles using the Bayesian $\mathcal{D}$-optimal experimental design approach described in Kessels et al. (2009) and implemented in the statistical software package JMP. We divided the design into two surveys of 12 choice sets, where every respondent evaluated one survey. We spread the two surveys equally over a total of 448 respondents who participated in the questionnaire. For the construction of the Bayesian design, we assumed that, all other things being equal, the third level of each attribute was preferred over the middle level, and that the middle level was preferred over the first level. More specifically, we assumed a normally distributed prior utility with a mean of -0.6 for the first level of each attribute, a mean of 0 for the second level, and a mean of $+0.6$ for the third level. We specified a variance of 0.1 around every prior mean value. To ensure that the variance for the third level of each attribute also equals 0.1, we followed Kessels et al. (2008) in specifying a covariance of -0.05 between the prior utilities for the first and middle level of each attribute.

3.3 MNL model estimation

We analyzed the data from the compensation study first by aggregating the data from all 448 respondents and estimating the MNL model assuming identical part-worths for each respondent. We then analyzed the data from each respondent separately by estimating 448 MNL models, one for each respondent. We used the traditional ML method and Firth's method for estimating the MNL models. We performed all analyses using JMP.

For the aggregate data analysis, we estimated all $k = 8$ part-worths using the traditional ML method and Firth's method. Both methods led to the same estimates so that $\hat{\boldsymbol{\beta}}_{\text{ML}} = \hat{\boldsymbol{\beta}}_{\text{FIRTH}}$. This is due to the large number of respondents, choice sets and profiles, providing a total of 10,752 DF, which makes the bias of the ML estimates negligible. The second column of Table 2 contains the part-worth estimates from the aggregate data analysis, where the implied estimates for the last level of each attribute corresponding to effects-type coding are also shown. Bonus and salary increase are the most preferred forms of compensation, followed by extra vacation and flexible working time.

<Insert Table 2 about here>

We also estimated a MNL model for each of the 448 respondents separately using the traditional ML method and Firth's method. Each individual dataset provides a total of 24 DF for the estimation of $k = 8$ individual-level part-worths. In doing so, the ML method resulted in separation for 386 respondents, i.e. in 86% of the cases. Because of this high probability of separation, the traditional ML method is not a viable method for individual-level estimation. On the other hand, Firth's method yielded individual-level part-worth estimates for every respondent. The last three columns of Table 2 contain the mean individual-level estimates obtained using Firth's method, the standard deviations of the individual-level estimates, and the standard errors of their means. In accordance with effects-type coding, the mean estimates for the last level of each attribute can be calculated as minus the sum of the other mean estimates for that attribute. We conclude from the table that the mean estimates lie close to the estimates from analyzing the aggregate data, within approximately two standard errors, and that the individual-level estimates make sense overall. Note that, in general, there is no reason to expect that the mean individual-level estimates converge to the estimates from the aggregate data analysis. This is because of Jensen's inequality. That is, a nonlinear function of the expectation of a random variable is generally not equal to the expectation of the nonlinear function of the random variable. This applies here because the parameter estimates in a MNL model are a nonlinear function of the responses.

Finally, Figure 1 plots the estimates from analyzing the aggregate data together with the Firth individual-level estimates. The latter estimates are displayed using the 95% confidence intervals of their means. As the summary statistics in Table 2 already revealed, the confidence intervals are fairly narrow and either contain or come close to the estimates from the aggregate data analysis, illustrating that the Firth individual-level estimates make sense.

<Insert Figure 1 about here>

3.4 MNL model comparison

The motivation for estimating individual-level MNL models is that we assume that preferences are heterogeneous across respondents. Therefore, an individual-level model specification should mimic reality better and provide a better fit to the data than the MNL model assuming homogeneous preferences. It is interesting to investigate whether this is indeed the case by comparing the homogeneous and heterogeneous MNL model specifications for the compensation study. We therefore perform a LR test using Equation (15). More specifically, we compare the restricted model specification, that is, the MNL model with homogeneous part-worth estimates, to the unrestricted model specification, allowing for 448 individual-level vectors of part-worth estimates. The ordinary log-likelihood value for the restricted model is $-4,747.7$, whereas, for the unrestricted model, it is $-2,069.9$. The value for the LR test statistic is $5,355.6$.

Under the null hypothesis of equal part-worths across respondents, the LR test statistic is χ^2_ν distributed with $\nu = 3,576$. In total, there are $3,576$ restrictions on the part-worths, because the null hypothesis of homogeneous preferences across the 448 respondents requires 447 equalities for each of the eight part-worths. The p-value for the LR test statistic is essentially zero, so that we decisively reject the null hypothesis of equal part-worths. The individual-level MNL model specification is thus a better approximation of reality than the MNL model assuming homogeneous part-worths. This conclusion is similar to that drawn by Beggs et al. (1981) who studied potential consumer demand for electric cars. So, we have shown that there is significant respondent heterogeneity, which begs the question of segmentation. Uncovering the source of the respondent heterogeneity is, however, beyond the scope of this paper.

4 Simulation study

In this section, we present a simulation study to identify the situations in which Firth's bias-adjustment method proves most useful. We first discuss the setup of the simulation study revealing the various experimental conditions. We then compare the performance of Firth estimates with that of ML estimates obtained from aggregate data as well as individual-level data in each of the conditions.

4.1 Setup of the simulation study

We wanted to compare the performance of Firth's method with the traditional ML method for estimating the MNL model by simulating choices from different numbers of respondents in various experimental conditions. As in the compensation study discussed in Section 3, the experimental conditions all involve four three-level attributes. This allowed us to simulate data using the vector of part-worth estimates obtained from the aggregate data analysis of the compensation study. We assumed this vector to be the true vector of part-worths, that is, $\boldsymbol{\beta}_T = [-0.920, 0.186, -1.005, 0.200, -0.460, 0.114, -0.264, 0.096]'$, and used it to simulate a series of $1,000$ datasets with choices from $R = 1$, 6, 12, 24, 48 and 96 respondents. Note that $R = 1$ denotes the individual respondent case.

We originally designed the simulation experiment as a 2^3 factorial experiment, with factors type of design, number of choice sets, S, and number of profiles per choice set, J. The design type is either Bayesian $\mathcal{D}$-optimal or utility-neutral $\mathcal{D}$-optimal (see Kessels et al. (2011a) for a definition of $\mathcal{D}$-optimality in these cases), S is either 12 or 18, and J is either 2 or 3. This setup resulted in eight different designs. Later, in the analysis stage, we learned that the interaction of S and J, corresponding to the residual DF, provided a more natural explanation of our results than studying the effects of S and J separately. More precisely, the residual DF are defined by $S(J-1)-k$, with k the number of part-worths equal to 8. This factor has four levels equal to 4, 10, 16 and 28. Table 3 provides an overview of the characteristics of the eight designs. The designs themselves appear in Appendix C. Note that, for each design, we assume a single survey is used. We do not consider multiple surveys because, in our simulation study, this is equivalent to increasing the number of choice sets.

<Insert Table 3 about here>

As a prior distribution for the Bayesian designs, we chose the same normal distribution used to construct the Bayesian design for the compensation study (see Section 3.2).

Because the assumed part-worth values resulting from the prior parameter specification are reasonably small, the Bayesian designs do not outperform the utility-neutral designs by much in terms of the Bayesian $\mathcal{D}$-optimality criterion. More specifically, the utility-neutral designs 2, 4, 6 and 8 have relative efficiencies equal to 83.72%, 87.58%, 88.54% and 89.19%, respectively. The relative performance of the utility-neutral designs therefore increases with the residual DF.

For each of the simulated datasets in the experimental conditions, we used the traditional ML method and Firth's method to estimate the part-worths. We identified the cases involving data separation and quantified the estimates' bias and variance. To measure the overall quality of the estimates, we also computed their mean squared error (MSE) as the average of the squared differences between each estimate and the true value of the corresponding part-worth. The MSE then decomposes into a sum of the squared bias and variance. We used all part-worth estimates in the computations, including the implied part-worth estimates resulting from effects-type coding. In the next two sections, we discuss the performance of the part-worth estimates from aggregate data and from individual-level data.

4.2 Performance of the estimates from aggregate data

In this section, we compare the estimation performance of the traditional ML method with Firth's method using the series of $1,000$ simulated datasets with choices from $R =$ 6, 12, 24, 48 and 96 respondents generated with the eight designs in Table 3. Some of the cases resulted in data separation, though for every case, Firth's method was able to provide part-worth estimates. Table 4 shows the cases in which data separation occurred, as well as the frequency with which it happened. They mainly involve choice data from a small number of respondents (equal to 6, 12 and 24) generated using the utility-neutral designs with few residual DF (equal to 4 and 10). The worst case resulted in separation in 42.4% of the datasets and involved choices from 6 respondents generated using the utility-neutral design with 4 residual DF. For the smallest number of respondents and residual DF, the Bayesian design also resulted in separation, but only in 1.7% of the datasets. Because of the frequent occurrence of data separation, the use of traditional ML estimation for small datasets is not an option. Instead, we recommend using Firth's method as it is capable of providing part-worth estimates in all aggregate data scenarios considered.

<Insert Table 4 about here>

The simulation study that evaluated the likelihood of observing separation used a completely random creation of $1,000$ datasets. For each of the cases involving data separation in Table 4, we also generated $1,000$ datasets where the ML estimates existed by creating random datasets and then discarding those exhibiting separation. So, these $1,000$ datasets are random conditional on the existence of the ML estimates. For each of these $1,000$ datasets we obtained estimates using the traditional ML method and Firth's method. This allows for direct paired comparison of both the variance and the MSE of the ML and Firth estimates. For the bias comparison, however, we did not use the $1,000$ datasets generated conditional on the existence of the ML estimates to calculate the bias of the Firth estimates. We used the completely random sample of $1,000$ datasets instead. That is because making the sampling of datasets conditional on the existence of the ML estimates causes the Firth estimates to appear biased due to the restriction in the randomization. Using the completely random sampling of datasets demonstrates both that the Firth estimates always exist (even in the cases where ML estimation fails) and that the Firth estimates are unbiased even for small numbers of respondents and studies involving few residual DF (see Section 4.2.1 for further details).

To summarize, the variance and MSE comparisons use paired estimates by both the ML and Firth's method. The bias comparison uses independent samples for the ML and Firth's method so that a paired comparison is not possible. Normally, we would prefer to always make paired comparisons because such comparisons are less variable than comparisons made using independent samples. Due to the large number of datasets and the large magnitude of the difference in performance of the ML and Firth estimates with respect to bias, the lack of paired samples has no practical impact.

4.2.1 Bias of the ML and Firth estimates

To compare the bias of the ML estimates with that of the Firth estimates, we plotted the bias against the true part-worth values, since it turns out that the bias of the ML estimates increases with the absolute size of the true parameters. Figure 2 shows the bias of the estimates from analyzing choices from all five numbers of respondents generated using the Bayesian and utility-neutral designs 1-8 with all four values of the residual DF. The plots reveal that Firth's method removes the bias completely in all situations. The advantage of Firth's method is most pronounced when the true part-worth values are large in absolute magnitude, the number of respondents equals 6, 12 or 24, and the residual DF equal 4 or 10. The bias of the ML estimates is large in all these situations. It is generally even larger for the utility-neutral designs than for the Bayesian designs. On the other hand, the bias of the ML estimates is zero for zero true part-worth values in all situations and negligible for true part-worth values that are small in absolute magnitude. Also, the bias of the ML estimates disappears gradually as the number of respondents and residual DF increases.

<Insert Figure 2 about here>

These results are similar to the results that Heinze and Schemper (2002) obtained from a simulation study comparing the standard ML method with their implementation of Firth's method to estimate a logistic regression model. The authors also found that Firth's method always yields parameter estimates under separation and that these estimates have relatively little bias, even under extreme conditions. They confirmed the safe use of Firth's method for logistic regression and its superiority over standard ML particularly in situations with large parameter values. They also noted that the bias reduction causes the Firth estimates to be slightly smaller in absolute value than ML estimates. This is the shrinkage effect typical for any penalized likelihood approach.

4.2.2 Variance of the ML and Firth estimates

To compare the variance of the ML estimates with that of the Firth estimates, we plotted the variance against the squared true part-worth values. Figure 3 shows the variance of the estimates in all situations under study. The variance of the estimates increases with the squared true part-worth values, especially for numbers of respondents equal to 6, 12 or 24, and residual DF equal to 4, 10 or 16. For these small datasets, Firth's method reduces the variance compared with traditional ML. For larger datasets, the variances of the Firth estimates are not substantially smaller than those of the ML estimates. This result is in line with the results of Firth (1993) and Heinze and Schemper (2002), who observed for the logistic regression model that the bias reduction of the estimates in small to moderate sample settings has a beneficial impact on the variance too.

<Insert Figure 3 about here>

Figure 3 visualizes the variances of the ML and Firth estimates independently, but our simulation study also allowed us to pair the variances and thus to compute the difference in variance between the ML and Firth estimates. Figure 4 shows these differences for the Bayesian and utility-neutral designs. All differences are positive meaning that Firth's method reduces the variance in all situations, but this reduction is only substantial for the smaller datasets and for the true part-worth values that are large in absolute magnitude, which are the ones that matter.

<Insert Figure 4 about here>

4.2.3 MSE of the ML and Firth estimates

To further quantify the quality of the ML estimates compared with that of the Firth estimates, we computed the MSE or the average of the squared differences between each estimate and the true value of the corresponding part-worth for all situations under study. We used the datasets randomly generated conditional on the existence of the ML estimates allowing for paired comparisons of the MSE values. In Figure 5, we plotted the differences in MSE between the ML and Firth estimates generated using the Bayesian and utility-neutral designs against the squared true part-worth values. For each situation, the difference in MSE is positive, meaning that the Firth estimates are uniformly better than the ML estimates. This is especially so for the smaller studies and for the true part-worth values that are large in absolute magnitude. The ML estimates are inadmissible then since the Firth estimates outperform the ML estimates in terms of MSE for every situation.

<Insert Figure 5 about here>

4.3 Performance of the individual-level estimates

In this section, we compare the estimation capabilities of Firth's method to traditional ML for the interesting case where we simulated choices from $R = 1$ respondent for each of the eight designs in Table 3. Using traditional ML, we observed many instances where the estimation failed due to data separation in each of the design situations. In contrast, Firth's method always provided individual-level part-worth estimates. Table 5 shows the frequency with which data separation occurred. The worst scenario involves the designs with 4 residual DF. In that scenario, almost all datasets exhibit separation. On the other hand, for the designs with 28 residual DF, the frequency of data separation is smaller, but still substantial. It equals 15% for the Bayesian design and twice as much for the utility-neutral design. For all four values of the residual DF, the Bayesian designs resulted in data separation less often than the utility-neutral designs.

<Insert Table 5 about here>

Using the separation data in Table 5, we modeled the probability of separation as a function of the type of design, either Bayesian or utility-neutral, and the residual DF. We obtained the best regression fit using the probit model which predicts the probability of separation as

$$\hat{\pi} = \Phi\left(1.849 + 0.222 \text{ Design[Utility-Neutral]} - 0.095 \text{ Residual DF}\right), \tag{17}$$

where Φ denotes the standard normal cumulative distribution function and the factor design type is coded using a $+1$ for a utility-neutral design and a -1 for a Bayesian design. Figure 6 visualizes the model showing that the probability of separation decreases with the residual DF and is smaller for the Bayesian designs than for the utility-neutral designs.

<Insert Figure 6 about here>

To conclude, the high probabilities of data separation in Table 5 and Figure 6 imply that the traditional ML method is not a viable method for the estimation of individual-level part-worths, because it fails whenever there is separation. In contrast, Firth's method overcomes the challenge of separation as it permits the estimation of individual-level part-worth estimates in all design situations under study. We therefore limit our investigation of the bias, variance and MSE of individual-level part-worth estimates to Firth estimates.

4.3.1 Bias of the Firth estimates

Figure 7 shows the bias of the Firth individual-level estimates obtained using the Bayesian and utility-neutral designs 1-8 with all four values of the residual DF. The plots present the bias against the true part-worth values as the bias of the Firth estimates increases with the absolute size of the true parameters. In contrast with the Firth estimates from aggregate data, which are unbiased (see Figure 2), the Firth individual-level estimates are still somewhat biased. The bias generally decreases with the residual DF and is smaller for the Bayesian designs than for the utility-neutral designs. However, the bias of the individual-level estimates from the Bayesian and utility-neutral designs with 28 residual DF is close to zero, such as that from for the Bayesian designs with 10 and 16 residual DF. Also, the bias of the individual-level estimates is zero for zero true part-worth values in all situations.

<Insert Figure 7 about here>

The nonzero bias of the individual-level estimates from the Bayesian and utility-neutral designs with 4 residual DF and from the utility-neutral designs with 10 and 16 residual DF is most likely due to the higher-order bias terms in Equation (7). Firth's method does not tackle this higher-order bias, as explained in Section 2.3.

4.3.2 Variance of the Firth estimates

Figure 8 shows the variance of the Firth individual-level estimates obtained using the Bayesian and utility-neutral designs 1-8 with all four values of the residual DF. As opposed to the variance plots of the Firth estimates from aggregate data, which are presented against the squared true part-worth values in Figure 3, the plots in Figure 8 present the variance against the residual DF because it turns out that the variance is independent of the squared true part-worth values here. Surprisingly, we obtained the counterintuitive result that the variance increases with the residual DF. Also, the variance is larger for the Bayesian designs than for the utility-neutral designs.

<Insert Figure 8 about here>

4.3.3 MSE of the Firth estimates

By evaluating the overall quality of the Firth individual-level estimates using the MSE, we obtained more intuitive results. Figure 9 shows the MSE of the Firth individual-level estimates obtained using the Bayesian and utility-neutral designs 1-8 with all four values of the residual DF. We plotted the MSE against the squared true part-worth values as the MSE increases with those values. In contrast with the variance of the estimates, the MSE decreases with the residual DF. Also, the MSE for the utility-neutral designs is generally larger than for the Bayesian designs, especially for cases with few residual DF and for true part-worth values that are large in absolute magnitude. As a result, we conclude that Bayesian designs should be used in combination with Firth's method for individual-level estimation.

<Insert Figure 9 about here>

5 Summary and discussion

We adapted the bias-adjustment method of Firth (1993, 1995) for estimating the MNL model using stated or revealed choice data. Compared to traditional ML estimation, Firth's method removes the first-order bias of the parameter estimates by penalizing the likelihood function with the Jeffreys prior. Through a real-life application in employee compensation and an extensive simulation study, we have shown that the method proves useful for three reasons. First, Firth's method always yields parameter estimates for the MNL model, whereas the traditional ML method fails in the case of data separation.

Second, Firth's method removes the bias of the ML estimates for studies with small to moderate numbers of choice sets evaluated by few respondents. This bias removal goes along with a reduction of the variance. Third, by applying Firth's procedure to the MNL model, it is possible to estimate individual-level parameters with relatively little bias when the number of choice sets evaluated by each respondent permits.

In conclusion, we observed that Firth's method always provides parameter estimates for the MNL model in the case of data separation, regardless of whether the data are aggregated or not, as long as the number of DF is at least as large as the number of parameters to estimate. We can therefore model the preferences of individuals directly, and construct empirical distributions of the individuals' preferences allowing us to detect preference heterogeneity. The real utility of individual preference measuring is the potential to group respondents according to similar preferences rather than just assume that they are all the same. In our application, we have shown that respondents are not behaving like a homogeneous population, which begs the question of segmentation. In subsequent research, we plan to group individuals into different market segments and determine how big the difference in market segments has to be before it is possible to detect them using differences in individual-level estimates.

Inspired by Louviere et al. (2008), we classify Firth's method for estimating individual-level MNL models as a bottom-up approach since it allows for direct estimation of individual-level parameters. This is in contrast to estimating individuals' preferences indirectly using mixed logit models, hierarchical Bayes models, latent class models and convex optimization techniques. Louviere et al. (2008) call such modeling approaches top-down approaches as they are based on prior assumptions about preference distributions across samples of people. In case these assumptions are incorrect, the inferences from top-down models will be compromised. Firth's method avoids this problem and is also computationally simpler than the current top-down models.

In our simulation study of the performance of the Firth estimates from aggregate data, we found that the advantages of Firth's method are most pronounced for studies with a small number of respondents (around 24 or smaller) and few residual DF (around 10 or fewer), and for parameters that are large in absolute magnitude, which are the ones that matter. Using utility-neutral designs rather than Bayesian designs for small studies results in more data separation and more biased ML estimates. Firth's method is therefore especially useful for poor experimental designs. For large numbers of respondents and residual DF, and for parameters that are small in absolute magnitude, the Firth estimates converge to the ML estimates. As a result, there are no reasons to avoid the use of Firth's method: either it outperforms ML estimation or it performs equally well. We provided strong evidence for this statement using a paired comparison of MSE values for Firth and ML estimates. For every situation under study, the MSE values of the ML estimates are larger than the MSE values of the Firth estimates. The ML estimates are therefore essentially inadmissible.

Acknowledgements

The research described in this paper was carried out while Roselinde Kessels was a postdoctoral fellow of the Research Foundation – Flanders. The authors are grateful to Rob Reul from Isometric Solutions for collecting the data described in Section 3 and to Bas Donkers and Dennis Fok for providing rigorous comments on a presentation of the paper at Erasmus Universiteit Rotterdam.

Appendix A. Derivation of Firth's modified score function for the MNL model

The modified score function proposed by Firth (1993) for a general statistical model is

$$\frac{\partial LL_{\text{FIRTH}}(\boldsymbol{\beta})}{\partial \beta_i} = \frac{\partial LL(\boldsymbol{\beta})}{\partial \beta_i} + \frac{1}{2}\frac{\partial \ln|\mathbf{M}(\boldsymbol{\beta})|}{\partial \beta_i}, \quad i = 1, \ldots, k,$$

which consists of two parts: 1) the traditional ML score function and 2) the first-order derivative of the logarithm of the penalty function with respect to β_i. In this appendix, we derive Firth's modified score function for the MNL model. For each of the two parts, we first provide the derivations for $S = 1$ choice set and $R = 1$ respondent and then generalize to the situation where R respondents evaluate a design involving S choice sets.

1) The traditional ML score function for $S = 1$ choice set and $R = 1$ respondent is

$$\frac{\partial LL(\boldsymbol{\beta})}{\partial \beta_i} = \frac{\partial}{\partial \beta_i}\sum_{j=1}^{J} y_{jsr} \ln\left(p_{js}(\mathbf{X}_s, \boldsymbol{\beta})\right), \tag{A1}$$

$$\begin{aligned} &= \sum_{j=1}^{J} y_{jsr} \frac{\sum_{t=1}^{J} e^{\mathbf{x}'_{ts}\boldsymbol{\beta}}}{e^{\mathbf{x}'_{js}\boldsymbol{\beta}}}\left(\frac{e^{\mathbf{x}'_{js}\boldsymbol{\beta}} x_{jsi} \sum_{t=1}^{J} e^{\mathbf{x}'_{ts}\boldsymbol{\beta}}}{\left(\sum_{t=1}^{J} e^{\mathbf{x}'_{ts}\boldsymbol{\beta}}\right)^2}\right) \\ &\quad - \sum_{j=1}^{J} y_{jsr} \frac{\sum_{t=1}^{J} e^{\mathbf{x}'_{ts}\boldsymbol{\beta}}}{e^{\mathbf{x}'_{js}\boldsymbol{\beta}}}\left(\frac{e^{\mathbf{x}'_{js}\boldsymbol{\beta}}\left(\sum_{t=1}^{J} e^{\mathbf{x}'_{ts}\boldsymbol{\beta}} x_{tsi}\right)}{\left(\sum_{t=1}^{J} e^{\mathbf{x}'_{ts}\boldsymbol{\beta}}\right)^2}\right), \end{aligned} \tag{A2}$$

$$= \sum_{j=1}^{J} y_{jsr} x_{jsi} - \sum_{j=1}^{J} y_{jsr} \frac{\sum_{t=1}^{J} e^{\mathbf{x}'_{ts}\boldsymbol{\beta}} x_{tsi}}{\sum_{t=1}^{J} e^{\mathbf{x}'_{ts}\boldsymbol{\beta}}}, \tag{A3}$$

$$= \sum_{j=1}^{J} y_{jsr} x_{jsi} - \sum_{j=1}^{J} y_{jsr} \sum_{t=1}^{J} p_{ts} x_{tsi}, \tag{A4}$$

$$= \sum_{j=1}^{J} y_{jsr} x_{jsi} - \sum_{t=1}^{J} p_{ts} x_{tsi}, \tag{A5}$$

$$= x_{msri} - \mathbf{p}'_s \mathbf{x}^*_{si}, \tag{A6}$$

where x_{jsi} is the ith entry of vector $\mathbf{x}_{js}$, x_{msri} is the ith entry of vector $\mathbf{x}_{msr}$ denoting the profile that respondent r chooses from choice set s and $\mathbf{x}^*_{si}$ is the ith column vector of choice set matrix $\mathbf{X}_s$.

Then, the traditional ML score function for S choice sets and R respondents is

$$\frac{\partial LL(\boldsymbol{\beta})}{\partial \beta_i} = \frac{\partial}{\partial \beta_i} \sum_{r=1}^{R} \sum_{s=1}^{S} \sum_{j=1}^{J} y_{jsr} \ln\left(p_{js}(\mathbf{X}_s, \boldsymbol{\beta})\right), \tag{A7}$$

$$= \sum_{r=1}^{R} \sum_{s=1}^{S} \left(x_{msri} - \mathbf{p}'_s \mathbf{x}^*_{si}\right). \tag{A8}$$

2) The first-order derivative of the logarithm of the penalty function with respect to β_i for $S = 1$ choice set and $R = 1$ respondent is

$$\frac{1}{2}\frac{\partial \ln|\mathbf{M}|}{\partial \beta_i} = \frac{1}{2|\mathbf{M}|}\frac{\partial|\mathbf{M}|}{\partial \beta_i}, \tag{A9}$$

$$= \frac{1}{2}\operatorname{tr}\left(\mathbf{M}^{-1}\frac{\partial \mathbf{M}}{\partial \beta_i}\right), \tag{A10}$$

$$= \frac{1}{2}\operatorname{tr}\left(\mathbf{M}^{-1}\frac{\partial\left(\mathbf{X}'_s\left(\mathbf{P}_s - \mathbf{p}_s\mathbf{p}'_s\right)\mathbf{X}_s\right)}{\partial \beta_i}\right), \tag{A11}$$

$$= \frac{1}{2}\operatorname{tr}\left(\mathbf{M}^{-1}\mathbf{X}'_s\frac{\partial\left(\mathbf{P}_s - \mathbf{p}_s\mathbf{p}'_s\right)}{\partial \beta_i}\mathbf{X}_s\right), \tag{A12}$$

$$= \frac{1}{2}\operatorname{tr}\left(\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}'_s\right)\frac{\partial\left(\mathbf{P}_s - \mathbf{p}_s\mathbf{p}'_s\right)}{\partial \beta_i}\right), \tag{A13}$$

$$= \frac{1}{2}\operatorname{tr}\left(\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}'_s\right)\frac{\partial \mathbf{P}_s}{\partial \beta_i}\right) - \frac{1}{2}\operatorname{tr}\left(\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}'_s\right)\frac{\partial\left(\mathbf{p}_s\mathbf{p}'_s\right)}{\partial \beta_i}\right). \tag{A14}$$

Here, we obtain Equation (A10) using Jacobi's formula for the invertible matrix $\mathbf{M}$

$$\frac{\partial|\mathbf{M}|}{\partial \beta_i} = |\mathbf{M}|\operatorname{tr}\left(\mathbf{M}^{-1}\frac{\partial \mathbf{M}}{\partial \beta_i}\right).$$

Also, Equation (A13) is made possible due to the cyclic property of the trace

$$\mathbf{M}^{-1}\mathbf{X}'_s\frac{\partial\left(\mathbf{P}_s - \mathbf{p}_s\mathbf{p}'_s\right)}{\partial \beta_i}\mathbf{X}_s = \mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}'_s\frac{\partial\left(\mathbf{P}_s - \mathbf{p}_s\mathbf{p}'_s\right)}{\partial \beta_i},$$

which holds because the matrix product on the left-hand side of the identity yields a square matrix and the matrix product on the right-hand side exists.

The result in Equation (A14) consists of two terms, the second of which can be rewritten using the fact that the trace of the product of the two $J \times J$ matrices, $\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}'_s$ and

$\frac{\partial(\mathbf{p}_s\mathbf{p}_s')}{\partial\beta_i}$, equals the sum of the entry-wise products of their elements:

$$\begin{aligned}
\frac{1}{2}\,\mathrm{tr}\left(\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}_s'\right)\frac{\partial\left(\mathbf{p}_s\mathbf{p}_s'\right)}{\partial\beta_i}\right) &= \frac{1}{2}\sum_{u=1}^{J}\sum_{v=1}^{J}\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}_s'\right)_{uv}\frac{\partial\left(p_{us}p_{vs}\right)}{\partial\beta_i}, &\text{(A15)}\\
&= \frac{1}{2}\sum_{u=1}^{J}\sum_{v=1}^{J}\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}_s'\right)_{uv}2\frac{\partial p_{us}}{\partial\beta_i}p_{vs}, &\text{(A16)}\\
&= \mathrm{tr}\left(\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}_s'\right)\frac{\partial\mathbf{p}_s}{\partial\beta_i}\mathbf{p}_s'\right), &\text{(A17)}\\
&= \mathrm{tr}\left(\mathbf{p}_s'\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}_s'\right)\frac{\partial\mathbf{p}_s}{\partial\beta_i}\right), &\text{(A18)}\\
&= \mathbf{p}_s'\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}_s'\right)\frac{\partial\mathbf{p}_s}{\partial\beta_i}. &\text{(A19)}
\end{aligned}$$

Equations (A16) and (A17) are made possible because the terms

$$\frac{\partial\left(p_{us}p_{vs}\right)}{\partial\beta_i} = \frac{\partial p_{us}}{\partial\beta_i}\,p_{vs} + p_{us}\,\frac{\partial p_{vs}}{\partial\beta_i}$$

can be grouped in a matrix such that $2p_{vs}\frac{\partial p_{us}}{\partial\beta_i}$ is on the uth row and $2p_{us}\frac{\partial p_{vs}}{\partial\beta_i}$ is on the vth row.

To obtain Equation (A18), we have used again the cyclic property of the trace. This results in the trace of a scalar, which is a scalar itself, as shown in Equation (A19).

Combining Equations (A14) and (A19) yields the following expression for the first-order derivative of the logarithm of the penalty function with respect to β_i for $S = 1$ choice set and $R = 1$ respondent:

$$\frac{1}{2}\frac{\partial\ln|\mathbf{M}|}{\partial\beta_i} = \frac{1}{2}\,\mathrm{tr}\left(\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}_s'\right)\frac{\partial\mathbf{P}_s}{\partial\beta_i}\right) - \mathbf{p}_s'\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}_s'\right)\frac{\partial\mathbf{p}_s}{\partial\beta_i}. \quad \text{(A20)}$$

For the situation where R respondents evaluate S choice sets, Equation (A20) becomes

$$\frac{1}{2}\frac{\partial\ln|\mathbf{M}|}{\partial\beta_i} = R\sum_{s=1}^{S}\left[\frac{1}{2}\,\mathrm{tr}\left(\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}_s'\right)\frac{\partial\mathbf{P}_s}{\partial\beta_i}\right) - \mathbf{p}_s'\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}_s'\right)\frac{\partial\mathbf{p}_s}{\partial\beta_i}\right]. \quad \text{(A21)}$$

Finally, by adding up Equations (A8) and (A21), we obtain Firth's modified score function for the MNL model using choices from R respondents for S choice sets:

$$\begin{aligned}
\frac{\partial LL_{\text{FIRTH}}(\boldsymbol{\beta})}{\partial\beta_i} = &\sum_{r=1}^{R}\sum_{s=1}^{S}\left(x_{msri} - \mathbf{p}_s'\mathbf{x}_{si}^{*}\right)\\
&+ R\sum_{s=1}^{S}\left[\frac{1}{2}\,\mathrm{tr}\left(\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}_s'\right)\frac{\partial\mathbf{P}_s}{\partial\beta_i}\right) - \mathbf{p}_s'\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}_s'\right)\frac{\partial\mathbf{p}_s}{\partial\beta_i}\right].
\end{aligned} \quad \text{(A22)}$$

Appendix B. Motivation of the LR test with Firth estimates

We show that the LR test statistic using the traditional log-likelihood function of the Firth estimates

$$-2\left[LL\left(\hat{\boldsymbol{\beta}}^{R}_{\text{FIRTH}}\right) - LL\left(\hat{\boldsymbol{\beta}}^{U}_{\text{FIRTH}}\right)\right],$$

is asymptotically equivalent to the LR test statistic using the traditional log-likelihood function of the ML estimates

$$-2\left[LL\left(\hat{\boldsymbol{\beta}}^{R}_{\text{ML}}\right) - LL\left(\hat{\boldsymbol{\beta}}^{U}_{\text{ML}}\right)\right],$$

if the ML estimates $\hat{\boldsymbol{\beta}}^{R}_{\text{ML}}$ and $\hat{\boldsymbol{\beta}}^{U}_{\text{ML}}$ exist.

We begin by writing the traditional log-likelihood function of the Firth estimates in terms of the first three terms of its Taylor series expansion around the ML estimates:

$$\begin{aligned} LL\left(\hat{\boldsymbol{\beta}}_{\text{FIRTH}}\right) \approx{}& LL\left(\hat{\boldsymbol{\beta}}_{\text{ML}}\right) + \left(\hat{\boldsymbol{\beta}}_{\text{FIRTH}} - \hat{\boldsymbol{\beta}}_{\text{ML}}\right) \frac{\partial LL\left(\hat{\boldsymbol{\beta}}_{\text{ML}}\right)}{\partial \hat{\boldsymbol{\beta}}_{\text{ML}}} \\ &+ \frac{1}{2}\left(\hat{\boldsymbol{\beta}}_{\text{FIRTH}} - \hat{\boldsymbol{\beta}}_{\text{ML}}\right)' \frac{\partial^2 LL\left(\hat{\boldsymbol{\beta}}_{\text{ML}}\right)}{\partial \hat{\boldsymbol{\beta}}_{\text{ML}} \partial \hat{\boldsymbol{\beta}}'_{\text{ML}}} \left(\hat{\boldsymbol{\beta}}_{\text{FIRTH}} - \hat{\boldsymbol{\beta}}_{\text{ML}}\right). \end{aligned} \tag{B1}$$

We can simplify Equation (B1) as follows. First, Firth's bias-adjustment method removes the first-order bias of the ML estimates so that we have

$$\hat{\boldsymbol{\beta}}_{\text{FIRTH}} - \hat{\boldsymbol{\beta}}_{\text{ML}} \approx O\left(N^{-1}\right). \tag{B2}$$

Second, given the ML estimates, it holds that

$$\frac{\partial LL\left(\hat{\boldsymbol{\beta}}_{\text{ML}}\right)}{\partial \hat{\boldsymbol{\beta}}_{\text{ML}}} = \mathbf{0}_k, \tag{B3}$$

and that

$$\frac{\partial^2 LL\left(\hat{\boldsymbol{\beta}}_{\text{ML}}\right)}{\partial \hat{\boldsymbol{\beta}}_{\text{ML}} \partial \hat{\boldsymbol{\beta}}'_{\text{ML}}} \approx O\left(N\right). \tag{B4}$$

As a result, Equation (B1) becomes

$$LL\left(\hat{\boldsymbol{\beta}}_{\text{FIRTH}}\right) \approx LL\left(\hat{\boldsymbol{\beta}}_{\text{ML}}\right) + O\left(N^{-1}\right). \tag{B5}$$

Using Equation (B5), we can write the LR test statistic using the traditional log-likelihood function of the Firth estimates as

$$-2\left[LL\left(\hat{\boldsymbol{\beta}}^{R}_{\text{FIRTH}}\right) - LL\left(\hat{\boldsymbol{\beta}}^{U}_{\text{FIRTH}}\right)\right] \approx -2\left[LL\left(\hat{\boldsymbol{\beta}}^{R}_{\text{ML}}\right) - LL\left(\hat{\boldsymbol{\beta}}^{U}_{\text{ML}}\right)\right] + O\left(N^{-1}\right). \quad \text{(B6)}$$

This equation shows that the LR test statistic using the traditional log-likelihood function of the Firth estimates is asymptotically equivalent to the LR test statistic using the traditional log-likelihood function of the ML estimates if $\hat{\boldsymbol{\beta}}^{R}_{\text{ML}}$ and $\hat{\boldsymbol{\beta}}^{U}_{\text{ML}}$ exist.

Appendix C. Designs used in the simulation study

<Insert Tables C1 to C4 about here>

References

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models, *Biometrika* **71**: 1–10.

Allison, P. D. (2008). Convergence failures in logistic regression, SAS Global Forum, Technical Paper 360-2008, http://www2.sas.com/proceedings/forum2008/360-2008.pdf [visited on 11 May 2013].

Andrews, R. L., Ainslie, A. and Currim, I. S. (2002). An empirical comparison of logit choice models with discrete versus continuous representations of heterogeneity, *Journal of Marketing Research* **39**: 479–487.

Beggs, S., Cardell, N. S. and Hausman, J. (1981). Assessing the potential demand for electric cars, *Journal of Econometrics* **17**: 1–19.

Bull, S. B., Mak, C. and Greenwood, C. M. T. (2002). A modified score function estimator for multinomial logistic regression in small samples, *Computational Statistics and Data Analysis* **39**: 57–74.

Cardell, N. S. (1993). A modified maximum likelihood estimator for discrete choice models, *Journal of the American Statistical Association: Proceedings of the Statistical Computing Section*, 118–123.

Chapman, R. G. (1984). An approach to estimating logit models of a single decision maker's choice behavior, *Advances in Consumer Research* **11**: 656–661.

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B. and Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression, *Journal of the American Statistical Association* **86**: 68–78.

Evgeniou, T., Pontil, M. and Toubia, O. (2007). A convex optimization approach to modeling consumer heterogeneity in conjoint estimation, *Marketing Science* **26**: 805–818.

Firth, D. (1993). Bias reduction of maximum likelihood estimates, *Biometrika* **80**: 27–38.

Firth, D. (1995). Amendments and corrections: Bias reduction of maximum likelihood estimates, *Biometrika* **82**: 667.

Goos, P. and Gilmour, S. G. (2012). A general strategy for analyzing data from split-plot and multistratum experimental designs, *Technometrics* **54**: 340–354.

Heinze, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data, *Statistics in Medicine* **25**: 4216–4226.

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression, *Statistics in Medicine* **21**: 2409–2419.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems, *Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences* **186**: 453–461.

Kessels, R., Jones, B., Goos, P. and Vandebroek, M. (2008). Recommendations on the use of Bayesian optimal designs for choice experiments, *Quality and Reliability Engineering International* **24**: 737–744.

Kessels, R., Jones, B., Goos, P. and Vandebroek, M. (2009). An efficient algorithm for constructing Bayesian optimal choice designs, *Journal of Business and Economic Statistics* **27**: 279–291.

Kessels, R., Jones, B. and Goos, P. (2011a). Bayesian optimal designs for discrete choice experiments with partial profiles, *Journal of Choice Modelling* **4**: 52–74.

Kessels, R., Jones, B., Goos, P. and Vandebroek, M. (2011b). The usefulness of Bayesian optimal designs for discrete choice experiments, *Applied Stochastic Models in Business and Industry* **27**: 173–188.

Kessels, R., Jones, B., Goos, P. and Vandebroek, M. (2011c). Rejoinder: The usefulness of Bayesian optimal designs for discrete choice experiments, *Applied Stochastic Models in Business and Industry* **27**: 197–203.

King, E. N. and Ryan, T. P. (2002). A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression, *The American Statistician* **56**: 163–170.

Lenk, P. J., DeSarbo, W. S., Green, P. E. and Young, M. R. (1996). Hierarchical Bayes conjoint analysis: recovery of partworth heterogeneity from reduced experimental designs, *Marketing Science* **15**: 173–191.

Lenk, P. and Orme, B. (2009). The value of informative priors in Bayesian inference with sparse data, *Journal of Marketing Research* **46**: 832–845.

Lesaffre, E. and Albert, A. (1989). Partial separation in logistic discrimination, *Journal of the Royal Statistical Society Series B* **51**: 109–116.

Louviere, J. J., Street, D., Burgess, L., Wasi, N., Islam, T. and Marley, A. A. J. (2008). Modeling the choices of individual decision-makers by combining efficient choice experiment designs with extra preference information, *Journal of Choice Modelling* **1**: 128–163.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior, in Zarembka, P., *Frontiers in Econometrics*, New York: Academic Press, 105–142.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series, *Journal of the Royal Statistical Society Series B* **11**: 68–84.

Quenouille, M. H. (1956). Notes on bias in estimation, *Biometrika* **43**: 353–360.

Santner, T. J. and Duffy, D. E. (1986). A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models, *Biometrika* **73**: 755–758.

Train, K. (2003). *Discrete Choice Methods with Simulation*, Cambridge, U.K.: Cambridge University Press.

Woods, D. C. and van de Ven, P. (2011). Blocked designs for experiments with correlated non-normal response, *Technometrics* **53**: 173–182.

Tables

Table 1: Attributes and attribute levels used in the compensation study.

Attribute	Level 1	Level 2	Level 3
Salary Increase	No Raise	Small Raise (3%)	Large Raise (6%)
Bonus	No Bonus	Small Bonus (5%)	Large Bonus (10%)
Extra Vacation	No Extra Week	1 Extra Week	2 Extra Weeks
Flexible Time	No Flexibility	2 Days / Week	4 Days / Week

Table 2: Estimates obtained from the aggregate choice data of the compensation study using either traditional maximum likelihood or Firth's method, and summary statistics of the Firth individual-level estimates.

Attribute Level	Aggregate	Individual-Level		
		Mean	Std Dev	Std Err
No Raise	-0.920	-0.841	0.691	0.033
Small Raise	0.186	0.181	0.489	0.023
Large Raise	0.734	0.660	0.672	0.032
No Bonus	-1.005	-0.951	0.710	0.034
Small Bonus	0.200	0.219	0.448	0.021
Large Bonus	0.805	0.732	0.636	0.030
No Extra Vacation	-0.460	-0.471	0.621	0.029
1 Extra Week	0.114	0.148	0.416	0.020
2 Extra Weeks	0.346	0.323	0.567	0.027
No Flex	-0.264	-0.269	0.656	0.031
2 Days Flex	0.096	0.090	0.495	0.023
4 Days Flex	0.168	0.179	0.514	0.024

Table 3: Eight designs used in the simulation study.

Design	Type	Number of Choice Sets	Number of Profiles per Choice Set	Residual Degrees of Freedom (DF)
1	Bayesian	12	2	4
2	Utility-Neutral	12	2	4
3	Bayesian	18	2	10
4	Utility-Neutral	18	2	10
5	Bayesian	12	3	16
6	Utility-Neutral	12	3	16
7	Bayesian	18	3	28
8	Utility-Neutral	18	3	28

Table 4: Occurrence of separation when analyzing aggregate choice data.

Design	Type	Residual Degrees of Freedom (DF)	Number of Respondents	Cases of Separation (%)
1	Bayesian	4	6	1.7
2	Utility-Neutral	4	6	42.4
2	Utility-Neutral	4	12	6.4
2	Utility-Neutral	4	24	0.2
4	Utility-Neutral	10	6	0.3

Table 5: Occurrence of separation when analyzing individual-level choice data.

Design	Type	Residual Degrees of Freedom (DF)	Cases of Separation (%)
1	Bayesian	4	95.6
2	Utility-Neutral	4	99.7
3	Bayesian	10	64.0
4	Utility-Neutral	10	79.6
5	Bayesian	16	58.2
6	Utility-Neutral	16	70.2
7	Bayesian	28	15.0
8	Utility-Neutral	28	30.7

Table C1: Bayesian Design 1 and utility-neutral Design 2 used in the simulation study as described in Table 3.

	Design 1				Design 2			
Choice	Attributes				Attributes			
Set	1	2	3	4	1	2	3	4
1	2	1	2	3	2	1	3	2
1	3	2	1	1	1	3	1	3
2	2	3	2	1	3	2	2	3
2	1	2	3	3	1	3	3	1
3	3	2	2	3	2	3	1	2
3	1	3	3	2	3	1	2	3
4	1	3	3	1	2	1	1	3
4	2	2	1	2	3	3	2	2
5	2	3	1	2	2	2	2	1
5	3	1	2	1	3	3	1	3
6	3	3	1	3	2	3	2	3
6	2	1	3	2	3	2	1	1
7	2	2	2	1	2	2	1	1
7	3	1	1	2	1	1	3	2
8	2	1	3	3	1	1	1	2
8	1	2	2	2	2	3	3	3
9	3	1	3	2	1	1	2	1
9	1	3	1	3	3	2	3	2
10	3	3	2	2	1	2	3	1
10	1	2	3	3	2	1	1	2
11	1	1	2	3	1	2	2	2
11	2	2	3	1	3	1	3	1
12	1	2	2	2	1	2	3	3
12	2	1	1	1	3	3	2	1

Table C2: Bayesian Design 3 and utility-neutral Design 4 used in the simulation study as described in Table 3.

	Design 3				Design 4			
Choice	Attributes				Attributes			
Set	1	2	3	4	1	2	3	4
1	1	2	3	3	3	1	2	3
1	3	1	2	1	1	2	3	1
2	1	1	3	3	3	2	1	3
2	3	2	1	2	2	1	3	2
3	3	1	2	1	2	1	1	1
3	1	2	1	2	3	3	3	3
4	1	2	2	3	3	3	2	2
4	2	3	3	1	1	1	1	3
5	2	2	2	2	1	1	3	2
5	1	1	3	1	3	2	2	3
6	1	3	2	3	3	2	3	1
6	3	1	3	2	1	3	1	3
7	1	1	2	3	1	3	3	3
7	2	2	1	1	2	2	1	2
8	3	3	2	2	1	2	1	2
8	2	2	3	1	2	1	2	1
9	2	3	1	3	1	1	2	3
9	3	2	2	1	2	3	1	1
10	3	2	1	3	1	1	1	1
10	2	3	2	1	2	2	3	2
11	2	2	2	3	1	3	3	1
11	1	3	3	2	3	1	1	2
12	2	2	3	1	1	2	2	1
12	3	1	1	2	2	3	3	3
13	1	2	3	2	2	2	1	3
13	2	3	1	3	3	1	2	1
14	2	1	2	2	3	3	1	1
14	3	3	1	1	1	2	2	2
15	2	3	2	2	3	3	1	2
15	3	1	3	3	2	2	2	3
16	1	3	1	1	3	2	3	1
16	2	1	3	2	2	3	2	2
17	3	2	2	1	3	1	3	2
17	2	1	1	2	2	3	2	1
18	2	1	1	3	2	1	3	3
18	1	3	3	2	1	3	2	2

Table C3: Bayesian Design 5 and utility-neutral Design 6 used in the simulation study as described in Table 3.

	Design 5				Design 6			
Choice	Attributes				Attributes			
Set	1	2	3	4	1	2	3	4
1	2	1	2	2	2	3	2	2
1	1	2	3	3	3	2	3	3
1	3	3	1	1	1	1	1	1
2	1	2	2	2	1	3	3	3
2	3	3	3	1	2	1	1	2
2	2	1	1	3	3	2	2	1
3	1	3	1	2	3	3	1	1
3	2	1	3	3	2	2	3	2
3	3	2	2	1	1	1	2	3
4	2	3	1	3	1	2	1	2
4	3	1	2	1	2	1	2	3
4	1	2	3	2	3	3	3	1
5	1	1	2	3	3	1	3	2
5	3	3	1	2	1	3	1	3
5	2	2	3	1	2	2	2	1
6	1	1	3	3	2	3	3	1
6	2	3	1	1	3	1	2	3
6	3	2	2	2	1	2	1	2
7	1	3	2	3	2	1	1	1
7	3	1	1	2	3	2	2	3
7	2	2	3	1	1	3	3	2
8	3	1	3	2	2	2	3	1
8	1	2	1	3	3	3	1	3
8	2	3	2	1	1	1	2	2
9	3	1	3	1	3	2	1	2
9	1	2	1	3	2	1	3	3
9	2	3	2	2	1	3	2	1
10	3	1	2	3	1	2	3	3
10	1	3	3	1	3	1	1	1
10	2	2	1	2	2	3	2	2
11	2	1	3	2	1	2	2	1
11	1	3	2	1	2	3	1	3
11	3	2	1	3	3	1	3	2
12	2	2	2	3	3	3	2	2
12	1	3	3	2	1	1	3	1
12	3	1	1	1	2	2	1	3

Table C4: Bayesian Design 7 and utility-neutral Design 8 used in the simulation study as described in Table 3.

	Design 7				Design 8			
Choice	Attributes				Attributes			
Set	1	2	3	4	1	2	3	4
1	2	1	2	3	2	3	2	3
1	3	2	3	1	1	2	1	1
1	1	3	1	2	3	1	3	2
2	2	2	3	2	1	2	2	1
2	1	3	2	1	3	3	3	3
2	3	1	1	3	2	1	1	2
3	2	2	1	3	2	1	1	1
3	3	1	2	1	1	3	2	2
3	1	3	3	2	3	2	3	3
4	3	2	1	2	2	3	2	1
4	1	1	2	3	3	2	3	2
4	2	3	3	1	1	1	1	3
5	1	3	2	2	1	1	3	3
5	3	2	1	1	2	3	1	1
5	2	1	3	3	3	2	2	2
6	2	2	2	3	1	1	1	2
6	3	1	1	2	3	2	2	1
6	1	3	3	1	2	3	3	3
7	1	1	3	3	2	2	2	3
7	2	2	1	2	1	3	1	2
7	3	3	2	1	3	1	3	1
8	3	3	1	2	3	2	1	1
8	2	2	2	1	2	1	3	2
8	1	1	2	3	1	3	2	3
9	1	2	1	3	2	2	1	3
9	2	3	2	2	1	1	2	2
9	3	1	3	1	3	3	3	1

Cont'd	Design 7				Design 8			
Choice	Attributes				Attributes			
Set	1	2	3	4	1	2	3	4
10	1	2	2	3	3	1	2	3
10	3	3	1	1	2	3	3	1
10	2	1	3	2	1	2	1	2
11	1	3	2	3	2	2	1	3
11	3	1	1	2	3	3	2	2
11	2	2	3	1	1	1	3	1
12	1	2	2	2	3	2	1	2
12	2	3	1	1	2	1	2	3
12	3	1	3	3	1	3	3	1
13	3	1	2	3	1	3	1	3
13	2	3	1	2	3	1	2	1
13	1	2	3	1	2	2	3	2
14	3	2	1	3	2	3	2	2
14	2	3	2	1	1	2	3	1
14	1	1	3	2	3	1	1	3
15	2	1	2	2	2	2	3	2
15	1	2	3	3	3	3	1	3
15	3	3	1	1	1	1	2	1
16	1	3	3	2	2	1	3	1
16	2	1	1	3	3	3	1	2
16	3	2	2	1	1	2	2	3
17	1	2	3	1	3	1	1	3
17	2	3	1	3	1	3	3	2
17	3	1	2	2	2	2	2	1
18	3	2	2	2	3	3	1	1
18	2	1	3	1	2	1	2	2
18	1	3	1	3	1	2	3	3

Figures

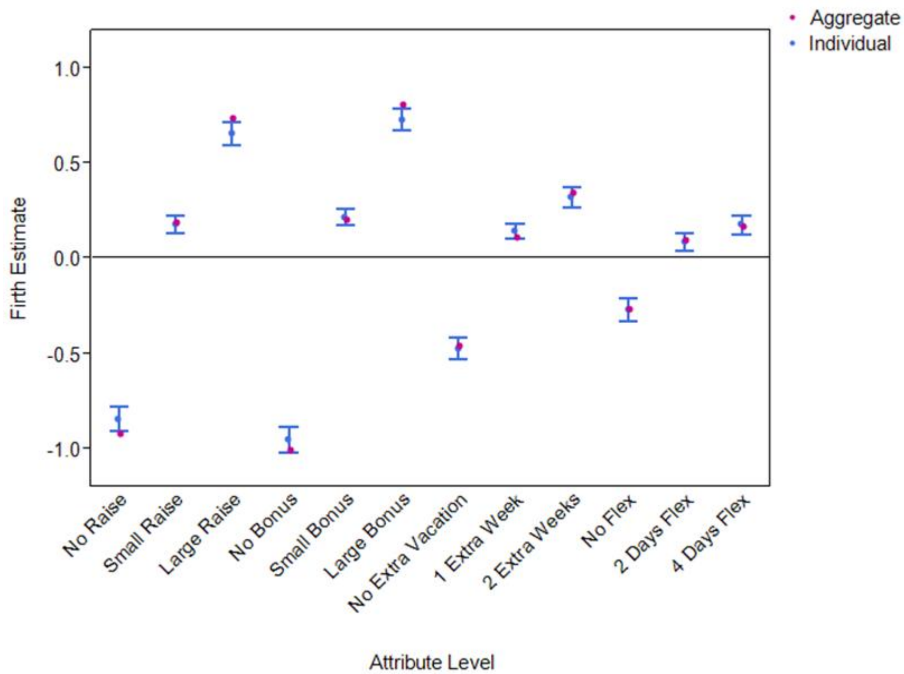

Figure 1: Estimates obtained from the aggregate choice data of the compensation study using either traditional maximum likelihood or Firth's method and 95% confidence intervals of the means of the Firth individual-level estimates.

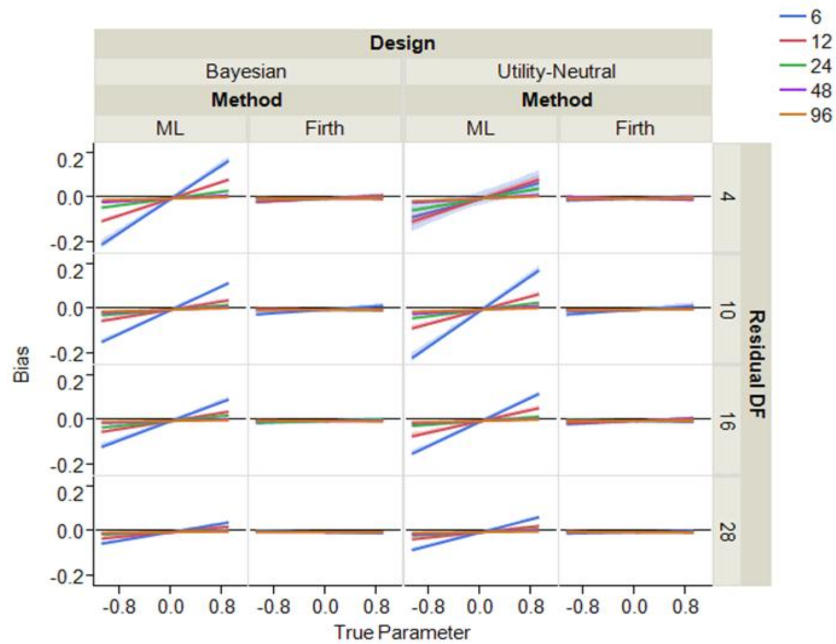

Figure 2: Bias of the traditional maximum likelihood (ML) and Firth estimates obtained from the Bayesian and utility-neutral designs 1-8 in Table 3 and aggregate choice data from 6, 12, 24, 48 and 96 respondents.

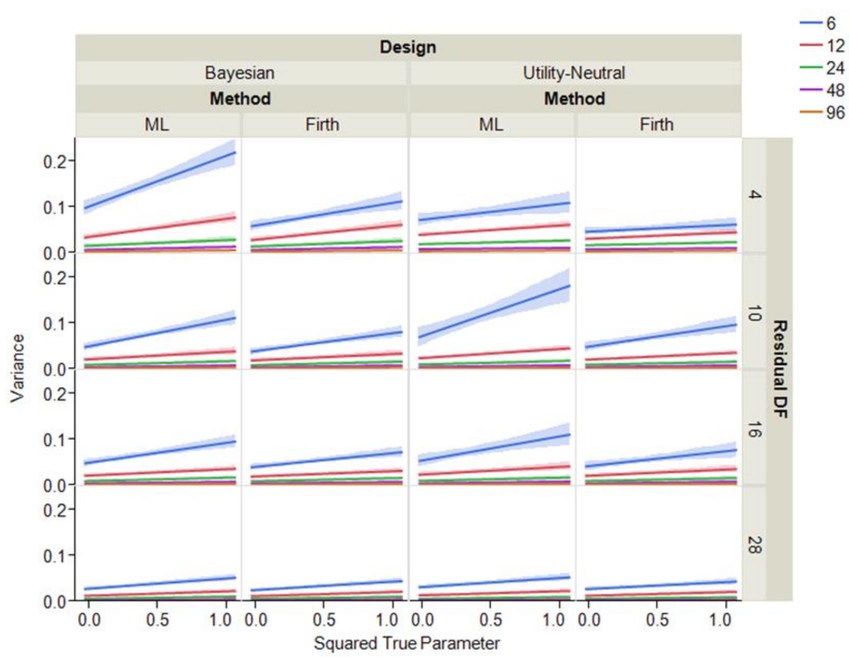

Figure 3: Variance of the traditional maximum likelihood (ML) and Firth estimates obtained from the Bayesian and utility-neutral designs 1-8 in Table 3 and aggregate choice data from 6, 12, 24, 48 and 96 respondents.

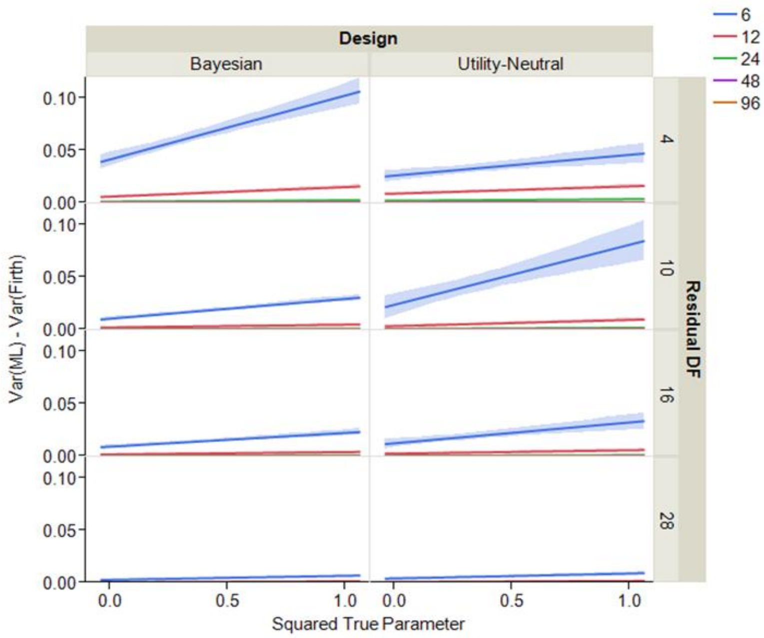

Figure 4: Difference in variance between the traditional maximum likelihood (ML) and Firth estimates obtained from the Bayesian and utility-neutral designs 1-8 in Table 3 and aggregate choice data from 6, 12, 24, 48 and 96 respondents.

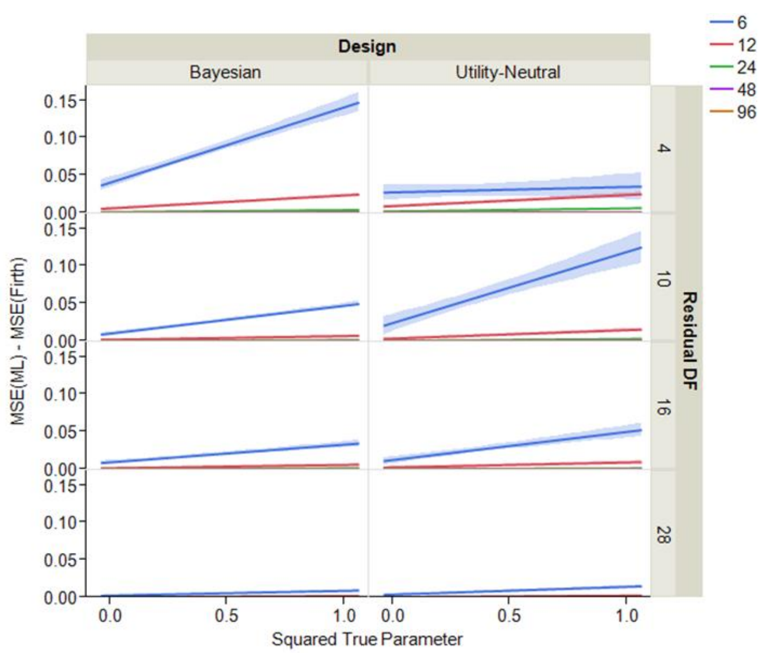

Figure 5: Difference in mean squared error (MSE) between the traditional maximum likelihood (ML) and Firth estimates obtained from the Bayesian and utility-neutral designs 1-8 in Table 3 and aggregate choice data from 6, 12, 24, 48 and 96 respondents.

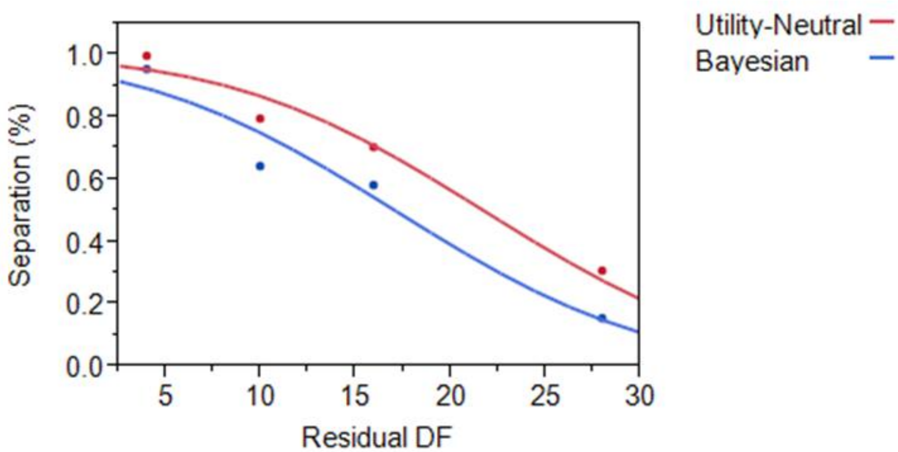

Figure 6: Predicted probability of separation when analyzing individual-level choice data as a function of the type of design, Bayesian or utility-neutral, and the residual degrees of freedom (DF).

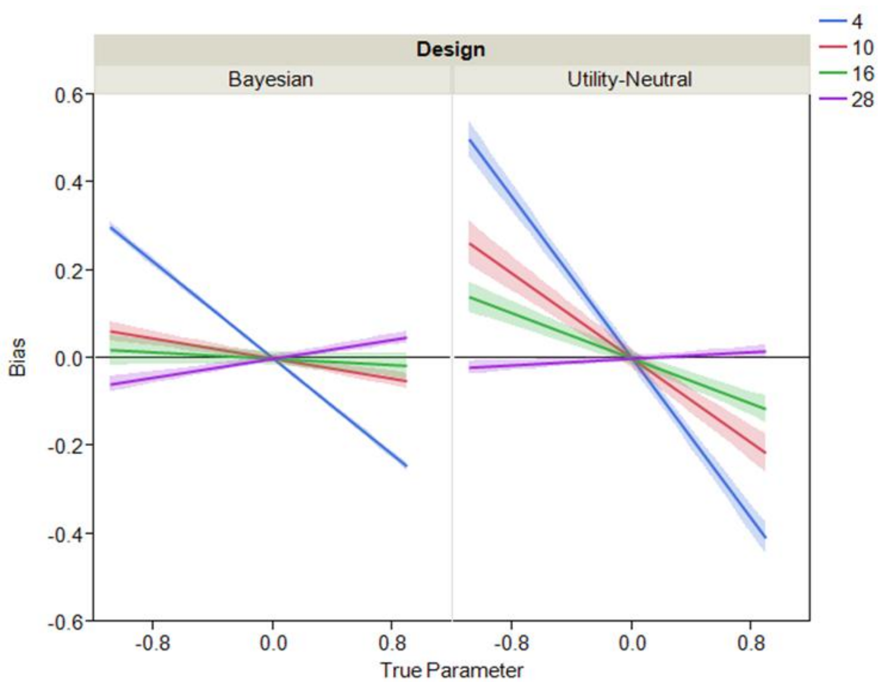

Figure 7: Bias of the Firth individual-level estimates obtained from the Bayesian and utility-neutral designs 1-8 in Table 3 with 4, 10, 16 and 28 residual degrees of freedom.

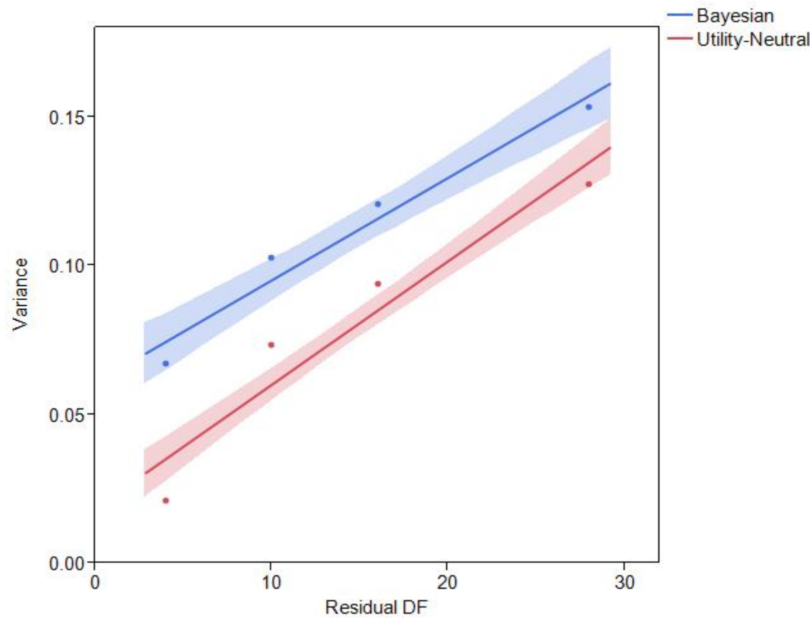

Figure 8: Variance of the Firth individual-level estimates obtained from the Bayesian and utility-neutral designs 1-8 in Table 3 with 4, 10, 16 and 28 residual degrees of freedom (DF).

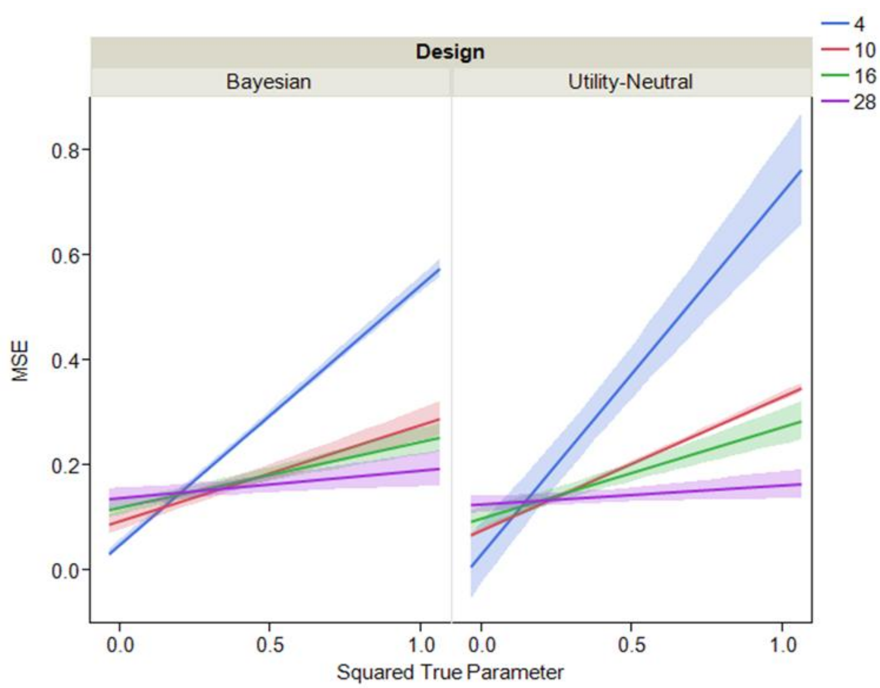

Figure 9: Mean squared error (MSE) of the Firth individual-level estimates obtained from the Bayesian and utility-neutral designs 1-8 in Table 3 with 4, 10, 16 and 28 residual degrees of freedom.