Using Firth's method for model estimation and market segmentation based on choice data

Roselinde Kessels*†

University of Antwerp

University of Amsterdam

Bradley Jones‡

SAS Institute

Peter Goos§

University of Antwerp

KU Leuven

*Roselinde Kessels is a research fellow of the Flemish Research Foundation (FWO) at the University of Antwerp, Faculty of Business and Economics & StatUa Center for Statistics, Prinsstraat 13, 2000 Antwerp, Belgium, and a lecturer in econometrics at the University of Amsterdam, School of Economics, PO Box 15867, 1001 NJ Amsterdam, The Netherlands. Tel.: +32 (0)486 28 98 27. E-mail: roselinde.kessels@uantwerp.be

†Corresponding author.

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

§Peter Goos is a full professor in statistics at the University of Antwerp, Faculty of Business and Economics & StatUa Center for Statistics, Prinsstraat 13, 2000 Antwerp, Belgium, and at the KU Leuven, Faculty of Bioscience Engineering & Leuven Statistics Research Centre, Kasteelpark Arenberg 30, 3001 Heverlee, Belgium. Tel.: +32 (0)3 220 40 59. E-mail: peter.goos@uantwerp.be

Using Firth's method for model estimation and market segmentation based on choice data

Abstract

Using maximum likelihood (ML) estimation for discrete choice modeling of small datasets causes two problems. The first problem is that the data may exhibit separation, in which case the ML estimates do not exist. Also, provided they exist, the ML estimates are biased. In this paper, we show how to adapt Firth's penalized likelihood estimation for use in discrete choice modeling. A powerful advantage of Firth's estimation is that, unlike ML estimation, it provides useful estimates in the case of data separation. For aggregates of six or more respondents, Firth estimates have negligible bias. For preference estimates on an individual level, Firth estimates show little bias as long as each person evaluates a sufficient number of choice sets. Additionally, Firth's individual-level estimation makes it possible to construct an empirical distribution of the respondents' preferences without imposing any *a priori* population distribution and to effectively predict people's choices and detect market segments. Segment recovery may even be better when individual-level estimates are obtained using Firth's method instead of hierarchical Bayes estimation under a normal prior. We base all findings on data from a stated choice study on various forms of employee compensation.

Keywords: discrete choice modeling, data separation, Firth's penalized maximum likelihood, hierarchical Bayes estimation, individual-level estimates, market segmentation

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 stated or discrete choice experiments (DCEs), where respondents state their choices in hypothetical situations, or via observational studies, where respondents reveal their actual choices made. 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 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.

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)). 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.

A commonly used procedure to fit discrete choice models is maximum likelihood (ML) estimation, which guarantees that the estimator is unbiased with an infinite sample size. However, for many applications, the sample data collected are small. One consequence of finite samples is that the probability of separation is always strictly positive. In the event of separation, the ML estimator does not exist. Therefore, for finite samples and logistic models such as discrete choice models, the expectation of the ML estimator does not exist (Le Cam, 1990). That is, the integral defining their expected value does not converge. This is because the probability of data separation is never nonzero.

In practical applications, data may or may not exhibit separation. When data separation occurs, computer implementations of ML estimation often show the likelihood estimates converging 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. When attempting ML estimation for individuals, data separation occurs so frequently as to make such an approach infeasible. For small numbers of respondents, a lesser problem with ML estimation is that it tends to overestimate the utility of strongly preferred attribute levels. Similarly, undesirable attribute levels are modeled as being even less desirable than their true utilities would indicate. This bias is often far from negligible and can have practical implications in the decisions that practitioners make.

To overcome these two weaknesses of ML estimation, we introduce the penalized ML method of Firth (1993, 1995) for estimating the multinomial logit (MNL) model in the literature on the analysis of choice data. As we show in this paper, a major advant-

age of the method is that it allows fitting a MNL model to individual-level data, and subsequently, exploring the heterogeneity in the respondents' preferences and segmenting the market. Bull et al. (2002) were the first to propose Firth's method to estimate the MNL model, but they applied it to small sample clinical trials outside a choice modeling context, and did not consider individual-level data.

Firth's method was originally developed as a general bias reducing technique of ML estimation, but it was also shown to provide finite parameter estimates in the case of separation (see, for binomial and trinomial logistic regression on clinical data, Bull et al. (2002), Heinze and Schemper (2002) and Heinze (2006)). Separation is to occur more often in small samples and in larger experiments where a design is used in which the success probability of every observation is near 0 or 1 (Woods and van de Ven, 2011). In a DCE context, Kessels et al. (2011a,b) describe an example of an orthogonal design involving eight choice sets of two alternatives. Such design is also called a utility-neutral 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 the example 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 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 (Sándor and Wedel, 2001; Bliemer and Rose, 2010; Kessels et al., 2011a). 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.

Recently, to overcome the data separation challenge and estimate choice models for each individual separately, Frischknecht et al. (2014) presented a penalized ML method where the penalty function corresponds to a Bayesian approach that augments the limited data with prior beliefs about the behavior of the data (Geweke, 2005). The authors built on the tradition of penalty methods proposed by Clogg et al. (1991) and Cardell (1993). Typical of any penalized ML approach, including Firth's method, is that it shrinks the estimates toward zero. Another approach proposed by Dumont et al. (2015) to estimate individual-level choice models uses a Bayesian framework that draws priors from a sample-level model in which the data are pooled. Both approaches by Frischknecht et al. (2014) and Dumont et al. (2015) require weaker prior distributional assumptions compared to hierarchical Bayes (HB) estimation, thereby attempting to simplify the computations.

Nevertheless, a large body of literature focuses on obtaining individual-level preference estimates from sample-level models such as mixed logit models (Train, 2009), HB models (Lenk et al., 1996), latent class models (Andrews et al., 2002) and convex optimization techniques (Evgeniou et al., 2007). These individual-level estimates are subsequently used for market segmentation (see, for a case study, Allenby and Ginter (1995)). All techniques involve making distributional assumptions about the respondents usually treating them as coming from a single multivariate distribution. In the common case of segmented mar-

kets, the assumption of a single population is also impractical. Moreover, it is impossible *a priori* to rely on guesses about the number or multivariate location and variability of market segments.

Along the lines of the penalized ML method proposed by Frischknecht et al. (2014), Firth's method fits a model to each individual's choices separately with no prior distributional assumptions imposed on the parameters. We can therefore construct an empirical distribution of the respondents' preferences. Firth's method differs from the method of Frischknecht et al. (2014) in that it uses a prior that depends on the estimated model itself rather than on artificial data augmentation, making it much simpler to implement.

Inspired by Louviere et al. (2008), we classify Firth's method to obtain individual-level parameters for the empirical distribution of sample preferences as a "bottom-up" approach. We call an approach that makes use of prior distributional assumptions a "top-down" approach. Also, Louviere et al. (2008) 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.

The remainder of this paper is organized as follows. Section 2 reviews the MNL model and explains the ML and Firth's estimation techniques for this model. In Section 3, we illustrate Firth's method for individual-level preference estimation using an application in employee compensation and compare its performance to HB estimation of the panel mixed logit model. To provide an overview of the situations in which Firth's method proves most effective for aggregate and individual-level estimation, we describe the results of a simulation study in Section 4. In another simulation study presented in Section 5, we compare the performance of Firth individual-level estimates for market segmentation to that of HB individual-level estimates. Finally, in Section 6, we summarize and discuss the results.

2 Model estimation

In this section, we define the MNL model for analyzing choice data, and discuss the ML estimation approach. Next, we explain how to adapt the penalized ML method of Firth (1993, 1995) for estimating the MNL 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 in this paper 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.

The panel mixed logit (PML) model is a flexible version of the MNL model (2) that allows the parameter value associated with each attribute level to vary randomly across respondents according to an *a priori* continuous distribution. Typically, the distribution for $\boldsymbol{\beta}$ is a single multivariate distribution $f(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. The PML probability for respondent n is then the integral

$$p_{njs}(\mathbf{X}_s, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int_{\boldsymbol{\beta}} \frac{\exp\left(\mathbf{x}'_{js}\boldsymbol{\beta}\right)}{\sum_{t=1}^{J} \exp\left(\mathbf{x}'_{ts}\boldsymbol{\beta}\right)} f(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\boldsymbol{\beta}. \tag{3}$$

The PML model captures the unobserved heterogeneity among respondent preferences by taking into account the correlation of the probabilities for a single respondent in all S choice sets. The model combines the individual logit models into a population-level model. Following common practice, we assume a normal distribution $\mathcal{N}(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for the parameters with no correlation between the attributes. We estimated the PML model using HB following Train's (2009) approach combined with a normal diffuse prior $\mathcal{N}(\boldsymbol{\beta_0}|\mathbf{0}_k, 1000\mathbf{I}_k)$. This approach has been implemented in the Choice Modeling platform of the statistical software package JMP Pro 13 (SAS Institute, Cary, NC, USA). For the problems discussed in this paper, we performed 10,000 iterations, where the first half were removed for convergence and the other half were used for estimation.

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

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

or, alternatively, by maximizing the log-likelihood function

$$LL(\boldsymbol{\beta}) = \sum_{r=1}^{R} \sum_{s=1}^{S} \sum_{j=1}^{J} y_{jsr} \ln(p_{js}) \tag{5}$$

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:

$$\frac{\partial LL(\boldsymbol{\beta})}{\partial \beta_i} = \sum_{r=1}^{R} \sum_{s=1}^{S} \left(x_{msri} - \mathbf{p}_s' \mathbf{x}_{si}^* \right), \quad i = 1, \ldots, k, \tag{6}$$

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 $\mathbf{X}_s$.

For finite data samples, the ML estimator, if it exists, is known to be biased away from the zero vector. The asymptotic bias of the ML estimator $\hat{\boldsymbol{\beta}}_{\text{ML}}$ for the parameter vector $\boldsymbol{\beta}$ can then be expressed as

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

where $b_1, b_2, b_3, \ldots$ are $O(1)$ functions of $\boldsymbol{\beta}$, 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 $b_1(\boldsymbol{\beta})/N$ 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, $b_1(\boldsymbol{\beta})/N$, 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 that applies for exponential family nonlinear models. 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

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

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 likelihood function $L(\boldsymbol{\beta})$ is given by Equation (4). 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 log-likelihood function $LL(\boldsymbol{\beta})$ is given by Equation (5).

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 the ordinary ML score function of Equation (6) and the first-order derivative of the logarithm of the penalty function with respect to β_i. In Appendix A, we show that the latter part corresponds to

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

We denote Firth's penalized ML estimator resulting from the modified score function by $\hat{\boldsymbol{\beta}}_{\text{FIRTH}}$. Firth's estimation procedure has been incorporated in the Choice Modeling platform of JMP, as well as the following inferential issues.

2.4 Inferential issues

Once a model has been estimated, it is usually desirable to make inferences about its parameters. 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

$$\mathbf{Var}_{\text{FIRTH}} \left(\hat{\boldsymbol{\beta}}_{\text{FIRTH}}\right) = \mathbf{M}^{-1}_{\text{FIRTH}} \left(\hat{\boldsymbol{\beta}}_{\text{FIRTH}}\right), \tag{13}$$

$$= \left(- \frac{\partial^2 LL_{\text{FIRTH}} \left(\hat{\boldsymbol{\beta}}_{\text{FIRTH}}\right)}{\partial \hat{\boldsymbol{\beta}}_{\text{FIRTH}} \partial \hat{\boldsymbol{\beta}}'_{\text{FIRTH}}} \right)^{-1}. \tag{14}$$

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 ordinary log-likelihood function values of the two models computed using Equation (5) 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 ordinary 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.

3 An application in employee compensation

This section presents an application using stated choice data related to employee compensation to illustrate Firth's penalized ML method for estimating the MNL model, in particular for estimating individual-level MNL models. We first describe the design of the study with the attributes and attribute levels of interest. We proceed with the estimation of different MNL models using the traditional ML method and Firth's method, which we compare to HB estimation of the PML model. In this regard, we score the models on their in-sample and out-of-sample prediction performance.

3.1 Design, attributes and levels

We commissioned a choice experiment to investigate preferences for various forms of compensation of employees. A compensation scheme or profile combined levels of four attributes: salary increase, bonus, extra vacation and flexible working time. Each of these attributes had three levels, which appear in Table 1. We assume all attributes are categorical and used effects-type coding for the attribute levels. This means that we coded the three levels of each attribute as [1 0], [0 1] and $[-1\ -1]$, respectively. In this way, the part-worths associated with each of the attributes sum to zero.

<Insert Table 1 about here>

We generated a Bayesian $\mathcal{D}$-optimal design of 24 choice sets of three profiles using Kessels et al.'s (2009) algorithm implemented in JMP's Choice Design platform. We divided the design into two surveys of 12 choice sets, where every respondent evaluated one survey. We distributed the surveys equally over a total of 448 respondents who participated in the questionnaire.

3.2 MNL and PML model estimation

We analyzed the data from the compensation study first by aggregating the data from all 448 respondents and estimating the 8 part-worths of the MNL model. The ML method and Firth's method led to part-worth estimates that are the same up to the third or fourth decimal place. 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 and third columns of Table 2 contain 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>

Second, we estimated a MNL model for each of the 448 respondents separately. Each individual dataset provides a total of 24 DF for the estimation of 8 part-worths. In doing so, the ML method resulted in separation for 386 respondents, i.e. in 86% of the cases, which makes it unfeasible for individual-level estimation. On the other hand, Firth's method yielded individual-level part-worth estimates for every respondent, which we compared to the individual-level estimates obtained from HB estimation of the PML model.

The last six columns of Table 2 contain summary statistics of the Firth and HB individual-level estimates. Figure 1 plots the estimates from analyzing the aggregate data together with the 95% confidence or credible intervals of the mean individual-level estimates. For the Firth estimates, the mean individual-level estimates lie close to the estimates from analyzing the aggregate data, within approximately two standard errors, and the confidence intervals are fairly narrow. This illustrates that the Firth 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 estimator in a MNL model is a nonlinear function of the responses.

<Insert Figure 1 about here>

Compared to the Firth mean individual-level estimates, the HB estimates are larger in absolute magnitude, especially for effect sizes that matter. This is because the Firth estimates are shrunk toward zero. The credible intervals of the HB means are of the same size or slightly narrower than the confidence intervals of the Firth means.

3.3 MNL and PML model prediction

To examine which of the estimated models performs better for predicting the individual choices, we computed hit rates or fractions of correct predictions for each model. We did so based on the full sample used for estimation as well as on four randomly selected holdout samples each consisting of 4, 3, 2 or 1 choice set(s) of the 12 choice sets presented to each respondent. We thus used 8, 9, 10 or 11 choice sets per respondent for estimation or training in case of the out-of-sample prediction.

The in-sample hit rates equal 61% using the estimates from the aggregate analysis, 82% using the HB individual-level estimates and 93% using the Firth individual-level estimates. On the other hand, the four holdout hit rates all equal 61% using the estimates from the aggregate analyses, 67%, 68%, 68% and 70% using the HB individual-level estimates, and 61%, 63%, 64% and 67% using the Firth individual-level estimates. We observe that the individual-level estimates mostly lead to better predictions than the estimates from the aggregate analyses, so that respondents seem to be heterogeneous in their preferences. For the individual-level estimates, the in-sample hit rates are much larger than the holdout hit rates, especially using Firth's method. Hence, some overfitting of the data takes place. Furthermore, based on the training samples of this example involving 8, 10, 12 and 14 residual DF from each respondent (given by $2S - 8$, with S the number of choice sets in the training sample), the HB estimates outperform the Firth estimates for prediction, but the difference in performance decreases with the residual DF.

The relatively good performance of HB individual-level estimates for small residual DF is due to the fact that these estimates borrow information from other individuals in the population. This is inherent to the top-down nature of the HB approach. The Firth individual-level estimates are obtained independently for every respondent, and their quality increases substantially with the residual DFs.

3.4 MNL and PML model evaluation

By performing a LR test using Equation (15), we can establish whether an individual-level model specification provides a better fit to the compensation data than the aggregate MNL model. More specifically, we compare the restricted or aggregate MNL model to the unrestricted MNL model, allowing for 448 individual-level vectors of part-worth estimates, and to the PML model. The ordinary log-likelihood value for the restricted MNL model is -4,747.7, whereas for the unrestricted MNL model, it is -2,069.9, and for the PML model, it is -3,181.3 on average. The value for the LR test statistic is then 5,355.6 for the unrestricted MNL model and 3,132.8 for the PML model.

For the comparison between the aggregate and individual-level MNL models, under the null hypothesis of equal part-worths across respondents, the LR test statistic is χ^2_ν distributed with $\nu = 3{,}576$ DF ($8 \times 448 - 8$). For the comparison between the aggregate MNL and PML model, the LR test statistic is χ^2_ν distributed with $\nu = 12$ DF ($20 - 8$, where 20 is the sum of 8 posterior means, 8 posterior variances and 4 posterior within-

attribute covariances). In both cases, the p-value for the LR test statistic is essentially zero, so that we decisively reject the null hypothesis of equal part-worths. So, we have shown that there is significant respondent heterogeneity, which begs the question of segmentation. Uncovering the source of the respondent heterogeneity is the topic of Section 5.

4 Simulation study

In this section, we present a simulation study to identify the situations in which Firth's penalized ML method proves most useful. We first discuss the setup of the simulation study revealing the various experimental conditions. We then compare the empirical performance of Firth estimates to that of standard 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 study the empirical performance of Firth's method for estimating the MNL model by simulating choices from different numbers of respondents in various experimental conditions. Similar to the compensation study discussed in Section 3, the experimental conditions all involve four three-level attributes. Also, we assumed the 8 part-worth estimates from the aggregate analysis of the compensation data to be the true 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 them 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. However, we learned that the interaction of S and J, corresponding to the residual DF, provides a more natural explanation of our results than studying the effects of S and J separately. The residual DF are defined by $S(J-1)-k$, with k the number of part-worths. 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 are available from the authors upon request.

<Insert Table 3 about here>

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 bias and variance of all estimates obtained. 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.

4.2 Performance of the estimates from aggregate data

We simulated series of 1,000 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 4a 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.

<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 4a, 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).

4.2.1 Bias of the ML and Firth estimates

To compare the bias of the ML estimates to 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 their absolute size. 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. 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 those obtained by Heinze and Schemper (2002) and Bull et al. (2002) from simulation studies comparing the standard ML method to their implementation of Firth's method to estimate logistic regression models. 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 of any penalized ML approach.

4.2.2 Variance and MSE of the ML and Firth estimates

To compare the variance and MSE of the ML estimates to those of the Firth estimates, we computed the paired differences in the variance and MSE. Figures 3a and 3b plot these differences for the Bayesian and utility-neutral designs against the squared true part-worth values. As shown in Figure 3a, the differences in the variance are all positive, meaning that Firth's method reduces the variance in all situations. This reduction is, however, only substantial for the smaller studies involving 6 or 12 respondents and 4, 10 or 16 residual DF, and for the true part-worth values that are large in absolute magnitude, which are the ones that matter. This result is in line with the results of Firth (1993), Heinze and Schemper (2002) and Bull et al. (2002), who observed for logistic regression models 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>

Regarding the differences in MSE, Figure 3b shows a similar picture as Figure 3a, confirming that the Firth estimates are uniformly better than the ML estimates. The ML estimates are inadmissible then since the Firth estimates outperform the ML estimates in terms of MSE for every situation.

4.3 Performance of the individual-level estimates

We now study 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 4b 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.

Using the separation data in Table 4b, we modeled the probability of separation as a function of the design type, 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 4 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 4 about here>

To conclude, Firth's method overcomes the challenge of separation as it permits the estimation of individual-level parth-worths 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 5 shows the bias of the Firth individual-level estimates obtained using the Bayesian and utility-neutral designs. Also here, we plotted the bias against the true part-worth values as the bias increases with their absolute size. 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 5 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 6 shows the variance of the Firth individual-level estimates obtained using the Bayesian and utility-neutral designs. The plots present the variance against the residual DF because 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 6 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 7 shows the MSE of the Firth individual-level estimates obtained using the Bayesian and utility-neutral designs. 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, Firth individual-level estimates have the highest quality overall when generated from Bayesian designs.

<Insert Figure 7 about here>

5 Market segmentation

As a last study item, we demonstrate the usefulness of Firth individual-level estimates in a "bottom-up" approach to detect market segments in a population. For comparison, we also gauge the performance of HB individual-level estimates of the PML model in a "top-down" approach.

5.1 Setup of the segmentation study

Using a full factorial simulation experiment involving four factors, we explore various segmentation scenarios to determine whether the segments can be recovered using Firth and HB individual-level estimates. These scenarios differ in the distance between the segment mean part-worth vectors, the within-segment heterogeneity, the segment size and the design used for the data simulation. We simulated respondent data assuming the setting of the compensation study (see Section 3).

In a first step, we quantified the respondent heterogeneity in the data from the compensation study using an established segmentation method. As shown by Crabbe et al. (2013), a useful segmentation method, especially in terms of segment recovery, is the use of the forces as a basis for hierarchical clustering. The forces are individual-level gradient values of the likelihood function of the MNL model expressing the respondents' individual effects on the aggregate MNL model estimates. They can be obtained using JMP's Choice Modeling platform. We applied Ward's hierarchical clustering procedure to the forces from the aggregate MNL model analysis of the compensation data and identified two distinct segments. Segment S1 contains 79% of the respondents who prefer a good balance between work and personal life. Besides the financial remuneration in terms of salary increase and bonus, they also value the non-financial compensation of extra vacation and flexible working time. Segment S2 contains 21% of the respondents who are attracted by a financial reward only. They do not care much for extra vacation or flexible working time. Table 5 shows the mean part-worth estimates of the two segments, obtained by

estimating a MNL model using Firth's method for each segment separately. The mean of segment S2 is quite extreme compared to the mean of segment S1, in the sense that the former is much further away from the zero vector than the latter (4.13 versus 1.20 in terms of Euclidean distance).

<Insert Table 5 about here>

The means of segments S1 and S2 served as input for constructing the levels of the first factor of our factorial simulation experiment, which is the distance between the segment means or the mean distance. This factor has four levels: a distance of 3.11 between the means of segments S1 and S2 and smaller distances of 2.01, 1.00 and 0.50 between the mean of segment S1 and three less extreme mean vectors that we defined for segment S2. These segment means appear in Table 5 and are referred to as S2', S2'' and S2'''. The second factor in our simulation experiment is the within-segment heterogeneity. To account for it, we added individual-specific values to the mean part-worths of the segments which we randomly drew from a normal distribution with mean $\mathbf{0}_8$, the 8-dimensional zero vector, and variance $\sigma^2\mathbf{I}_8$, where $\mathbf{I}_8$ is the identity matrix. Like Crabbe et al. (2013) and references therein, we set the levels for σ^2 equal to 0, 0.05 and 0.10, where the zero represents the case of homogeneous segments. The third factor in our simulation experiment is the number of respondents in each segment or the segment size, which we assume to be equal for the two segments and set to either 100 or 300.

Combining the levels of these three factors in our full factorial experiment, there are $4 \times 3 \times 2 = 24$ scenarios involving two segments. We simulated the segments using each of the four Bayesian designs, 1, 3, 5 and 7 in Table 3 providing 4, 10, 16 and 28 residual DF, respectively. The residual DF is therefore the fourth factor in our experiment resulting in $24 \times 4 = 96$ scenarios. We made use of Firth and HB individual-level analysis to recover the segments, and performed 100 simulations for each scenario and analysis method and averaged our results over the simulations.

5.2 Segmentation results using Firth and HB individual-level estimates

Tables C1 to C4 in Appendix C contain the simulation results for Bayesian designs 1, 3, 5 and 7, respectively. We studied three responses for all scenarios. The first response is the segment recovery, as measured by the percentage of subjects classified in segment S1. We classified respondents into segments based on the smallest distance between each subject's estimates and either one of the segment means. The second response is the mean distance between the recovered or simulated segments for S1 and S2, which, ideally, equals the true mean distance (i.e., 3.11, 2.01, 1.00 or 0.50 depending on the scenario). This response is denoted by "Sim d(S1, S2)" in the tables. The third response is the mean distance between the simulated and true segment, which, ideally, equals zero. This response is denoted by "d(Sim S1, True S1)" for S1 and by "d(Sim S2, True S2)" for S2 in the tables. The numbers in Tables C1 to C4 are the mean values over all 100 simulations with standard errors of the means in parentheses for the segment recovery,

as these standard errors are substantial. Figures 8 to 10 visualize the main results from all mean response values. To provide more detail, we also regressed the mean response values on the four factors in our experiment. We discuss the results only briefly.

<Insert Figures 8 to 10 about here>

For the first response, the percentage of subjects classified in segment S1, aimed at 50%, Tables C1 to C4 and the boxplots in Figure 8 show that the mean percentages vary much more over the simulations for the HB analysis than for the Firth analysis. All Firth mean percentages are larger than 50%, whereas the HB mean percentages take all possible values. Also, for 21 of the 96 cases (22%) using HB analysis assigns all the respondents to segment S1 whereas the Firth analysis never assigns all the respondents to one segment. For all 21 cases it is therefore impossible to allot an estimated distance between segments or to provide data on S2, so this becomes missing data in our experiment.

The boxplots in Figure 8 reveal that the Firth mean percentages come closest to 50% for large residual DF and small true mean distances between S1 and S2. Also, regression analysis showed that a small within-segment variance matters to a minor extent, and that the effect of the mean distance on the Firth mean percentages is smaller for large residual DF. We can explain this result as follows. A design that provides many residual DF (around 10 or larger) results in Firth individual-level estimates with small bias and variance (see Section 4.3), which enhances segment recovery. Also, segments that lie close to each other have small segment mean sizes (see the mean sizes of segments S1, and S2'' and S2''' in Table 5). They therefore do not suffer from the shrinkage effect of the Firth estimates, which again makes segment recovery easier.

The boxplots in Figure 8 and the contour plots in Figure 9 indicate that the HB mean percentages come closest to 50% for small true mean distances, a large within-segment variance, a small segment size and large residual DF, where these variables are ranked in decreasing order of effect following from regression analysis. Also, the regression results revealed that the effect of the mean distance on the HB mean percentages is smaller for a large within-segment variance and large residual DF. The contour plots in Figure 9 shed some light on the situations where the HB method outperforms the Firth method for market segmentation and vice versa, based on the analysis that comes closest to the 50% segment recovery. The HB method turns out to be the better method for small residual DF and a small segment size. In the opposite cases, the Firth method proves better. The within-segment variance is not decisive in this matter, although the Firth method seems to perform slightly better for recovering homogeneous segments and the HB method for recovering heterogeneous segments.

The motivation for our observations is as follows. HB estimation of the PML model requires pooling the choice data from all respondents so that the residual DF from the individual designs hardly matters. As a result, the HB method outperforms the Firth method for small residual DF, all other things being equal. Also, HB individual-level estimates show shrinkage toward the overall mean, especially when the number of respondents gets

large (Kruschke and Vanpaemel, 2015). They flatten out with large segment sizes so that the Firth method performs better in that case.

Figure 10 shows to what extent the estimated mean distance between S1 and S2, i.e. "Sim d(S1, S2)", deviates from the true mean distance, where the latter explains the deviation. The estimated mean distance is generally closer to the true one for the Firth analysis than for the HB analysis, except for the smallest true mean distance of 0.5, where the deviation is the smallest for the HB analysis. For larger true mean distances, the deviations for the HB analysis increase substantially. Similarly, regression analysis revealed that the true mean distance largely explains the mean distance between the simulated and true segment for both S1 and S2, shortened as "d(Sim S, True S)". This result is even more pronounced for the HB analysis than for the Firth analysis. Also here, the HB method outperforms the Firth method for the smallest true mean distance of 0.5.

The reason why the true mean distance has a large impact on the overall segmentation performance of the HB analysis is most likely due to the single normal prior parameter distribution we assumed instead of, for example, a mixture of normal priors to better mimic the true segments. However, the true segmentation structure is rarely known in advance so that the HB analysis has to rely on distributional assumptions about the respondents, which may lead to biased results in case they are incorrect. Further research is needed to expand the segmentation study to different prior parameter distributions as input for the HB analysis.

To summarize the study, the use of Firth individual-level estimates proves to be effective for market segmentation, especially in the case of large residual DF (around 10 or larger) and small segment mean sizes (segments S2'' and S2''' in Table 5). Compared to using HB individual-level estimates for market segmentation, segment recovery based on Firth estimates is more precise overall. The Firth method outperforms the HB method for large residual DF, a large segment size (around 300 respondents per segment), large segment mean sizes (segments S2, S2' and S2'' in Table 5) and to a very minor extent, homogeneous segments. Perhaps if we were to further increase the within-segment variance, the effect of this variable might have been more pronounced.

6 Summary and discussion

We adapted the penalized ML method of Firth (1993, 1995) for estimating the MNL model using choice data. Through a real-life application in employee compensation and subsequent simulation studies, we have shown that the method proves useful for three reasons. First, Firth's method yields parameter estimates for the MNL model that are reasonable in the case of data separation, whereas the traditional ML method fails to do so. Second, if the ML estimates exist, Firth's method removes their bias 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. These individual-level estimates can effectively be used for predicting people's choices and market segmentation as long as the number of choices per person is sufficiently large. For the problems described in this paper, this comes down to having more than 10 residual DF resulting in individual-level estimates of good overall quality.

In our simulation study to compare the empirical performance of Firth estimates to that of standard ML estimates obtained from aggregate data, we obtained the following results. 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 compared to standard ML: 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.

A more important advantage of Firth's method is that it provides individual-level estimates in a computationally simple manner. Using the compensation data, we compared the prediction performance of the Firth individual-level estimates to that of the HB individual-level estimates from the PML model. All individual-level estimates overfit the data, and this is especially so for the Firth estimates. The Firth individual-level estimates have a much better model fit than the HB individual-level estimates, but this does not translate into better prediction performance. However, the prediction performance of the Firth individual-level estimates increases substantially with the residual DFs. The larger the number of choices per person available for estimation, the greater the ability of Firth's method to predict individual choices in the holdout sample.

In our simulation study to compare the performance of Firth and HB individual-level estimates for market segmentation, we observed once more that the number of choices per person matters a great deal. In case of large residual DF (around 10 or larger), the Firth method outperforms the HB method for segment recovery. This is also the case when the segments are large in size (around 300 respondents per segment), because the HB individual-level estimates then show shrinkage toward the overall mean. Another important factor is the distance between the segment means. Assuming a single normal prior distribution, it is no surprise that the HB individual-level estimates lose ground to the Firth estimates when the mean distance gets larger. This is because Firth's method does not require imposing an a priori preference distribution. This is important since it is not at all clear what an appropriate a priori preference distribution would be when markets are segmented.

Lastly, on a broader scale, it is still an empirical question as to how well market segmentation based on individual-level estimates describes the situation for a particular product or service to provide input to managerial decisions. We therefore propose a comparison of different types of individual-level estimates for predicting external validity or real world segmentation performance as an avenue for future research.

Acknowledgements

The research described in this paper was carried out while Roselinde Kessels was a postdoctoral fellow of the Flemish Research Foundation (FWO). The authors are grateful to Rob Reul from Isometric Solutions for collecting the data described in Section 3 and to Melinda Thielbar and Chris Gotwalt for providing assistance with hierarchical Bayes modeling in JMP Pro 13. We also express our gratitude to two anonymous reviewers and the editor for their extensive comments and suggestions, which have led to substantial improvements. The paper also benefited from comments made by participants in the Interdisciplinary Choice Workshop held in Santiago de Chile in August 2018.

References

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

Allenby, G.M. and Ginter, J.L. (1995). Using extremes to design products and segment markets, *Journal of Marketing Research* **32**: 392–403.

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.

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.

Bliemer, M.C.J. and Rose, J.M. (2010). Construction of experimental designs for mixed logit models allowing for correlation across choice observations, *Transportation Research B* **44**: 720–734.

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.

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.

Crabbe, M., Jones, B. and Vandebroek, M. (2013). Comparing two-stage segmentation methods for choice data with a one-stage latent class choice analysis, *Communications in Statistics – Simulation and Computation* **42**: 1188–1212.

Dumont, J., Giergiczny, M. and Hess, S. (2015). Individual level models vs. sample level models: contrasts and mutual benefits, *Transportmetrica A: Transport Science* **11**: 465–483.

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.

Frischknecht, B.D., Eckert, C., Geweke, J. and Louviere, J.J. (2014). A simple method for estimating preference parameters for individuals, *International Journal of Research in Marketing* **31**: 35–48.

Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics*, Hoboken, NJ: John Wiley & Sons.

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. (2009). An efficient algorithm for constructing Bayesian optimal choice designs, *Journal of Business and Economic Statistics* **27**: 279–291.

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

Kruschke J.K. and Vanpaemel, W. (2015). Bayesian estimation in hierarchical models, in Busemeyer, J.R., Wang, Z., Townsend, J.T. and Eidels, A. (Eds.), *The Oxford Handbook of Computational and Mathematical Psychology*, Oxford, UK: Oxford University Press, 279–299.

Le Cam, L. (1990). Maximum likelihood: An introduction, *International Statistical Review* **58**: 153–171.

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.

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.

Sándor, Z. and Wedel, M. (2001). Designing conjoint choice experiments using managers' prior beliefs, *Journal of Marketing Research* **38**: 430–444.

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. (2009). *Discrete Choice Methods with Simulation*, 2nd Edition, 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 traditional ML and Firth's method, and summary statistics of the Firth and HB individual-level estimates.

Attribute Level	Aggregate		Individual-Level					
			Mean		Std Dev		Std Err*	
	ML	Firth	Firth	HB	Firth	HB	Firth	HB
No Raise	-0.921	-0.920	-0.841	-1.345	0.691	0.613	0.033	0.029
Small Raise	0.186	0.186	0.181	0.307	0.489	0.133	0.023	0.006
Large Raise	0.735	0.734	0.660	1.038	0.672	0.652	0.032	0.031
No Bonus	-1.006	-1.005	-0.951	-1.457	0.710	0.574	0.034	0.027
Small Bonus	0.200	0.200	0.219	0.313	0.448	0.098	0.021	0.005
Large Bonus	0.806	0.805	0.732	1.144	0.636	0.589	0.030	0.028
No Extra Vacation	-0.461	-0.460	-0.471	-0.663	0.621	0.375	0.029	0.018
1 Extra Week	0.114	0.114	0.148	0.197	0.416	0.066	0.020	0.003
2 Extra Weeks	0.347	0.346	0.323	0.466	0.567	0.398	0.027	0.019
No Flex	-0.264	-0.264	-0.269	-0.378	0.656	0.233	0.031	0.011
2 Days Flex	0.096	0.096	0.090	0.171	0.495	0.111	0.023	0.005
4 Days Flex	0.168	0.168	0.179	0.207	0.514	0.276	0.024	0.013

*called posterior std dev in the case of HB

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 (a) aggregate and (b) individual-level choice data.

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

Table 5: Means of the two segments S1 and S2('/''/''') in the compensation segmentation study, with S2 getting closer to S1, the Euclidean distance between them and the Euclidean distance from the zero vector.

Attribute Level	S1	S2	S2'	S2''	S2'''
No Raise	-0.683	-2.677	-1.862	-1.162	-0.869
Small Raise	0.060	0.711	0.611	0.371	0.171
No Bonus	-0.769	-2.904	-2.089	-1.389	-1.096
Small Bonus	0.067	0.890	0.790	0.550	0.350
No Extra Vacation	-0.519	-0.334	-0.334	-0.334	-0.414
1 Extra Week	0.149	0.004	0.004	0.004	0.084
No Flex	-0.280	-0.194	-0.194	-0.194	-0.274
2 Days Flex	0.108	0.086	0.086	0.086	0.086
Distance from Mean S1		3.11	2.01	1.00	0.50
Distance from Zero	1.20	4.13	3.00	1.97	1.54

Figures

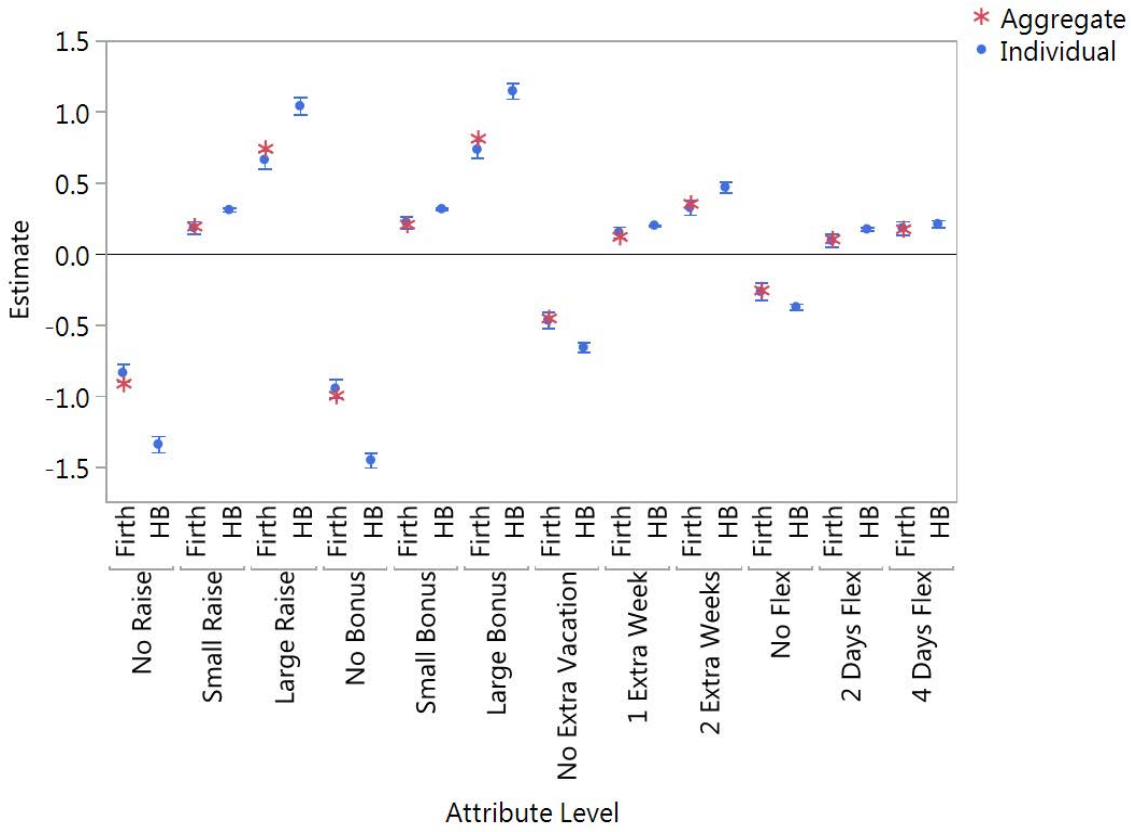


Figure 1: Estimates obtained from the aggregate choice data of the compensation study using Firth's method and 95% confidence and credible intervals of the means of the Firth and HB individual-level estimates, respectively.

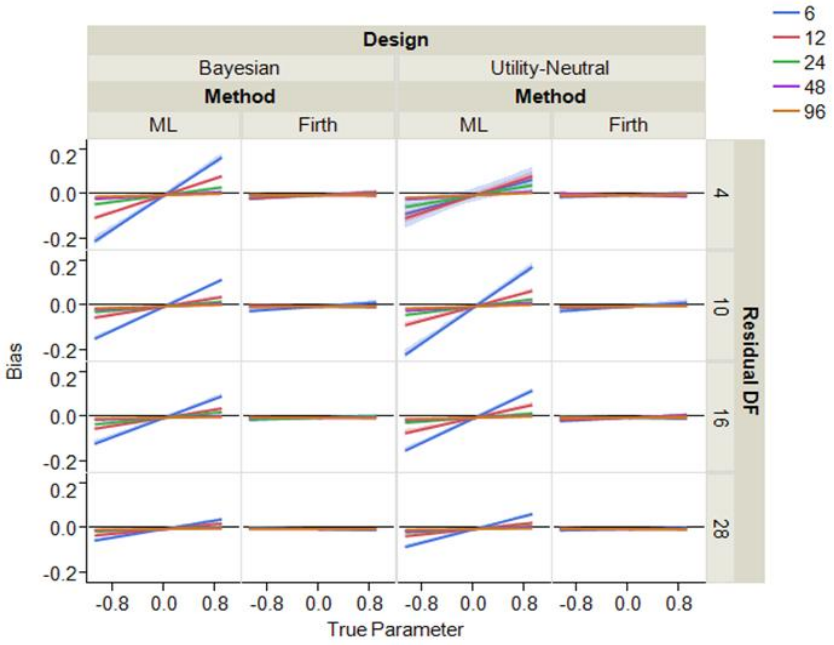
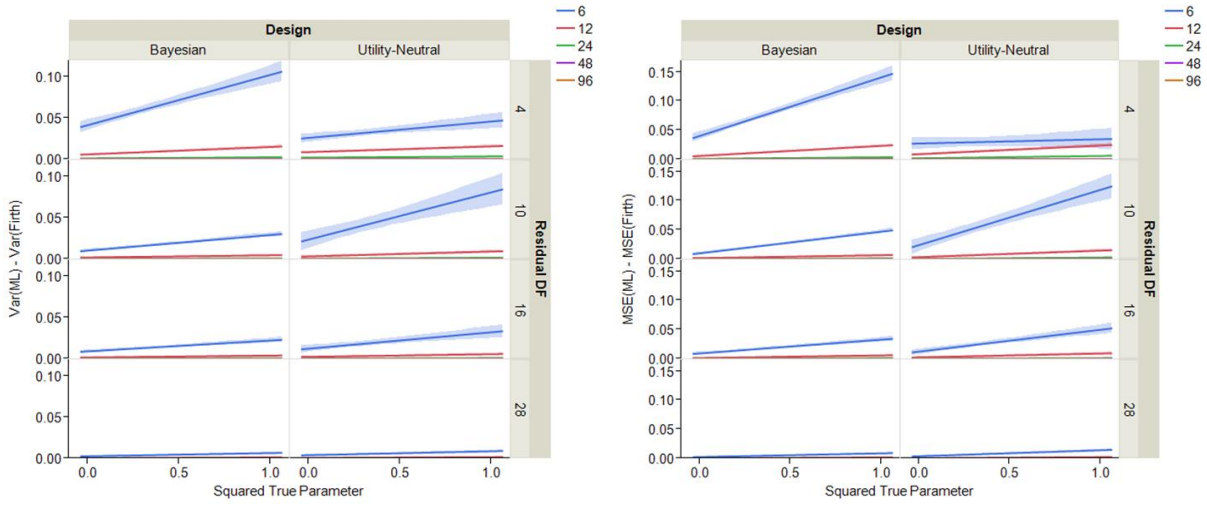


Figure 2: Bias of the traditional maximum likelihood (ML) and Firth estimates obtained from the Bayesian and utility-neutral designs in Table 3 and aggregate choice data from 6, 12, 24, 48 and 96 respondents.



(a) Difference in variance

(b) Difference in MSE

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

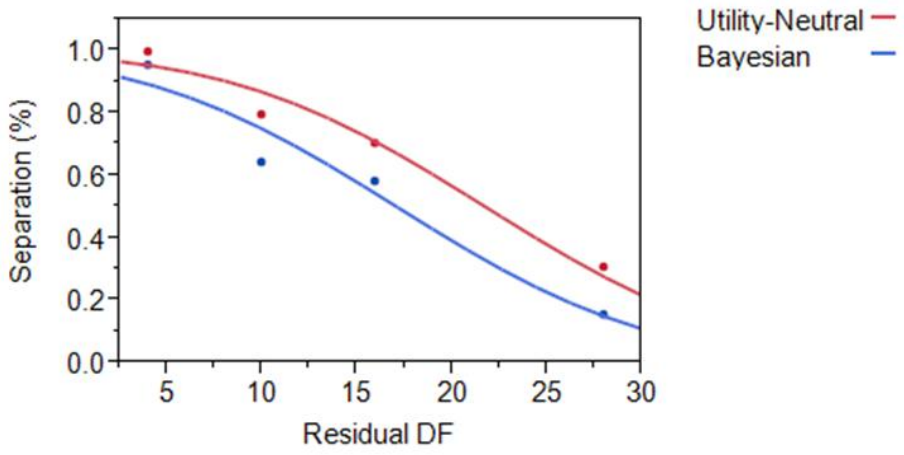


Figure 4: Predicted probability of separation when analyzing individual-level choice data as a function of the design type, Bayesian or utility-neutral, and the residual degrees of freedom (DF).

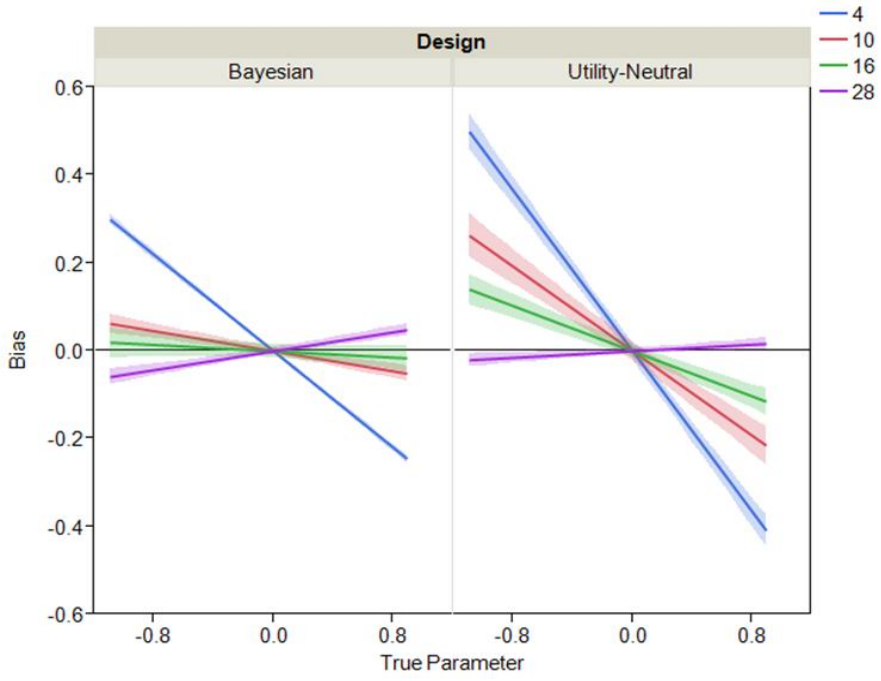


Figure 5: Bias of the Firth individual-level estimates obtained from the Bayesian and utility-neutral designs in Table 3 with 4, 10, 16 and 28 residual degrees of freedom.

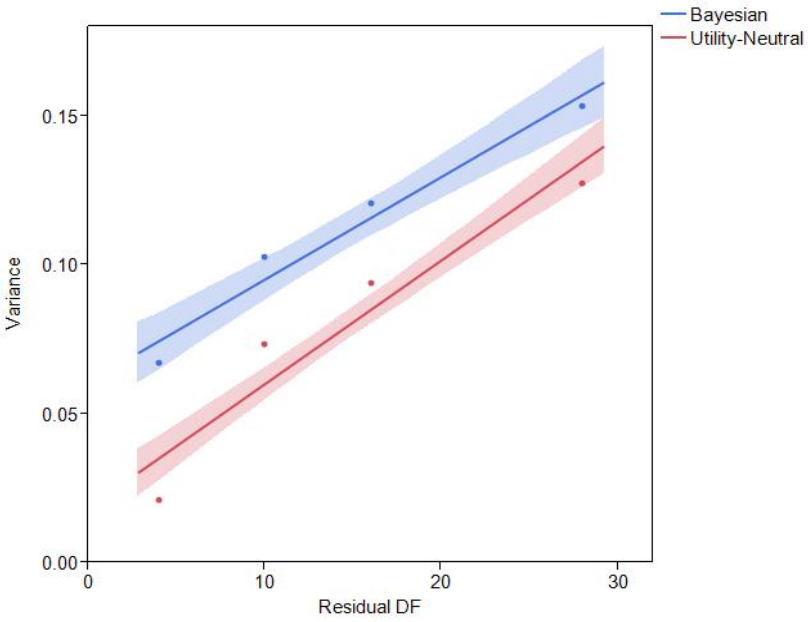


Figure 6: Variance of the Firth individual-level estimates obtained from the Bayesian and utility-neutral designs in Table 3 with 4, 10, 16 and 28 residual degrees of freedom (DF).

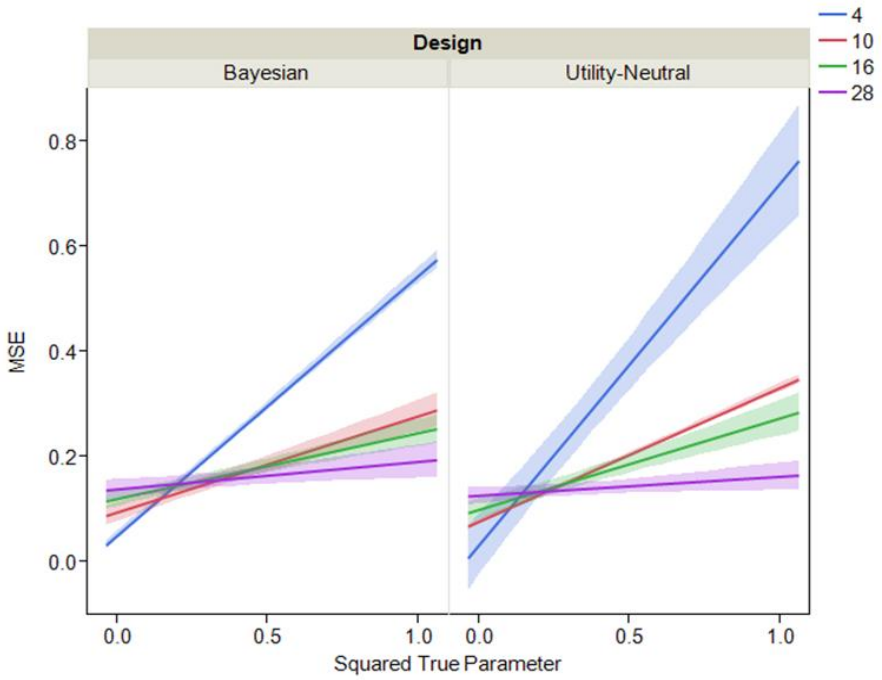


Figure 7: Mean squared error (MSE) of the Firth individual-level estimates obtained from the Bayesian and utility-neutral designs in Table 3 with 4, 10, 16 and 28 residual degrees of freedom.

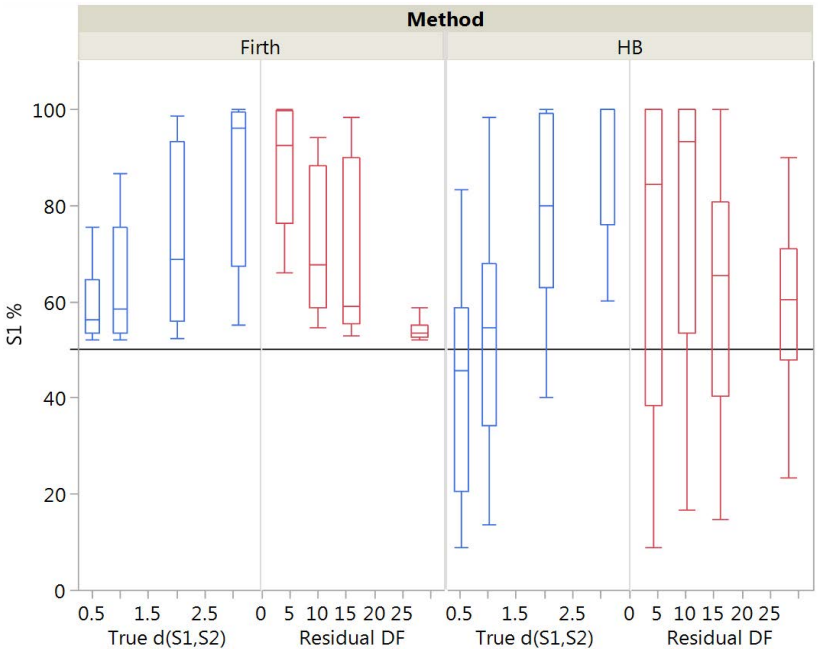


Figure 8: Boxplots of the percentage of subjects classified in segment S1, aimed at 50%, using the Firth and HB method as a function of the true mean distance between S1 and S2 and the residual degrees of freedom (DF).

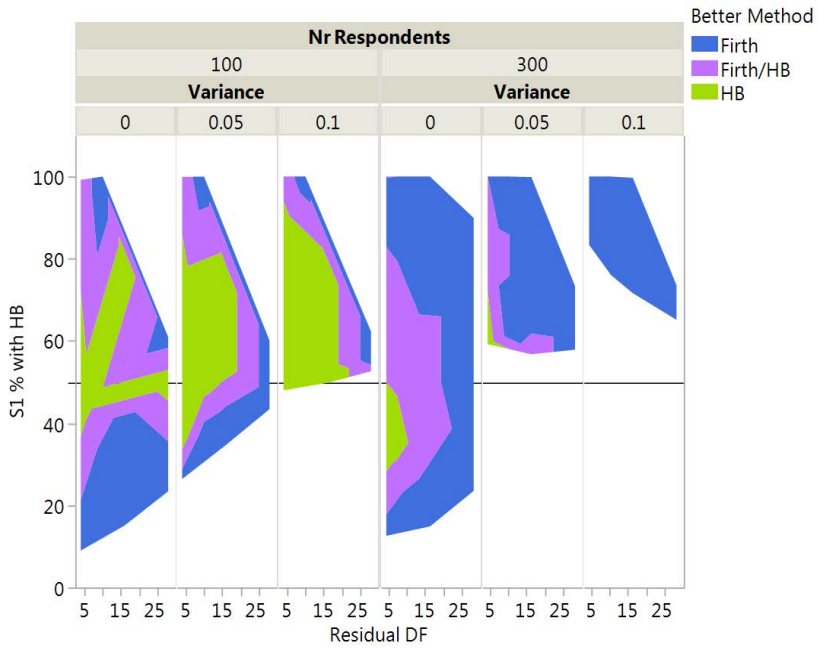


Figure 9: Contour plots of the percentage of subjects classified in segment S1, aimed at 50%, using the HB method as a function of the residual degrees of freedom (DF), the number of respondents in each segment and the within-segment variance. Regions are colored by the better method for segmentation (Firth or HB) that comes closest to the true 50%–50% segmentation for S1 and S2.

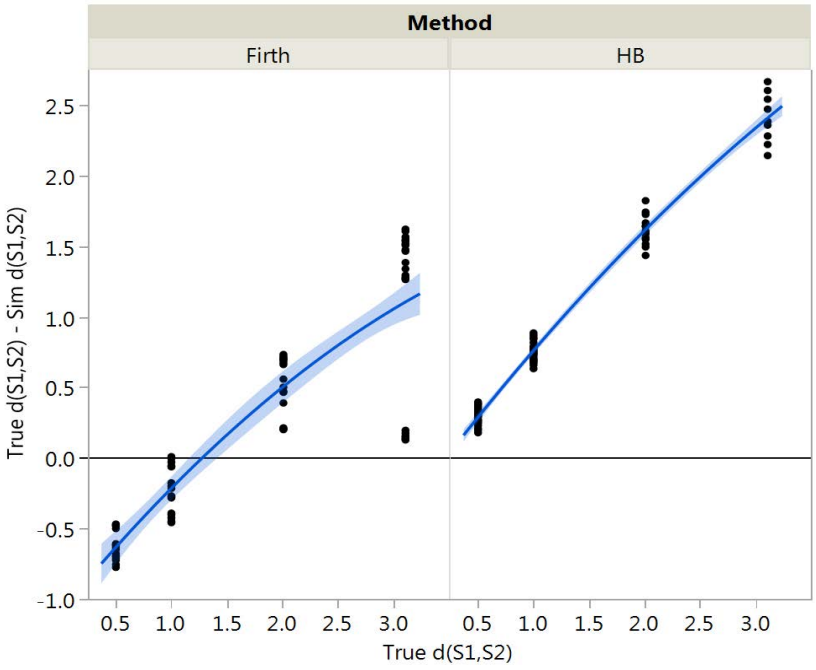


Figure 10: Deviation of the mean distance between the simulated segments for S1 and S2 from the true mean distance between S1 and S2 as a function of the true mean distance, with segments simulated using the Firth and HB method.

Appendix A. Firth's modification to the score function for the MNL model

In this appendix, we derive the first-order derivative of the logarithm of Firth's penalty function for the MNL model with respect to β_i, $i = 1, \ldots, k$. 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.

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

$$\begin{aligned}
\frac{1}{2}\,\frac{\partial \ln|\mathbf{M}|}{\partial \beta_i} &= \frac{1}{2\,|\mathbf{M}|}\,\frac{\partial\,|\mathbf{M}|}{\partial \beta_i}, && \text{(A1)}\\
&= \frac{1}{2}\,\mathrm{tr}\left(\mathbf{M}^{-1}\,\frac{\partial \mathbf{M}}{\partial \beta_i}\right), && \text{(A2)}\\
&= \frac{1}{2}\,\mathrm{tr}\left(\mathbf{M}^{-1}\,\frac{\partial\left(\mathbf{X}_s'\left(\mathbf{P}_s - \mathbf{p}_s\mathbf{p}_s'\right)\mathbf{X}_s\right)}{\partial \beta_i}\right), && \text{(A3)}\\
&= \frac{1}{2}\,\mathrm{tr}\left(\mathbf{M}^{-1}\mathbf{X}_s'\,\frac{\partial\left(\mathbf{P}_s - \mathbf{p}_s\mathbf{p}_s'\right)}{\partial \beta_i}\,\mathbf{X}_s\right), && \text{(A4)}\\
&= \frac{1}{2}\,\mathrm{tr}\left(\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}_s'\right)\frac{\partial\left(\mathbf{P}_s - \mathbf{p}_s\mathbf{p}_s'\right)}{\partial \beta_i}\right), && \text{(A5)}\\
&= \frac{1}{2}\,\mathrm{tr}\left(\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}_s'\right)\frac{\partial \mathbf{P}_s}{\partial \beta_i}\right) - \frac{1}{2}\,\mathrm{tr}\left(\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}_s'\right)\frac{\partial\left(\mathbf{p}_s\mathbf{p}_s'\right)}{\partial \beta_i}\right). && \text{(A6)}
\end{aligned}$$

Here, we obtain Equation (A2) using Jacobi's formula for the invertible matrix $\mathbf{M}$

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

Also, Equation (A5) 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 (A6) 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:

$$\frac{1}{2}\,\mathrm{tr}\left(\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}_s'\right)\frac{\partial\left(\mathbf{p}_s\mathbf{p}_s'\right)}{\partial\beta_i}\right) = \frac{1}{2}\sum_{u=1}^{J}\sum_{v=1}^{J}\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}_s'\right)_{uv}\frac{\partial\left(p_{us}p_{vs}\right)}{\partial\beta_i}, \tag{A7}$$

$$= \frac{1}{2}\sum_{u=1}^{J}\sum_{v=1}^{J}\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}_s'\right)_{uv} 2\frac{\partial p_{us}}{\partial\beta_i}p_{vs}, \tag{A8}$$

$$= \mathrm{tr}\left(\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}_s'\right)\frac{\partial\mathbf{p}_s}{\partial\beta_i}\mathbf{p}_s'\right), \tag{A9}$$

$$= \mathrm{tr}\left(\mathbf{p}_s'\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}_s'\right)\frac{\partial\mathbf{p}_s}{\partial\beta_i}\right), \tag{A10}$$

$$= \mathbf{p}_s'\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}_s'\right)\frac{\partial\mathbf{p}_s}{\partial\beta_i}. \tag{A11}$$

Equations (A8) and (A9) 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 (A10), 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 (A11).

Combining Equations (A6) and (A11) 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:

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

For the situation where R respondents evaluate S choice sets, Equation (A12) becomes

$$\frac{1}{2}\,\frac{\partial\ln|\mathbf{M}|}{\partial\beta_i} = R\sum_{s=1}^{S}\left[\frac{1}{2}\,\mathrm{tr}\left(\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}_s'\right)\frac{\partial\mathbf{P}_s}{\partial\beta_i}\right) - \mathbf{p}_s'\left(\mathbf{X}_s\mathbf{M}^{-1}\mathbf{X}_s'\right)\frac{\partial\mathbf{p}_s}{\partial\beta_i}\right]. \tag{A13}$$

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

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

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

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

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

Appendix C. Detailed results of the compensation segmentation study

<Insert Tables C1 to C4 about here>

Table C1: Results of 100 simulations (means and standard errors of the means in parentheses) of the compensation segmentation study using Bayesian design 1 with 4 residual DF.

Scenario	True d(S1, S2)	Variance	Nr Resp	S1 Classification (%)		Sim d(S1, S2)		d(Sim S1, True S1)		d(Sim S2, True S2)	
				Firth	HB	Firth	HB	Firth	HB	Firth	HB
1	3.11	0	100	**100***(0.00)	**100***(0.00)	2.93	/	**0.32**	1.19	/	/
2	3.11	0.05	100	**100***(0.00)	**100***(0.00)	2.95	/	**0.35**	1.06	/	/
3	3.11	0.10	100	**100***(0.00)	**100***(0.00)	2.98	/	**0.36**	0.99	/	/
4	3.11	0	300	**100***(0.00)	**100***(0.00)	2.91	/	**0.32**	0.95	/	/
5	3.11	0.05	300	**100***(0.00)	**100***(0.00)	2.96	/	**0.34**	0.83	/	/
6	3.11	0.10	300	**100***(0.00)	**100***(0.00)	2.98	/	**0.36**	0.76	/	/
7	2.01	0	100	98 (0.09)	**45** (1.72)	1.54	0.36	**0.31**	0.88	1.19	**0.91**
8	2.01	0.05	100	99 (0.08)	**71** (1.88)	1.54	0.37	**0.34**	0.80	1.16	**0.97**
9	2.01	0.10	100	99 (0.08)	**88** (1.34)	1.62	0.41	**0.36**	0.74	1.18	**1.01**
10	2.01	0	300	**98** (0.05)	100 (0.00)	1.45	/	**0.30**	0.82	1.10	/
11	2.01	0.05	300	**99** (0.05)	100 (0.00)	1.51	/	**0.33**	0.69	1.11	/
12	2.01	0.10	300	**99** (0.05)	100 (0.00)	1.55	/	**0.36**	0.61	1.13	/
13	1.00	0	100	**79** (0.27)	14 (0.95)	0.99	0.34	**0.37**	0.53	**0.37**	0.39
14	1.00	0.05	100	84 (0.27)	**36** (1.61)	1.03	0.31	**0.42**	0.44	**0.38**	0.40
15	1.00	0.10	100	87 (0.24)	**55** (1.70)	1.06	0.32	0.48	**0.38**	**0.40**	0.44
16	1.00	0	300	79 (0.18)	**34** (1.54)	1.00	0.15	**0.37**	0.47	**0.35**	0.45
17	1.00	0.05	300	**84***(0.16)	**85***(1.36)	1.03	0.18	0.43	**0.40**	**0.36**	0.50
18	1.00	0.10	300	**86** (0.12)	98 (0.38)	1.06	0.25	0.47	**0.35**	**0.37**	0.55
19	0.50	0	100	**66** (0.29)	9 (0.71)	0.97	0.32	0.49	**0.38**	**0.22**	0.26
20	0.50	0.05	100	**71** (0.35)	26 (1.35)	0.98	0.29	0.54	**0.27**	0.22	**0.19**
21	0.50	0.10	100	75 (0.34)	**48** (1.52)	1.00	0.29	0.60	**0.20**	0.24	**0.20**
22	0.50	0	300	**66** (0.18)	13 (0.96)	0.97	0.15	0.49	**0.27**	**0.20***	**0.20***
23	0.50	0.05	300	71 (0.17)	**59** (1.76)	0.97	0.15	0.54	**0.19**	**0.19**	0.22
24	0.50	0.10	300	**75** (0.18)	83 (1.18)	0.97	0.19	0.59	**0.18**	**0.20**	0.28

Note: Values in bold highlight the better method (Firth or HB) yielding a segmentation that comes closest to the true 50%–50% segmentation for S1–S2. Pairwise comparisons indicated by a star (*) are not significant at the 5% level; all others are.

Table C2: Results of 100 simulations (means and standard errors of the means in parentheses) of the compensation segmentation study using Bayesian design 3 with 10 residual DF.

Scenario	True d(S1, S2)	Variance	Nr Resp	S1 Classification (%)		Sim d(S1, S2)		d(Sim S1, True S1)		d(Sim S2, True S2)	
				Firth	HB	Firth	HB	Firth	HB	Firth	HB
1	3.11	0	100	**91** (0.19)	100 (0.00)	1.50	/	**0.43**	0.94	1.44	/
2	3.11	0.05	100	**93** (0.16)	100 (0.00)	1.56	/	**0.40**	0.85	1.46	/
3	3.11	0.10	100	**94** (0.17)	100 (0.00)	1.59	/	**0.38**	0.82	1.49	/
4	3.11	0	300	**92** (0.12)	100 (0.00)	1.49	/	**0.43**	0.79	1.43	/
5	3.11	0.05	300	**93** (0.10)	100 (0.00)	1.54	/	**0.40**	0.72	1.44	/
6	3.11	0.10	300	**94** (0.10)	100 (0.00)	1.57	/	**0.38**	0.68	1.45	/
7	2.01	0	100	**71** (0.34)	93 (1.33)	1.35	0.28	**0.23**	0.81	**0.64**	1.01
8	2.01	0.05	100	**76** (0.29)	98 (0.45)	1.34	0.40	**0.26**	0.73	**0.68**	1.04
9	2.01	0.10	100	**78** (0.26)	99 (0.35)	1.35	0.46	**0.27**	0.67	**0.72**	1.05
10	2.01	0	300	**71** (0.15)	100 (0.00)	1.33	/	**0.22**	0.69	0.64	/
11	2.01	0.05	300	**75** (0.16)	100 (0.00)	1.33	/	**0.25**	0.61	0.67	/
12	2.01	0.10	300	**78** (0.14)	100 (0.00)	1.34	/	**0.26**	0.56	0.69	/
13	1.00	0	100	**58** (0.31)	30 (1.24)	1.22	0.24	**0.26**	0.46	**0.26**	0.42
14	1.00	0.05	100	61 (0.31)	**50** (1.40)	1.18	0.28	**0.30**	0.39	**0.22**	0.42
15	1.00	0.10	100	65 (0.31)	**62** (1.46)	1.18	0.32	**0.34***	**0.35***	**0.23**	0.44
16	1.00	0	300	**58***(0.19)	**59***(1.78)	1.20	0.11	**0.25**	0.44	**0.21**	0.47
17	1.00	0.05	300	**61** (0.18)	89 (0.86)	1.18	0.18	**0.29**	0.38	**0.20**	0.50
18	1.00	0.10	300	**64** (0.19)	94 (0.49)	1.17	0.26	**0.33**	0.34	**0.20**	0.53
19	0.50	0	100	**55** (0.25)	17 (1.00)	1.14	0.23	0.37	**0.28**	0.38	**0.20**
20	0.50	0.05	100	**57** (0.34)	37 (1.41)	1.12	0.25	0.41	**0.21**	0.34	**0.18**
21	0.50	0.10	100	61 (0.34)	**52** (1.18)	1.12	0.29	0.44	**0.17**	0.30	**0.19**
22	0.50	0	300	**55** (0.20)	20 (1.25)	1.13	0.12	0.35	**0.25**	0.38	**0.20**
23	0.50	0.05	300	**57** (0.20)	62 (1.38)	1.12	0.15	0.40	**0.18**	0.32	**0.22**
24	0.50	0.10	300	**60** (0.17)	76 (0.84)	1.11	0.20	0.43	**0.16**	0.29	**0.25**

Note: Values in bold highlight the better method (Firth or HB) yielding a segmentation that comes closest to the true 50%–50% segmentation for S1–S2. Pairwise comparisons indicated by a star (*) are not significant at the 5% level; all others are.

Table C3: Results of 100 simulations (means and standard errors of the means in parentheses) of the compensation segmentation study using Bayesian design 5 with 16 residual DF.

Scenario	True d(S1, S2)	Variance	Nr Resp	S1 Classification (%)		Sim d(S1, S2)		d(Sim S1, True S1)		d(Sim S2, True S2)	
				Firth	HB	Firth	HB	Firth	HB	Firth	HB
1	3.11	0	100	98 (0.10)	**80** (1.61)	1.83	0.44	**0.61**	1.22	1.83	**1.53**
2	3.11	0.05	100	98 (0.09)	**76** (1.52)	1.77	0.57	**0.58**	1.12	1.78	**1.50**
3	3.11	0.10	100	98 (0.10)	**77** (1.44)	1.73	0.64	**0.55**	1.06	1.74	**1.49**
4	3.11	0	300	**98** (0.06)	100 (0.00)	1.64	/	**0.60**	1.15	1.64	/
5	3.11	0.05	300	**98** (0.06)	100 (0.00)	1.60	/	**0.57**	1.11	1.62	/
6	3.11	0.10	300	**98** (0.06)	100 (0.00)	1.64	/	**0.54**	1.06	1.62	/
7	2.01	0	100	**59** (0.29)	40 (1.02)	1.29	0.36	**0.19**	0.86	**0.74**	0.85
8	2.01	0.05	100	63 (0.26)	**54** (1.04)	1.31	0.41	**0.23**	0.79	**0.74**	0.88
9	2.01	0.10	100	66 (0.30)	**64** (1.18)	1.32	0.45	**0.25**	0.74	**0.75**	0.90
10	2.01	0	300	**59** (0.13)	73 (1.45)	1.28	0.19	**0.17**	0.88	**0.74**	0.97
11	2.01	0.05	300	**63** (0.16)	86 (0.95)	1.29	0.27	**0.20**	0.80	**0.73**	0.98
12	2.01	0.10	300	**66** (0.16)	90 (0.77)	1.30	0.34	**0.24**	0.74	**0.74**	0.98
13	1.00	0	100	**54** (0.29)	17 (1.00)	1.28	0.25	**0.26**	0.47	**0.18**	0.37
14	1.00	0.05	100	**57** (0.34)	41 (1.48)	1.28	0.27	**0.30**	0.41	**0.18**	0.40
15	1.00	0.10	100	59 (0.38)	**54** (1.48)	1.27	0.31	**0.34**	0.38	**0.19**	0.41
16	1.00	0	300	**53** (0.19)	24 (1.42)	1.27	0.13	**0.24**	0.47	**0.14**	0.43
17	1.00	0.05	300	**56** (0.20)	67 (1.30)	1.27	0.16	**0.28**	0.42	**0.13**	0.46
18	1.00	0.10	300	**59** (0.18)	81 (0.85)	1.27	0.22	**0.32**	0.37	**0.16**	0.47
19	0.50	0	100	**54** (0.36)	15 (1.30)	1.18	0.21	0.36	**0.27**	0.38	**0.19**
20	0.50	0.05	100	**55** (0.31)	35 (1.39)	1.20	0.23	0.41	**0.22**	0.36	**0.18**
21	0.50	0.10	100	57 (0.33)	**50** (1.49)	1.20	0.27	0.45	**0.19**	0.34	**0.19**
22	0.50	0	300	**53** (0.18)	15 (1.08)	1.18	0.12	0.35	**0.24**	0.37	**0.19**
23	0.50	0.05	300	**55***(0.18)	**57***(1.43)	1.19	0.14	0.40	**0.19**	0.34	**0.21**
24	0.50	0.10	300	**57** (0.19)	72 (0.99)	1.19	0.20	0.44	**0.16**	0.32	**0.23**

Note: Values in bold highlight the better method (Firth or HB) yielding a segmentation that comes closest to the true 50%–50% segmentation for S1–S2. Pairwise comparisons indicated by a star (*) are not significant at the 5% level; all others are.

Table C4: Results of 100 simulations (means and standard errors of the means in parentheses) of the compensation segmentation study using Bayesian design 7 with 28 residual DF.

Scenario	True d(S1, S2)	Variance	Nr Resp	S1 Classification (%)		Sim d(S1, S2)		d(Sim S1, True S1)		d(Sim S2, True S2)	
				Firth	HB	Firth	HB	Firth	HB	Firth	HB
1	3.11	0	100	**55** (0.19)	61 (0.82)	1.85	0.75	**0.18**	0.99	**1.18**	1.44
2	3.11	0.05	100	**57** (0.21)	60 (0.62)	1.83	0.89	**0.23**	0.89	**1.16**	1.39
3	3.11	0.10	100	**59** (0.23)	62 (0.66)	1.84	0.97	**0.24**	0.84	**1.16**	1.37
4	3.11	0	300	**55** (0.10)	90 (1.09)	1.83	0.51	**0.17**	1.12	**1.17**	1.54
5	3.11	0.05	300	**57** (0.13)	73 (0.67)	1.82	0.73	**0.21**	0.96	**1.16**	1.48
6	3.11	0.10	300	**59** (0.13)	72 (0.63)	1.82	0.83	**0.23**	0.89	**1.14**	1.45
7	2.01	0	100	52 (0.22)	**50** (0.66)	1.81	0.42	**0.14**	0.79	**0.24**	0.85
8	2.01	0.05	100	**53** (0.21)	57 (0.70)	1.80	0.52	**0.15**	0.70	**0.27**	0.86
9	2.01	0.10	100	**55** (0.26)	62 (0.75)	1.81	0.58	**0.19**	0.64	**0.29**	0.86
10	2.01	0	300	**53** (0.13)	72 (0.90)	1.80	0.27	**0.10**	0.82	**0.22**	0.96
11	2.01	0.05	300	**54** (0.13)	72 (0.72)	1.80	0.40	**0.12**	0.72	**0.23**	0.93
12	2.01	0.10	300	**55** (0.15)	74 (0.57)	1.80	0.49	**0.14**	0.66	**0.25**	0.92
13	1.00	0	100	**52** (0.28)	31 (1.12)	1.40	0.23	**0.20**	0.44	**0.37***	**0.38***
14	1.00	0.05	100	**52** (0.31)	47 (1.10)	1.46	0.30	**0.23**	0.38	**0.38***	**0.39***
15	1.00	0.10	100	**54** (0.27)	56 (0.97)	1.46	0.37	**0.25**	0.33	**0.36**	0.39
16	1.00	0	300	**52** (0.15)	40 (1.03)	1.39	0.13	**0.18**	0.45	**0.35**	0.44
17	1.00	0.05	300	**53** (0.18)	65 (0.82)	1.43	0.21	**0.20**	0.39	**0.34**	0.45
18	1.00	0.10	300	**53** (0.17)	68 (0.62)	1.45	0.29	**0.23**	0.34	**0.33**	0.43
19	0.50	0	100	**53** (0.29)	23 (1.24)	1.15	0.18	0.28	**0.24**	0.46	**0.18**
20	0.50	0.05	100	**53** (0.32)	43 (1.34)	1.22	0.24	0.32	**0.19**	0.49	**0.17**
21	0.50	0.10	100	**53***(0.36)	**53***(1.18)	1.28	0.30	0.36	**0.16**	0.50	**0.18**
22	0.50	0	300	**52** (0.20)	24 (1.32)	1.14	0.10	0.27	**0.23**	0.45	**0.20**
23	0.50	0.05	300	**53** (0.20)	58 (1.20)	1.21	0.16	0.31	**0.18**	0.47	**0.20**
24	0.50	0.10	300	**53** (0.21)	65 (0.84)	1.26	0.25	0.34	**0.14**	0.47	**0.21**

Note: Values in bold highlight the better method (Firth or HB) yielding a segmentation that comes closest to the true 50%–50% segmentation for S1–S2. Pairwise comparisons indicated by a star (*) are not significant at the 5% level; all others are.