Generalized Linear Mixed Effect Models with Crossed Random Effects for Experimental Designs having Non-Repeated Items: Model Specification and Selection

By

Woo-yeol Lee

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

May 11, 2018

Nashville, Tennessee

Approved:

Sonya K. Sterba, Ph.D.

Andrew J. Tomarken, Ph.D.

Kristopher J. Preacher, Ph.D.

David Lubinski, Ph.D.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Researchers in psychology often rely on an experimental design for their studies to infer a relationship between dependent variables and independent variables in a controlled situation. In many psychological studies, the attributes are indirectly measured with multiple trials or items. The multiple trials are often scored as binary variables, such as a force-choice variable, an accuracy variable, and choice in production. For instance, if a researcher wants to measure memory, he or she might present a group of words (i.e., trials) to persons and record the accuracy of the responses in a test phase to measure memory. When multiple items are given to the persons, the dependent variables are *cross-classified* by items and persons because every item is offered to all persons and every person responds to all items.

In an experimental design called the within-subjects design, the dependent variables are measured from a group of persons at every level of an independent variable of interest. A strategy for designing an experiment is to present items repeatedly at every level such that the total number of trials is the number of items times the number of levels of the independent variable. The advantage of using the within-subjects design is to increase the power to detect the experimental condition effect because the variation across items is controlled between the levels. As another strategy, an item used in level A may not appear repeatedly in level B. Items may be randomly selected from a large item pool, or cannot be matched across the levels because of procedural constraints. As an example, researchers might want to control the learning effect or the exposure effect which indicates that persons are affected by exposure to merely repeated items (Gordon & Holyoak, 1983). In this dissertation, a *non-repeated items* (NRI) design is defined as the experimental design where the items are not repeated across the levels of an independent variable.

In many experimental studies in psychology, the main purpose of conducting an experimental study is to examine whether the dependent variables have the same *mean* between the levels of the independent variable (i.e., the experimental condition effect). To test a null hypothesis of equal means between the levels, repeated measures analysis of variance (RM-ANOVA) is most widely used for the within-subjects design. The RM-ANOVA takes individual differences into account in the model because the responses from within a person are more likely to be more correlated than the responses between persons. However, an NRI design has characteristics that the assumptions of the RM-ANOVA may not satisfy: the independence of errors and equal variances across levels. Errors are assumed to be independent of each other when all sources of variation are accounted for by the model. When responses are cross-classified with persons and items as in the NRI design, the independence assumption is likely to be violated if the responses are also influenced by the items. In addition, the items are not repeated across levels in the NRI design. Thus, it is possible that the influence of the items may not be the same between the levels, if it exists.

In some areas such as psycholinguistics, the problems of RM-ANOVA for cross-classified designs are known as "language-as-fixed-effect fallacy," and remedies have been discussed (Baayen et al., 2008; Barr et al., 2013; Clark, 1973; Raaijmakers, 2003; Raaijmakers et al., 1999). For example, Clark (1973) suggested using quasi-F to test the experimental condition effect. It is known that the quasi-F is generally robust to the violations of normality and sphericity (Maxwell & Bray, 1986; Santa, Miller, & Shaw, 1979). However, the limitation of the quasi-F is that this statistic is difficult to compute in cases of missing data (Raaijmakers, 2003; Raaijmakers et al. 1999). Instead, in psycholinguistics, person and item ANOVAs (called $F1/F2$ analysis, respectively) have been conventionally reported (Clark, 1973; Raaijmakers, 2003; Raaijmakers et al., 1999). According to this approach, the mean difference between levels is considered significant only when the person analysis ($F1$) and the item analysis ($F2$) are significant. However, $F1/F2$ analysis produces inflated type I error rates and a deflated power rate, meaning that researchers are more likely to have

false positive results and miss the detection of target effects (e.g., experimental condition effects) (Baayen et al., 2008; Barr et al., 2013; Raaijmakers et al., 1999).

The generalized linear mixed-effect model (GLMM) with crossed random effects has been applied as an alternative method to the ANOVA framework for binary responses.[1] Several papers introduced GLMM to researchers in experimental psychology. Baayen et al. (2008) suggested the linear mixed-effect model with crossed random effects for an alternative to ANOVA in the special issue of *Journal of Memory and Language* (JML). The GLMM has been applied in other areas of psychology, such as social personality (e.g., Judd et al., 2012) and cognitive (e.g., Trippas et al., 2017) areas. The mixed-effect model consists of fixed effects and random effects. The crossed random effects indicate that this model has person random effects and item random effects, which are crossed random effects (instead of nested random effects). Therefore, a response can be explained by the sum of the fixed effects, person random effects, and item random effects in the GLMM with the crossed random effects.

In the case of the NRI design, item random effects are specified as many as the number of levels to consider a full random structure in the GLMM. An advantage of the level-specific item random effect model is that this model enables researchers to test heterogeneity of random effects across levels by comparing this model with a simpler model, that is, an item random effect across the levels. It is also possible to consider person random slopes (to model variability in the experimental condition effect across persons) and level-specific item random effects in a model. Barr et al. (2013) suggested using a random-intercept and random-slope model to consider the full structure of random effects in the NRI design. The meaning of the (person) random intercept is the heterogeneity among persons relative to a baseline level. The (person) random slope refers to the heterogeneity among persons in the magnitude of the fixed effects (i.e., the experimental condition effect). In addition,

[1] Population averaging methods such as robust standard errors or the generalized estimation equation (GEE) can also be used to explain correlated data. The main difference from the GLMM is the sources of correlation, such as item variability, cannot be identified in population averaging methods.

González et al. (2014) considered models that have all possible combinations of the random effect structures that have either person random slopes or level-specific item random effects in the linear mixed effect model for continuous dependent variables. When different items are used for each condition as in the NRI design, it is unclear whether the person-specific effect differs between the conditions even if the variance of the random slope is significantly different from 0 in the presence of the uncontrolled variances of level-specific item random effects.

The inference for fixed effects of an experimental condition effect using the GLMM is trustworthy when random effects are specified correctly (e.g., Barr et al., 2013; Gurka, Edward, & Muller, 2011; Jacqmin-Gadda et al., 2007). Previous studies that employed GLMM for an experimental study focused less on an NRI design compared to other experimental designs, such as fully crossed, Latin-square, or split-plot designs (e.g., Baayen et al., 2008). Thus, the random effects may not be appropriately specified when the existing model specification of the GLMM is applied to the NRI design. However, to our knowledge, the GLMM has not been specified for the full structure of the random effects, illustrated, and evaluated for the NRI design that has binary dependent variables.

1.1 Goal and Outline

The first purpose of this dissertation is to specify, illustrate, and evaluate the GLMM for within-subjects designs with the NRI, in testing a (fixed) experimental condition effect of experimental data. As discussed, selecting a model among the GLMMs that have different kinds of random effects is important for the valid inference of the experimental condition effect. Thus, the second purpose of this dissertation is to show the relative inferential performance of the testing and model selection approaches for the experimental condition effect when an approximate marginal maximum likelihood estimation (Laplace approximation) is used. For the model selection approach, a novel contribution of this dissertation is to derive the asymptotic null distribution for likelihood ratio test (LRT) statistics

based on Self and Liang's (1987) and Zhang and Lin's (2008) results when the variance of the person random slope in the GLMM with the NRI design is tested. Further, the differential performance of information criteria commonly used in practice is evaluated regarding random effect structures in the GLMM, which has not been investigated. In addition, the consequence of the misspecification of the level-specific item random effect in the NRI design is presented to show the necessity of the level-specific item random effect in terms of parameter recovery and the inferential qualities in detecting fixed effects. Results of the current study are expected to provide a guideline for applying the GLMM for the NRI design when binary outcomes are collected.

This dissertation is organized as follows. First, the details of the ANOVA framework and the GLMM are introduced for the NRI design. Second, an illustrative example of the ANOVA framework and the GLMM for the NRI design is provided. Third, the parameter recovery of the newly specified GLMMs and the performance of the hypothesis testing and model comparison approaches are investigated with a simulation study so that the GLMM's effectiveness is verified in simulations commonly encountered in practice. Last, a summary and a discussion of the current study are provided.

CHAPTER 2

METHODS

2.1 Non-Repeated Items Design

As evidence for the claim that an NRI design is common in psychology, a literature review was conducted. One hundred eleven papers were reviewed published in the 146th volume of *Journal of Experimental Psychology: General* in 2017. The frequency of the papers with at least one NRI design study was surveyed. In addition, the frequency of binary dependent variables, the statistical analysis, and the sample sizes for persons and items were also examined. A summary of the literature review is presented in Table 2.1. The survey results showed that 24 papers out of 111 papers (22%) were based on the NRI design. Ten papers included at least one binary dependent variable out of the 24 papers in which the NRI design was used. Sixteen papers used RM-ANOVA only and two papers reported Bayes factors in addition to the RM-ANOVA results, revealing the RM-ANOVA was the most frequently used method for statistical analysis. Six papers used the GLMM framework as a statistical method. Only four papers included both persons and items as random factors, but none of these papers took the heterogeneity of the item random effect into account. The number of persons ranged from 16 to 255, and the number of items per level ranged from 4 to 144.

Throughout this dissertation, an NRI design is explained with the following situation for illustrative purposes: The dependent variable is measured at each level k of a factor that has two levels ($K = 2$). J persons ($j = 1, \ldots, J$) are exposed to both levels. The responses are measured with $n_k = I/2$ items ($i = 1, \ldots, I$) at each level k. The items are not repeated over levels; therefore, the total number of items is I. Taken together, the total number of responses is $N = J \times I$.

Table 2.1: Survey Results

	N
Num. of articles	111
Num. of NRI design	24
Binary outcome	10
RM-ANOVA	16
RM-ANOVA + Bayes factor	2
GLMM	6
Num. of persons (median[range])	41.5[16-255]
Num. of items (median[range])	20[4-144]

2.2 ANOVA Framework

RM-ANOVA model specification

Let y_{jik} be the response of person j and item i at the kth level of an experimental condition. When the data from an NRI design are analyzed using RM-ANOVA, the responses are collapsed over the items at the level of the independent variable. Define a new dependent variable y^*_{jk} for each k,

$$y^*_{jk} = \bar{y}_{j.k} = \frac{\sum_{i=1}^{n_k} y_{jik}}{n_k}. \tag{2.1}$$

For the binary responses, y^*_{jk} is the proportion for person j at the kth level.

The RM-ANOVA (e.g., Maxwell & Delaney, 2004, pp. 533) that uses the proportion is specified as follows:

$$y^*_{jk} = \mu + \alpha_k + \pi_j + (\alpha\pi)_{kj} + e_{jk}, \tag{2.2}$$

where μ denotes the overall mean, α_k denotes the main effect of A (i.e., the experimental condition), π_j denotes the person effect, $(\alpha\pi)_{kj}$ denotes the interaction effect of the experimental condition and the person, and e_{jk} denotes the error. By adding to Equation 2.2 the constraint that the person effect does not interact with the experimental condition effect,

the model becomes

$$y^*_{jk} = \mu + \alpha_k + \pi_j + e_{jk}. \tag{2.3}$$

In this dissertation, the constraint, $(\alpha\pi)_{kj} = 0$, is assumed in the RM-ANOVA (i.e., the $F1$ analysis) as in many analyses in practice. The random components π_j and e_{jk} follow the normal distribution $N(0, \sigma^2_\pi)$ and $N(0, \sigma^2_e)$, respectively.

Hypothesis testing in RM-ANOVA

The idea of ANOVA is to compare the variances (mean squares) induced by the independent variable against the variance from some error term. To test the main effect of A (i.e., the experimental condition), the ratio of two variances, MS_A and MS_E, are compared using an F-statistic:

$$F_{K-1,(K-1)(J-1)} = \frac{MS_A}{MS_E} = \frac{MS_A}{MS_{A\times P}}, \tag{2.4}$$

where $MS_A = \frac{J\sum_{k=1}^{K}(\bar{y}_{.k}-\bar{y}_{..})^2}{K-1}$, $MS_{A\times P} = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K}(y^*_{jk}-\bar{y}_{j.}-\bar{y}_{.k}+\bar{y}_{..})^2}{(K-1)(J-1)}$, $\bar{y}_{.k} = \frac{\sum_{j=1}^{J} y^*_{jk}}{J}$, $\bar{y}_{j.} = \frac{\sum_{k=1}^{K} y^*_{jk}}{K}$, and $\bar{y}_{..} = \frac{\sum_{k=1}^{K}\sum_{j=1}^{J} y^*_{jk}}{KJ}$.

Assumptions, limitations, and alternatives

RM-ANOVA has the following assumptions:

1. Independence: The responses are conditionally independent of one another.

2. Normality: The errors follow the normal distribution.

3. Sphericity: The population variances of all difference scores are equal.

When the responses are obtained from an NRI design, using the RM-ANOVA may entail assumption violation. First, the RM-ANOVA is vulnerable to the violation of independence assumptions due to the cross-classification by persons and items. In the NRI design, the responses of a person are collapsed over items for each level of an experimental condition. The independence assumption is met if the responses are explained by only one source of variation, which is expressed as an error. Aggregating responses over items

might be a solution to control multiple sources of variability, treating the item effect as a controlling factor. If variability in the item effect is large in the population, across sample to sample, the variability of the dependent variable will also be large. When the RM-ANOVA is used, sampling variability will be ignored and the inference about the mean difference will also be inaccurate. Clark (1973) showed that the F-statistic to test experimental condition effects is larger than it should be if we ignore the item variability.

In some subfields in experimental psychology, researchers try to reduce the sampling error by obtaining responses hundreds to thousands of times from a person and averaging them (Luck, 2005). The standard error of the averaged response is reduced by the squared quantity of the number of items at each level of an experimental condition. However, increasing the number of items may not always be the solution to reduce the sampling error. When the scale of the dependent variable is binary, the variance of responses depends on the mean response of the dependent variable. For example, the dependent variable includes a larger sampling error when the mean response is close to 0.5 on the proportion scale than it is far from 0.5. The item effect is also expected to vary around the mean response. Because it is intended to have differences in the mean response between the levels in an experimental condition by the researchers, ignoring item effect results in heterogeneity in responses on the proportion scale as well as incorrect estimation of standard errors.

Quasi-F uses the original dependent variable instead of averaging the responses over items. It adjusts the numerator and the denominator of the F-statistic to have the equal expected value of the variances when the effect of the independent variable is 0 (Clark, 1973). The disadvantage of Quasi-F is that the calculation is complicated. Similar to other types of ANOVA, Quasi-F cannot be calculated in the presence of non-balanced group sizes and missing data. More practically, $F1$ (the by-person F-test) $F2$ (the by-item F-test) analysis has been used more widely. The null hypothesis that the main effect of A does not exist is rejected when the $F1$ and $F2$ analyses reveal a significant result. The $F1$ analysis is the same as in the RM-ANOVA. The $F2$ analysis is performed after the responses are

collapsed over persons. For the $F2$ analysis, ANOVA can be implemented.

Second, using RM-ANOVA violates the normality assumption if the dependent variable is binary. When a binary dependent variable is collapsed over items in an NRI design, the proportion is reported as a dependent variable, and this variable is often treated as continuous. Treating proportion as a continuous variable should be avoided for the following reasons. One reason is that the standard error is incorrect because of the non-linear relationship between the dependent and independent variables (Jaeger, 2008). When the dependent variable is associated with the independent variable with a non-linear relationship, the confidence interval is not symmetric in a linear model such as ANOVA. Thus, if we treat the proportion as a continuous variable, the inference about the experimental condition effect may not be correct. Another reason is that the responses are bounded between 0 and 1. When we construct a confidence interval for the fixed effect, it is possible that the confidence interval contains a range less than zero or greater than one. Then, the confidence interval loses interpretability. A logistic regression which is a generalized linear model can be used to avoid the two problems when the dependent variable is binary.

Third, the sphericity assumption indicates that all the variances of the level differences are equal, while the compound symmetry assumption requires that the variances and covariances of the different levels of the repeated-measures factor are homogeneous. Compound symmetry is a sufficient but not necessary condition. That is, if compound symmetry is not satisfied, then sphericity is not met. In an NRI design, the sphericity may be violated if the variability in the item effect is heterogenous across the levels of an experimental condition. If the variances of the item effect differs across the levels of an experimental condition, the variability in the dependent variable averaged over items is also influenced by the heterogenous item effect. The difference scores would also have different variances when the variances of the dependent variable differ across the levels.

2.3 GLMM Specification for an NRI Design

Model specification

The GLMM specification with a full random structure for an NRI design can be written as

$$\text{logit}[P(y_{ji[k]} = 1 | s_{0j}, s_{1j}, w_{i[k]})] = \beta_0 + \beta_1 x_{ji} + s_{0j} + s_{1j} x_{ji} + w_{i[k]}, \tag{2.5}$$

where j is an index for a person ($j = 1, \ldots, J$); i is an index for an item ($i = 1, \ldots, I$); k is an index for a level of the experimental condition ($k = 1, \ldots, K$); $i[k]$ represents that an item i is nested with a level k; $y_{ji[k]}$ is the response from person j and item i at the kth level of a factor; x_{ji} is the independent variable (i.e., the experimental condition) for the levels of the factor, respectively; β_0 and β_1 are the fixed effects for the intercept and the slope, respectively; s_{0j} and s_{1j} are the person random effects for the intercept and the slope, respectively (called the person random intercept and the person random slope, respectively, in this dissertation); and $w_{i[k]}$ is the item random effect at the kth level.

The random effects are assumed to be distributed as follows:

$$\begin{bmatrix} s_{0j} \\ s_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_0^2 & \\ \tau_{10} & \tau_1^2 \end{bmatrix} \right) \tag{2.6}$$

and

$$w_{i[1]} \sim N(0, \omega_{[1]}^2); w_{i[2]} \sim N(0, \omega_{[2]}^2). \tag{2.7}$$

Estimation

Maximum likelihood (ML) estimation is a widely used method for estimating the parameters of mixed-effect models. The ML estimation for binary data is challenging because there is no closed form for the integral involved. When the random effects are nested, the integrals are also nested, keeping the computational burden low (e.g., Rabe-Hesketh et al., 2005). However, the computational burden worsens when a model contains two types of

random effects such as person and item random effects because the likelihood requires the integration of person random effects and item random effects.

Several approximation methods have been proposed to overcome the challenge of the high dimensional integration. One way to approximate likelihood is a quasi-likelihood approach including penalized quasi-likelihood (PQL; Breslow & Clayton, 1993) and marginal quasi-likelihood (MQL; Goldstein, 1991). Quasi-likelihood is simple to implement, but it is known that the estimates are biased downward (Rodiguez & Goldman, 1995; Goldstein & Rasbash, 1996). Although several improved methods have been proposed such as bias-corrected PQL (Breslow & Lin 1995) and PQL with the second order improvement (PQL-2, Goldstein & Rasbash, 1996), MQL and PQL yield biased estimates for binary responses, especially when the cluster size is small (Rodríguez & Goldman, 2001, Raudenbush et al. 2000).

The distribution of random effects can be simulated via data generation rather than mathematical derivations. Monte Carlo expectation-maximization (MCEM) algorithm (Wei & Tanner, 1990) maximizes the likelihood by alternating the two following processes. In the first E-step, the random effects are sampled from their posterior distribution using a Gibbs sampler. In the second M-step, the parameters of the GLMM are estimated. The alternating imputation-posterior (AIP) algorithm (Clayton & Rasbash, 1999) with adaptive quadrature (Cho & Rabe-Hesketh, 2011) alternately samples item random effects when given person random effects and person random effects when given item random effects (in the I-step), and involves the exact ML estimates using adaptive quadrature (in the P-step). The MCEM and AIP algorithms provide approximate ML estimates. However, the two algorithms are computationally expensive because they require many draws of the random effects and the estimation of the exact ML estimates.

In Bayes estimation, the parameters are sampled from the posterior distribution. The posterior distribution is generated by combining the prior distribution of each parameter with the data through the Markov Chain Monte Carlo (MCMC) method. The MCMC

method iteratively generates samples for parameters via sampling techniques such as Gibbs sampling (Geman & Geman, 1984) or the Metropolis-Hastings algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953). The Bayes estimates show little difference from the ML estimates if prior information is weak and datasets are highly informative. However, if the number of higher-level units is small, the estimates from the Bayes estimation are highly dependent on the prior distribution (Lambert, 2006).

In this dissertation, the `glmer` function (Bates et al., 2015) in `R` (R Core Team, 2017) is chosen to estimate model parameters, considering computational efficiency and accuracy. The `glmer` function relies on the ML estimation implementing Laplace approximation for binary outcomes, which is a special case with one quadrature point in the Gauss Hermite quadrature method in the `glmer` function. Therefore, the efficiency is relatively good because the quadrature point does not increase when the number of random effects increases. It has been shown that Laplace approximation provides accurate estimates for the GLMM with the crossed random effects for binary responses unless the cluster sizes (e.g., the number of items and persons) are small (e.g., 10 items), and the intraclass correlation[1] is high (e.g., Cho & Rabe-Hesketh, 2011; Joe, 2008).

Model-building strategy

The main interest is often in the inference about an experimental condition effect. However, when the random effects are not correctly specified, the standard error of the estimates is distorted (Barr et al., 2013; Gurka, Edwards, & Muller, 2011; Jacqmin-Gadda et al., 2007). Thus, it is important to find a model with the correct specification of the random effects before the fixed effects are tested.[2] In an NRI design, two kinds of variance testing

[1] See Cho and Rabe-Hesketh (2011) for the derivation of ICC in a GLMM with crossed random effects. ICCs are defined based on a latent response formulation. Let y^*_{ji} be a latent response such that the observed response is 1 if $y^*_{ji} > 0$ and 0 otherwise. For Model 1 as example, an ICC for persons is defined as the correlation among latent responses for the same person, conditional on the items, $Corr(y^*_{ji}, y^*_{ji'}|w_i, w_{i'}) = \frac{\tau_0}{\tau+\frac{\pi^2}{3}}$. Likewise, the ICC for items is defined as $Corr(y^*_{ji}, y^*_{j'i}|s_j, s_{j'}) = \frac{\omega}{\omega+\frac{\pi^2}{3}}$.

[2] For non-experimental data, it is not always possible to find the correct specification of the random effect because all the relevant predictors may not be included in the model. One can compare the estimates and the standard error of the fixed effects across models that have different random effects, as well as the inference about the random effects.

or model comparisons are required to find the best-fitting random effect specification. The first kind of variance testing or model comparison is to investigate whether the variance of the item random effect is homogenous across levels of an experimental condition. If the item random effect is homogenous, it is more efficient to model one parameter for the common variance of the item random effect than to model level-specific variances of item random effects. The second kind of variance testing or model comparison is to choose the number of person random effects. This is often done by comparing the person random intercept-only model with the person random slope model in the GLMM.

Consider a case in which there is an experimental condition that has the two levels in an NRI design. Four models that have different random effects can be considered.

- Model 1: Person random intercept only + homogenous item random effect

$$\text{logit}[P(y_{ji[k]} = 1|s_{0j}, w_i)] = \beta_0 + \beta_1 x_{ji} + s_{0j} + w_i, \tag{2.8}$$

  where $s_{0j} \sim (0, \tau_0^2)$ and $w_i \sim N(0, \omega^2)$.

- Model 2: Person random slope + homogenous item random effect

$$\text{logit}[P(y_{ji[k]} = 1|s_{0j}, s_{1j}, w_i)] = \beta_0 + \beta_1 x_{ji} + s_{0j} + s_{1j} x_{ji} + w_i, \tag{2.9}$$

  where $s_{1j} \sim N(0, \tau_1^2)$.

- Model 3: Person random intercept only + level-specific item random effects

$$\text{logit}[P(y_{ji[k]} = 1|s_{0j}, w_{i[k]})] = \beta_0 + \beta_1 x_{ji} + s_{0j} + w_{i[k]}, \tag{2.10}$$

  where $w_{i[k]} \sim N(0, \omega_{[k]}^2)$.

- Model 4: Person random slope + level-specific item random effects (Equation 2.5)

The best-fitting model can be found through forward or backward strategies (Barr et al., 2013). The forward strategy starts with a simple model and adds model complexity whereas the backward strategy begins with the most complex model and removes complexity. Figure 2.1 shows an example of the model building strategy for an NRI design. The final model is selected through a sequential procedure. In Step 1, the homogeneity of the item random effects is tested. If the forward strategy is used, Model 1 and Model 3 which have a person random intercept are compared. Under the backward strategy, Model 2 and Model 4 which have a person random intercept and a person random slope are compared. In Step 2, the null hypothesis of the zero variance of the person random slope is tested. Depending on the model selected in Step 1, two models that have homogenous item random effects (i.e., Model 1 and Model 2) or level-specific item random effects (i.e., Model 3 and Model 4) can be compared to explore whether the person random slope model must be specified.

Hypothesis testing and model comparison methods

In this section, hypothesis testing and model comparison methods for an NRI design are discussed. To simplify the discussion, a single condition that has the two levels mentioned in the previous section was considered. According to the model building strategy we discussed earlier, selecting a model in terms of the random effects precedes the inference about fixed effects. When the GLMM is applied for an NRI design, there are two types of testing for variances of random effects: (a) testing the variance of level-specific item random effects and (b) testing the variance of a person random slope. Below, each type is described in detail.

Testing variance of item random effects. The LRT can be used when a model with the level-specific item random effect is compared with a model with the homogenous item random effect (i.e., Model 1 vs. Model 3; Model 2 vs. Model 4). The LRT compares the change in deviance (i.e., -2 times the log-likelihood) between the null model and the

alternative model. The test statistic is expressed as follows:

$$T_{LR} = -2[l(\hat{\theta}_0) - l(\hat{\theta}_1)], \tag{2.11}$$

where θ_0 is the parameter set of the null model, θ_1 is the parameter set of the alternative model, $l(\hat{\theta}_0)$ is the log-likelihood of the null model, and $l(\hat{\theta}_1)$ is the log-likelihood of the alternative model. Because the parameter space is not constrained at zero variance under the null and alternative hypotheses, the appropriate reference distribution for the LRT is a chi-square distribution with the degree of freedom of the difference in the number of parameters. That is, the test statistic is compared to $\chi^2(1)$.

Testing the variance of a person random slope effect. Testing whether or not there is between-person variation for an experimental condition effect is equivalent to testing the variance of a person random slope (i.e., τ_1^2) equal to 0. Suppose that an observation $y_{ji[k]}$ is explained by q random effects in the null model. Null and alternative hypotheses can be set as follows:

$$H_0 : D = \begin{bmatrix} D_{11} & 0 \\ 0 & 0 \end{bmatrix} \qquad \text{versus} \qquad H_1 : D = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix}, \tag{2.12}$$

where D_1 is a q-by-q positive-definite matrix. We can test H_0: D_{11} is positive definite, $D_{12} = D_{21} = 0$, and $D_{22} = 0$ versus H_1: D is positive definite. For the comparison between Model 1 and Model 2, as an example, D_{11} is expressed as

$$\begin{bmatrix} \tau_0^2 & 0 \\ 0 & \omega^2 \end{bmatrix} \tag{2.13}$$

with a 2-by-2 diagonal matrix, and D is expressed as

$$\begin{bmatrix} \tau_0^2 & 0 & \tau_{01} \\ 0 & \omega^2 & 0 \\ \tau_{10} & 0 & \tau_1^2 \end{bmatrix} \tag{2.14}$$

with a 3-by-3 positive-definite matrix.

In general, the LRT statistic (T_{LR}) follows a chi-square distribution with degree of freedom equal to the difference between the two nested models (e.g., Casella & Berger, 2002). However, using this degree of freedom for zero-variance testing is too conservative because of the boundary of the model parameter space. Suppose the sampling distribution of the LRT statistic is obtained through numerous experiments from a population. If all the variance estimates are greater than zero across all experiments, the LRT statistic follows the standard chi-square distribution. When a zero-variance is tested, however, the variance estimates are either zero or positive values from sample to sample. The sampling distribution of the LRT statistic differs between the cases where the variance estimates are zero and positive values. Self and Liang (1987) formulated the asymptotic null distribution of the LRT statistic (T_{LR}) for testing a null hypothesis when the true parameter is possibly on the boundary of the model parameter space. Stram and Lee (1994[3], 1995) applied Self and Liang's (1987) results to investigate the asymptotic null distribution of the LRT statistic for variance testing in linear mixed models. Because Self and Liang's (1987) results are for a general parametric model, Zhang and Lin (2008) and Molenbergh and Verbeke (2007) extended use of the mixture of chi-square distribution for the variance testing in the GLMM. However, Zhang and Lin (2008) and Molenbergh and Verbeke (2007) did not provide the derivation of the mixture chi-square distribution for GLMM with crossed random effects.

In the below, the derivation for the GLMM with crossed random effects is provided. Denote by $\theta = [\beta, \psi]'$, the vector of the fixed effects and variance-covariance parameters

[3]Stram and Lee (1994) incorrectly specified constraints in their derivation, which was noted by Stram and Lee (1995).

of random effects in a model, respectively. Chernoff (1954) formulated the asymptotic null distribution of the LRT statistic (T_{LR}) for testing

$$H_0 : \theta_0 \in \Omega_0 \quad \text{versus} \quad H_1 : \theta_0 \in \Omega_1,$$

where the true value θ_0 is on the boundary of the model parameter space Ω. Assume that the parameter space Ω_0 under H_0 and the parameter space Ω_1 under H_1 can be approximated at θ_0 by cones C_{Ω_0} and C_{Ω_1} with vertex θ_0, respectively. Under suitable regularity conditions, Self and Liang (1987, p. 607) showed that the asymptotic representation of the T_{LR} may be written as

$$\inf_{\psi \in C_{\Omega_0} - \theta_0} \{(U-\theta)' I(\theta_0)(U-\theta)\} - \inf_{\psi \in C_{\Omega} - \theta_0} \{(U-\theta)' I(\theta_0)(U-\theta)\}, \tag{2.15}$$

where C_Ω is the cone approximating Ω with a vertex at θ_0, $C_\Omega - \theta_0$ and $C_{\Omega_0} - \theta_0$ are translated cones of C_Ω and C_{Ω_0} such that their vertices are the origin, $I(\theta_0)$ is the Fisher information matrix at θ_0, and U is a random vector (assumed to follow $N(0, I^{-1}(\theta_0))$.

For the comparison of Model 1 and Model 2, the parameter vector θ can be partitioned into three components, θ_1, θ_2, and θ_3. Denote the fixed effects and the unique elements of D_{11} by $\theta_1 = (\beta_0, \beta_1, \tau_0^2, \omega^2)$. In addition, denote by $\theta_2 = D_{12} = D_{21} = \tau_{01} = \tau_{10}$ and $\theta_3 = D_{22} = \tau_1^2$. Under H_0, the translated approximating cone at θ_0 is $C_{\Omega_0} - \theta_0 = R^4 \times \{0\} \times \{0\}$. Under $H_0 \cup H_1$, D_{11} is positive definite and D is positive semidefinite, which is equivalent to D_{11} is positive definite and $D_{22} - D_{12}' D_{11}^{-1} D_{12} \geq 0$. Because the boundary defined by $D_{22} - D_{12}' D_{11}^{-1} D_{12} = 0$ is a smooth surface for any given positive definite D_{11}, the translated approximating cone at θ_0 under $H_0 \cup H_1$ is $C_\Omega - \theta_0 = R^4 \times R^1 \times [0, \infty)$.

Decompose U and $I^{-1}(\theta_0)$ (in Equation 2.15) into $U = [U_1, U_2, U_3]'$ and $I_{..}$ for θ_1, θ_2, and θ_3, respectively. Note that U_1 is a 2×1 random vector. Zhang and Lin (2008) showed

that

$$\inf_{\psi \in C_{\Omega_0} - \theta_0} \{(U-\theta)' I(\theta_0)(U-\theta)\} = [U_2, U_3] \begin{bmatrix} I_{22} & I_{23} \\ I_{32} & I_{33} \end{bmatrix}^{-1} \begin{bmatrix} U_2 \\ U_3 \end{bmatrix} \tag{2.16}$$

and

$$\inf_{\psi \in C_{\Omega} - \theta_0} \{(U-\theta)' I(\theta_0)(U-\theta)\} = I_{33} U_3^2 I(U_3 \leq 0). \tag{2.17}$$

Because of $U = [U_1, U_2, U_3]^{'} \sim N(0, I^{-1}(\theta_0))$, the distribution difference (Equation 2.15) is a 50% and 50% mixture of $\chi^2(1)$ and $\chi^2(2)$. Thus, the significance level α can be compared to the LRT p-value

$$0.5P[\chi^2(1) \geq T_{obs}] + 0.5P[\chi^2(2) \geq T_{obs}], \tag{2.18}$$

where T_{obs} is the observed T_{LR}. The p-value based on the mixture of $\chi^2(1)$ and $\chi^2(2)$ is always smaller than incorrect p-value (i.e., $P[\chi^2(2) \geq T_{obs}]$). Thus, the decision based on the incorrect p-value is conservative.

To illustrate an example of the mixture chi-square distribution, a histogram of the deviance between Model 1 and Model 2 is presented in Figure 2.2. Five hundred datasets were simulated from Model 1 with 100 persons and 50 items. The parameters used for the simulation were as follows: [β_0, β_1, τ_0^2, ω^2] = [1, 0.5, 0.5, 0.5]. Model 1 and Model 2 were fitted to each dataset and the deviance was calculated. As shown in Figure 2.2, the distribution of the deviances passes between $\chi^2(1)$ and $\chi^2(2)$.

For the comparison between Model 1 and Model 2, $\chi^2(1)$ and $\chi^2(2)$ are used as the reference distribution to test $\tau_1^2 = 0$ and $\tau_{01} = 0$.

For the comparison between Model 3 and Model 4, the mixture chi-square distribution is $0.5(\chi^2(2) + \chi^2(1))$ in the same way.

Information criteria. Information criteria can be considered to compare models with different structures of random effects along with the LRT. In general, information criteria consist of the deviance and the penalizing term for model complexity. The best-fitting

model is the one with the smallest value among the competing models. The Akaike information criterion (AIC; Akaike, 1973) includes only the number of parameters in the penalizing term. The marginal AIC based on the marginal likelihood is a commonly used model selection method for linear mixed-effect models (e.g., Greven & Kneib, 2010), and some extensions have been suggested for the GLMM (Saefken et al., 2014). In this dissertation, the marginal AIC was chosen as calculated in the `glmer` function:

$$AIC = -2l(\hat{\theta}) + 2(P+Q), \tag{2.19}$$

where P is the number of fixed parameters, and Q is the number of unique variance and covariance parameters of the random effects. The Bayesian information criterion (BIC; Schwarz, 1978) includes the number of parameters and the sample size in the penalty term:

$$BIC = -2l(\hat{\theta}) + ln(N)(P+Q). \tag{2.20}$$

N in the BIC calculation is the total sample size calculated by the number of persons $\times$ the number of items for the GLMM with crossed random effect models (see the derivation of N in Cho and De Boeck, 2018).[4] When the true model is infinitely dimensional, the AIC is asymptotically less efficient in the sense that the criterion selects a model with nearly minimum risk; however, when the candidate models contain a true model with a finite dimension, the BIC can select the true model consistently (e.g., Burnham & Anderson, 2002).

Testing fixed effects. The fixed effects can be tested based on the final model based on a model selection result regarding random effects. The Wald test is commonly used for hypothesis testing of the fixed effect because of the convenience of being able to obtain the result based on the model being evaluated (e.g., Baayen et al., 2008). The following null and alternative hypotheses regarding the experimental condition effect β_1 in Models $1-4$

[4]The BIC value from the `glmer` function is based on the total sample size.

can be tested using the Wald test:

$$H_0 : \beta_1 = 0 \qquad \text{versus} \qquad H_1 : \beta_1 \neq 0. \tag{2.21}$$

The test statistic is $T_{Wald} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$, where $\hat{\beta}_1$ is the estimate, and $SE(\hat{\beta}_1)$ is the standard error of the estimate. If the sample size is large enough, the test statistic is assumed to follow a normal distribution.

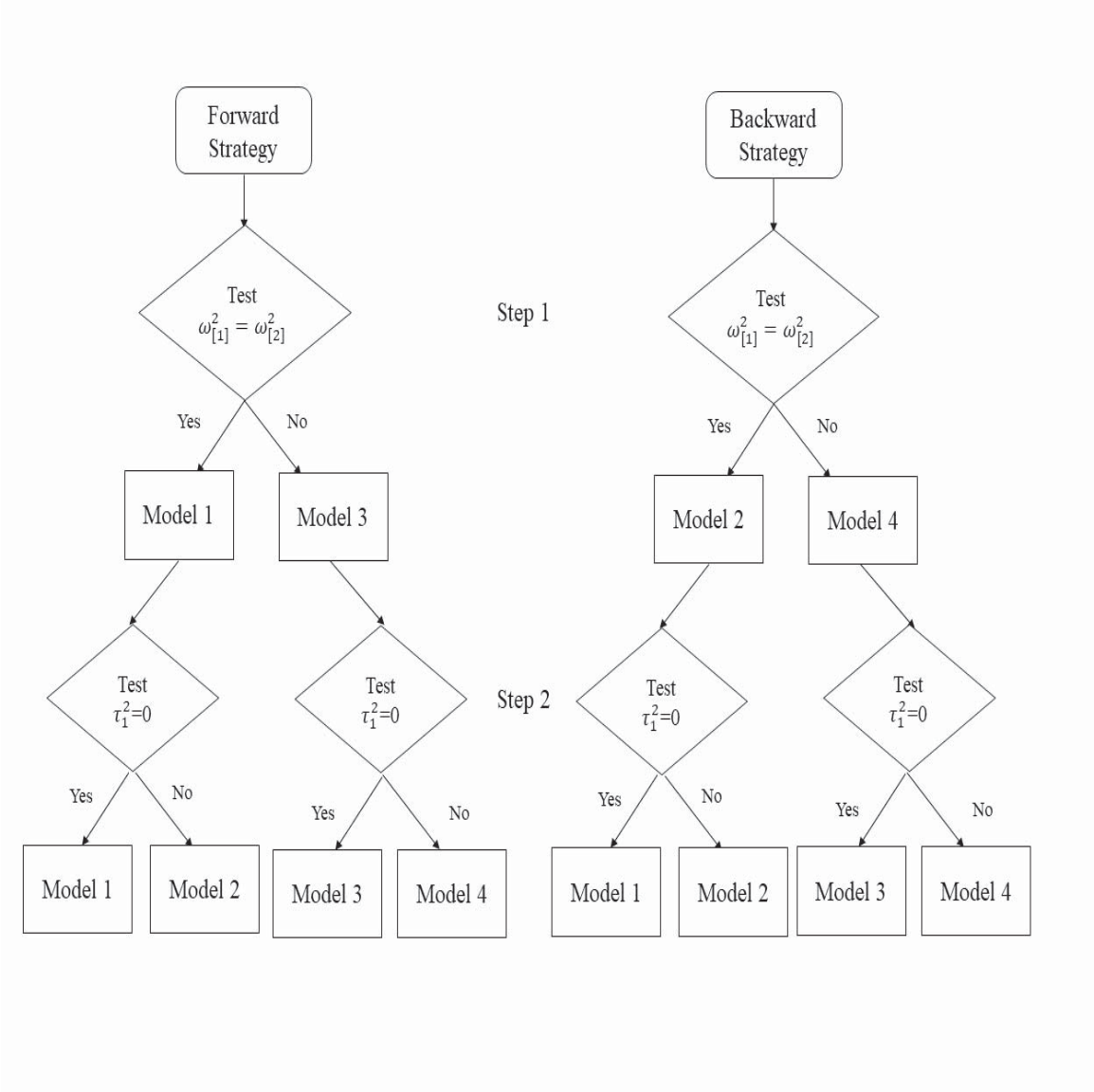

Figure 2.1: Tree diagram for the model building strategy

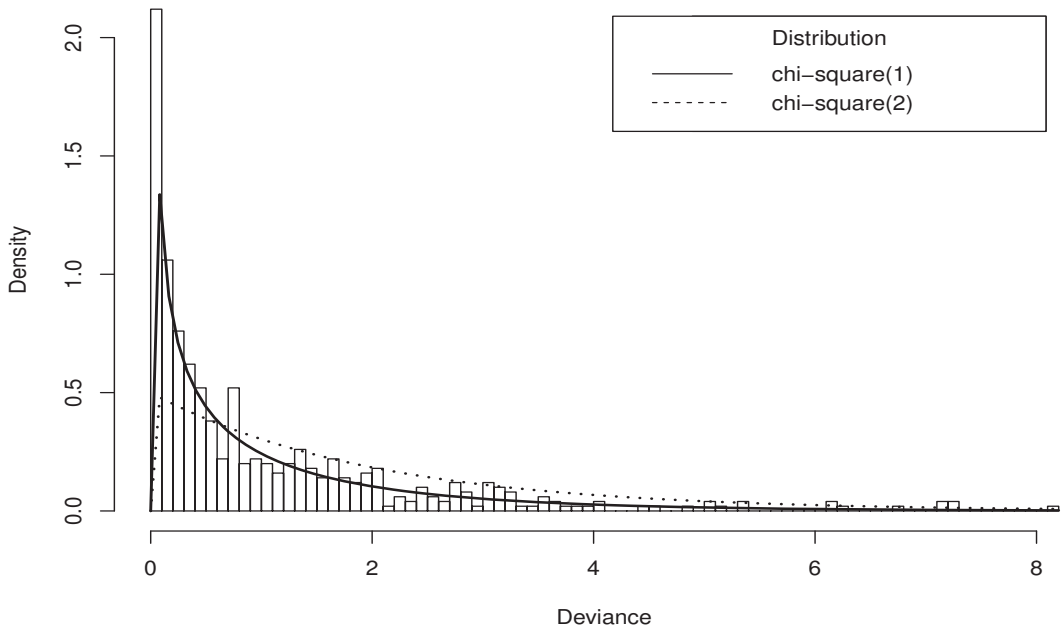

Figure 2.2: Histogram of mixture chi-square distribution

CHAPTER 3

EMPIRICAL STUDY

In this section, $F1/F2$ analysis and GLMMs with crossed random effects (Models $1-4$) with the model building strategy are illustrated and compared. The R code to implement $F1/F2$ analysis and GLMMs with crossed random effects is shown in Appendix.

3.1 Data Description

Nosek and Banaji (2001) designed a method for measuring the effect of implicit social cognition. The authors hypothesized that a human responds to words faster and more accurately if the valence of the context matches the words than when it does not. The raw dataset can be accessed via the Dataverse website (https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/11157). As a task, 26 persons were asked to respond to a word if the valence (e.g., good and bad) or the category (e.g., fruit and bugs) matched the context where one of the valences and one of the attributes were presented. The target words had one of the valences or were from one of the categories. Across 16 blocks that each has 40 trials, the words were presented in half of the trials in the congruent context and in the incongruent context. The authors reported behavioral responses such as the accuracy (d-prime) and response times.

In this study, the original study was reframed in the following steps for illustrative purposes. First, the dataset from Experiment 2A was chosen. Second, we chose the accuracy data coded as 0 for an incorrect response and 1 for a correct response. Only trials where the target words were presented in the congruent context were included. Therefore, all the correct responses were 'hit' responses and all the incorrect responses were 'miss' responses. Third, the responses in the fruit category were compared with the responses in the bugs category when the categories were presented with the good valence in the context.

Therefore, the fruit category always matched the valence while the bugs category always mismatched the valence. Fourth, the responses were included only in the single-category condition where the distracter items were also in one of the four (fruit, bugs, good, and bad) types. In the generic contexts condition, in contrast, the distracter items were not taken from these types (e.g., table, potato, and car). As a result, the responses from 40 trials out of 640 (16 blocks $\times$ 40 trials) were included in the analysis per person (640 $\times\frac{1}{2}$ [target items only] $\times\frac{1}{2}$ [fruit and bugs categories only] $\times\frac{1}{2}$ [single-category condition only] $\times\frac{1}{2}$ [congruent context only] = 40). The total number of responses was 1040 (20 trials $\times$ 2 levels $\times$ 26 persons).

The data structure of the study is presented in Table 3.1. The stimuli category was used as an independent variable, and the accuracy was compared between the two levels ('fruit' and 'bugs') of the category factor (i.e., the experimental condition). All persons were exposed to both categories such that the category was a within-subjects factor. There were 24 target items at each category level. The experiment had an NRI design in which the items at the fruit level were not overlapped with the items in the bugs level, as shown in Table 3.1. Twenty items were randomly presented to a person for each category without replacement within a block. The primary goal of the analysis was to test the experimental condition effect (i.e., different kinds of categories), while accounting for all possible random effects in the design.

3.2 $F1/F2$ Analysis

In Nosek and Banaji (2001), the authors conducted only an $F1$ analysis with d-prime as a dependent variable and found a significant experimental condition effect. However, the result was not reported with any assumption-checking procedure. In the ANOVA framework, $F1/F2$ can be considered to control the item effect. For the by-person ($F1$) analysis, the mean responses were computed across items within a level per each person. A total of 52 responses (2 levels $\times$ 26 persons) were obtained from 24 target items. The mean pro-

Table 3.1: Data Structure of Nosek and Banaji (2001)

	Category	
Subject	Fruit	Bugs
1	10, 15, 19,..., 1	14, 13, 20,... ,7
2	1, 2, 13,..., 14	23, 16, 10,...,19
⋮	⋮	⋮
26	6, 16, 13,..., 3	19, 6, 13,...,17

Note. The numbers denote the item ID.

Category	ItemID	Item
Fruit	fruit1	apple
Fruit	fruit2	apricot
Fruit	⋮	⋮
Fruit	fruit24	watermelon
Bugs	bugs1	aphid
Bugs	bugs2	ants
Bugs	⋮	⋮
Bugs	bugs24	wasp

portion was 0.927 ($SD = 0.079$) at the fruit level and 0.779 ($SD = 0.137$) at the bugs level. An RM-ANOVA yielded a significant effect of the category, $F(1,25) = 33.74$, $p < .001$. For the by-item ($F2$) analysis, the responses were averaged over persons. The number of responses was 24 at each level. The mean proportion was 0.923 ($SD = 0.089$) and 0.778 ($SD = 0.105$) at the fruit level and at the bugs level, respectively. For an NRI design, because items are not overlapped over levels, one-way ANOVA is the proper analysis. A one-way ANOVA revealed that the mean responses differed significantly between the fruit level and the bugs level, $F(1,46) = 26.61$, $p < .001$. The category effect was significant according to the $F1$ and $F2$ analyses.

3.3 The GLMM for an NRI Design

Models $1-4$ we specified earlier were fit to the same data we used for the $F1$ and $F2$ analyses. Table 3.2 shows the number of parameters, the specification of random effects, the deviance, the AIC, and the BIC of the models that correspond to the experimental design. For the random effect specifications, the existence of the heterogeneity of the item random effect and person random slope were tested. With a forward strategy, item heterogeneity was tested assuming only the person random effect exists by comparing Model 1

with Model 3. **T**he model with item heterogeneity (Model 3) was significant according to the LRT, $\chi^2(1) = 3.964,\ p < .05$. Next, given the level-specific item random intercept model (Model 3), the model was compared with the person random slope model (Model 4). The person random slope model (Model 3) was selected based on the LRT because the observed chi-square value (0.576) was smaller than the critical value on a mixture chi-square distribution, $0.5(\chi^2(1,.95) + \chi^2(2,.95)) = 4.916$. With a backward strategy, item heterogeneity was tested assuming the person random slope effect exists by comparing Model 2 with Model 4. **T**he model with item heterogeneity (Model 4) was significant according to the LRT, $\chi^2(1) = 4.050,\ p < .05$. Based on the result, the model with the person random slope and level-specific item random effect (Model 4) was compared with the model with the level-specific item random intercept (Model 3). Model 3 showed a significantly better model fit compared with Model 4 (observed chi-square value=0.576; the critical value=$0.5(\chi^2(1,.95) + \chi^2(2,.95)) = 4.916$). To summarize, the forward and backward strategies both results in Model 3 as the best-fitting model regarding random effect structures. **T**he AIC also indicated that the person random intercept-only with level-specific item random effect (Model 3) is the best-fitting model. However, the BIC selected the person random intercept-only with homogenous item random effect (Model 1) for the best-fitting model.

Table 3.3 shows the estimates for Models $1-4$. The category fixed effect, β_1, was tested based on the model selected by the AIC, the BIC, and the LRT, that is, the person random slope and the homogenous item effect model (Model 3). The category fixed effect was significant by a Wald test, $T_{Wald} = 5.070,\ p < .001$. With an effect coding, β_0 represents the overall mean, and β_1 represents the mean difference of a category effect from the overall mean. The fixed effect at the fruit and bugs levels can be found with $\hat{\beta}_0 + 0.5(\hat{\beta}_1)$ and $\hat{\beta}_0 - 0.5(\hat{\beta}_1)$, respectively. The coefficient of the fixed effect was 2.960 at the fruit level and 1.368 at the bugs level on the logit scale. Those coefficients can be transformed into 0.951 and 0.797 on the probability scale (calculated based on $\frac{1}{1+exp[-\hat{\beta}]}$), respectively.

Table 3.2: The number of parameters, the specification of random effects, the AIC, the BIC, and the deviance of the GLMMs (up) and the LRT procedure (bottom)

Model	#parameters	p.int	p.slp	i.homo	i.hetero	Deviance	AIC	BIC
Model 1	4	✓		✓		800.2	808.2 (2)	827.9 (1)
Model 2	6	✓	✓	✓		799.7	811.7 (4)	841.3 (3)
Model 3	5	✓			✓	796.2	806.2 (1)	830.9 (2)
Model 4	7	✓	✓		✓	795.6	809.6 (3)	844.2 (4)

Note. p.int: person random intercept; p.slp: person random slope;
i.int: homogenous item random effect; i.hetero: level-specific item random effect
The numbers in the parentheses indicate the rank of the AIC and the BIC values across the models.

Test	Observed χ^2	DF	Critical value	Result
[Forward strategy]				
Model1 vs. Model3	3.964	1	3.841	Select Model3
Model3 vs. Model4	0.576	2	4.916	Select Model3
[Backward strategy]				
Model2 vs. Model4	4.050	1	3.841	Select Model4
Model3 vs. Model4	0.576	2	4.916	Select Model3

Last, the fixed effect testing results were compared among the four models. The estimate of the category effect was comparable across all models (ranging from 1.380 to 1.671), and all models yielded a significant result because of the large effect size, ps $<.0001$. However, the standard errors of the fixed effects varied across the models. The standard error of the fixed effects in Model 1 was smaller than that of Model 3 with the level-specific item random effects. In contrast, the standard error increased when the person random slope was redundantly specified as in Model 4.

3.4 Comparisons of the Results

A significant category effect was found with the $F1/F2$ analysis and the GLMM. In addition, the coefficient from the GLMM on the logit scale was comparable to the mean proportion when the scale was transformed (probability based on the GLMM results: [0.951, 0.797]; the mean proportion: [0.927, 0.779] for the fruit level and the bugs level, respec-

Table 3.3: The estimates of the fixed effects and the random effects of the GLMMs

	Model 1		Model 2		Model 3		Model 4	
	EST	SE	EST	SE	EST	SE	EST	SE
Fixed effect								
Overall mean[β_0]	2.092	0.180	2.127	0.194	2.164	0.201	2.203	0.215
Category[β_1]	1.380	0.248	1.443	0.306	1.592	0.314	1.671	0.365
Random effect								
Person								
Var(Intercept[τ_0^2])	0.364		0.384		0.363		0.392	
Var(Slope[τ_1^2])	-		0.234		-		0.254	
Corr(Intercept,Slope)	-		0.242		-		-0.299	
Item								
Var(Item[ω^2])	0.220		0.219		-		-	
Var(Item1[$\omega_{[1]}^2$])	-		-		0.734		0.745	
Var(Item2[$\omega_{[2]}^2$])	-		-		0.049		0.047	

tively). However, the ANOVA framework revealed limitations. Although the $F1/F2$ analyses were designed to reduce the inflated Type I error rate, this cannot be the remedy for the violation of the independence assumption. The responses in the dependent variable for either the $F1$ or $F2$ analysis were still not independent from one another because the mean responses across one source of effect (person or item) were not controlled for the other effect (item or person). In addition, the dependent variable was not treated for the non-normality. The skewness of the distribution due to the boundary of the proportion would affect the correct inference about the main effect of the category and the error variance. Last, the heterogeneity in the variance of the responses was not considered in the $F1$ analysis. A single error variance was compared against the category effect, but the variance of the responses was larger in the bugs level than in the fruit level.

The GLMM overcame the limitations of the ANOVA framework. The GLMM simultaneously took into account both person and item effects; thus, the independence assumption of the errors could be satisfied after the responses were explained by all sources of variations. The non-normality of the responses was overcome by using the logit link function.

The inference about the category effect was done on the logit scale. The confidence interval based on the GLMM was free from the problem of the boundary and the asymmetry of the distribution of the proportion. In the GLMM, the heterogeneity of the random item variance was tested with model comparison methods. In the current example, the model with the level-specific item random effect was selected over the model with the homogenous item random effect. The result suggests that the heterogeneity in the mean responses between the levels was explained by the difference in the variance of the item random effects, rather than an individual difference on the category effect which was quantified with the person random slope.

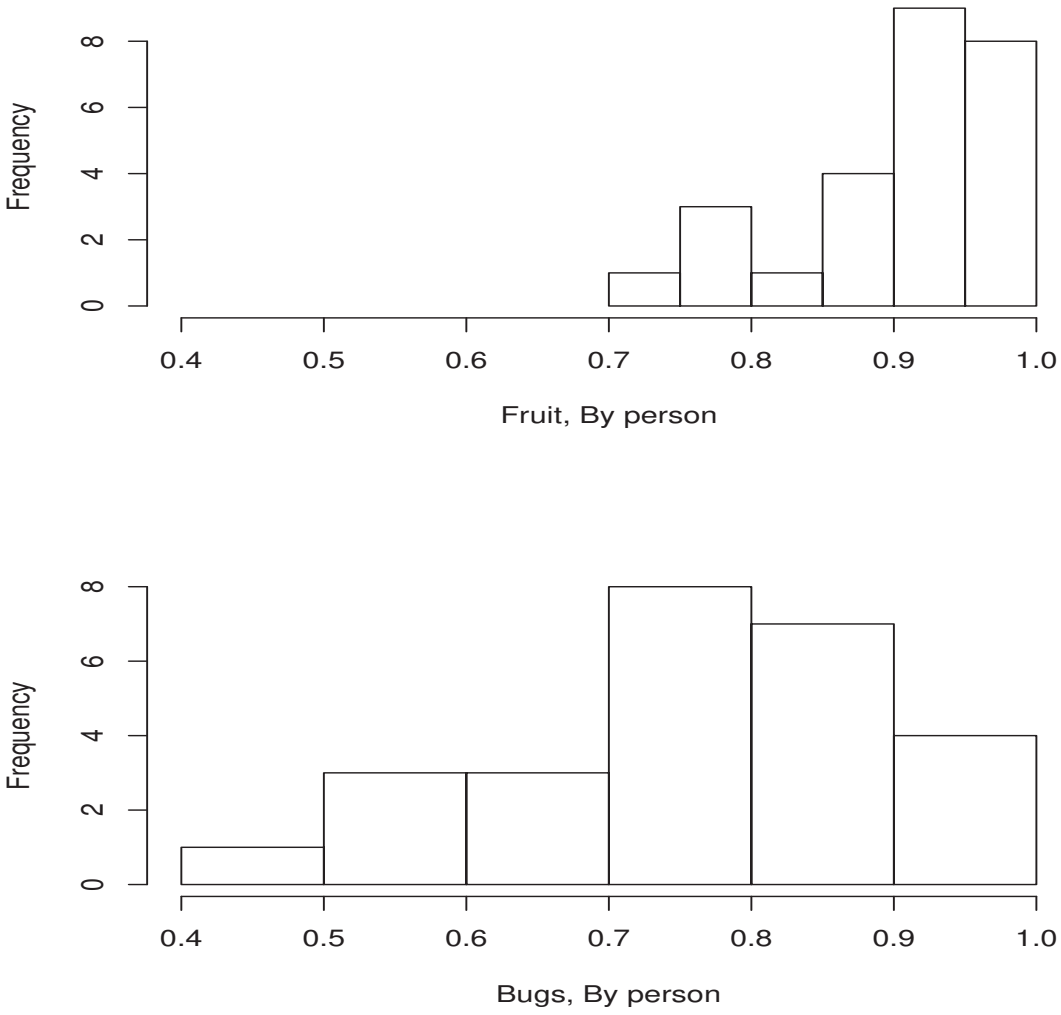

Figure 3.1: Histogram of accuracy at the fruit level (top) and at the bugs level (bottom) for the $F1$ analysis

CHAPTER 4

SIMULATION STUDY

The first purpose of the simulation study was to investigate the parameter recovery of the GLMM in various conditions found in empirical studies to achieve the first purpose of this dissertation. The second purpose of the simulation study was to evaluate the inferential qualities (the Type I error rate and power) of the model comparison methods (LRT, AIC, and BIC) in selecting the random effects and detecting a fixed effect (i.e., the experimental condition effect) using the Wald test to achieve the second purpose of the dissertation. In addition, the consequence of the misspecification of the level-specific item random effect in an NRI design is presented to examine the necessity of the effect in terms of parameter recovery and the inferential qualities in detecting a fixed effect.

4.1 Simulation Conditions

The simulation conditions that may influence precision and power for the fixed effect (the experimental condition effect) in the GLMM were considered. They include (a) the number of persons, (b) the number of items, (c) the magnitude of the fixed effect, (d) the magnitude of the person random effects, and (e) the type and the magnitude of item random effects.

The number of persons

The number of persons was selected as J = 20, 50, 300, and 1000. These numbers were chosen to mimic the minimum (16), median (41.5), and maximum (255) of the number of persons based on the survey results (see Table 2.1). The level of 1000 persons was included to show a large sample property of the hypothesis testing methods and the consistency of model selection in a large sample size.

The number of items

The 10, 30, 70, and 100 items were chosen as the number of items in each level of a condition (n_k). The 10-item level was chosen as the minimum number with which there was no convergence problem in using Laplace approximation (e.g., Cho & Rebe-Hesketh, 2011). The 100-item level were adapted from the maximum (144) of the number of items on the survey. The 30 items and the 70 items were chosen to interpolate between the 10-item and 100-item levels.

Magnitude of the fixed effect

Three levels of the magnitude of the fixed effect (β_1) were chosen: 0, 0.2, and 0.5. The effect coding was used for each level of the condition; thus, the three levels of the magnitude of the fixed effect were 0 (no effect), -0.1 versus 0.1, and -0.25 versus 0.25 for a covariate that had two levels. The two nonzero magnitudes, 0.2 and 0.5, were chosen to reflect small and large effects relatively.[1]

Magnitude of person random effects

The magnitudes of the person random intercept and slope ($[\tau_0^2, \tau_1^2]$) were set to [0, 0] or [0.3, 0.3]. In addition, the covariance of the person random effects (τ_{01}) was set to 0.15. This magnitude corresponds to the correlation of 0.5. The variance of 0.3 and the correlation of 0.5 were considered as the structure of the person random effects based on the results from a person random-slope model for an empirical study in Jaeger (2008).

Type and magnitude of item random effects

The following four levels were considered for the type (homogenous vs. level-specific) and magnitude of the item random effects: $\omega^2 = 0$, $\omega^2 = 0.2$, $\omega^2 = 1.3$, $[\omega^2_{[1]}, \omega^2_{[2]}]$=[0.2, 0.4] and $[\omega^2_{[1]}, \omega^2_{[2]}]$=[0.2, 1.3]. To manipulate the different degrees of the variances of item random effects, the magnitude of 0.2 was selected as a small effect based on survey results on the variance of item random effects in Cho et al. (2017), and the magnitude of 1.3 was followed by the results of the variance of the item random effect in De Boeck (2008). The

[1] Small and large effects should be interpreted relatively between the two levels.

levels of 0.2 and 1.3 were considered to examine small and large homogenous item random effects. The [0.2, 0.4] and [0.2, 1.3] levels were included for level-specific item random effects with small and large differences between the levels.

The five varying conditions were fully crossed, such that the total number of simulation conditions was 480 ($= 4 \times 4 \times 3 \times 2 \times 5$). For each condition, 500 replications were simulated. For the simulation conditions that had level-specific random item effects (54 conditions), results from the GLMM with homogenous random item effects and the $F1/F2$ analysis were investigated to show the consequence of the misspecification of the level-specific item random effect in detecting a fixed effect. For model comparisons, Models 1 to 4 were considered the candidate models. The final model was chosen based on the model building strategies described in the previous section, and then the significance of the fixed effect was tested.

4.2 Evaluation Measures

The bias and root mean square error (RMSE) were calculated to evaluate the parameter recovery of the data-generating model and the consequences of model misspecification. For model comparisons using the LRT, the AIC, and the BIC, the proportion that the true model was selected based on the model building strategy we described earlier was calculated out of 500 replications. Strictly speaking, the notion of power does not apply for the AIC and the BIC. The proportion refers to a true positive. The proportion is the conceptual analogue of power. In addition, the proportion of 500 replications was calculated regarding an inference about the fixed effect β_1. The Type I error rate was defined as the proportion that the fixed effect was incorrectly identified by the Wald test in the $\beta_1 = 0$ condition. Power was defined as the proportion that the fixed effect was correctly identified by a hypothesis testing method in the $\beta_1 = 0.2$ and 0.5 conditions. The nominal alpha level $\alpha = .05$ was used for the statistical tests. For the Type I error rate, the values close to the nominal alpha level .05 are considered the indicator of a good statistical test. As a rule of

thumb, power higher than .80 was considered satisfactory.

4.3 Simulation Result Hypotheses

In this section, the result hypotheses about parameter recovery, inferential quality, and consequences of misspecification are described.

Parameter recovery

A GLMM has coefficients for fixed effects and the variance and covariance of random effects as parameters. In this study, the primary interest is in the parameter recovery of a fixed effect for an experimental condition. The bias and the RMSE of the fixed effects are expected to decrease as the number of persons and items increases. The RMSE is expected to be smaller as the variances of random effects decrease.

Inferential quality

The inferential quality of the model comparisons regarding random effects is evaluated based on power (or a true positive for the AIC and the BIC). Between the AIC and the BIC, the BIC tends to select the more complicated model because the BIC penalizes more than the AIC when the sample size becomes larger than 8 ($ln(8) = 2.08$) (e.g., Burnham & Anderson, 2002). The AIC is an efficient criterion that minimizes the mean squared error of prediction, whereas the BIC is a consistent criterion in the sense that the probability of selecting the true model approaches 1 as N increases (if the true model is among the candidate models (e.g., Burnham & Anderson, 2002). Thus, it is expected that the power and the Type I error of BIC are satisfactory as the number of items and persons increases.

For nested models, the performance of the LRT, the AIC, and the BIC can be compared by quantifying the critical value for the LRT and how much the complex model is penalized for the AIC and the BIC. For example, when a complex model has one more parameter than the simpler model (i.e., $df = 1$ for Model 1 vs. Model 3; Model 2 vs. Model 4), the LRT selects the complex model when the deviance exceeds the critical value of the chi-square

distribution (3.84 at $\alpha = 0.05$), while the AIC does when the deviance is greater than 2, and the BIC does when the deviance is greater than the natural log of the sample size ($ln(20 \times 10) = 5.3$ for 20 persons and 10 items as the smallest number in our simulation condition). Therefore, the AIC is expected to have a higher power than the BIC and the power of the LRT is expected to be placed between the true positive of the AIC and the BIC. This order relation holds for $df = 2$ (i.e., Model 1 vs. Model 2; Model 3 vs. Model 4).

The power to detect a fixed effect using the Wald test is influenced by the magnitude of the fixed effect, the number of sample sizes, and the variance of the random effects. It is expected that the power increases as the number of persons, the number of items, and the magnitude of the fixed effects increase and the variances of the random effects decrease. The Type I error rate deviates from the nominal level if the accuracy and precision of the estimate are problematic. Accordingly, an inflated Type I error rate is expected when the number of persons and items is small, and the variance of the person and item random effects is large.

Consequences of misspecification

When level-specific item random effects are not considered in an NRI design, a model does not correctly specify all possible dependencies among the outcomes. When random effects are not fully specified in the misspecified models, biased estimates of the fixed effects are expected. The degree of bias is expected to be different depending on the kinds of random effects (i.e., the person random intercept, the person random slope, and the item random intercept) in the misspecified models. The Type I error rate and power by the Wald test for testing $\beta_1 = 0$ are expected to be satisfactory when the fixed effect and its standard error are correctly estimated under a true model in the condition of a larger number of persons and items and a smaller variance of random effects. Thus, a higher Type I error and lower power are expected in the misspecified models than in the true model in such conditions. This consequence is expected to be more severe when the difference in the

variances of the item random effects are large.

4.4 Results

No convergence problem occurred during the estimation process. The results are presented in the order of the parameter recovery of a fixed effect β_1 (for an experimental condition), inferential quality, and the consequence of ignoring level-specific item random effects.

Parameter recovery

Figures 4.1 and 4.2 show the bias and the RMSE of $\hat{\beta}_1$ for each true model, respectively. Under a true model, the bias quickly approached to 0, and the RMSE decreased as the number of persons and items increased. The magnitude of bias was not greater than 0.05 in all conditions and was lower than 0.025 in the medium number of persons (50) and items (30). Other factors in the simulation design, such as the magnitude of β_1 and the variance of the item random effects, did not affect the magnitude of the bias. For all models, the patterns in the RMSE were associated with the variance of the random effects, the number of persons, and the number of items. The RMSE decreased with a smaller variance of the item random effects and a larger number of persons and items.

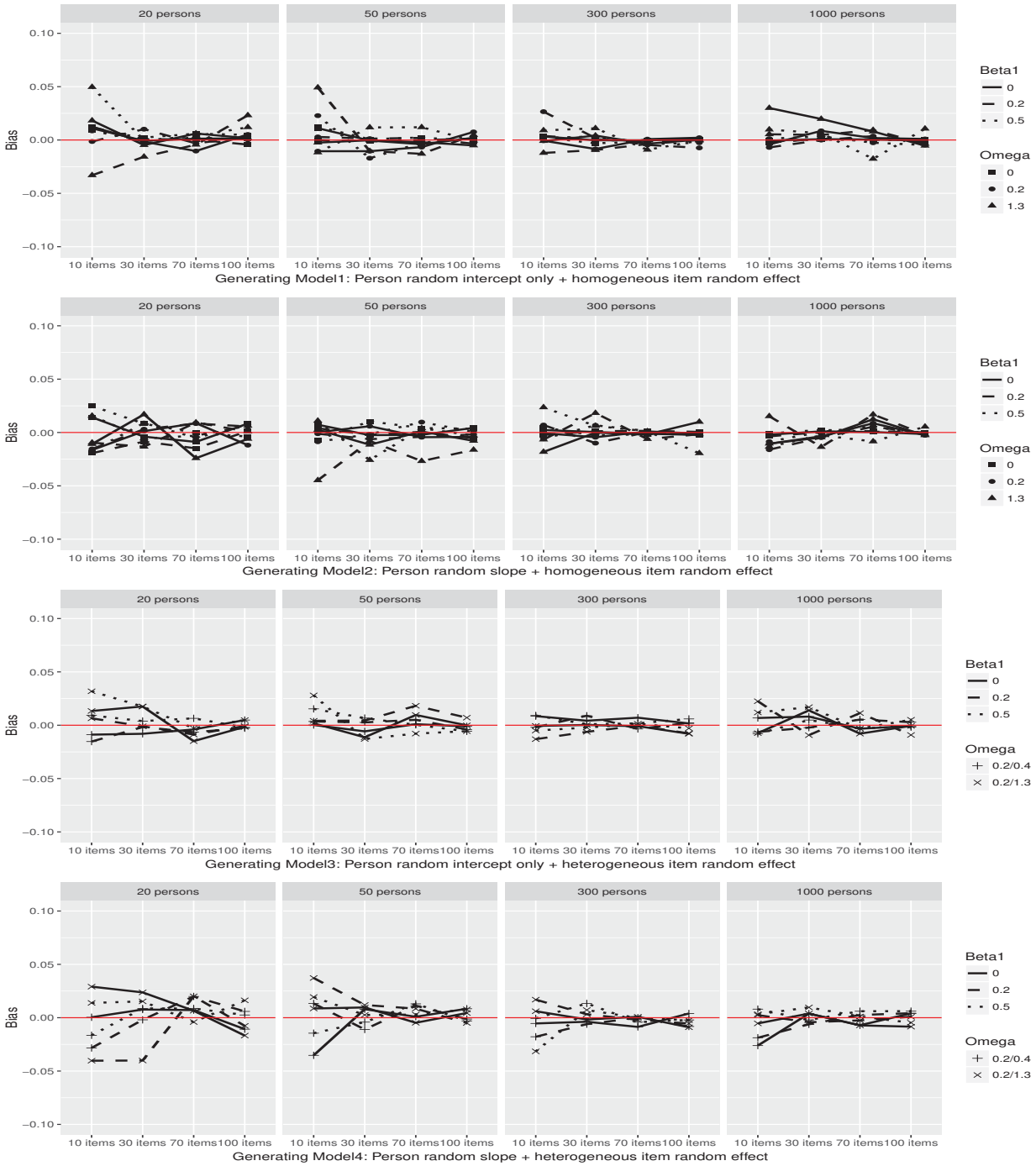

Figure 4.1: Bias of $\hat{\beta}_1$ when the true model is specified.

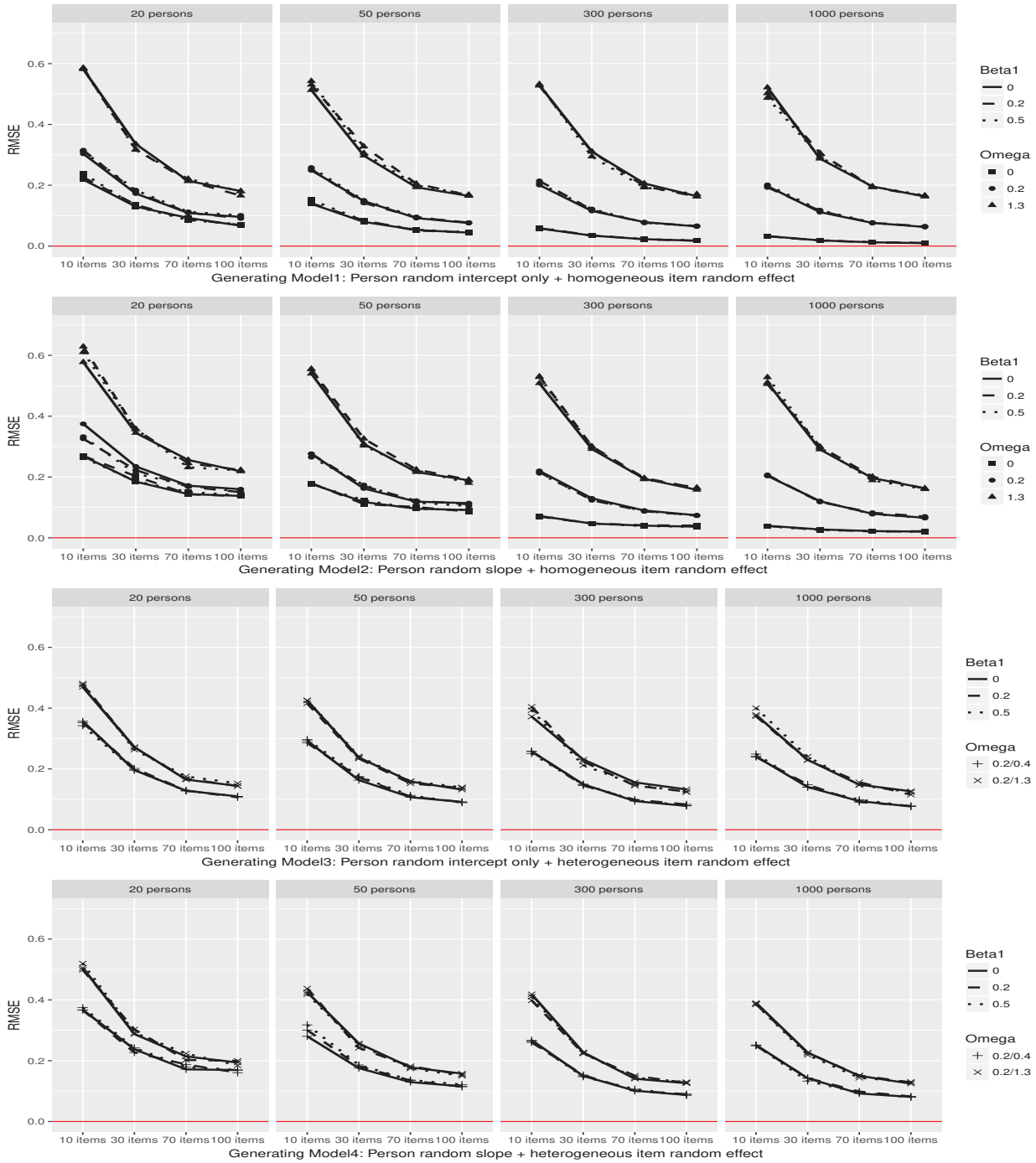

Figure 4.2: RMSE of $\hat{\beta}_1$ when the true model is specified.

Inferential quality

In this section, the power to detect the correct specification of random effects through the LRT, the AIC, and the BIC is reported, followed by the results of testing the fixed effect β_1 with the Wald test based on the selected model regarding random effects. The results are presented in Figure 4.3 for the power to select the true random effect specification using the LRT, the AIC, and the BIC; and in Figures 4.4 and 4.5 for the Type I error rate and the power for testing the fixed effect when the true model is Models 1 to 4, respectively. Because the performance of the LRT did not differ between the forward approach and the backward approach, only the results for the LRT with the forward approach are reported.

As shown in Figure 4.3, the power to detect the correct specification of random effects differed depending on the true model. For all models, the power was not associated with the magnitude of the fixed effect. When there were no person random slope and homogenous item random effects (i.e., Model 1 is a true model), power higher than .80 was reached in all conditions for the LRT and the BIC, which had an average power of 0.997 and 0.944, respectively. However, the power of the AIC was .80 on average when the variance of the item random effects was greater than 0, implying that the AIC tended to select a more complicated model. A target power of .80 was not reached mostly at the level of 10 and 30 of the number of items by the AIC. When the person random slope existed and the item random effects were homogenous (i.e., Model 2 is a true model), The AIC, the BIC, and the LRT successfully selected the true model in the conditions in which the number of persons was 50 or larger and the number of items was 70 or larger. The power was not associated with the magnitude of the fixed effect and was higher at the zero level of the variance of the item random effects than at the 0.2 or 1.3 level. When the variance of the person random slope was 0 and the variances of the item random effects were not equal between the levels (i.e., Model 3 is a true model), a power of 0.8 was reached for the LRT and the AIC in the medium size of the number of persons (50) and the small number of items (30) for the large difference (0.2/1.3) in the variance of the item random effects, whereas a large number of

persons (300) and items (70) was for the small difference (0.2/0.4) in the variance of the item random effects. The performance of the BIC was satisfactory with the large number of items (70). However, the BIC never successfully selected the true model at the 0.2/0.4 level of the variance of item random effects. A similar pattern of power was shown as in Model 3 when there was a person random slope in Model 4 in the true model. However, the power was lower than 0.3 by all methods when both the number of persons and the number of items were small (10 items and 20 persons) compared to the power under Model 3.

Figure 4.4 presents the Type I error rate using the Wald test based on results for the AIC, the BIC, and the LRT in the case of $\beta_1 = 0$ in the true models. The pattern of the Type I error rate differed across the true models. When the person random slope did not exist in the true model as in Model 1 and Model 3, the Type I error rate was satisfactory (ranged from 0.04 to 0.08) based on the final model selected by the AIC, the BIC and the LRT except for the following two conditions. First, the Type I error rate was conservative ($<$0.03) with zero variance of the item random effects, 1000 persons and 100 items. Second, the Type I error rate was higher than 0.08 when the variance of the item random effects was 1.3, the number of items was 10, and the number of persons was 300 or larger. When the person random effect existed in Model 2 and Model 4, the Type I error rate by the models based on the AIC and the LRT was lower than 0.08 when the number of persons was 50 or larger and the number of items was 70 or larger. The Type I error rate after model selection using the BIC was inflated when either the number of persons or the number of items was small. The Type I error rate based on the BIC was often higher than 0.10 in the following combinations of the number of persons and items: (20, 10), (20, 70), (20, 100), (50,10), (50, 30), and (300,10).

Regarding the power to detect non-zero β_1 presented in Figure 4.5, the LRT, the AIC, and the BIC revealed roughly equal performance.

For all models, the power was associated with the number of persons, the number of items, and the structure of the random effects. The power increased as the number of

persons and items increased and the variance of random effects decreased. When there were no person random slope and homogeneous item random effects (i.e., under Model 1 as the true model), a small number of persons (20) and a medium number of items (30) were required to detect β_1 of 0.5 successfully ($>$.80) when the variance of the item random effects was 0; and a medium number of items (50) was needed if the variance of the item random effects was 0.2. However, the power reached .80 only at the maximum level of the number of persons (1000) and items (100) when the variance of the item random effect was 1.3. The power to detect β_1 of 0.2 was reached at 0.8 with more persons and items: 50 persons and 70 items when the variance of the item random effect was zero, and 300 persons and 100 items when the variance of the item random effect was 0.2. A similar pattern was observed in the presence of the person random slope and the homogeneous item random effect (under Model 2 as the true model), but the power was lower especially to detect β_1 of 0.2. For example, the power was below .33 when the number of persons was 20 and the number of items was 100. The power of 0.8 to detect β_1 of 0.2 could be satisfied in 300 persons, 30 items, and zero variance of the item random effect, while it was achieved only the maximum number of persons (1000) and items (100) when the variance of the item random effect was 0.2. When the item heterogeneity existed in the true model (Model 3 and Model 4), the power was higher in the small difference (0.2/0.4) than in the large (0.2/1.3) difference in the variance of the item random effects. The power to detect β_1 of 0.5 was acceptable with 20 persons, 70 items, and 0.2/1.3 of the variance of the item random effects under Model 3, and 50 persons, 70 items, and 0.2/1.3 of the variance item random effects under Model 4, respectively. However, the power to detect a small magnitude of the fixed effect (0.2) did not reach 0.8 in any simulation conditions with the item heterogeneity.

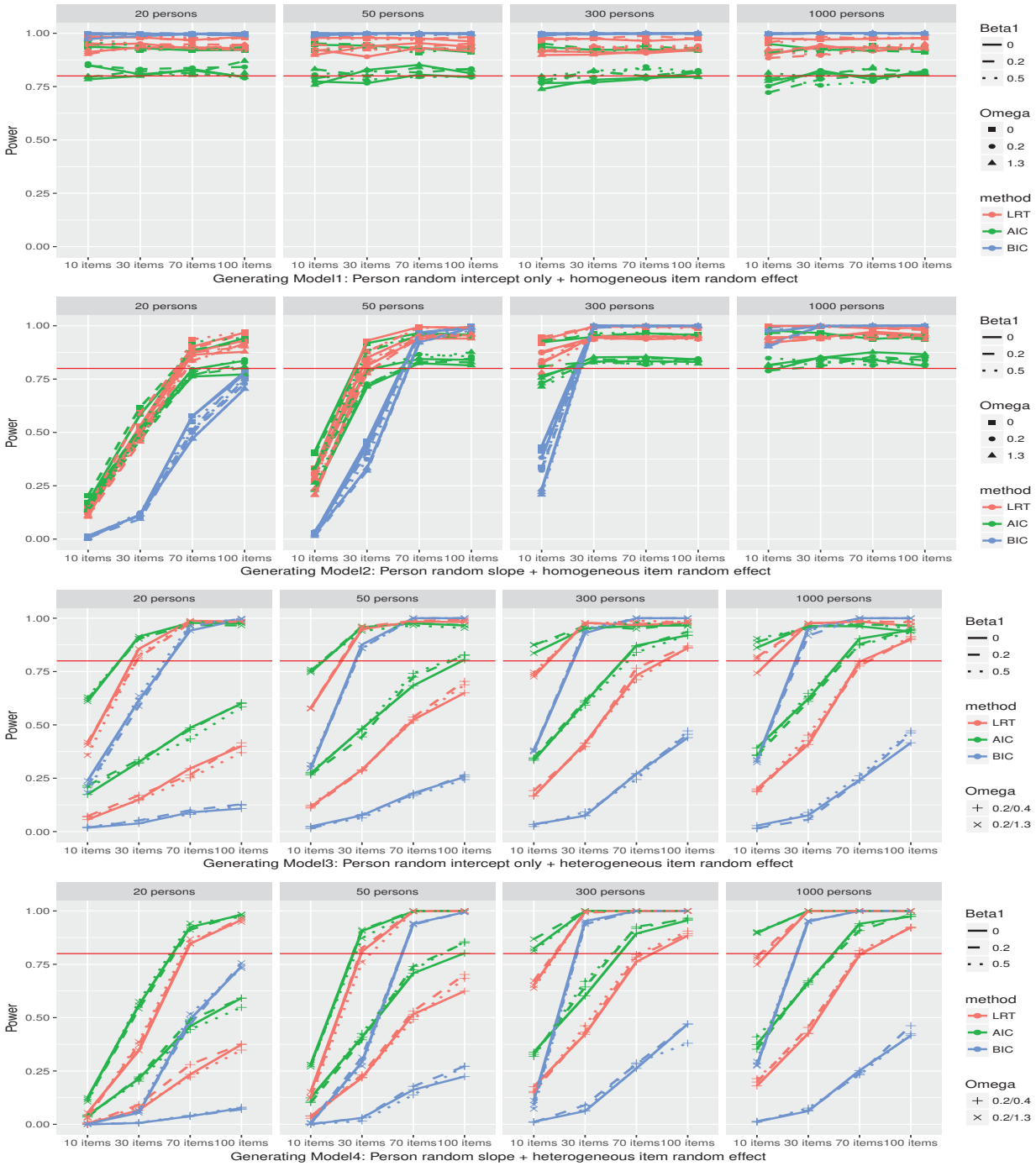

Figure 4.3: Power to select the true specification of random effects

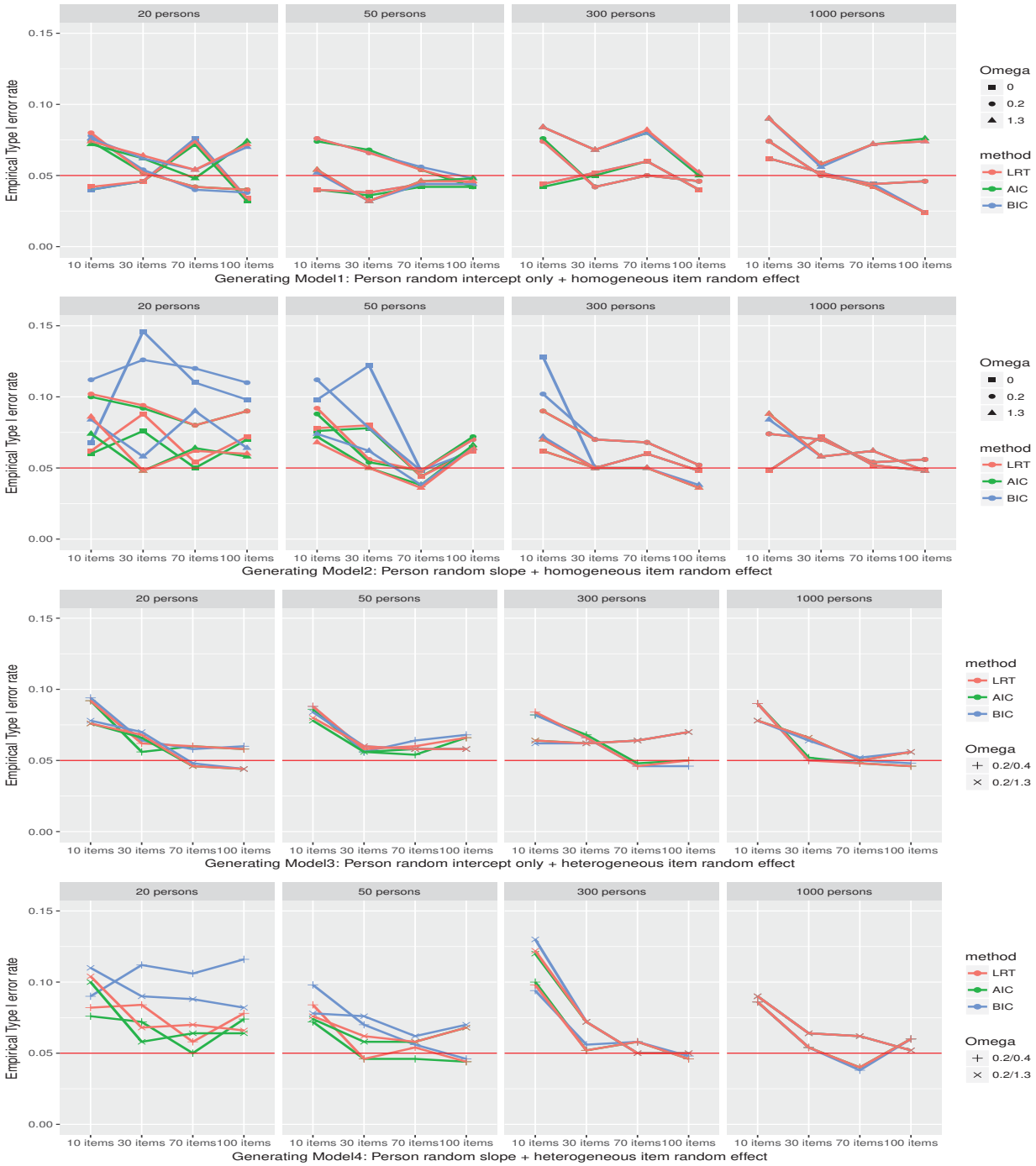

Figure 4.4: Type I error rate of the Wald test after model selection

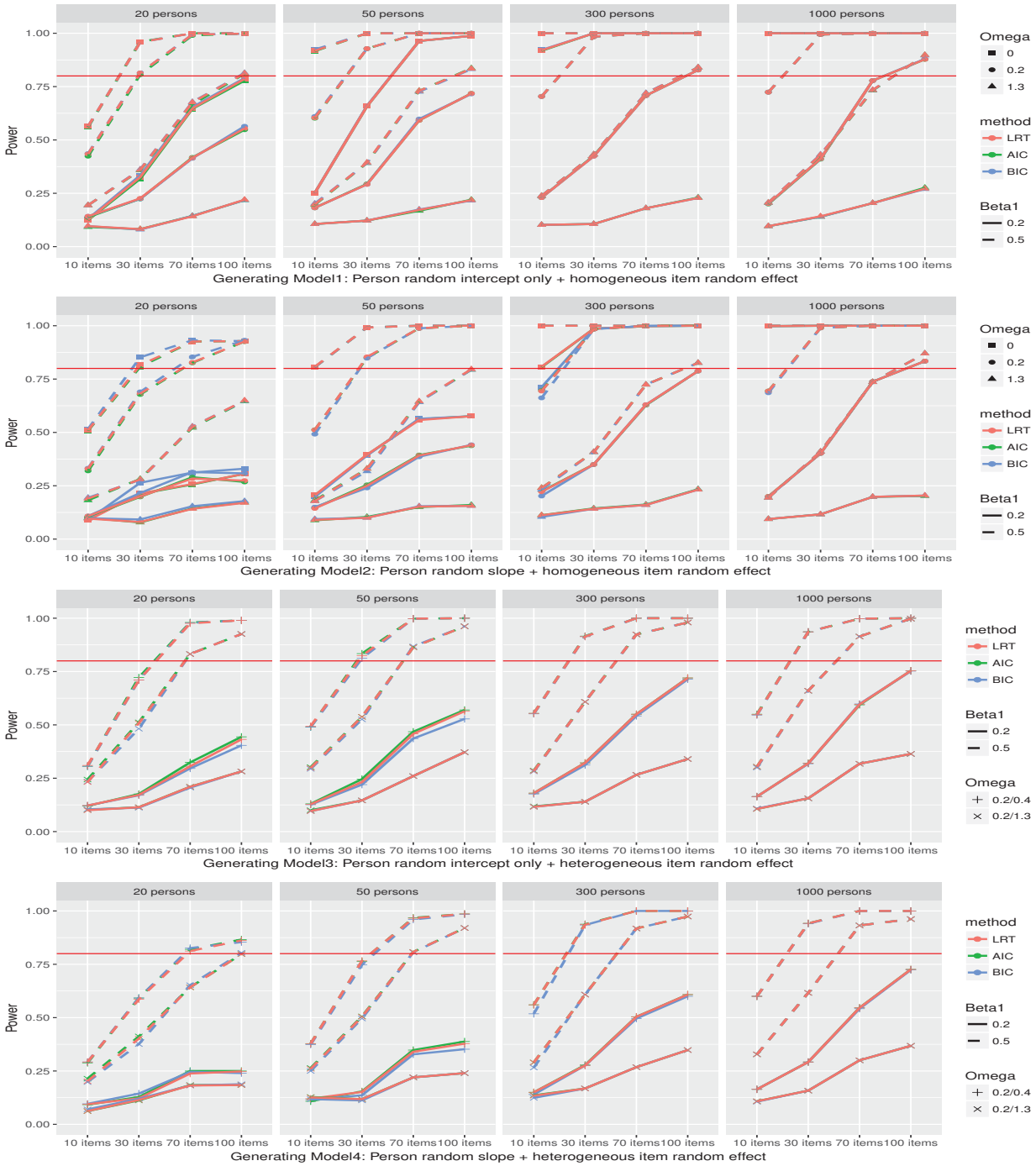

Figure 4.5: Power of the Wald test after model selection

Consequences of misspecification

In this section, the results of the parameter recovery and the inference about the fixed effect using the Wald test are reported when the model is misspecified regarding random effects. In addition to the misspecified GLMMs, the fixed effect was tested with $F1/F2$ analyses.

Figure 4.6 presents the bias of $\hat{\beta}_1$ in the case of misspecification. The pattern of bias in case of misspecification differed depending on the true models, except for the pattern of Model 4 of which the estimate was consistently unbiased. Under Model 1 which had only the person random intercept and the homogenous item random effect, the estimate was nearly unbiased for all fitting models if the number of items was greater than 10. When the person random slope existed as Model 2, however, the estimate was consistently biased with the homogenous item random effect as Model 1 and Model 3. The magnitude of the bias was not affected by the variance of the item random effect. However, the variance of the item random effects was positively associated with the precision (quantified with the RMSE) of the estimate. When the true model had a level-specific item random effect and the person random intercept only under Model 3 as a true model, Model 1 and Model 2 yielded biased estimates for 20 and 50 persons, but the magnitude of the bias approached zero as the number of persons increased. The magnitude of the bias was larger at the 0.2/1.3 level of the variance of the item random effects than 0.2/0.4 level. Under Model 4 which had a level-specific item random effect and a person random slope, all the misspecified models revealed biased estimates for 20 and 50 persons. As the number of persons increased, the estimate by Model 2 became unbiased whereas the estimate by Model 1 and Model 3 was consistently biased. Larger heterogeneity in the item random effects yielded a larger magnitude in the bias.

Figure 4.7 shows the difference in the Type I error rate to test β_1 between the true model and the misspecified models in addition to the $F1/F2$ results. Under Model 1, all the misspecified models and $F1/F2$ showed no difference in the Type I error rate from the

true model. In contrast, under Model 2 having a person random slope, Model 1, Model 3, and $F1/F2$ analysis revealed inflated Type 1 error rates. The Type 1 error rate increased with the increasing number of items and persons. The Type I error rate was noticeably higher when the variance of the item random effect was 0 than when it was 0.2 or 1.3. For example, at the maximum level of the number of persons (1000) and items (100), the Type 1 error rates increased by at least 0.75 by Model 1 and Model 3, and the $F1/F2$ at the zero level of the item random effects whereas the rates increased by less than 0.1 at the level of 0.2 and 1.3. Under Model 3 that had item heterogeneity and the person random intercept only, the misspecified models showed Type I error rates comparable to those of the true model across all simulation conditions except for Model 1 at 20 persons and 100 items. $F1/F2$ analysis revealed high Type I error rates when the number of items was 70 or greater and the level-specific item variances were 0.2/1.3. Under Model 4 that had item heterogeneity and a person random slope, Model 1 and Model 3 which ignored person random slope yielded increasing Type 1 error rates with increasing number of items. In contrast, Model 2 which ignored only item heterogeneity showed inflated Type I error rates only in the small number of items (10). The pattern of the Type I error rate by the $F1/F2$ analysis was similar to that under Model 3.

The power of the Wald test in the case of misspecification was compared to the power in the correct specification. Figure 4.8 shows the difference in power from the true model. The power of Model 4 was comparable with that of the true model across all conditions. In the case of Model 1 for the true model, all the misspecified GLMMs yielded power comparable to that of power to the true model when the number of persons was 50 and the number of items was 30 or larger. Under this condition, the decreased power was less than 0.1. The power by $F1/F2$ was lower than the power of the GLMM. When there was a person random slope in the true model (Model 2), the person random intercept models (Model 1 and Model 3) showed higher power than the true model in conditions where the number of items was greater than the number of persons. However, the person random intercept

models were underpowered in the conditions that had 300 or more persons. In contrast, the power of Model 4 was as high as that of the true model across all conditions. The power of the $F1/F2$ was always lower than that of the true model. The difference increased as the number of items increased. When the true model was Model 3, the power difference between the true model and the models with item homogeneity (Model 1 and Model 2) was remarkable in the conditions where the number of items was larger than the number of persons. However, the power difference disappeared when the number of persons was 300 and 1000. The power of the $F1/F2$ was lower than that of the GLMM. The power difference increased as the number of items increased with one exception. The power of the $F1/F2$ approached the power of the true model as the number of items increased if the variance of the item random effect was 0.2/0.4 and the magnitude of the fixed effect was 0.5. When the true model was Model 4, Model 1 and Model 3 that had a person random intercept showed a lower power than the true model as the number of persons increased. The power difference was larger with $\beta_1 = 0.2$ and $[\omega^2_{[1]}, \omega^2_{[2]}] = [0.2, 0.4]$ conditions. The pattern of Model 2 and the $F1/F2$ in item heterogeneity and the person random slope was similar to the pattern in item heterogeneity and person intercept only.

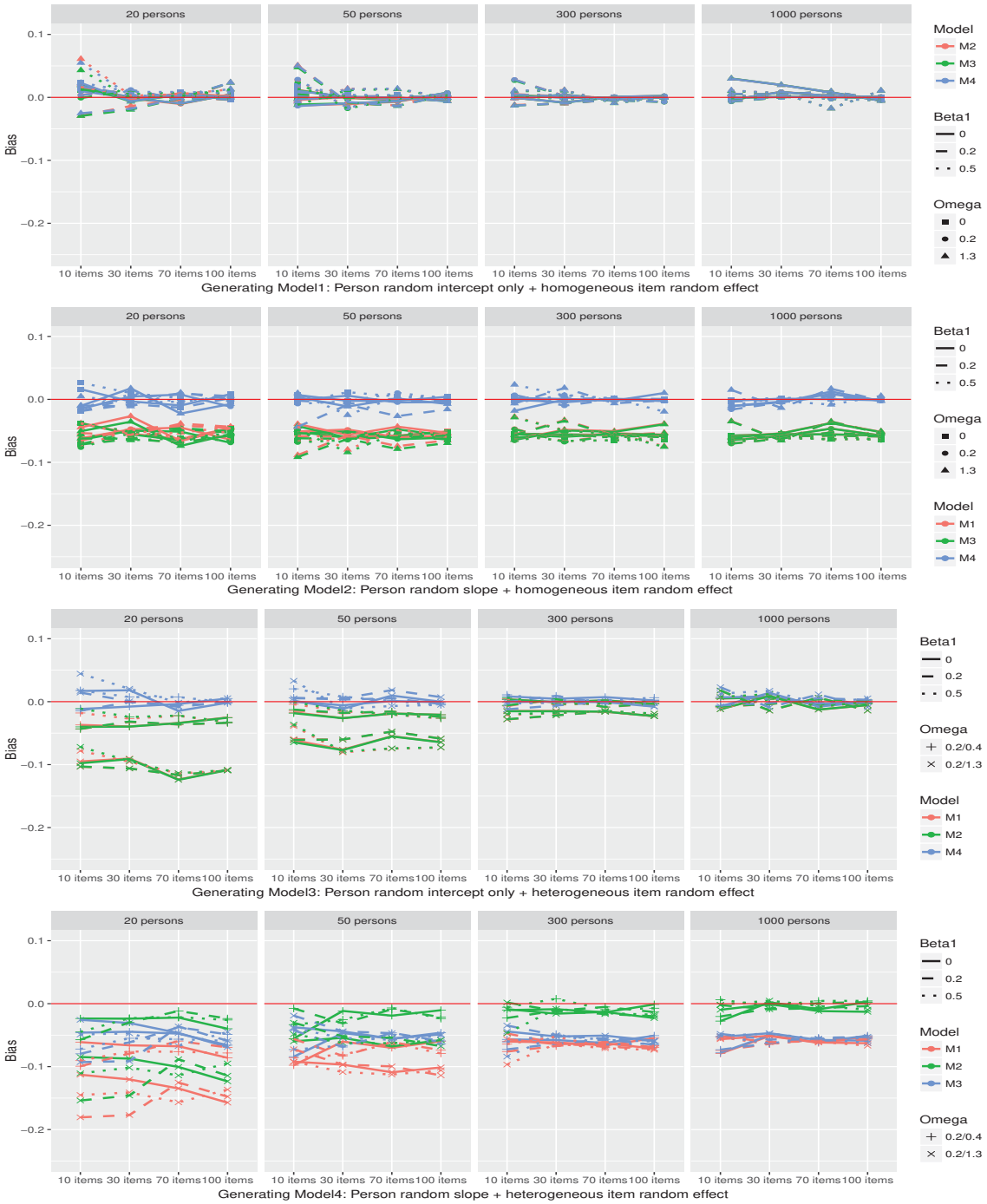

Figure 4.6: Bias of $\hat{\beta}_1$ in the case of misspecification

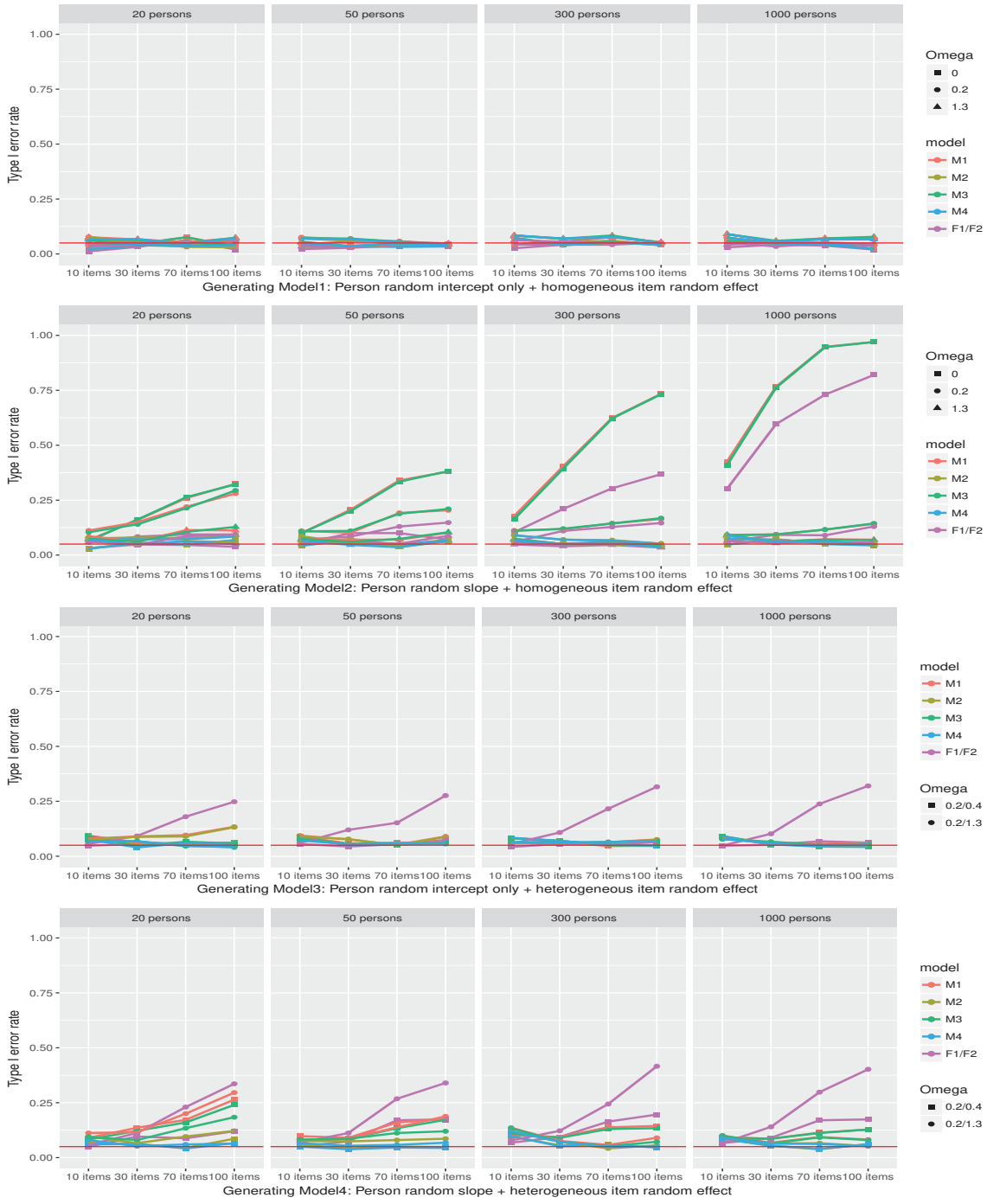

Figure 4.7: Type I error rates of the Wald test in the case of misspecification

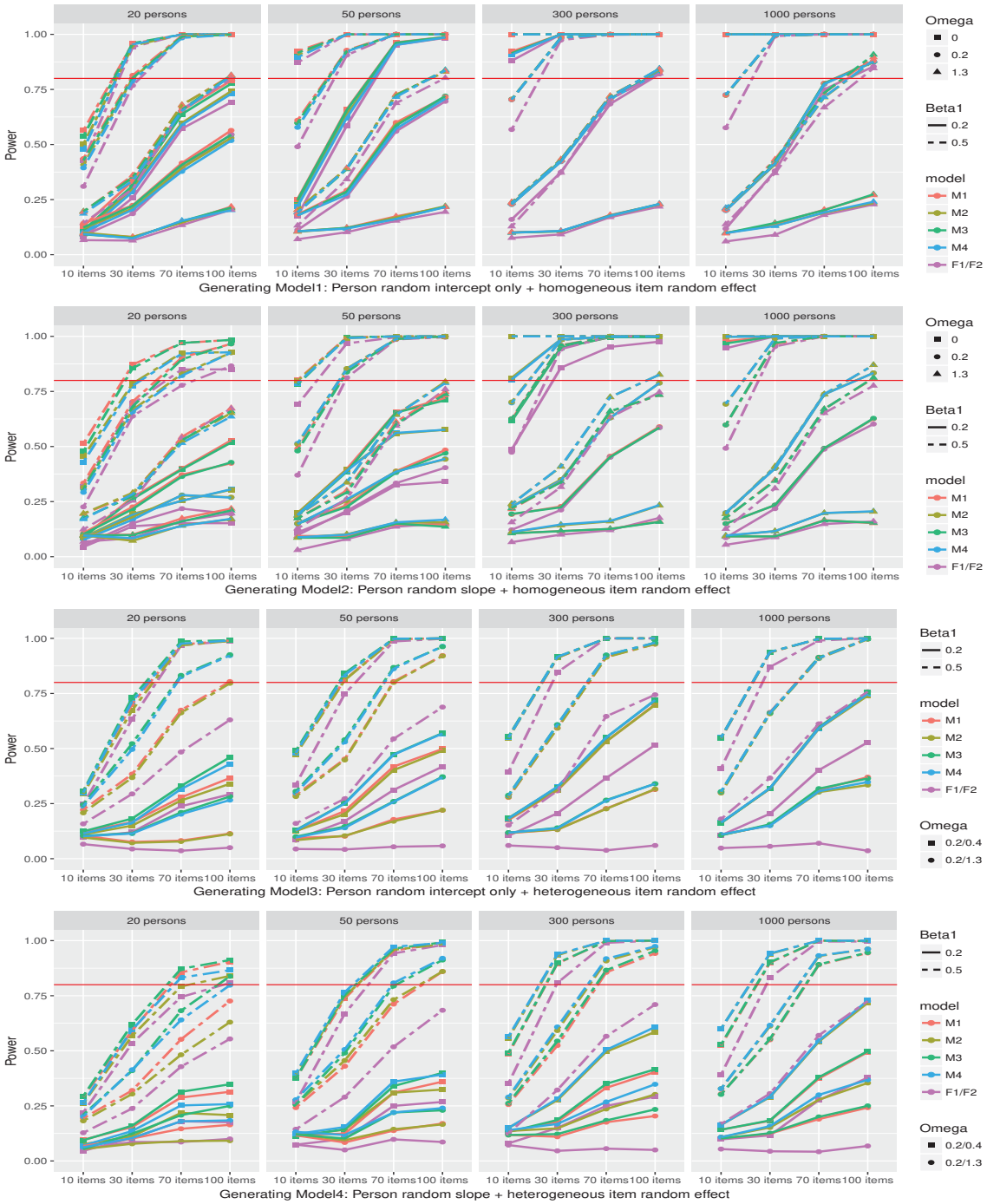

Figure 4.8: Power of the Wald test in the case of misspecification

CHAPTER 5

SUMMARY AND DISCUSSION

The GLMM has been applied in complex psychological experiments with binary responses in lieu of the ANOVA framework (e.g., Barr, 2008; Grodner et al., 2010; Jaeger, 2008; Rowland et al., 2012). In the current applications, GLMMs with crossed random effects used only one variance parameter to account for item variability in the model for an NRI design, such that those models do not account for item heterogeneity between the levels of a condition in the NRI design. In this dissertation, the GLMM with crossed random effects for the NRI design was specified, illustrated, and evaluated. In addition, this dissertation showed the model building strategy and model comparison methods such as the optimal conditions in testing an experimental condition effect. The derivation of the mixture chi-square distribution was provided for the LRT between two GLMMs with crossed random effects. Because item heterogeneity is ignored in many applications in the NRI design (as we showed in Table 2.1), the consequences of ignoring item heterogeneity were investigated in testing the experimental condition effect to show the necessity of modeling it.

5.1 Summary

Parameter recovery

The GLMM is applicable for an NRI design. If the structure of random effects is correctly specified, the fixed effect parameter was well recovered in all random structures we considered when there was medium number of persons (50) and items (30) per level of a condition.

Inferential quality: Model selection

The structure of the random effect is successfully specified by the model selection meth-

ods even with 10 items and 20 persons when the true structure is simple such as person random intercept-only and has a variance of 0.2. As the complexity of the structure of the random effects increases, more persons and items are required to select the correct specification. The performance of the AIC was comparable to the performance of the LRT, but the AIC often selected an overfitting model when the true specification is simple. The BIC did not perform at a satisfactory level in selecting the correct specification of random effects in an NRI design, especially with a small number of persons and items.

Inferential quality: Testing the fixed effect

The inferential performance for detecting the fixed effect was influenced by the model selection quality. With the complexity of the structure of the random effects, Type I error inflation and low power were found in small sample sizes. The Type I error rate of the BIC was inflated when the number of persons and items was small. Regarding the optimal conditions for detecting an experimental condition effect, the sample sizes are recommended depending on the structure of the random effects to detect a large effect size (e.g., 0.5) with adequate power (e.g., 0.80). If the person random intercept only and the homogenous item random effects are expected, the fixed effect can be detected in a small number of persons (e.g., 20) and a medium number of items (e.g., 30) per level. If a person random slope is expected, a larger number of persons (e.g., 50) are needed to detect the same magnitude of the fixed effect. If unequal variances of the item random effects are expected, one needs to consider a larger number of items (e.g., 70) for each level to the same number of persons as in the homogenous item random effect models. If the size of the fixed effect is smaller (0.2) than the variance of the person random effect (0.3), a very large sample (e.g., 300 persons or larger) is needed to successfully detect the fixed effect even in the person intercept-only and homogenous item random effect model.

Consequences of misspecification

Fitting a model with the homogenous item random effect to data that have item heterogeneity yields biased estimates of the fixed effect. The consequences depend on the random

effect specification of the true model. When there was no person random slope, the fixed effect was biased with a small number of persons and a large difference in the variances of level-specific item random effects, but increasing the number of persons reduces the bias. However, when there are item heterogeneity and a person random slope in the true model, a person random intercept-only model produces a biased fixed effect estimate regardless of the number of persons. A person random slope model can still be considered in the case of a large number of persons because the bias of the fixed effect decreases when the number of persons is greater than 300, although this number of persons is impractically large for experimental research in psychology. When the variance of the item random effects is level-specific but there is no person random effect, a person random slope model maintains an acceptable Type I error rate and comparable power to the true model if the number of persons is 50 or larger. With item heterogeneity and a person random slope in the true model, the inferential quality is problematic when the number of items is larger than the number of persons. Using $F1/F2$ analysis can be problematic when it is suspected that the item effects are not constant. Ignoring the item random effects resulted in an inflated Type I error rate and deflated power to detect the target effect. This finding is consistent with Baayen et al. (2008).

5.2 Discussion

Considerations for Practice

Based on the findings in the simulation study, the following recommendations are made. First, it is recommended using a large enough number of persons and items for GLMM specifications such as 50 persons and 30 items per level of a condition. The fixed effect parameter in the GLMM is well recovered if the model is correctly specified, but the estimate for the fixed effect may not be precise if the number of items in a level is too small (e.g., 10 items). Second, the LRT is recommended instead of information criteria for selecting the random effect specification if the researchers have a medium number of persons

(50) and items (30) or larger. Third, considering the level-specific item random effect is recommended when the number of items is large compared with the number of persons. In contrast, the heterogeneity of the item random effect can be ignored when there is a small number of items, but the number of persons is large. In such a condition, a person random slope model performs as well as the model with level-specific item random effects.

Limitations and Future Directions

In this dissertation, simulation results are limited to the two levels of the experimental condition with the fixed effect magnitudes (0.2 vs. 0.5 with effect coding). To generalize our findings, other levels such as more than two levels of the experimental condition and more varying magnitudes should be examined in future studies.

Another limitation of the current study is that the fixed effect was tested with the Wald test as a common testing method. Because the Wald test is based on an approximation to the likelihood at the maximum likelihood estimate, it is possible that this approximation is poor in the case of a smaller number of persons and items. The performance of the Wald test can be compared with other alternative hypothesis testing methods such as LRT, a score test, bootstrapping, and Bayesian hypothesis testing methods, especially in conditions of smaller numbers of persons and items. Further comparison studies may facilitate researchers' informed choices among testing methods for the experimental condition.

Third, in the simulation study, a (multivariate) normal distribution was assumed for the random effects. When this normality assumption of the random effects is violated, a fixed effect estimate can be biased (e.g., Verbeke & Lesaffe, 1997). Further research is required to investigate the degree of bias in the fixed effect of the GLMM with crossed random effects in the case of non-normality for random effects and to consider alternative estimation methods for the non-normality when non-ignorable bias is found.

Lastly, robust standard errors may be used instead of the results of the GLMM in small sample sizes, in testing experimental condition effects. The robust standard errors are used to calibrate the distribution of residuals when it does not satisfy the independence assump-

tion of a model (Huber, 1967; White, 1980). When there are two sources of dependency in responses, it is possible that one source is modeled as a random effect and the dependency due to the other source is corrected using a robust standard error (e.g., SAS Institute Inc., 2017). For example, GLIMMIX procedure allows to compute the empirical covariance estimators of the covariance matrix of the fixed effects for GLMMs (SAS Institute Inc, 2017, pp. 3350). Alternatively, multiway robust standard errors can be considered for more than one sources of correlations to correct the standard errors of the estimates (Cameron, Gelbach, & Miller, 2011). However, the robust standard errors also have limitations in small sample sizes such as Type I error inflation (Morel & Neerchal, 2006). Further, simulation studies revealed that the robust standard error is either too liberal or too conservative depending on whether the bias of the variance estimates was corrected or not (Stroup, 2013). To our knowledge, relative performances between robust standard errors with the GLMM and multiway robust standard errors has not been shown. In future studies, systematic comparisons of the two methods is necessary for considering multiple sources of correlations in small sample sizes for the GLMM.

BIBLIOGRAPHY

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Caski (Eds.), Second International Symposium on Information Theory. Akademiai Kiado, Budapest, 267-281.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390-412.

Barr, D. J. (2008). Analyzing 'visual world eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59*, 457-474.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255-278.

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

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88*, 9-25.

Breslow, N. E. & Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of a dispersion. *Biometrika, 82*, 81-91.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach.* New York, NY: Springer.

Cameron, A. C., Galbach, J. B., & Miller, D. L. (2011). Robust inference with multiway clustering. *Journal Business and Economic Statistics 29*, 238-249.

Casella, G., & Berger, R.L. (2002). *Statistical inference* (2nd Edition). Pacific Grove, CA: Wadsworth & Brooks/Cole.

Chernof, H. (1954). On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics, 25*, 573-578.

Cho, S.-J., & Rabe-Hesketh, S. (2011). Alternating imputation posterior estimation of models with crossed random effects. *Computational Statistics and Data Analysis, 55*, 12-25.

Cho, S.-J., De Boeck, P., & Lee, W.-y. (2017). Evaluating testing, profile likelihood confidence interval estimation, and model comparisons for item covariate effects in linear logistic test models. *Applied Psychological Measurement, 41,* 353-371.

Cho, S.-J., & De Boeck., P. (2018). [Brief Reports] A note on *N* in Bayesian information criterion (BIC) for item response models. *Applied Psychological Measurement, 42*, 169-172.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335-359.

Clayton, D. G. & Rasbash, J. (1999). Estimation in large crossed random-effect models by data augmentation. *Journal of the Royal Statistical Society, Series A, 162*, 425-436.

item DeBoeck, P. (2008). Random item IRT models. *Psychometrika, 73*, 533-559.

Geman, D. & Geman, S. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*, 721-741.

Goldstein, H. (1991). Nonlinear multilevel models with an application to discrete response data. *Biometrika, 78*, 45-51.

Goldstein, H. & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A, 159*, 505-513.

González B., J., De Boeck, P., Tuerlinckx, F. (2014). Linear mixed modelling for data from a double mixed factorial design with covariates: a case-study on semantic categorization response times. *Journal of the Royal Statistical Society: Series C, 63*, 289-302.

Gordon, P. C., & Holyoak, K. J. (1983). Implicit learning and generalization of the "mere exposure" effect. *Journal of Personality and Social Psychology, 45*, 492-500.

Greven, S. & Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika, 97*, 773-789.

Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). "Some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition, 116*, 42-55.

Gurka, M. J., Edwards, L. J., & Muller, K. E. (2011). Avoiding bias in mixed model inference for fixed effects. *Statistics in Medicine, 30*, 2696-2707.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1*, 221-233. University of California Press, Berkeley.

Jacqmin-Gadda, H., Sibillot, S., Proust, C., Molina, J.-M., & Thiébaut, R. (2007). Robustness of the linear mixed model to misspecified error distribution. *Computational Statistics and Data Analysis, 51*, 5142-5154.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434-446.

Joe, H. (2008). Accuracy of Laplace approximation for discrete response mixed models. *Computational Statistics and Data Analysis, 52*, 5066-5074.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*, 54-69.

Lambert, P. (2006). Comment on article by Browne and Draper. *Bayesian Analysis, 1*, 543-546.

Luck, S. J. (2005). Ten simple rules for designing ERP experiments. In T. C. Handy (Ed.), *Event-related potentials: A methods handbook* (pp. 17-32). Cambridge, MA: MIT Press.

Maxwell, S. E., & Bray, J. H. (1986). Robustness of the quasi *F* statistic to violation of sphericity. *Psychological Bulletin, 99*, 416-421.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective.* Psychology Press.

Metropolis, N., Rosenbluth, A. W. Rosenbluth, M. N., & Teller, A. H. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics, 21*, 1087-1092.

Molenberghs, G., & Verbeke, G. (2007). Likelihood ratio, score, and Wald tests in a constrained parameter space. *The American Statistician, 61*, 22-27.

Morel, J. G. & Neerchal, N. K. (2006). *Comparison of four small sample bias corrections of the empirical covariance estimators.* (Report No. 2006-02). Baltimore, MD: Department of mathematics and statistics, University of Maryland.

Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. *Social Cognition, 19*, 625-664.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raaijmakers, J. G. W. (2003). A further look at the "language-as-fixed-effect fallacy". *Canadian Journal of Experimental Psychology, 57*, 141-151.

Raaijmakers, J. G. W., Schrinemakers, J. M. C., & Gremmen, F. (1999). How to deal with "the language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language, 41*, 416-426.

Rabe-Hesketh, S., Skrondal, A. & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics, 128*, 301-323.

Raudenbush, S. W. Yang, M., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics, 9*, 141-157.

Rodríguez, G. & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A, 158*, 73-89.

Rodríguez, G. & Goldman, N. (2001). Improved estimation procedures for multilevel models with binary response: a case study. *Journal of the Royal Statistical Society, Series A, 164*, 339-355.

Rowland, C. F., Chang, F., Ambridge, B., Pine, J. M., & Lieven, E. V. M. (2012). The development of abstract syntax: Evidence from structural priming and the lexical boost. *Cognition, 125*, 49-63.

Saefken, B., & Kneib, T. (2014). A unifying approach to the estimation of the conditional Akaike information in generalized linear mixed models. *Electronic Journal of Statistics, 8*, 201-225.

Santa, J. L., Miller, J. J., & Shaw, M. L. (1979). Using quasi F to prevent alpha inflation due to stimulus variation. *Psychological Bulletin, 86*, 37-46.

SAS Institute Inc. (2017). *SAS/STAT Users Guide* (Version 14.3). Cary, NC: SAS Institute Inc.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

Self, S. G., & Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association, 82*, 605-610.

Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics, 50*, 1171-1177.

Stram, D. O., & Lee, J. W. (1995). Corrections: Variance components testing in the longitudinal mixed effects model. *Biometrics, 51*, 1196.

Stroup, W. W. (2013). *An introduction to generalized linear models*. Boca Raton, FL: CRC Press.

Trippas, D., Thompson, V. A., & Handley, S. J. (2017). When fast logic meets slow belief: Evidence for a parallel-processing model of belief bias. *Memory and Cognition, 45*, 539-552.

Verbeke, G., & Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics and Data Analysis, 23*, 541-556.

Wei, G. C. G. & Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association, 85*, 699-704.

White, H. (1980). A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrika, 48*, 817-838.

Zhang, D., & Lin, X. (2008). Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. In D.B. Dunson (Ed.) *Random effect and latent variable model selection.* (pp.19-36). New York, NY: Springer.

Appendix. R code to Test the Category Effect with $F1/F2$ and GLMMs for an Empirical Study

```
######  ANOVA framework
#### 1. by person
dat<- read.table("behavioral.txt", header=T)
dat1<-aggregate(dat$accu, list(dat$cond, dat$person), mean)
colnames(dat1)<- c("cond","person","accu")
bysub<- aov(accu~ factor(cond)+ Error(factor(person)), data=dat1)
summary(byperson)
shapiro.test((dat$accu))
leveneTest(dat1$accu ~ factor(dat1$cond))

#### 2. by item
dat1<-aggregate(dat$accu, list(dat$cond, dat$item), mean)
colnames(dat1)<- c("cond","item","accu")
byitem<- aov(accu~ factor(cond), data=dat1)
summary(byitem)
shapiro.test((dat1$accu))


####################glmm
library(lme4)

dat$cond[dat$fruit==1] <- 0.5  #'dat$fruit' is dummy coded.
dat$cond[dat$bugs==1] <- -0.5  #'dat$bugs' is dummy coded.

model1<- glmer(accu ~ cond  + (1|person)+(1|item), family=binomial)
model2<- glmer(accu ~ cond  + (1+cond|person)+(1|item), family=binomial)
model3 <- glmer(accu ~ cond + (1|person)+(-1+fruit|item)+(-1+bugs|item),family=binomial)
model4 <- glmer(accu ~ cond + (1+cond|person)+(-1+fruit|item)+(-1+bugs|item),family=binomial)
```