Published in final edited form as:
Stat Med. 2011 August 30; 30(19): 2451–2466. doi:10.1002/sim.4302.

Simulation study of power and sample size for repeated measures with multinomial outcomes: an application to sound direction identification experiments (SDIE)‡

Dingfeng Jiang[*,†] and **Jacob J. Oleson**
Department of Biostatistics, University of Iowa, Iowa City, IA, 52242-1009, USA

Abstract

This study focuses on sample size determination in repeated measures studies with multinomial outcomes from multiple factors. In settings where multiple factors have repeated measures, a single subject could have hundreds of observations. Sample size selection may then refer to the number of subjects, the number of levels within a factor, or the number of repetitions within the level. We simulate multinomial data through a generalized linear mixed model (GLMM) with and without overdispersion, compute the empirical power of detecting group difference for several analytical methods and contrast their performance in group comparison studies with repeated multinomial data. We use four spatial functions to model the spatial correlation structures among observations. We evaluate the factors affecting the power under various scenarios. We also present a dataset typical in hearing studies for sound localization, in which a spatially distributed array of audio loudspeakers plays multiple sounds in order to compare two programming schemes for a hearing aid device.

Keywords

GLMM; multinomial distribution; sample size; sound localization; spatial correlation

1. Introduction

Statistical power and sample size are two major issues during any study design. Most designs require a desired sample size, that is, the number of subjects in order to achieve a specified level of power, typically 80% or 90%. Sample size calculation tools like web applets and commercial software are available to help determine the number of subjects required to obtain a specified level of power. Many of the tools include specific types of mixed models such as repeated measures analysis of variance (ANOVA). When considering repeated measures, the repeated observations can come in a variety of ways, which are not necessarily over time. Although traditional longitudinal studies frequently have a relatively small number of time points that are measured, we consider a setting with hundreds of observations per subject arising from more than one repeated level. In such cases, there are repetitions within a subject from multiple levels, and we may require fewer subjects to obtain the desired power level. Then, the term ‘sample size’ could be referring to (i) how many levels of a factor there should be, (ii) how many repetitions within each level of a factor there could be, or (iii) how many subjects to enroll. This paper examines how these

‡Supporting information may be found in the online version of this article.

*Correspondence to: Dingfeng Jiang, Department of Biostatistics, University of Iowa, Iowa City, IA, 52242-1009 USA.
†dingfeng-jiang@uiowa.edu

three meanings of sample size interplay in sound direction identification experiments (SDIEs).

Multinomial outcomes are common in repeated measures studies. Kuss and Mclerran [1] demonstrated various analysis approaches that keep the multinomial data intact in the form of a generalized linear mixed model (GLMM) using SAS (SAS Institute Inc. 100 SAS Campus Drive Cary NC 27513-2414 USA). For repeated measurements, summary statistics have long been a tool for the analysis [2–4]. Along these lines, analysts often modify repeated multinomial outcomes to create summary statistics for analysis. In some cases, we dichotomize the outcome, rephrasing the research question in which the multinomial data are instead treated as binomial, that is, percent correct. In other cases, we can convert the multinomial data into data that are approximately normal and analyze as such. For example, we can apply a logit or probit transformation to the percent correct from a multinomial outcome, thus creating a continuous variable. We see other means of creating normal data, such as the root mean square (RMS), introduced in the following sections. Because of the transformed outcomes, analysts frequently use statistical models other than GLMM for analysis purposes. We will consider the GLMM setting and give a specific example where the original data are multinomial and compare a GLMM analysis with several frequently used alternative analytic approaches.

Sample size computations for repeated measures have long been an area of research [5–8]. Thompson [9] provided a procedure for selecting sample sizes when the goal is to simultaneously estimate the parameters of a multinomial distribution. Hilton and Mehta [10] developed a sample size determination algorithm for categorical data using linear rank statistics. Nonetheless, sample size computations for the multinomial distribution under the repeated measures setting remain scarce in the literature, which may be because of the fact that a closed form of statistical power of a GLMM is extremely difficult, or impossible, to calculate without further assumptions. The use of summary statistics could reduce the difficulty in sample size determination because the transformation may produce a continuous outcome. Dawson [11] studied how the choice of summary statistics affects the required sample size in a two-sample comparison of slopes in a longitudinal setting. Hedeker, Gibbons, and Waternaux [12] explored sample size estimation when comparing time-related contrasts between two groups under different attrition patterns. Heo and Leon [13, 14] also examined the sample size requirements when the main interest was the intervention effect over time with possible correlation among subjects under different randomization schemes. These studies focused on the intervention effect over time where they assumed only a random effect for subjects. The sample size for mean group difference in a longitudinal setting with multiple random effects requires further study.

We begin by presenting the motivating dataset and field of study. We also introduce two summary statistics commonly used in the proposed field. Then we give general formulas and settings for GLMMs with different correlation structures. We use four common spatial correlation functions to model the factor having spatial correlation among the levels. We apply the proposed statistical tests on two summary statistics, some of which ignore important assumptions. Assume each subject is measured under multiple conditions, and the primary goal is to test for a significant difference between the conditions (groups) in which each condition has multiple factor levels and repeated measurements within each factor level. We compare the summary statistics under various settings and different analytic techniques by focusing on how the empirical power is impacted via different sources of errors through simulation. Simulation study allows for a full examination of the empirical performance of many tests even under small sample size scenarios. The purpose of these comparisons is to compare and contrast their effectiveness in capturing appropriate variability in the data and in detecting significant differences between groups. We present a

table of empirical power under different settings as Supplementary Material, which provides a reference of sample size estimation for appropriate applications. We close with a discussion of the results and implications.

2. Hearing localization studies

The ability to recognize the direction and source of a sound is critical for people in day-to-day living. People with a hearing deficiency or hearing loss may struggle in their ability to localize sound. Hearing aids and cochlear implants are capable of improving hearing dramatically, but there remain questions concerning how well the user can determine the location the sound came from, termed localization. Evaluation of subjects' localization ability with these devices requires a way to quantify the accuracy. A sound direction identification experiment (SDIE) is such an experiment designed to measure the participants' ability to localize sound.

An SDIE is carried out in a quiet room with specific acoustic requirements specified by the American Speech-Language-Hearing Association [15]. In an SDIE, researchers place a total of K audio loudspeakers in the frontal horizontal plane (Figure 1). The speakers are evenly spanned, and the investigator specifies the angle between the adjacent two speakers. Note that it is not necessary for the K speakers to span a half circle. Investigators sequentially assign the numbers 1, 2, …, K to each speaker in a clockwise (or counter clockwise) manner to indicate its relative location in the speaker array. A participant must listen to M sounds from each of the K speakers, that is, $K \times M$ sounds in total. We call the speaker playing the sound the 'expected speaker' for brevity in this paper. Often the sounds represent daily noises such as a bell, phone call, or thunder. The orders of the noises and the expected speakers are both randomized. After each sound is played, the participant localizes the expected speaker. We term the speaker indicated by the participants as the source as the 'observed speaker'.

The relative position of speakers makes it reasonable to expect that the speakers closer to the expected speaker have a higher chance of being reported as the observed speaker by the participants. That is, the greater the distance from the expected speaker, the smaller the probability of being the observed speaker. We refer to this feature that the correlation between the expected speaker and the observed speaker decays as the distance between them grows as spatial correlation. Additionally, the boundary speakers 1 and K behave differently from the non-boundary speakers, and those edge effects also play a role in localization. Localization bias in the results is invariably subject specific. If the listeners cannot determine the expected speaker, some listeners will consistently choose the same speaker as the observed speaker, whereas others will randomly vary their observed speakers. This individual bias is particularly noticeable when the listener's localization ability is low. Therefore, we need a deeper exploration to better understand how various influences and spatial correlation affect the efficiency of an SDIE.

An intuitive measurement for an SDIE is the percentage of correct identification (PCI) [16], which simply calculates the percentage of correct localization. Although the measure ignores proximity to the expected speaker, it is a rather simple measure to compute, and we would like to know how more sophisticated measures perform in relation to PCI. A more common measurement is RMS [15, 17–19]. RMS quantifies the fact that a higher localization accuracy corresponds to a smaller deviation between observed speakers and expected speakers by taking the average squared deviation between response and source. Lower RMS corresponds to better localization precision. We can calculate both measurements at either the speaker level or the subject level. We wish to compare these two summary statistics of the raw multinomial data because, to our knowledge, the comparison

of PCI versus RMS is not done in the literature. We prefer the measurement with higher efficiency in detecting group differences.

Most SDIE studies involve comparisons between groups of participants who may have different hearing aid devices or different settings. We then phrase the research question as 'is there a significant group difference in terms of localization accuracy?' The motivating example uses a crossover design, in which the same subjects take the same tests under two programming schemes within the same device. The two schemes under contrast are the bilaterally mismatched gain reduction scheme and the unaltered bilaterally linear time-invariant amplification scheme. We randomly assign the participants to one scheme and then the other to assess the programming schemes. The random allocation of two schemes balances the possible carry-over effect or learning effect. The replication of subjects in both groups, however, adds another level of correlation in addition to the multiple measurements from each subject.

In the context of an SDIE, the empirical power not only depends on the sample size but also on the combination of speakers, K, and sounds per speaker, M. Given a fixed number of total sounds, there are multiple combinations of K and M, and the power may vary according to the combination of K and M. In practice, the investigator should balance K and M in order to maximize the power. Thus, it is highly desirable to investigate how the number of speakers and the number of sounds per speaker affect the empirical power of detecting the difference of localization ability of groups in an SDIE.

3. Simulation study

In practice, researchers used a multitude of statistical methods for group comparison in factorial experiments with multiple repeated factors. The methods used in SDIE studies are the Wilcoxon rank sum test [20], paired t-test [21], two-way ANOVA [22], and repeated measures ANOVA (RM ANOVA) [15, 23–25]. Therefore, in this paper, the following methods will be compared by simulation: the aforementioned four tests, a two- sample t-test, a Wilcoxon signed rank test, linear mixed models (LMM), three GLMMs with binomial and multinomial distributions, and a generalized estimating equation (GEE) model.

We simulate data from a hierarchical model incorporating three main features of interest in an SDIE: a random effect of speaker (factor A), spatial correlation between the levels of factor A, and a random subject effect. After simulating the data, we apply the statistical methods mentioned in the previous paragraph to both the original multinomial form and the transformed forms. By computing the empirical power as the percentage of a detected difference at the 5% significance level, we assess the performance of the statistical methods used in practice when the underlying correlations are unknown. The specific aspects studied in this paper are: (i) the impact of spatial correlations on the performance of proposed tests; (ii) the performance of two summary statistics under different spatial correlation structures; (iii) the empirical power of the proposed tests; and (iv) the relationship between the empirical power and the parameters of interest, for example, the localization accuracy, the variance components, and the number of repeated values from two different sources, that is, K and M.

3.1. Hierarchical model of generating data

We break the process of generating simulated data into three main components as described in the following subsections. Note that in the data-generating step, we use four spatial correlation structures to capture the possible spatial correlation among the speaker array. In the analysis step, however, such prior information is purposely blocked to imitate the

analysis in practice because the true underlining spatial correlation is unknown to the analyst.

3.1.1. Multinomial data process—Let i be group index ($i = 1, \ldots, I$), j be participant index ($j = 1, \ldots, J$), k be the level of factor A ($k = 1, \ldots, K$), and m be repeated observations within factor A ($m = 1, \ldots, M$). X_{ijkl} denotes the total number of times level l ($l = 1, \ldots, K$) of factor A is reported as the multinomial outcome by subject j in group i when level k is the true outcome. Because level k is repeated M times, we have $X_{ijkl} \in [0, M]$. Assuming independence between responses, X_{ijkl} is distributed as a multinomial distribution, that is,

$$\boldsymbol{X_{ijk}} = (X_{ijk1}, \ldots, X_{ijkK})' \sim Multinomial(M, \boldsymbol{\theta_{ijk}} = (\theta_{ijk1}, \ldots, \theta_{ijkK})'), \quad (1)$$

where θ_{ijkl} denotes the probability of level l being reported as the outcome by subject j in group i with level k being the true outcome. For example, suppose we played eight sounds from expected speaker 1 (six speakers in total), and a participant reported the observed speaker as (1, 3, 2, 1, 2, 5, 1, 2), then $X_{ij11} = 3$, $X_{ij12} = 3$, $X_{ij13} = 1$, $X_{ij14} = 0$, $X_{ij15} = 1$, $X_{ij16} = 0$.

3.1.2. Spatial correlation—We expect the outcomes X_{ijkl} to be correlated with one another. Let d_{kl} denote the distance between levels k and l of factor A. The correlation between observations would decay as the distance between k and l, d_{kl}, increases. For example, in the SDIE, d_{kl} would be the distance of observed speaker l from the expected speaker k, $|l - k|$. That distance could be in terms of speaker angles or in terms of simply numbering speakers from 1 to K around the listener. We define below how that correlation can be factored into the multinomial model.

With the constraint $\sum_{l=1}^{K} \theta_{ijkl} = 1$ in the multinomial distribution, only the probability of a correct response, θ_{ijkk}, denoted as η_{ijk}, needs to be specified to simulate θ_{ijkl} if the following functional form is used:

$$\theta_{ijkl} = \begin{cases} \eta_{ijk}, & l = k \\ \frac{f(d_{kl})}{\sum_l f(d_{kl})}(1 - \eta_{ijk}), & l \neq k \end{cases}. \quad (2)$$

The function $f(\cdot)$ is a general representation of spatial functions that depends on a distance measure. We assume, for simulation purposes, equal spacing among levels. This study considers four common functions for spatial correlation in the speaker array as listed below:

- Exponential model: $f(x) = exp(-|x|)$,
- Gaussian model: $f(x) = exp(-x^2)$,
- Inverse distance model: $f(x) = |x|^{-1}$,
- Power model: $f(x) = \rho^{|x|}$, $\rho = 1/2$ in the study to present a median level of correlation.

3.1.3. Generalized linear mixed model—Given the structure defined in equation (2), the probability of level l being the response depends solely on η_{ijk} and the distance d_{kl}. Thus, only η_{ijk} and the total number of levels, K, are further needed to generate the multinomial distribution of X_{ijkl} in equation (1). To simulate η_{ijk}, a GLMM is assumed such that

$$g(\eta_{ijk})=\alpha_i+s_j+a_k+\varepsilon_{ijk}, \quad (3)$$

where g is the logit link function, α_i is the ith group mean and treated as a fixed effect; s_j and a_k denote the random effects of subjects and factor A, that is, speakers, respectively, which we assume to have normal distributions, that is, $s_j \sim N(0, \sigma_s^2)$, $a_k \sim N(0, \sigma_a^2)$; ε_{ijk} denotes the measurement error with a normal distribution, $\varepsilon_{ijk} \sim N(0, \sigma^2)$.

Note that under a canonical link, α_i can be viewed as $\alpha_i = \text{logit}(\delta_i)$, with δ_i being the average ability of localization accuracy for group i, which is of primary interest in an SDIE application. The random effect s_j aims to capture individual characteristics of each participant, whereas random effect a_k aims to capture the individual characteristics of each speaker. The error component ε_{ijk} aims to simulate the case of overdispersion, which we commonly see in categorical data.

To simulate the experiment with known K and M using the hierarchical model, only the probability of correct response for groups, δ_i, $i = 1, 2, \ldots, I$, as well as the three error components, σ_s^2, σ_a^2, and σ^2, need to be specified.

3.2. Statistical analysis

Recall that the purpose of the proposed tests is to detect a significant group difference. In terms of SDIE, we are interested in the difference of average localization accuracy, that is, $\delta_1 - \delta_2$. Because the true underlying correlation among the speaker array is unknown to the analyst, we deliberately ignore the information of the true spatial model generating the data in the analysis step. Similarly, we also ignore the information regarding whether we simulate the data with or without overdispersion in the analysis. Therefore, we know the models to be mis-specified as we examine the test procedures to find robustness against spatial correlation and overdispersion. We introduce first the tests based on the multinomial form followed by the tests based on the summary statistics.

A GLMM assuming an ordinal multinomial distribution is the most appropriate model for an SDIE experiment where ordinal refers to the relative distance from observed speaker to expected speaker. However, the extremely low rate of convergence during simulations makes the model difficult to use. We consider two alternatives: a GLMM with nominal multinomial distribution and a GLMM with a binomial distribution. We can write both models as

$$\eta_{ijk}=h(\mu_i+a_j+b_k) \quad (4)$$

for level k, subject j, and group i, h is the canonical link function, and μ_i is the fixed effect of group mean; a_j and b_k are the independent random effects of subject j and level k, respectively, that is, $a_j \sim N(0, \sigma_a^2)$ and $b_k \sim N(0, \sigma_b^2)$. The incorporation of the random effect of level k accounts partially for the spatial correlation imposed by the underlying model, although the true spatial correlation is not included in the analysis. The nominal multinomial GLMM is fit assuming a Poisson distribution with a log link function because the likelihoods of the two models are the same as shown by Chen and Kuo [26]. Both the binomial GLMM and nominal multinomial GLMM use a Laplace approximation of the likelihood for model fitting [27]. The null hypothesis is H_0: $\mu_1 = \mu_2$, that is, no group difference.

Generalized estimating equation is another popular tool for repeated measures data. It produces a consistent parameter estimate regardless of the working correlation [28,29]. The

population-averaged model accounts for the correlation among repeated measures by using different working correlation structures. The subject-specific models, such as LMM and GLMM, not only account for the correlation of repeated measures but can also explain the source of said correlation. The working correlation matrix models the correlation among speakers.

The multinomial outcome is often difficult to analyze, so analysts have used various approaches to accommodate for this difficulty [1]. Summary statistics are a common choice. In some cases, we dichotomize the outcome, and a percentage of correct responses becomes the outcome of interest. We denote the logit of percentage of correctly identified answers as LPCI for the remainder of the paper. In other cases, we manipulate the data to make them continuous and assume that they follow a normal distribution. One method is to compute a relative measure of correctness such as RMS. For example, consider the SDIE study. Given the multinomial distribution of X_{ijkl}, we can calculate the LPCI and RMS of speaker k by participant j in group i as follows:

$$\text{LPCI}_{ijk}=\text{logit}\left(\frac{X_{ijkk}}{M}\right);\quad \text{RMS}_{ijk}=\sqrt{\frac{\sum_{l=1}^{K}X_{ijkl}(l-k)^2}{M}}.$$

Using the example given in Section 3.1.1, the LPCI is $\text{LPCI}_{ij1}=\text{logit}(\frac{X_{ij11}}{M})=\text{logit}(\frac{3}{8})=-0.511$. The RMS value is

$$\text{RMS}_{ij1}=\sqrt{\frac{3*(1-1)^2+3*(2-1)^2+\ldots+1*(5-1)^2+0*(6-1)^2}{8}}=\sqrt{23}=4.796.$$

The LPCI and RMS are the main summary statistics discussed in this paper. Simulation results suggest that among the tests using summary statistics, repeated measures ANOVA and LMM are superior to the two-sample t-test, paired t-test, Wilcoxon rank sum test, Wilcoxon signed rank test, and two-way ANOVA; thus, contents related to these latter models are not reported in this paper. Therefore, comparisons are made among GEE, binomial and multinomial GLMMs, repeated measures ANOVA, and LMM. We outline next how the statistical models using two summary statistics are written under the aforementioned testing procedures. Each of those procedures can be written in terms of a linear model where the outcome of interest is a function of the true outcome, $f(Y)$. For example, $f(Y)$ = logit(PCI) or $f(Y)$ = RMS.

Assuming fixed effects of groups and factor levels and accounting for the correlation of multiple measurements from the same subjects, we can use repeated measures ANOVA, assuming the sphericity condition is satisfied. The model is

$$f(Y)_{ijk}=\mu_i+a_j+\beta_k+\varepsilon_{ijk} \tag{5}$$

for level k, subject j, and group i; μ_i is the ith group effect; a_j is the normally distributed random effect of subject j; β_k is the kth level effect; ε_{ijk} is the normally distributed measurement error. The null hypothesis is H_0: $\mu_1 = \mu_2$, that is, no group difference. We apply the same model to the ranks of the data for a non-parametric analysis.

Linear mixed model is fitted due to its ability to handle spatial correlation and the correlation of multiple measurements from the same subject and factor level. The test assumes random effects for the levels and random effects for the subjects. To increase the speed of the analyses of the simulated data, we use an independence correlation structure for the error component. The mixed model is

$$f(Y)_{ijk}=\mu_i+a_j+b_k+\varepsilon_{ijk}, \quad (6)$$

for level k, subject j, and group i; μ_i is the fixed effect of the group mean; a_j and b_k are the normally distributed random effects of subject j and level k, respectively; ε_{ijk} is the measurement error with a normal distribution. The null hypothesis is H_0: $\mu_1 = \mu_2$, that is, no group difference. We also apply The same model to the ranks of the raw data.

We define a clinically meaningful effect size to occur when the difference between two groups' localization ability is larger than random guessing, that is, $|\delta_1 - \delta_2| \geq 1/K$. For a particular SDIE, a participant is considered to have good localization ability if he/she can differentiate the source from two adjacent speakers. We report the results based on 2,000 replications in simulation. We analyze the GEE under the multinomial distribution using SAS 9.2 [30]. The remaining models are fitted by R 2.11.1 [31], including the LMM and GLMM, which were fit using the 'lme4' package by Bates and Maechler [32].

4. Simulation results

Regardless of the test procedures used, we calculate the empirical power presented below by counting the percentage of rejecting the null hypotheses of no group difference ($\delta_1 - \delta_2$) for each model. The correlation coefficient in the power spatial function is fixed at $\rho = 0.5$ to present a median level of correlation during the simulation. We choose parameters reported below for the sake of illustration. We test other values to ensure that the patterns and relationships are not dependent on the choice of parameters.

4.1. Comparison of spatial models

In simulations, we set K to be 7, which represents the median number of speakers typically used in an SDIE. The remaining parameters are also common in SDIE settings with $J = 5$, $M = 9$, $\delta_1 = 2/7$, $\delta_2 = 3/7$, $\sigma_s^2=\sigma_a^2=0.1$, and $\sigma^2 = 1.0$. Type I error rate is set to 0.05.

Table I presents the empirical power of various tests under the four true models that generated the data. The GEE model operates best under the Gaussian model, whereas the other three spatial models performs similarly. The empirical power of the binomial GLMM under all four spatial models are close to each other, indicating that the spatial model has very little impact on the binomial GLMM model. Such a phenomenon is reasonable because the binomial GLMM only counts the correct responses as an index for the localization accuracy, and the correct responses are determined by δ_i rather than the postulated spatial functions. Therefore, the robustness of binomial GLMM to spatial correlation is expected. For the nominal multinomial GLMM, the spatial model has a prominent effect on performance. The multinomial GLMM performs best under the inverse distance model, followed by the power model, then the exponential model, and worst under the Gaussian model.

The two summary statistics also behave differently. Using LPCI, the empirical power under all four spatial models varies slightly for a given test procedure. This indicates that LPCI is robust to spatial correlation as the binomial GLMM. Additional simulations under different settings give similar results. Therefore, the choice of spatial correlation model does not

impact the behavior of LPCI. In contrast to LPCI, RMS is indeed sensitive to the spatial correlation of speakers as the empirical power changes. The Gaussian model outperforms the remaining three spatial models within the same test procedure, whereas the other three spatial correlation models perform similarly. Due to these characteristics, it is sufficient for further simulations to discuss the behavior of LPCI under the exponential model and the RMS under both the exponential and Gaussian models.

4.2. Comparison of tests

In order to provide a general comparison of the tests, we simulate scenarios both with and without overdispersion by using model 3. To compare the performance of proposed tests in practice, we deliberately choose the error parameters close to estimates from the SDIE example. For the model with overdispersion, the parameters are σ_s^2=0.05, σ_a^2=0.15, σ^2 = 1.5, J = 5, K = 7, M = 9, δ_1 = 2/7, and δ_2 = 3/7. For the model without overdispersion, the parameters are σ_s^2=0.06, σ_a^2=0.13, σ^2 = 0, J = 3, K = 7, M = 9, δ_1 = 2/7, and δ_2 = 3/7.

Table II displays the empirical power of the proposed tests under both scenarios. Note that the spatial models and overdispersion referred to in Table II are the true models generating the data. For the overdispersed case, the binomial GLMM achieves the highest power, followed by the multinomial GLMM, GEE, and then tests using summary statistics. Among the tests using the summary statistics, LPCI operates better than RMS within the same test procedure. Among the tests using LPCI, the non-parametric versions have higher power than the corresponding parametric versions, whereas for the tests using RMS, parametric and nonparametric versions of the tests operate similarly. For LPCI, LMM achieves the highest power by using ranks; for RMS, LMM achieves the highest power using raw data, which is very close to LMM using ranks.

For the case without overdispersion, binomial GLMM achieves the highest power. In contrast to the overdispersion case, GEE demonstrates better performance than the multinomial GLMM model. For the tests using summary statistics, LPCI is still more sensitive than RMS to detect a difference, and ranking LPCI improves the power under all tests. Similar to the overdispersion model, among the tests based on LPCI, the highest power is achieved by LMM using ranks; among those using RMS, LMM using raw data is the best for the exponential model, and LMM using rank data is the best for the Gaussian model.

4.3. Comparison of empirical type I error rate

The empirical type I error rate measures the false positive rate of a test, that is, reporting a group difference although there is none. The setup of the simulation is similar to Section 4.2, with the change of no group effect, $\delta_1 = \delta_2 = 2/7$.

Table III displays the empirical type I error for the tests under various situations given the true models generating the data. For the overdispersed model, the empirical type I error rates of the binomial GLMM is considerably higher than the other two. The type I error rate from the multinomial GLMM model is also high under the exponential, power, and inverse distance spatial models. The remaining type I error rates are approximately the nominal 0.05. Note that the type I error of the test using LPCI is always higher than those using RMS, and the type I error rate for the tests using RMS under the Gaussian model is consistently higher than the same test using RMS under the exponential model.

For the model without overdispersion, the empirical type I error is close to the nominal 0.05. In contrast to the model with overdispersion, the LPCI type I error is nearly uniformly lower than RMS. However, among the tests using RMS, the type I error under a Gaussian model is still higher than those under the exponential model.

Based on the comparison of power and type I error rates, in the remainder of the paper, we focus on the multinomial GLMM under a Gaussian spatial model, LMM using ranks of LPCI under an exponential model, and LMM using RMS under exponential and Gaussian models for the overdispersion case. For the case without overdispersion, the primary focus is on the binomial GLMM under an exponential model, LMM using ranks of LPCI under an exponential model, and LMM using RMS under exponential and Gaussian models.

4.4. Evaluating empirical power with variance components

To study the impact of variance components, that is, σ_s^2, σ_a^2, and σ^2 on the overdispersed model and σ_s^2 and σ_a^2 on the model without overdispersion, we sequentially set the value of the component being studied to be (0.1, 1, 10), whereas the others are fixed at 1. The remaining parameters are $K = 7$, $M = 9$, $\delta_1 = 2/7$, and $\delta_2 = 3/7$ for both models and $J = 5$ for the overdispersed model and $J = 3$ for the model without overdispersion.

Figure 2 shows the impact of variance components on the empirical power for the overdispersed case. The three error components impact the multinomial GLMM (Gaus) similarly by comparing the scale of change across the three plots. For the LMM using summary statistics, the measurement error, that is, overdispersion error, has the greatest effect on the empirical power. Among the three tests using summary statistics, LMM using RMS under an exponential model is the most robust to the change of error components, whereas LMM using the ranks of LPCI is the most sensitive.

Figure 3 shows the impact of variance components on the empirical power for the case without overdispersion. As seen in the overdispersed model, the impacts of σ_s^2 and σ_a^2 are comparable. Among the four tests, LMM using ranks of LPCI may be the most sensitive test to the change of variance components, whereas LMM using RMS under the exponential model is the most robust to those changes. Increasing the values of the variance components draws the same impact whether using the LMM with RMS under a Gaussian model or the binomial GLMM under an exponential model.

4.5. Evaluating empirical power with localization accuracy

We study the relationship between the empirical power and the localization accuracy by considering two scenarios: (1) a fixed probability of localization for one group and (2) a fixed difference between the probability of localization for two groups. Scenario (1) imitates the comparison between an experimental device to a standard one, whereas scenario (2) investigates the change of empirical power under the same difference between localization accuracy but with increasing localization ability for both groups. The parameters are set as $J = 3$, $K = 7$, $M = 9$, σ_s^2=0.1, and σ_a^2=0.5, $\sigma^2 = 5$ for scenario (1) and $J = 5$, $K = 7$, $M = 9$, σ_s^2=0.1, σ_a^2=0.1, and $\sigma^2 = 1$ for scenario (2).

In scenario (1), that is, δ_1 fixed at 1/7, δ_2 ranging over 2/7 – 6/7, the power increases as δ_2 gets larger (Figure 4 left panel). A similar pattern is also observed for the model without overdispersion.

In scenario (2), that is, $\delta_2 - \delta_1$ fixed at 1/7, δ_1 ranging over 1/7 – 5/7, the empirical power for the four tests first decreases then increases as δ_1 increases (Figure 4 right panel). For the tests based on the summary statistics, the value 0.5 is the cutoff point for power change. Such a nonlinear trend of power reflects the Bernoulli nature of the outcome given the fact that the highest variance is achieved when the probability is 0.5. Note that LMM using ranks of LPCI (Expo) and LMM using RMS (Gaus) have a more obvious ‘v’ shape than the other two tests. This indicates that the logit transformation in equation (3) has greater impact on

these two tests. Also note that LMM using ranks of LPCI (Expo) and LMM using RMS (Gaus) have a higher power than the other two when δ_i s are small. This suggests that these two tests are more appropriate for testing the difference of populations with lower localization accuracy. For the model without overdispersion, we also detect the 'v' shape for scenario (2).

4.6. Evaluating empirical power with K and M

For the purpose of illustration, we choose different combinations of K and M to give approximately $K \times M = 96$ sounds in total for the case with overdispersion and approximately $K \times M = 50$ sounds for the case without overdispersion. In the simulations, the level of factor A, that is, the number of speakers, K, ranges from 5 to 16. We set the parameters as $J = 3$, $\sigma_s^2 = \sigma_a^2 = 0.1$, $\sigma^2 = 1$, $\delta_1 = 2/7$, and $\delta_2 = 3=7$ for the overdispersed model and the same but without σ^2 for the model without overdispersion.

The left panel of Figure 5 shows the power for the four tests under different combinations of K and M for the overdispersed model. For the multinomial GLMM (Gaus), the power varies slightly as K changes. This indicates that the efficiency of detecting a group difference for an SDIE will not increase much by increasing the levels of factor A, that is, number of speakers, if the correlation structure of the speaker array is close to a Gaussian spatial model and the multinomial GLMM model is used. However, for the tests using summary statistics, the power increases as the number of levels increases. This is particularly obvious for LMM using ranks of LPCI (Expo) and LMM using RMS (Gaus). A useful message is that we can increase the efficiency of SDIE by increasing the number of speakers when adopting the LMM using ranks of LPCI or LMM using RMS under a Gaussian spatial model.

Contrary to the model with overdispersion, we do not observe the same increasing pattern of tests using summary statistics without overdispersion as the right panel of Figure 5 shows. Such consistency of the empirical power with respect to K suggests the efficiency of detecting the group difference should rely on other mechanisms other than increasing the number of speakers.

4.7. Summary of simulation results

Spatial correlation models have greater impact on the multinomial GLMM and GEE models than on the binomial GLMM. In addition, LPCI is more robust to spatial correlation than RMS. Note that ranking the raw data improves the performance of LPCI summary statistic for both overdispersed and not overdispersed data.

For overdispersed data, the type I error is too high for the binomial GLMM, and the multinomial GLMM only performed well when the underlying spatial correlation is Gaussian. Thus, GEE and LMM using LPCI are better analysis choices for overdispersed data where the error component in the analysis models protects the empirical power against the overdispersed data. Increasing the number of speakers K also improves the performance of LMM based on the summary statistics.

For the data without overdispersion, binomial GLMM is preferred, although GEE and multinomial GLMM generally perform well. LMM using LPCI is a viable alternative. The variance components of speakers and subjects have similar effects on the power. In this case, increasing the number of speakers has little effect on the performance of the test procedures.

5. Application

We apply the proposed tests to an example SDIE to illustrate how the results change with the choice of tests and summary statistics. For the example study, we also assess the design of the study, that is, to examine whether the current setup can achieve the required power and/or how many subjects are needed to achieve that power.

This SDIE study recruited 24 bilateral hearing aid users with age older than 20. All the participants had bilateral sensorineural hearing loss and had at least 1 year experience with hearing aids. We designed the study to address whether the localization performance in the frontal horizontal plane would be negatively affected when we would give hearing aid users a bilaterally mismatched gain reduction scheme as compared with an unaltered bilaterally linear time-invariant amplification scheme. We randomly gave the participants one scheme and then the other to exclude possible learning effect and carry-over effect. During the test, we placed the subjects in the center of an arc with diameter 1 m; nine speakers were evenly spanned from −60° to 60° azimuth. Ten stimulus sounds including telephone ring, buzzer, child laughing etc, were played randomly from the speakers.

Using summary statistics requires some attention to check the goodness of fit of the model used. As in the simulation, the residuals of LMM using summary statistics appear nearly normal, and considering the robustness of LMM to the non-normally distributed residuals, the results hold.

Table IV presents *p*-values for a group difference reported by various tests applicable to the example dataset. The results show that a significant difference would be reported for GLMMs and GEE, except the binomial GLMM model assuming overdispersion. Note that the binomial GLMM assuming overdispersion is fit by using the residual subject-specific pseudo-likelihood technique (RSPL) in SAS 9.2, whereas other GLMMs are fit by Laplace approximation technique, which is more accurate than RSPL in terms of numerical precision [33]. For tests using summary statistics, all the tests using LPCI would report a significant difference, whereas only the paired *t*-test and Wilcoxon signed rank test would report a significant difference for those using RMS.

The mean probabilities of correct localization for the two groups are $\hat{\delta}_1 = 0.2898$ and $\hat{\delta}_2 = 0.3444$. The estimated error components are $\widehat{\sigma}^2_{sub}=0.044$, $\widehat{\sigma}^2_{spk}=0.141$, and $\hat{\sigma^2} = 1.497$ for the overdispersed model and $\widehat{\sigma}^2_{sub}=0.057$ and $\widehat{\sigma}^2_{spk}=0.146$ for the model without overdispersion. We use these values to assess the empirical power of tests for the current setting, that is, $J = 24$, $K = 9$, and $M = 10$. Based on the simulation results, for the model without overdispersion, the empirical power of four tests (binomial GLMM (Expo), LMM using ranks of LPCI (Expo), and LMM using RMS (expo and Gaus)) are all greater than 0.95. For the model with overdispersion, the empirical power of LMM using RMS (Expo) is around 0.88, whereas the remaining three (multinomial GLMM (Gaus), LMM using ranks of LPCI (Expo), and LMM using RMS (Gaus)) are also greater than 0.95. This suggests that the current design might be overpowered if these test procedures are used.

This paper also provides the empirical power under different designs for SDIE studies, which serves as a guide for sample size estimation in the practice of SDIEs (see Supplementary Material, which presents the empirical power of an SDIE under different settings). For generalization, we only consider the overdispersed model. As shown in Section 4.2, within the same test, LPCI is more sensitive for testing the group difference. Thus, we only consider LMM using ranks of LPCI as a representative for tests using summary statistics. We also only consider the Gaussian spatial function to speed up the simulation for two reasons. The first is that LPCI is robust to the specification of spatial

function so it should have similar performance for four spatial models. The second is that a multinomial GLMM only operates well under a Gaussian distribution model. In other words, the Gaussian model has higher power and a reasonable type I error rate when compared with the other three spatial functions. We choose the parameters as $J = 1, 3, 5, 7$, $K = 7, 9, 11$, and $M = 8, 10, 12$. The error components are the estimates from the SDIE example.

If crude estimates of δ_1 and δ_2, the group localization accuracy, are available from experience, the investigator can use the table to find the recommended sample size. Another way to use the table is to choose an appropriate combination of K and M in order to attain a desired level of power if only a limited number of subjects are available, which is often the case in practice. The values for $J = 1$ are important for those researchers interested in a single-subject design.

6. Discussion

In this paper, we compare the localization ability of participants by focusing on different statistical procedures' capabilities to detect group differences. An ideal statistical procedure would be sensitive to the difference of localization ability but robust to the setting of the experiment such as the correlation structure, overdispersion of the data, and variance components.

Results confirm that RMS is more sensitive to the choice of spatial models than LPCI because LPCI only accounts for the point probability, η_{ijk}, rather than the probability vector $\boldsymbol{\theta}_{ijk}$, which varies according to the postulated spatial function. Four spatial models are proposed to model the correlation of factor A, but, in practice, it may be difficult to determine which model fits the data best. Therefore, summary statistics with robustness to spatial correlation, like PCI, are recommended. Ranking PCI helps the performance of PCI, particularly for small sample sizes.

Because RMS uses both the correct and incorrect responses to construct the outcomes, the information from the incorrect responses biases and dilutes information from the correct responses. This may explain why PCI has higher empirical power than RMS if direct comparisons are made. However, RMS captures the proximity information, whereas PCI does not. If proximity localization is part of the research interest, then RMS would be a better measurement.

The choice of whether to fit a GLMM with overdispersion or not depends on the distribution of the data. Fitting an optimal GLMM would depend on that choice, and several tests are available to help such as those developed by Dean [34] or Kim and Margolin [35]. However, LMM produces an empirical power comparable to the highest GLMM with an acceptable type I error rate both with and without overdispersion as shown in Tables II and III. Thus, among the proposed analytical methods using summary statistics, LMM is recommended to analyze an SDIE with a crossover design.

An ordinal multinomial GLMM should be the best model for the multinomial longitudinal outcomes, specifically an SDIE experiment. GEE modeling would be an alternative if the marginal effects are the only interest. However, the sample sizes in practice are typically less than 20 [15,21,36], even one [19]. As previously stated, the convergence of the GLMM algorithm arises for small sample sizes. For GEE, the existing software programs also have limitations. For instance, SAS PROC GENMOD only provides an independent correlation structure for the multinomial data. Although the use of the 'sandwich' standard error helps, the lack of adequate correlation structures in statistical software certainly limits the usage of the methods.

Summary statistics are popular in the field of SDIE studies and provide an alternative solution given the technical difficulties in analyzing multinomial data. Using summary statistics enables exploration of more effects than factor A and subjects. Other research questions can also be asked within the SDIE experiment. For example, the replicated sound being played may be modeled as another random effect. It might be reasonable to expect that localization of a high pitch sound would be different from the localization of low pitch sound. In studies with background noise [37], we can also explore the interaction between the signal-to-noise and programming schemes. Because the motivating example concerns the group difference only, we did not explore these aspects.

In our modeling of spatial correlation, we assumed a symmetric localization ability. From a scientific perspective, an asymmetric localization is possible for hearing aid users. In that case, we may use anisotropic spatial structures to model the correlation. We did not consider the edge effect in the analysis step either. A possible way to handle such effects would be to create a separate random effect for the edge speakers. Incorporating asymmetric localization and edge effects would further complicate the design and analysis. Another aspect that the simulation neglects is the minimum angle that a subject can resolve. Theoretically, each person has a limit to the angular distance that he/she can resolve. Beyond that limit, the subject response may only be random guessing. Under that scenario, both PCI and RMS are essentially measuring the responses from random guessing. The phenomenon can happen if the speakers are too close together due to the large K and should be avoided by improving the design of an SDIE.

We may also apply the results of this study to other summary statistics found in literature. Usually, researchers use these measurements together with RMS and/or PCI. The azimuth error computes the absolute value of the difference between the source azimuth and the response azimuth [25, 38]. Mean absolute error is another measure used in an SDIE [39], which takes the absolute difference between the source and response numbers. Both measurements use the information from the incorrect responses as RMS does. Thus, we expect them to be sensitive to the postulated spatial functions and to have an inferior efficiency in testing the group difference.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank two referees and the associate editor for their helpful comments in improving the quality of the paper. This research was supported in part by research grant 5 P50 DC00242 from the National Institutes on Deafness and Other Communication Disorders, National Institutes of Health; grant MO1-RR-59, National Center for Research Resources, General Clinical Research Centers Program, National Institutes of Health; the Lions Clubs International Foundation; and the Iowa Lions Foundation. The authors would like to thank Hua Ou and Ruth Bentler for usage of the SDIE.

References

1. Kuss O, McLerran D. A note on the estimation of the multinomial logistic model with correlated responses in SAS. Computer Methods and Programs in Biomedicine. 2007; 87(3):262–269.10.1016/j.cmpb.2007.06.002 [PubMed: 17686544]
2. Frison L, Pocock SJ. Repeated measures in clinical trials: analysis using mean summary statistics and its implication for design. Statistics in Medicine. 1992; 11(13):1685–1704. [PubMed: 1485053]
3. Dawson JD, Lagakos SW. Size and power of two-sample tests of repeated measures data. Biometrics. 1993; 49(4):1022–1032. [PubMed: 7906957]

4. Weinberg JM, Lagakos SW. Linear rank tests under general alternatives, with application to summary statistics computed from repeated measures data. Journal of Statistical Planning and Inference. 2001; 96(1):109–127.10.1016/S0378–3758(00)00328–1
5. Rochon J. Sample size calculations for two-group repeated-measures experiments. Biometrics. 1991; 47:1383–1398.
6. Overall JE, Doyle SR. Estimating sample sizes for repeated measurement designs. Controlled Clinical Trials. 1994; 15(2):100–123.10.1016/0197–2456(94)90015–9 [PubMed: 8205802]
7. Rochon J. Application of GEE procedures for sample size calculations in repeated measures experiments. Statistics in Medicine. 1998; 17(14):1643–1658.10.1002/(SICI)1097–0258(19980730) [PubMed: 9699236]
8. Basagaña X, Spiegelman D. Power and sample size calculations for longitudinal studies comparing rates of change with a time-varying exposure. Statistics in Medicine. 2010; 29(2):181–192.10.1002/sim.3772 [PubMed: 19899065]
9. Thompson SK. Sample size for estimating multinomial proportions. The American Statistician. 1987; 41(1):42–46.
10. Hilton JF, Mehta CR. Power and sample size calculations for exact conditional tests with ordered categorical data. Biometrics. 1993; 49(2):609–616. [PubMed: 8369392]
11. Dawson JD. Sample size calculations based on slopes and other summary statistics. Biometrics. 1998; 54(1):323–330. [PubMed: 9544525]
12. Hedeker D, Gibbons RD, Waternaux C. Sample size estimation for longitudinal designs with attrition: comparing time-related contrasts between two groups. Journal of Educational and Behavioral Statistics. 1999; 24(1):70–93.10.3102/10769986024001070
13. Heo M, Leon AC. Statistical power and sample size requirements for three level hierarchical cluster randomized trials. Biometrics. 2008; 64(4):1256–1262.10.1111/j.1541–0420.2008.00993.x [PubMed: 18266889]
14. Heo M, Leon AC. Sample size requirements to detect an intervention by time interaction in longitudinal cluster randomized clinical trials. Statistics in Medicine. 2009; 28(6):1017–1027.10.1002/sim.3527 [PubMed: 19153969]
15. Neuman AC, Haravon A, Sislian N, Waltzman SB. Sound-direction identification with bilateral cochlear implants. Ear and Hearing. 2007; 28(1):73–82.10.1097/01.aud.0000249910.80803.b9 [PubMed: 17204900]
16. Firszt JB, Reeder RM, Skinner MW. Restoring hearing symmetry with two cochlear implants or on cochlear implant and a contralateral hearing aid. Journal of Rehabilitation Research and Development. 2008; 45(5):749–768.10.1682/JRRD.2007.08.0120 [PubMed: 18816424]
17. Hartmann WM. Localization of sound in rooms. Journal of Acoustical Society of America. 1983; 74(5):1380–1391.10.1121/1.390163
18. Hartmann WM, Rakerd B, Gaalaas JB. On the source-identification method. Journal of Acoustical Society of America. 1998; 104(6):3546–3557.10.1121/1.423936
19. Tyler RS, Noble W, Dunn C, Witt S. Some benefits and limitations of binaural cochlear implants and our ability to measure them. International Journal of Audiology. 2006; 45(7):113–119.10.1080/14992020600783095
20. Seeber BU, Baumann U, Fastl H. Localization ability with bimodal hearing aids and bilateral cochlear implants. Journal of Acoustical Society of America. 2004; 116(3):1698–1709.10.1121/1.1776192
21. Chung K, Neuman AC, Higgins M. Effects of in-the-ear microphone directionality on sound direction identification. Journal of Acoustical Society of America. 2008; 123(4):2264–2275.10.1121/1.2883744
22. Despres O, Boudard D, Candas V, Dufour A. Enhanced self-localization by auditory cues in blind humans. Disability and Rehabilitation. 2005; 27(13):753–759.10.1080/09638280400014865 [PubMed: 16096227]
23. Bosman AJ, Snik AFM, van der Pouw CTM, Mylanus EAM, Cremers CWRJ. Audiometric evaluation of bilaterally fitted bone-anchored hearing aids. Audiology. 2001; 40(3):158–167. [PubMed: 11465298]

24. Figueiredo JC, Abel SM, Papsin BC. The effect of the Audallion(R) BEAMformer noise reduction preprocessor on sound localization for cochlear implant users. Ear and Hearing. 2001; 22:539–547. [PubMed: 11770675]
25. D'Angelo WR, Bolia RS, Mishler PJ, Morris LJ. Effects of CIC hearing aids on auditory localization by listeners with normal hearing. Journal of Speech Language, and Hearing Research. 2001; 44(6):1209–1214.10.1044/1092–4388(2001/094)
26. Chen Z, Kuo L. A note on the estimation of the multinomial logit model with random effects. The American Statistician. 2001; 55(2):89–95.
27. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association. 1993; 88(421):9–25.
28. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 73(1):13–22.10.1093/biomet/73.1.13
29. Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. Biometrics. 1988; 44(4):1049–1060. [PubMed: 3233245]
30. SAS Institute Inc. SAS 9.2 help and documentation. Cary, NC: SAS Institute Inc; 2010.
31. R develop core team. R: a language and environment for statistical computing. 2010. URL http://www.R-project.org
32. Bates, D.; Maechler, M. lme4: linear mixed-effects models using s4 classes. 2009. URL http://CRAN.R-project.org/package=lme4, R package version 0.999375-32
33. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White JSS. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology & Evolution. 2009; 24(3):127–135.10.1016/j.tree.2008.10.008 [PubMed: 19185386]
34. Dean CB. Testing for overdispersion in poisson and binomial regression models. Journal of the American Statistical Association. 1992; 87(418):451–457.
35. Kim BS, Margolin BH. Testing goodness of fit of a multinomial model against overdispersed alternatives. Biometrics. 1992; 48(3):711–719.
36. Chan JCY, Freed DJ, Vermiglio AJ, Soli SD. Evaluation of binaural functions in bilateral cochlear implant users. International Journal of Audiology. 2008; 47(6):296–310.10.1080/14992020802075407 [PubMed: 18569102]
37. Good MD, Gilkey RH. Sound localization in noise: the effect of signal-to-noise ratio. The Journal of the Acoustical Society of America. 1996; 99(2):1108–1117.10.1121/1.415233 [PubMed: 8609294]
38. Litovsky RY, Parkinson A, Arcaroli J. Spatial hearing and speech intelligibility in bilateral cochlear implant users. Ear and Hearing. 2009; 30(4):419–431.10.1097/AUD.0b013e3181a165be [PubMed: 19455039]
39. Van Deun L, van Wieringen A, Van den Bogaert T, Scherf F, Offeciers FE, Van de Heyning PH, Desloovere C, Dhooge IJ, Deggouj N, De Raeve L, Wouters J. Sound localization, sound lateralization, and binaural masking level difference in yong children with normal hearing. Ear and Hearing. 2009; 30(2):178–190.10.1097/AUD.0b013e318194256b [PubMed: 19194296]

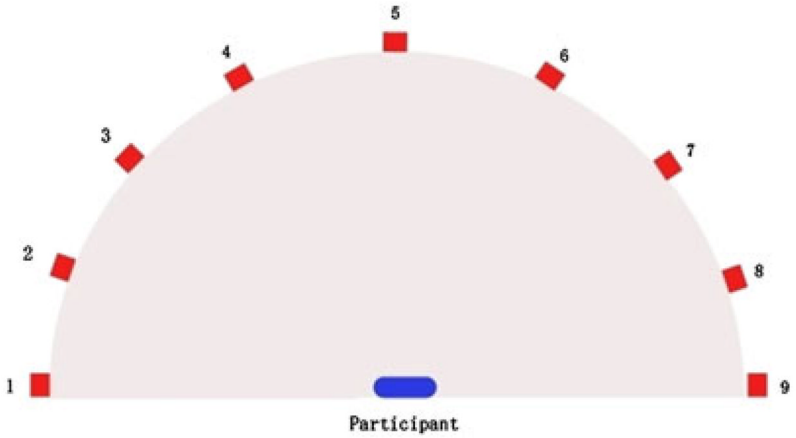

Figure 1.
Illustration of sound direction identification experiment (SDIE).

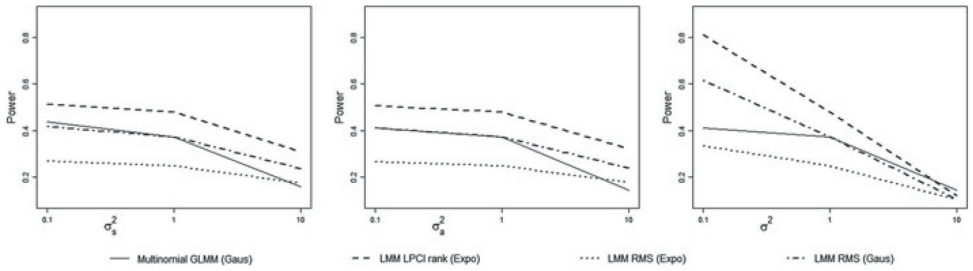

Figure 2.
Empirical power of tests versus variance components for model with overdispersion.

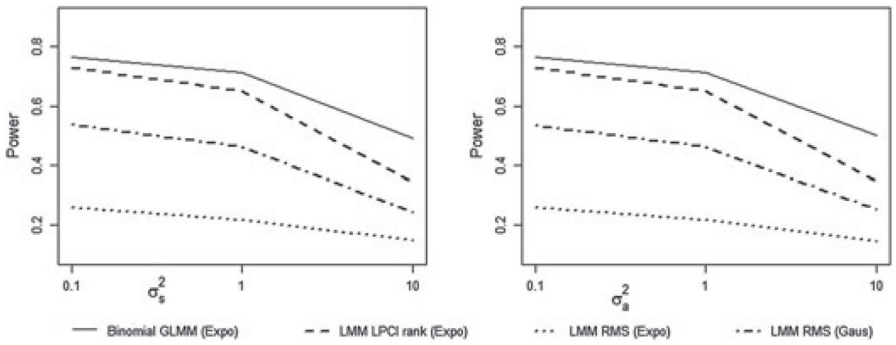

Figure 3.
Empirical power of tests versus variance components for model without overdispersion.

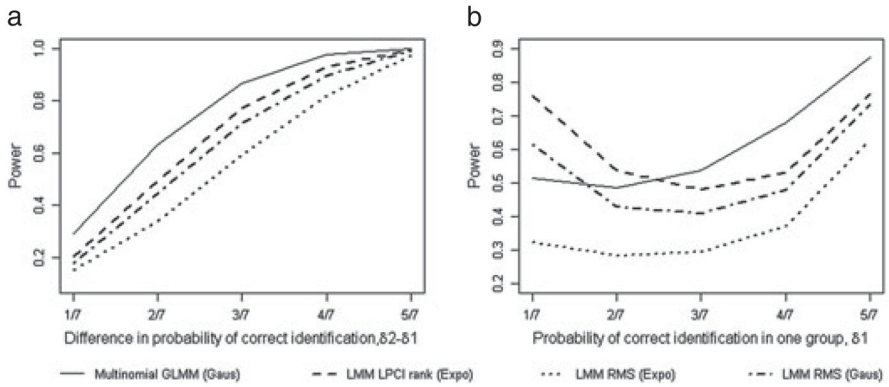

Figure 4.

Change of empirical power versus probability of correct identification; (a) shows the change of empirical power with respect to δ_2, given $\delta_1 = 1/7$; (b) shows the change of empirical power with respect to δ_1, given $\delta_2 - \delta_1 = 1/7$.

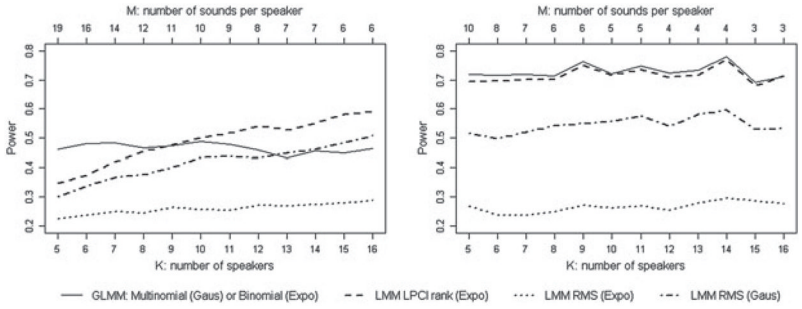

Figure 5.
Empirical power of tests versus number of speakers *K*.

Table I

Empirical power of tests to illustrate the impact of spatial models on the performance of test procedures.

Tests	**Exponential**	**Gaussian**	**Power**	**Inverse**	**Exponential**	**Gaussian**	**Power**	**Inverse distance**
GEE	0.4845	0.5155	0.4745	0.4745				
Binomial GLMM	0.7915	0.7880	0.7825	0.7895				
Multinomial GLMM (nominal)	0.6060	0.5055	0.6305	0.6750				
Summary statistics		LPCI				RMS		
RM ANOVA (raw data)	0.4390	0.4385	0.4370	0.4380	0.2670	0.4425	0.2740	0.2630
RM ANOVA (rank)	0.5305	0.5340	0.5290	0.5225	0.2590	0.4495	0.2620	0.2455
LMM (raw data)	0.4560	0.4540	0.4555	0.4585	0.2875	0.4630	0.2850	0.2785
LMM (rank)	0.5480	0.5550	0.5430	0.5415	0.2710	0.4650	0.2735	0.2580

Stat Med. Author manuscript; available in PMC 2012 April 29.

Table II

Empirical power of tests to compare the power of detecting group differences with/without overdispersion for various test procedures.

Models	**Overdispersion**				**No overdispersion**			
	Exponential	**Gaussian**	**Power**	**Inverse distance**	**Exponential**	**Gaussian**	**Power**	**Inverse distance**
GEE	0.3780	0.4020	0.3555	0.3660	0.6935	0.7560	0.6460	0.6440
Binomial GLMM	0.7235	0.7255	0.7295	0.7355	0.8205	0.8245	0.8230	0.8190
Multinomial GLMM (nominal)	0.5615	0.4615	0.5985	0.6390	0.5355	0.3640	0.5715	0.6290
Summary statistics	LPCI Exponential	RMS Exponential	RMS Gaussian		LPCI Exponential	RMS Exponential	RMS Gaussian	
RM ANOVA (raw data)	0.3525	0.2195	0.3455		0.6685	0.2660	0.6005	
RM ANOVA (rank)	0.3995	0.2210	0.3485		0.7895	0.2555	0.5935	
LMM (raw data)	0.3675	0.2385	0.3725		0.6995	0.2935	0.6235	
LMM (rank)	0.4210	0.2380	0.3710		0.8120	0.2870	0.6245	

Table III

Type I error rate of test procedures in detecting the group difference with/without overdispersion.

Models	**Overdispersion**				**No overdispersion**			
	Exponential	**Gaussian**	**Power**	**Inverse distance**	**Exponential**	**Gaussian**	**Power**	**Inverse distance**
GEE	0.0535	0.0475	0.0525	0.0570	0.0420	0.0470	0.0515	0.0515
Binomial GLMM	0.2395	0.2375	0.2420	0.2360	0.0405	0.0510	0.0475	0.0580
Multinomial GLMM (nominal)	0.1260	0.0520	0.1745	0.2225	0.0255	0.0020	0.0525	0.0895
Summary statistics	LPCI Exponential	RMS Exponential	RMS Gaussian		LPCI Exponential	RMS Exponential	RMS Gaussian	
RM ANOVA (raw data)	0.0545	0.0440	0.0565		0.0380	0.0440	0.0500	
RM ANOVA (rank)	0.0565	0.0465	0.0515		0.0480	0.0450	0.0525	
LMM (raw data)	0.0610	0.0495	0.0585		0.0435	0.0530	0.0595	
LMM (rank)	0.0595	0.0525	0.0540		0.0555	0.0550	0.0575	

Table IV

Reported *p*-values for the example data by applying the proposed test procedures.

Tests	*p*-value	Test using summary statistics	LPCI	RMS
GLMM (ordinal multinomial) *	<0.0001	Two-sample *t*-test	0.0084	0.1839
GLMM (nominal multinomial) *	0.0011	Wilcoxon rank sum	0.0087	0.1776
GLMM (binomial) **	0.0287	Paired *t*-test	0.0040	0.0373
GLMM (binomial, overdispersion) **	0.0764	Wilcoxon signed rank	0.0018	0.0476
GEE	0.0011	ANOVA (raw data)	0.0074	0.1625
		ANOVA (rank)	0.0052	0.1493
		RM ANOVA (raw data)	0.0063	0.1144
		RM ANOVA (rank)	0.0038	0.1144
		LMM (raw data)	0.0062	0.1136
		LMM(rank)	0.0037	0.1135

* denotes the tests fitting by SAS 9.2 using Laplace approximation for log-likelihood.

** denotes the tests fitting by SAS 9.2 using residual subject-specific pseudo-likelihood technique.