

# Power computation for likelihood ratio tests for the transition parameters in latent Markov models

**Dereje W. Gudicha, Verena D. Schmittmann,  
and Jeroen K. Vermunt**  
Tilburg University

Latent Markov (LM) models are increasingly used in a wide range of research areas including psychological, sociological, educational, and medical sciences. Methods to perform power computations are lacking, however. Power computations can help 1) to make an informed decision on the sample size or the number of measurement occasions required to achieve a pre-specified power level of a statistical test, and 2) to evaluate the ability of this test in detecting a statistically meaningful effect when indeed there is such an effect in a population. This paper presents methods for performing power analysis in LM models. Two cases of tests of hypotheses on the transition parameters of LM models are considered. The first case concerns the situation where the likelihood ratio test statistic follows a chi-square distribution, implying that also the power computation can be based on this theoretical distribution. In the second case, power needs to be computed based on empirical distributions constructed via Monte Carlo methods. Numerical studies are conducted to illustrate the proposed power computation methods and to investigate design factors affecting the power of this test.

*Keywords:* Latent Markov Model, Likelihood Ratio, Monte Carlo methods, Transition probabilities, Design Factors, Hypothesis Testing, Power computation.

## POWER FOR LIKELIHOOD RATIO TEST IN LM MODELS

---

Address for correspondence: Dereje W. Gudicha, P.O.Box 90153, 5000LE Tilburg, Department of Methodology and Statistics, Tilburg University, The Netherlands. E-mail: D.W.Gudicha@uvt.nl.

Models involving latent classes are receiving increasing interest from applied researchers, not only for the analysis of cross-sectional data but also in longitudinal studies, in which respondents are assumed to switch between classes during the period of observation. The occurrence of these transitions between latent classes (also called latent states) can be studied by using latent Markov (LM) models, which are also referred to as hidden Markov models or latent transition models (Van de Pol & De Leeuw, 1986; Rabiner, 1989; Poulsen, 1990; Collins & Wugalter, 1992; Vermunt, Langeheine, & Bockenholt, 1999; Visser, Raijmakers, & Molenaar, 2002; Bartolucci, 2006).

This growing interest in LM models is fueled by both the progress that has been achieved in extending the basic model (e.g., Wiggins (1973)) and the development of various statistical packages for analyzing data using the LM models. Extensions to the basic model include the use of time-constant and/or time-varying covariates (Reboussin, Reboussin, Liang, & Anthony, 1998; Vermunt et al., 1999; Chung, Park, & Lanza, 2005), multiple response variables (Langeheine & Van de Pol, 1993; Bartolucci, 2006; Wall & Li, 2009), and grouping variable(s) (Collins & Lanza, 2010). These extensions, together with the growing number of statistical packages (e.g., Latent GOLD (Vermunt & Magidson, 2013), Mplus (Muthén & Muthén, 1998-2012), the R-packages *dempixS4* (Visser & Speekenbrink, 2010), and the SAS procedure *PROC LTA* (Lanza & Collins, 2008)) make it possible to successfully apply LM models to many practical problems in longitudinal studies.

Despite these developments, methods to perform power computation in LM models have received no attention in the methodological literature, as far as we know. Likewise, in applications of LM models, hypotheses are typically tested using the likelihood ratio (LR) test without addressing power issues. Computation of the power of statistical tests (i.e., the probability of rejecting the null hypothesis when it is false) is, however, extremely important for various reasons. When planning a study, power computation can help to make an informed decision on the sample size and/or the number of measurement occasions required for achieving a pre-specified power level for the tests of interest. When testing a particular hypothesis, power calculation assesses the ability of a test to detect a statistically meaningful effect when indeed there is such an effect in a population. This is of interest when we wish to determine the usefulness of a test.

To perform a power calculation in LM models, we not only need to take into account the sample size, effect size, and the level of significance, but also several other design factors. For instance, in the latent class model,

which can be conceived of as a special case of the LM model, Gudicha et al. (under review) showed that a test can be underpowered when associations between latent classes and response variables are weak, that is, if the latent classes are poorly separated. In LM models, also the sizes of the transition probabilities are expected to affect the power. The objective of this paper is twofold: to provide power computation methods for hypotheses regarding the parameters of LM models and to identify design factors that affect the power.

In general, two kinds of statistical test are of interest when using LM models. The first kind pertains to hypotheses about the number of latent states (e.g., the test of a model with three latent states against a model with two latent states). The second kind concerns hypotheses for the parameters of the LM model, for example, for the transition probabilities. In this paper we focus on the latter type of test. More specifically, we assume that the number of states is known, and focus on equality and fixed value hypotheses for the model parameters. These include hypotheses stating that transition probabilities are constant across time points, that certain transition probabilities are equal to 0, or that transition probabilities are equal across two groups. We restrict ourselves to hypotheses on transition probabilities because in most longitudinal studies hypotheses concerning state transitions over time are the main focus of interest. Of course, in certain applications, hypotheses about initial state and conditional response probabilities (the measurement model) can be of particular interest as well. The power computation methods described in this paper can be directly transferred to those situations.

As we explain in detail below, for certain hypotheses on model parameters the standard asymptotic results for the LR hold, implying that power computation can be based on asymptotic distributions. For other hypotheses or, more specifically, for hypotheses stating that probabilities are equal to 0 or 1, these asymptotic results do not hold (Bartolucci, 2006). For this non-standard situation in which asymptotic distributions cannot be used for power computation, we propose constructing the empirical distribution of the LR statistic via Monte Carlo (MC) methods. Hereafter, we refer to the former and latter situations as power computation under the standard and non-standard case, respectively.

The remainder of the paper is organized as follows. We first introduce the LM model and outline the most common hypotheses about the transition parameters of this model. We then explain the LR test and its asymptotic properties briefly. Later, power computation is presented for both the stan-

dard and the non-standard case. In addition, we describe the design factors that are expected to affect the power of the LR test. We also present a numerical study to illustrate the proposed power computation methods, to examine the effect of the design factors on the power levels, and identify design configurations with acceptable power levels. In the final section, we provide a discussion of the different power computation methods, as well as recommendations for applied researchers and suggestions for future methodological studies.

### The LM MODEL

The LM model is a probability based model, in which the observed response patterns at a given time point is related with latent states producing these response patterns, analogous to the latent class model. In addition, the probability of being in a latent state at the current time point depends on the latent state of the previous time point. The model has two sub-parts. The first sub-part, the measurement model, enables us to identify latent states (also called latent classes or latent subgroups) based on the observed responses to the outcome variables. The second sub-part, the Markov model, describes the probabilities of switching between latent states over time. The latter model applies Markovian chains to account for the dependence between the latent states at successive measurement occasions.

The LM model relies on two assumptions. The first is the local independence assumption, which implies that the observed response patterns produced at time  $t$  depend only on the current state. The second is the first-order Markov assumption, which implies that the state occupied at time point  $t$  depends only on the state occupied at time point  $t - 1$  (Vermunt et al., 1999; Bartolucci, 2006). These two assumptions are specified on the measurement and the Markov model, respectively. Below, we first introduce some notation and then present the LM model and commonly used hypotheses about the parameters of this model.

Let  $y_{itj}$  be the response of subject  $i$  to the  $j$ th response variable measured at occasion  $t$ , for  $i = 1, 2, 3, \dots, n$ ,  $j = 1, 2, 3, \dots, p$ , and  $t = 1, 2, 3, \dots, T$ . We denote the vector of responses for subject  $i$  at occasion  $t$  by  $\mathbf{y}_{it}$ , and the vector of responses at all occasions by  $\mathbf{y}_i$ . Let us denote a discrete latent state at time point  $t$  by  $X_t$  and its possible value by  $x_t$  where  $x_t = 1, 2, 3, \dots, C$ . Then the probability of observing the response pattern  $\mathbf{y}_i$  can be defined as

$$\begin{aligned}
p(\mathbf{y}_i, \Phi) = & \sum_{x_1} \sum_{x_2} \dots \sum_{x_T} \overbrace{P(X_1 = x_1)}^{\text{initial state probabilities}} \prod_{t=2}^T \overbrace{P(X_t = x_t | X_{t-1} = x_{t-1})}^{\text{transition probabilities}} \\
& \prod_{j=1}^P \underbrace{y_{itj} | X_t = x_t}_{\text{conditional response probabilities}}, \quad (1)
\end{aligned}$$

where  $\Phi$  is the vector of model parameters.

As can be seen from equation (1), the LM model has three fundamental sets of parameters: The initial state probabilities,  $P(X_1 = x_1)$ , the transition probabilities,  $P(X_t = x_t | X_{t-1} = x_{t-1})$ , and the conditional response probabilities,  $P(y_{itj} | X_t = x_t)$ . The initial state probabilities show the state proportions (or sizes) at the first measurement occasion. The transition probabilities, that is, the probabilities of switching between the states from one measurement occasion to the next, can be conveniently collected in the so-called transition matrix  $\mathbf{A}^{(t)}$ , as shown below. The conditional response probabilities provide information on the type and the strength of association between states and the response variables.

If we set the number of states to 3, for example, the transition between latent states at time point  $(t - 1)$  and  $t$  can be expressed using a matrix of transition probabilities as

$$\mathbf{A}^{(t)} = \left[ \begin{array}{c|ccc} & \multicolumn{3}{c}{t} \\ \hline t-1 & \pi_{1|1} & \pi_{2|1} & \pi_{3|1} \\ & \pi_{1|2} & \pi_{2|2} & \pi_{3|2} \\ & \pi_{1|3} & \pi_{2|3} & \pi_{3|3} \end{array} \right], \quad (2)$$

where the principal diagonal elements of matrix  $\mathbf{A}^{(t)}$  represent the probability of staying in the same state between consecutive measurement occasions, and the off-diagonal elements are the probabilities for switching from a particular state at time  $t - 1$  to another particular state at  $t$ . For instance,  $\pi_{1|1} = P(X_t = 1 | X_{t-1} = 1)$  represents the probability of remaining in state 1 at the current measurement occasion, and  $\pi_{3|2} = P(X_t = 3 | X_{t-1} = 2)$  represents the probability of switching from state 2 at the previous measurement occasion to state 3 at the current measurement occasion.

In many research situations, we want to investigate the effect of one or more covariates  $Z$  on the transition probabilities, for instance, the effect of a dichotomous grouping variable (e.g.,  $Z = 0$  for the control group, and  $Z = 1$  for the treatment group). This can be done by inserting the covariate  $Z$  into equation (1) as follows:

$$p(\mathbf{y}_i, \Phi | z_i) = \sum_{x_1} \sum_{x_2} \dots \sum_{x_T} P(X_1 = x_1) \prod_t P(X_t = x_t | X_{t-1} = x_{t-1}, z_i) \prod_j P(y_{itj} | X_t = x_t), \quad (3)$$

When covariates are included, the transition probabilities are generally re-parameterized by specifying a multinomial logistic regression:

$$P(X_t = r | X_{t-1} = s, z_i) = \pi_{r|s} = \frac{\exp(\beta_{sr} + \gamma_{sr} z_i)}{\sum_{l=1}^c \exp(\beta_{sl} + \gamma_{sl} z_i)}. \quad (4)$$

That is, in this case, we estimate the logit coefficients  $\beta$  and  $\gamma$  as parameters, rather than the probabilities  $\pi$  directly.

In the next sub-section we present the most relevant hypotheses on the transition probabilities, which can be specified by imposing linear constraints on either the probabilities in equation (2) or the logit coefficients in equation (4).

### Hypotheses specified on transition parameters

We can distinguish several types of hypotheses on the transition parameters of LM models. Table 1 contains a classification of the most commonly used hypotheses. A first distinction concerns whether the hypothesis implies an equality constraint (say,  $\pi_{1|2} = \pi_{2|1}$ ), or a fixed value constraint (say,  $\pi_{1|2} = 0.3$ ); this is shown in the first column of Table 1. Fixed value constraints can be further distinguished into boundary constraints, where the parameter is fixed on a value on the boundary of the parameter space (i.e., zero or one for probabilities) and non-boundary constraints, where the parameter is fixed to a value inside the parameter space. This distinction is important, because fixed value boundary constraints require non-standard hypothesis testing and power calculation methods, as we address in detail below. Which testing methods can be used is shown in the last column of Table 1. Further distinctions concern whether the constraints are imposed

Table 1: Typical hypotheses formulated on transition parameters

constraint types	hypotheses		testing methods
	on selected transition parameters	on whole transition matrix	
equality	$H_0^1: \pi_{r s} = \pi_{s r}$ for some $r, s$	$H_0^4: A = A'$	standard
	$H_0^2: \pi_{r s} = \pi_{k s}$ for some $s$	$H_0^5: A^{(t)} = A$	
	$H_0^3: \pi_{r s} = \pi_{k l}$ for some $r, s, k, l$	$H_0^6: A_1 = A_2$	
fixed value	non-boundary $H_0^7: \pi_{r s} = v, v \in (0, 1)$		standard
	on boundary $H_0^8: \gamma_r = 0$		
	$H_0^9: \pi_{r s} = 0$ for some $r, s$	$H_0^{10}: A = \text{diag}\{\pi_{s s}\}$ , for $r = 1, 2, 3, \dots, c$	non-standard

Note that  $A'$  denotes the transpose of matrix  $A$ , and that  $A^{(t)}$  denotes the transition matrix from time point  $t - 1$  to time point  $t$ .

on a specific parameter or on the whole set of transition parameters, and whether the constraints are imposed on the transition parameters of the basic LM model (i.e., on the probabilities) or on the transition parameters of the LM model with covariates (i.e., on the logit coefficients).

We will now describe the hypotheses in Table 1 in more detail.  $H_0^1$  states the probability of switching from state  $s$  to state  $r$  is equal to the probability of switching from state  $r$  to state  $s$ ;  $H_0^2$  states that given state  $s$ , the probability of a transition to state  $r$  and state  $k$  is equal;  $H_0^3$  indicates that the probabilities in two cells of the transition matrix are equal (e.g.,  $\pi_{1|2} = \pi_{4|3}$ );  $H_0^4$  assumes the transition matrix is symmetric (i.e.,  $A$  equals its transpose);  $H_0^5$  implies that the transition matrix is time homogeneous;  $H_0^6$  sets the transition matrix of one group (e.g., the treatment group) equal to that of another group (e.g., the control group);  $H_0^7$  fixes the probability of switching from state  $s$  to state  $r$  to  $v$ , where  $v \in (0, 1)$  can be any user defined value;  $H_0^8$  defines the covariate to have no effect on the probability of switching to the state  $r$ ;  $H_0^9$  sets the probability of switching from state  $s$  to state  $r$  to 0; and  $H_0^{10}$  assumes the transition matrix is diagonal, meaning that there are no changes in state over time.

It should be noted that in certain applications, hypotheses about the initial state and the conditional response probabilities may be of interest as well (Visser et al., 2002). As for the transition probabilities, also for the

conditional response probabilities one may define equality and/or fixed-value restrictions, as discussed for latent class models by Goodman (1974) and Mooijaart and Van der Heijden (1992). This means that the distinction between different types of hypotheses of Table 1 (i.e., equality constraints, boundary fixed-value constraints, and non-boundary fixed-value constraints) can analogously be applied to hypotheses about the initial state and conditional response probabilities. This distinction determines the classification as standard or non-standard case, implying the methods of power computation we present for these two cases are also applicable with hypotheses about initial state and conditional response probabilities.

### Parameter estimation

Hypothesis testing using the LR requires estimating both the restricted model defined by the hypothesis of interest and the unrestricted model by means of maximum likelihood. Assuming that the responses of the individuals are identically and independently distributed, the log-likelihood for the model defined in equation (1) (also for equation (3)) can be specified as

$$l(\Phi) = \sum_i \log(p(\mathbf{y}_i)), \quad (5)$$

As in other latent class and mixture models, maximum likelihood estimates of the parameters of LM models can be obtained using the expectation maximization (EM) algorithm. This is an iterative method which alternates between the E step in which the expected value of the complete data log-likelihood – the log-likelihood if the latent states would be observed – conditional on the observed data and the current parameter estimates is computed, and the M step in which the parameter are updated by maximizing the expected complete data log-likelihood (McLachlan & Krishnan, 2007). For efficient computation with more than a few measurement occasions and with larger numbers of states, a special implementation of the E step can be used for LM models. This algorithm which is referred to as the forward-backward or Baum-Welch algorithm (Baum, Petrie, Soules, & Weiss, 1970) is implemented in software for LM models, such as in Latent GOLD program (Vermunt & Magidson, 2013) we used for our numerical studies.

Estimating the parameters of the LM model using the above mentioned procedure is, however, not always straightforward. Firstly, the log-likelihood

function in equation (5) may contain local maxima to which the optimization algorithm may converge (Visser et al., 2002). Inference based on such a local maximum may result in erroneous conclusions about the parameters and the fit of the LM model of interest. To prevent local maxima, one should therefore make sure that the model is re-estimated with multiple start sets. Secondly, in LM models, initial state, transition, and measurement model probabilities are mutually dependent. Because of this contingency, misspecification in one part of the model will affect the estimates other parts of the model; that is, imposing an incorrect constraint on one set of LM parameters (e.g., the transition parameters) affects the estimates of the other parameters (Visser et al., 2002). It should be noted that the latter feature is used by the power computation methods described below, which involve estimating the misspecified  $H_0$  model using (data generated from) the  $H_1$  population. Misspecification in the transition model will be partially absorbed by the (unrestricted) measurement model, which may have a strong effect on the power of the test regarding the transition parameters.

### THE LIKELIHOOD RATIO TEST

Once the parameter estimates and the corresponding log-likelihood values are obtained for the null (restricted) and the alternative (unrestricted) model, hypotheses such as those presented in Table 1 can be tested using the LR test statistic. The LR test statistic for the comparison of the restricted and unrestricted model is defined as

$$LR = -2(l(\Phi_0) - l(\Phi)),$$

where  $l(.)$  is the log-likelihood function as shown in equation (5), and  $\Phi_1$  and  $\Phi_0$  are the parameters of the model under the alternative and null hypotheses, respectively. It is also possible to obtain the LR statistic by taking the difference between the goodness-of-fit test statistics of the null and the alternative model,  $LR = LR_0 - LR_1$ , where  $LR_0$  and  $LR_1$  compare the model concerned with the saturated model. Under certain regularity conditions, under the null hypothesis, LR follows a (central) chi-square distribution with  $df$  degrees of freedom (Giudici, Ryden, & Vandekerckhove, 2000). The number of degrees of freedom of the test is determined by subtracting the number of parameters under the null from the number of parameters under the alternative hypothesis. The general principle of this test is to reject the null hypothesis if the observed value of the LR exceeds the  $(1 - \alpha)$  quantile value,

also called the critical value, of the central chi-square distribution with  $df$  degrees of freedom. Such testing procedure can be classified under what we referred to above as the standard case.

There are also, however, hypotheses for which the LR statistic does not necessary follow a chi-square distribution. This happens, for example, with hypotheses of the type  $H_0^9: \pi_{r|s} = 0$  and  $H_0^{10}: A = \text{diag}\{\pi_{s|s}\}$ , that is, when parameters are fixed on the boundary of the parameter space. The reason for this is that restricting the parameter(s) on the boundary of the parameter space breaks the regularity conditions under which the LR asymptotically follows a chi-square distribution (Bartolucci, 2006). In such non-standard testing situations, one can perform the test using the empirical distribution constructed by MC simulation, which is usually referred to as the bootstrap LR test (Aitkin, Anderson, & Hinde, 1981; Collins, Fidler, Wugalter, & Long, 1993).

## POWER COMPUTATION

The power of a test is the probability of rejecting the null hypothesis given that the alternative hypothesis is true. To compute the power, we should know or estimate the distribution of the test statistic under both the null and alternative hypothesis. The distribution under the null hypothesis, which is indicated with  $H_0$  in Fig.1, is required to compute the critical value,  $Q_{1-\alpha}$ , corresponding to the pre-defined type 1 error  $\alpha$ . The distribution under the alternative hypothesis, indicated with  $H_1$  in Fig.1, is required to compute the power, that is, the probability that the test statistic exceeds this cut-off value given that the alternative hypothesis is true. In Fig.1, this probability corresponds to the shaded area below the  $H_1$  curve to the right of the vertical dashed line at the critical value. The next sub-sections describe various procedures for computing this probability under the standard and non-standard testing cases.

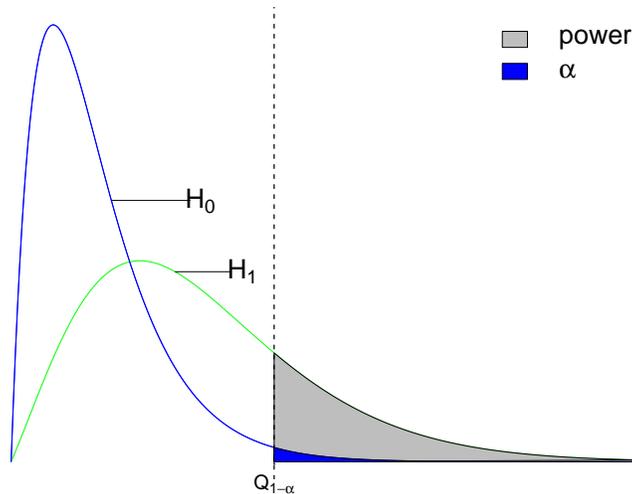


Figure 1: Distributions of LR under the null and alternative hypotheses.

### The standard case

As already mentioned above, in the standard case, the LR statistic follows a central chi-square distribution when the null hypothesis holds. When instead the alternative hypothesis holds, which is what we assume in power computation, the distribution of the LR becomes a non-central chi-square. One approach to power computation involves computing the non-centrality parameter  $\lambda$ , which quantifies how much the distribution of the LR under the alternative hypothesis deviates from its distribution under the null hypothesis.

But, first we describe a general power computation approach which does not require the computation of the non-centrality parameter. Instead, the empirical distribution of the LR under the alternative hypothesis is constructed using a Monte Carlo (MC) procedure. This procedure, which we refer to as MC-based power computation, proceeds as follows.

*Step 1.* A sample of a specified size,  $n$ , is repeatedly selected (say  $M$  times) from the population under the alternative hypothesis, and for each of these samples, the LR value is computed by estimating both the null model and the alternative model. We denote the LR value obtained with the  $m$  sample by  $LR_m$ .

*Step 2.* The actual power associated with a sample of size  $n$  is computed as the proportion of the simulated data sets in which the null hypotheses is rejected given the critical value  $Q_{1-\alpha}$ , which can be obtained from the central chi-square distribution with  $df$  degrees of freedom. More formally,

$$\text{power}_{MC_1} = \frac{\sum_{m=1}^M I(LR_m > Q_{1-\alpha})}{M}, \quad (6)$$

where  $I(LR_m > Q_{1-\alpha})$  is an indicator function taking the value 1 when the  $LR$  value of the  $m$ th sample exceeds the critical value, and otherwise being 0.

The second, more elegant and more standard, way of power computation involves obtaining an estimate of the non-centrality parameter and subsequently computing the power for a given  $n$  using the non-central chi-square distribution concerned. We discuss two methods to obtain the non-centrality parameter, which both require analyzing a single constructed data set. The first method, which we refer to as exemplary data method, uses a data file which is exactly in agreement with the population model under the alternative hypothesis (Self, Mauritsen, & Ohara, 1992). Power computation is implemented in four steps as follows.

*Step 1.* An 'exemplary' data set is created, which contains all possible response patterns with weights equal to the model expected proportions under the alternative hypothesis.

*Step 2.* Using this data set, the log-likelihood is computed for both the constrained null and the alternative model.

*Step 3.* The non-centrality parameter is approximated as

$$\lambda_1 = -2(l(\Phi_0) - l(\Phi_1)), \quad (7)$$

where  $l(\Phi_0)$  and  $l(\Phi_1)$  are the log-likelihood values under the null and the alternative hypothesis, respectively (see also (Williamson, Lin, Lyles, & Hightower, 2007)), and  $\lambda_1$  represents the noncentrality corresponding to a sample size of 1. Note that  $\lambda_1$  can also be computed as the difference between the goodness-of-fit tests for the null and alternative models. Since the latter equals 0 (the alternative model fits perfectly),  $\lambda_1$  equals the value of the likelihood-ratio goodness-of-fit statistic obtained when estimating the null model.

*Step 4.* The non-centrality parameter obtained in *Step 3* is rescaled to the sample size  $n$  of interest. This is achieved using the proportionality between

the sample size and the non-centrality parameter (Satorra & Saris, 1985; MacCallum, Browne, & Cai, 2006):  $\lambda_n = n \cdot \lambda_1$ , where  $\lambda_n$  denotes the non-centrality parameter for sample size  $n$ . The power can now be computed as follows:

$$\text{power} = P(LR > Q_{1-\alpha}(df)) = F_{\chi^2}(Q_{1-\alpha}, df, \lambda_n), \quad (8)$$

where  $F_{\chi^2}(df, \lambda)$  is a distribution function for non-central chi-square with  $df$  degrees of freedom and non-centrality parameter  $\lambda_n$ , and  $Q_{1-\alpha} = \chi^2_{(1-\alpha)}(df)$  is, as above, the  $(1 - \alpha)$  quantile value of the central chi-square distribution. Basically, the last step involves finding the probability that the non-central chi-square variate concerned exceeds the critical value.

The number of response patterns in the exemplary data set, which depends on the number of measurement occasions, the number of response variables, and the number of response categories, can quickly become very large. For instance, even in a relative small problem with four time points ( $T = 4$ ) and six response variables ( $P = 6$ ) with two categories, the number of possible response patterns is already larger than 16 million. This shows that the exemplary data method may quickly become impractical. We propose resolving this problem by using a large simulated data set from the population under the alternative hypothesis instead of an exemplary data set. We refer to this alternative to the exemplary data method as the 'large data' method. The steps that need to be taken for power computation are the following:

*Step 1.* Generate a large data set, say of size  $N = 100000$ , according to the model under the alternative hypothesis.

*Step 2.* Estimate the models under both the null and the alternative hypotheses based on the data obtained in *Step 1*. This yields the log-likelihood values for both models.

*Step 3.* Compute the non-centrality parameter as

$$\lambda_1 = \frac{-2(l(\Phi_0) - l(\Phi_1))}{N}, \quad (9)$$

where  $\lambda_1$  is again the noncentrality parameter for a sample size of 1. Note that now the likelihood-ratio goodness-of-fit test is not equal to 0 under the alternative model, which means that we have to estimate both models.

*Step 4.* As in the exemplary data method, get  $\lambda_n = n \cdot \lambda_1$  and obtain the power using equation (8).

### The non-standard case

In the non-standard case, the regularity conditions under which the LR follows asymptotically a  $\chi^2$ -distribution are not satisfied. This happens, for instance, if parameters are fixed on the boundary of the parameter space, as in hypotheses  $H_0^9$  and  $H_0^{10}$ . In this non-standard case, the cut-off value  $Q_{1-\alpha}$  is generally not equal to  $\chi_{(1-\alpha)}^2(df)$ . Thus, the critical value of the LR obtained from the central chi-square cannot be used in the subsequent power computation under the alternative hypothesis. Neither can we use the non-central chi-square to approximate the distribution of the LR under the alternative hypothesis, implying that the theoretical distributions mentioned above cannot be used here for power computation. However, with the advance in computing facility, instead of relying on these theoretical distributions, one may compute power by applying MC simulations. Two MC simulations are needed: one simulation is performed to obtain the value of  $Q_{1-\alpha}$  given the null hypothesis, and the other simulation is performed to compute power given the alternative hypothesis.

More specifically, in order to compute the value of  $Q_{1-\alpha}$ , the empirical distribution of the LR under the null hypothesis should be constructed first. That is, generate  $M$  data sets according to the model under the null hypothesis, and compute the LR statistic for each of these samples. For sufficiently large  $M$ , the distributions of these LR values approximate the population distribution of the LR statistic under the null hypothesis. Next, for the specified  $\alpha$ -level, this  $(1 - \alpha)$  quantile value is obtained as the value  $LR_{(1-\alpha)}$  that splits the sorted LR values into the following two sets: the  $100(1 - \alpha)$  percent smaller LR values and the top  $100\alpha$  percent large LR values. That is,

$$Q_{1-\alpha} = \{LR_{(1-\alpha)} : P(LR > LR_{(1-\alpha)} | H_0) = \alpha\}. \quad (10)$$

Similarly, the distribution of the  $LR$  under the alternative hypothesis is constructed using  $M$  samples, but now generated according to the model under the alternative hypothesis. Using this distribution the power is computed as the proportion of these LR values that exceeds the  $Q_{1-\alpha}$  value obtained from equation (9). That is,

$$\text{power}_{MC_{01}} = P(LR > Q_{1-\alpha} | H_1) = \frac{\sum_{m=1}^M I(LR_m > Q_{1-\alpha})}{M}, \quad (11)$$

where  $I(\cdot)$  is again an indicator function indicating whether the LR value (computed based on the  $H_1$  sample) exceeds the  $Q_{1-\alpha}$  value. Note that

$power_{MC_0}$  in equation (11) indicates that MC methods are applied under both the null and the alternative hypothesis, while  $power_{MC_1}$  in equation (6) indicated MC simulation is applied under the alternative hypothesis only.

The details of these power computation methods with the illustration using the Latent GOLD program (Vermunt & Magidson, 2013) syntax are provided in the appendix section.

### **Design factors**

As in any other standard statistical model, the power of a test also depends in LM models on the significance level  $\alpha$ , the effect size (difference between the parameter values under the null and alternative hypotheses), and the sample size. This can be explained using Fig.1. If the value of  $\alpha$  becomes larger,  $Q_{1-\alpha}$  shifts to the left, and consequently the region under the curve  $H_1$  to the right of  $Q_{1-\alpha}$  gets larger. This implies that the larger  $\alpha$ , the larger the power of a test. Now, for a fixed  $\alpha$ -level, if the effect size gets larger, the value of the non-centrality parameter gets larger, meaning that the curve indicated by  $H_1$  shifts to the right, and consequently the power becomes larger. The same applies to the sample size. If the sample size increases, the non-centrality parameter increases, meaning that the overlap between the distributions  $H_0$  and  $H_1$  decreases and thus the power increases.

In addition to those standard factors, the power of a test in LM models is expected to depend on aspects of the measurement part and the transition part of these models. From research on latent class models (which share the measurement part but not the transition part with LM models) we know that power is affected by uncertainty about the latent state occupied at a particular time point. This uncertainty depends on the number of states, the class proportions, the number of response variables, and the strength of the association between the latent states and the response variables (Gudicha et al., under review). The larger the number of response variables and the stronger the state-response associations, the better the separation between states will be. The better the states are separated, the less uncertain we are about the respondents state membership given his/her responses to the indicator variables, indicating a strong connection between the measurement and transition parts of the LM model. See also (Tein, Coxe, & Cham, 2013) who discussed statistical power to detect number of clusters in latent profile analysis. Additionally, specific to LM models, we expect that uncertainty about the latent state occupied at a particular time point is influenced by the transition probabilities. Also, each additional measurement occasion increases

the precision of the parameter estimates, which implies that the number of measurement occasions should also affect the power.

## NUMERICAL STUDY

The purpose of this numerical study is to 1) illustrate the power computation methods under the standard and non-standard case, 2) investigate how the study design factors and the population model characteristics mentioned above may affect the power, and 3) see which design configurations yield an acceptable power level (power  $\geq 0.8$ ). We focus on three of the hypotheses shown in Table 1. The first two hypotheses,  $H_0^4$ :  $\pi_{r|s} = \pi_{s|r}$  for some  $r, s$  in the basic LM model, and  $H_0^8$ :  $\gamma_r = 0$  in the LM model with covariates are examples of the standard case. The third hypothesis concerns testing for a zero entry in the transition probability matrix,  $H_0^9$ :  $\pi_{r|s} = 0$  for some  $s$  and  $r$ , is an example of the non-standard case.

### Design

In this numerical study, we restricted ourselves to LM models for dichotomous response variables (say with the categories negative and positive). Moreover, the  $\alpha$  value was always assumed to be 0.05.

We varied the sample size ( $n = 300, 500, \text{ or } 900$ ), the number of measurement occasions ( $T = 2 \text{ or } 4$ ), the number of latent states ( $C = 2 \text{ or } 3$ ), the number of response variables ( $p = 6 \text{ or } 10$ ), the initial states proportions (uniform or non-uniform), the strength of the association between latent states and response variables (weak, medium, or strong), and the stability of the state membership (unstable, moderately stable, and stable). The non-uniform initial state proportions were set to  $(0.7, 0.3)$  for  $C = 2$  and to  $(0.6, 0.3, 0.1)$  for  $C = 3$ . The settings for the association between states and responses were specified using response probabilities equal to 0.7, 0.8, and 0.9 (or .3, .2, and .1), respectively. For example, in weak association condition, the probability of a positive response was set to 0.7 for all variables in latent state one, to 0.3 in latent state two, and to 0.7 for the first half of the items and to 0.3 for the remaining items in state three. The basic settings for the latent transitions were obtained by setting the main diagonal elements of the transition matrix to  $\pi_{r|r} = 0.7, 0.8, \text{ or } 0.9$ , and the other elements to  $\frac{1-\pi_{r|r}}{C-1}$ , which corresponds to unstable, moderately stable, and highly stable state memberships, respectively.

1) For hypothesis  $H_0^1$ ;  $\pi_{r|s} = \pi_{s|r}$ , we arrived at the respective transition matrices under the alternative model (i.e., with differences in the off-diagonal elements) by specifying the transition odds-ratios comparing the transition from  $s$  to  $r$  with the transition from  $r$  to  $s$ , which is defined as  $\frac{\pi_{r|s}/\pi_{s|s}}{\pi_{s|r}/\pi_{r|r}}$ , to be equal to 1.3498, 1.8222 and 3.3201. These odds-ratios, which we hereafter refer as small, medium, and large effect size, correspond to differences in the transition probabilities ranging from 0.01 to 0.25.

2) For hypothesis  $H_0^8$ ;  $\gamma_r = 0$ , we added the effect of a dichotomous covariate on the transitions. The covariate effect was specified by setting its (effect coded) coefficient in the logistic regression model for the transitions to 0.25, 0.5, and 1. Or equivalently, by setting the transition odds-ratio to 1.648, 2.7182, and 7.389, which corresponds to a small, medium, and large effect size, respectively.

3) For hypothesis  $H_0^9$ ;  $\pi_{1|2} = 0$ , we restricted ourselves to the situation with  $T = 2$ ,  $p = 6$ ,  $\pi_{1|1} = 0.7$  or  $0.9$ ,  $C = 2$ , and equal initial state probabilities. This setting gives as a transition matrix of the type

$$\begin{bmatrix} 0.7 & 0.3 \\ \delta & 1 - \delta \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0.9 & 0.1 \\ \delta & 1 - \delta \end{bmatrix},$$

where  $\delta$  is the value of  $\pi_{r|s}$  under the alternative hypothesis. The value of  $\delta$ , which defines the effect size for the hypothesis that  $\pi_{r|s} = 0$ , was set to 0.05, 0.1, or 0.2. The association between states and response variables was set to weak, medium, and strong as defined earlier.

## Results

Table 2 and 3 present the power for the test of the hypothesis that  $\pi_{r|s} = \pi_{s|r}$ . As expected, the power depends on the association between the latent states and the response variables, the number of measurement occasions, the sample size, the size of initial states, the number of response variables, the transition probabilities, and the effect size. More specifically, the stronger the association between latent states and response variables, the larger the power. For a given number of measurement occasions, say  $T = 2$ , reasonable power levels are achieved by increasing the sample size (Table 2). Or, for a given sample size say  $n = 300$ , reasonable power levels are achieved by increasing the number of measurement occasions (Table 3). Comparison of these two factors shows that the power gain achieved by increasing number

of measurement occasions from 2 to 4 is larger than the power gain achieved by increasing the sample size from 300 to 900. Also, increasing the number of response variables, or sampling from a population with equal initial state probabilities increases the power with the current design and population model characteristics.

Another interesting observation is that, keeping the other design factors constant, the more unstable the state membership ( or the larger the transition probabilities), the larger the power to demonstrate differences between transition probabilities. One can also see from Table 3 that for the situation where the initial state proportions/probabilities are equal, the number of response variables is equal to 6, and the sample size is equal to 300, a power of 0.80 or higher is achieved 1) for low transition probabilities, when the effect size is large and the association between states and response variables is strong, 2) for moderate transition probabilities, when the effect size is large and the association between states and response variables is medium or strong, and 3) for high transition probabilities, when the effect size is large. For the situation when we have weak associations between states and response variables, highly stable transition probabilities, and a low or moderate effect size, such a power level is achieved only at the expense of increasing the sample size, or the number of measurement occasions. For 3-state LM model, we do not show the results of the power calculation, as they provide similar information with the results for 2-state LM model shown in Table 2 and 3. The power to demonstrate differences in the transition probabilities for the 3-state LM model is in general lower than its corresponding power value for the 2-state LM model.

The power of the LR test against the null hypothesis that the covariate has no effect on the transition probabilities is shown in Table 4 and 5. With respect to the design factors, the general trend found is similar to the results from Table 2, 3 and 8. That is, power increases with sample size and effect size. Also, for a fixed sample size and effect size, one can achieve a desired level of power by improving the measurement part of the LM model, for example, by increasing the associations between latent states and response variables from weak to strong, or by increasing the number of response variables. Increasing the number of measurement occasions could also greatly help in obtaining a desired power level. One can also see from these tables that power for the effect of the covariate on the transition probabilities becomes larger when the state membership is more unstable, say  $\pi_{r|r} = 0.7$ , than when we have highly stable states, say  $\pi_{r|r} = 0.9$ .

Table 2: Effect of sample size  $n$ , transition probabilities, and conditional response probabilities on the power for the test that  $\pi_{r|s} = \pi_{s|r}$  in the two state LM model.

		$P = 6, T = 2$ , and equal initial state proportions								
		$\pi_{r r} = 0.9$			$\pi_{r r} = 0.8$			$\pi_{r r} = 0.7$		
		state-response association			state-response association			state-response association		
effect size		weak	medium	strong	weak	medium	strong	weak	medium	strong
$n = 300$	small	.075	.094	.117	.096	.142	.176	.113	.199	.224
	medium	.166	.279	.306	.221	.444	.532	.280	.581	.622
	large	.397	.714	.873	.682	.954	.981	.795	.984	.996
$n = 500$	small	.093	.125	.164	.128	.206	.262	.156	.299	.341
	medium	.162	.425	.466	.335	.651	.750	.427	.798	.835
	large	.592	.903	.980	.881	.997	.999	.949	1.00	1.00
$n = 900$	small	.127	.187	.256	.192	.331	.426	.243	.485	.547
	medium	.254	.661	.711	.539	.883	.942	.663	.976	.976
	large	.837	.992	1.00	.988	1.00	1.00	.998	1.00	1.00

Power values computed using the large data set method (see also equation 8).

Table 3: Effect of the number of measurement occasions  $T$ , the number of response variables  $P$ , and the initial state proportions on the power for the test that  $\pi_{r|s} = \pi_{s|r}$  in the two state LM model.

		$\pi_{r r} = 0.9$			$\pi_{r r} = 0.8$			$\pi_{r r} = 0.7$		
effect size		state-response association			state-response association			state-response association		
		weak	medium	strong	weak	medium	strong	weak	medium	strong
Equal initial state proportions, and $n = 300$										
$T = 2, P = 6$	small	.075	.094	.117	.096	.142	.176	.113	.199	.224
	medium	.166	.279	.306	.221	.444	.532	.280	.581	.622
	large	.397	.714	.873	.682	.954	.981	.795	.984	.996
$T = 2, P = 10$	small	.075	.108	.127	.131	.151	.181	.153	.203	.238
	medium	.195	.285	.335	.388	.505	.544	.455	.622	.645
	large	.559	.836	.865	.874	.972	.988	.945	.995	.997
$T = 4, P = 6$	small	.193	.232	.267	.264	.403	.447	.247	.486	.556
	medium	.520	.686	.779	.757	.927	.944	.764	.970	.984
	large	.971	.999	1.00	.999	1.00	1.00	.999	1.00	1.00
Unequal initial state proportions, and $n = 300$										
$T = 2, P = 2$	small	.058	.095	.113	.091	.136	.149	.087	.148	.163
	medium	.108	.203	.302	.180	.358	.430	.219	.450	.586
	large	.269	.606	.787	.575	.898	.952	.683	.954	.984

Power values computed using the large data set method (see also equation 8).

Table 4: Effect of sample size  $n$ , transition probabilities, and conditional response probabilities on the power for testing the covariate effect on transition probabilities.

		$P = 6, T = 2$ , and equal initial state proportions								
		$\pi_{r r} = 0.9$			$\pi_{r r} = 0.8$			$\pi_{r r} = 0.7$		
effect size		state-response association			state-response association			state-response association		
		weak	medium	strong	weak	medium	strong	weak	medium	strong
$n = 300$	small	.085	.127	.187	.113	.239	.316	.146	.307	.395
	medium	.180	.428	.586	.352	.719	.874	.482	.860	.949
	large	.587	.961	.996	.889	.999	1.00	.966	1.00	1.00
$n = 500$	small	.109	.185	.289	.159	.376	.496	.217	.480	.606
	medium	.276	.649	.817	.547	.916	.983	.712	.979	.997
	large	.818	.998	1.00	.987	1.00	1.00	.999	1.00	1.00
$n = 900$	small	.162	.304	.485	.257	.616	.762	.363	.746	.863
	medium	.465	.894	.975	.813	.995	1.00	.932	1.00	1.00
	large	.975	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Power values computed using the large data set method (see also equation 8).

Table 5: Effect of the number of measurement occasions  $T$ , the number of response variables  $P$ , and the initial state proportions on the power for testing covariate effects.

effect size	$\pi_{r r} = 0.9$			$\pi_{r r} = 0.8$			$\pi_{r r} = 0.7$			
	state-response association			state-response association			state-response association			
	weak	medium	strong	weak	medium	strong	weak	medium	strong	
Equal initial state proportions, and $n = 300$										
$T = 2, P = 6$	small	.085	.127	.187	.113	.239	.316	.146	.307	.395
	medium	.180	.428	.586	.352	.719	.874	.482	.860	.949
	large	.587	.961	.996	.889	.999	1.00	.966	1.00	1.00
$T = 4, P = 6$	small	.224	.421	.523	.328	.584	.751	.396	.748	.860
	medium	.623	.942	.983	.893	.997	1.00	.951	1.00	1.00
	large	.998	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$T = 2, P = 10$	small	.139	.178	.193	.180	.277	.327	.223	.388	.418
	medium	.287	.538	.631	.562	.850	.885	.685	.928	.949
	large	.853	.992	.997	.990	1.00	1.00	.999	1.00	1.00
Unequal initial state proportions, and $n = 300$										
$T = 2, P = 6$	small	.078	.135	.187	.124	.198	.311	.163	.286	.403
	medium	.171	.446	.598	.380	.719	.865	.539	.870	.947
	large	.605	.968	.998	.912	1.00	1.00	.983	1.00	1.00

Power values computed using the large data set method (see also equation 8).

Table 6 shows the power for the LR test against the null hypothesis that the probability of switching from one state at time point  $(t - 1)$  to another state at time point  $t$  equals zero. As compared with these other two hypotheses, the roles of transition probability on power is small. Whereas the role of state-response association on the power is high. For example, in a sample of 100 observations, if the state-response association is weak, the power to detect that a true proportion of .05 is different from the null of 0 is lower than 20 %. If this state-response association is instead strong, this power becomes over 90 %. When the state-response associations are strong, that is the measurement model is strong, the state separation becomes high. In such a condition, the possibility of observing the expected pattern from state 1 while in state 2 becomes extremely small. Therefore, when the true underlying model generates some transitions that are impossible under the restricted model, the likelihood of this restricted model will decrease dramatically, because each impossible transition, say, a transition from state 1 to state 2, needs to be accommodated by assigning the observed pattern from state 1 to state 2, or vice versa. The decrease in the likelihood of the restricted model will be accompanied by biased parameter estimates: the estimates of the conditional response probabilities in the two states will be biased to be closer to each other, to increase the likelihood of observing a state 1 pattern given state 2 or a typical state 2 pattern given state 1.

The results presented in Table 2 to 5 concern the standard case, and were thus obtained using the large data method which assumes that the distributions under  $H_0$  and  $H_1$  are known. That is, we assume a central chi-square under the null and non-central chi-square under the alternative hypothesis, and moreover approximate the value of the non-centrality parameter based on a large data set. To investigate the quality of the latter approximation as well as the asymptotic approximation of the chi-square itself, we also performed power computation by the MC-based  $M_{01}$  method for some of the conditions of Table 2 to 5. The results presented in Table 7 show that both the large data set and the chi-square approximations are very good.

Table 6: Power for the hypothesis that  $\pi_{2|1} = 0$  in two state LM model

$\pi_{2 1}$	$n = 100$			$n = 200$			$n = 300$		
	state-response association			state-response association			state-response association		
	weak	medium	strong	weak	medium	strong	weak	medium	strong
$\pi_{2 2} = 0.7$									
.05	.182	.640	.906	.263	.810	.987	.344	.917	.999
.1	.381	.877	.991	.544	.978	1.00	.710	.997	1.00
.2	.533	.968	1.00	.795	.998	1.00	.999	1.00	1.00
$\pi_{2 2} = 0.9$									
.05	.181	.597	.894	.262	.806	.987	.330	.910	.999
.1	.362	.872	.992	.554	.979	1.00	.685	.996	1.00
.2	.535	.957	.999	.758	.999	1.000	.887	1.00	1.00

Power values computed using  $M_{01}$  method (see also equation (11)).

Table 7: Asymptotic approximation of the LR when estimating the non-centrality parameter with large data set: Testing the hypothesis that  $\pi_{r|s} = \pi_{s|r}$  in two state LM model.

		$n = 500, P = 6, T = 2, \text{ and equal initial state proportions}$								
		$\pi_{r r} = 0.9$			$\pi_{r r} = 0.8$			$\pi_{r r} = 0.7$		
effect size	theoretical	state-response association			state-response association			state-response association		
	versus empirical	weak	medium	strong	weak	medium	strong	weak	medium	strong
small	theoretical	.093	.125	.164	.128	.206	.262	.156	.299	.341
	empirical	.077	.134	.179	.126	.217	.256	.151	.293	.341
medium	theoretical	.162	.425	.466	.335	.651	.750	.427	.798	.835
	empirical	.175	.386	.474	.368	.658	.750	.438	.796	.846
large	theoretical	.592	.903	.980	.881	.997	.999	.949	1.00	1.00
	empirical	.5446	.907	.975	.875	.997	.999	.939	1.00	1.00

25

The empirical powers, computed using the  $M_{01}$  method of power computation, are presented below each entry of the theoretical LR powers. The empirical powers are computed as follows: First, the CV value is obtained by applying the MC under the null hypothesis. Next, based on this CV value, power is computed under the alternative hypothesis using the MC method. Thus, neither the CV nor the power computation is relied on theoretical chi-square. The theoretical powers are computed by assuming a central chi-square under the null and non-central chi-square under the alternative hypotheses, for which the non-centrality parameter is approximated by using a large data set.

## DISCUSSION AND CONCLUSION

This paper addressed power computation methods for testing hypotheses about transition parameters of LM models, which are the transition probabilities themselves in the basic LM model and the logistic regression coefficients in the LM model with covariate(s). We showed how the hypotheses of main interest can be specified by imposing equality constraints across parameters or fixing parameter(s) to some user defined value(s). We distinguished power computation for the standard case and power computation for the non-standard case, where the latter arises when probabilities are fixed to zero.

For the standard case, in which the likelihood-ratio statistic follows an asymptotic chi-square distribution, two power computation approaches were discussed. The first consists of approximating the distribution under the alternative hypothesis for a given sample size  $n$  using MC simulation (referred to as  $MC_1$ ). The second approach involves estimating the non-centrality parameter using either an exemplary data set or large simulated data set and subsequently obtaining the power for any sample size  $n$  from the non-central chi-square distribution. The advantage of the second approach is that it is computationally cheaper. However, when we have doubts that the distribution of the LR test statistic under the alternative is non-central chi-square, the  $MC_1$  simulation approach is the preferred option. The  $MC_1$  simulation approach can also be applied when the distribution under the null is known but the distribution under the alternative is unknown. We will come back to this issue when discussing topics for future research.

The non-standard case occurs when the likelihood-ratio does not follow a standard chi-square distribution. The most obvious example for this is when a parameter is fixed to the boundary of the parameter space, which equals zero or one for probabilities. In such situations, power computation by MC simulation is applicable (referred to as  $MC_{01}$ ). We use the  $MC_{01}$  method to compute both the critical value under the null hypothesis and the power under the alternative hypothesis given this critical value. Note that this procedure is similar to the  $MC_1$  simulation approach discussed under the standard case, with the only difference that the theoretical distribution under the null hypothesis is replaced by its empirical counterpart.

In our numerical study, we saw that the power to detect large effects can be small even with a not very small sample of say 500 observations. Based on the results of the numerical study, we therefore strongly recommend re-

searchers who apply LR tests in LM models to perform a power analysis prior to the data collection. Our findings point at several important issues that should be taken into account. Firstly, in addition to the usual design factors (i.e., effect size, sample size, and significance level  $\alpha$ ), a power analysis for LM models should also involve various other design factors, namely, the number of time points, the number of response variables, the strength of association between latent states and response variables, the number of states, the initial state probabilities, and the transition probabilities. Secondly, for a given effect size, a desired level of power can be achieved by increasing the number of measurement occasions, by increasing the number of response variables, and/or by using response variables that have strong associations with the latent states. Moreover, situations in which the transition probabilities are small need special care, since power may be low in such situations. Thirdly, when the association between states and response variables is weak and/or the effect size is small, a reasonable power level can be achieved at the expense of gathering more data, that is, by increasing the sample size and/or the number of measurement occasions. In the scenarios we studied, increasing the number of measurement occasions was more efficient than increasing the sample size. This is probably connected to the fact that we looked at hypotheses for the transition probabilities; that is, with more measurement occasions one has more information on the transition probabilities. When testing hypotheses on the initial state and/or the response probabilities, increasing the sample size is probably more effective.

In the MC-based power computation, the accuracy of the estimated power depends strongly on the number of replications used. This is especially the case in the  $MC_{01}$  method used in the non-standard case, in which not only the power but also the critical value under  $H_0$  was estimated by MC simulation. In our study, we used 5000 MC replications, which seemed to be large enough for the hypotheses we looked. However, the required number of MC replications may depend on the type of hypothesis and the model complexity, hence, further research might explore the required number of MC replications for LR tests in LM models.

While for the non-standard case we proposed approximating the distribution of the LR statistic under  $H_0$  by simulation, its asymptotic distribution has been shown to be chi-bar square (Bartolucci, 2006). This means one may also obtain the critical value under  $H_0$  from the chi-bar square distribution, which for multiple parameter hypotheses also requires performing some kind of MC simulation. But power computation using an asymptotic approach,

requires knowing also the distribution under the alternative. This is however a problem that has not yet been resolved. Another possible area for future research is investigating whether this distribution is, for instance, a certain type of non-central chi-bar square distribution (Shapiro, 1988).

Other possible areas for further research concern the application of the proposed power computation methods with other types of hypotheses relevant for LM modeling. It seems that both the standard and non-standard case methods can be directly transferred to hypotheses about other parameters of the LM model, namely, the initial probabilities and the conditional response probabilities. The MC method proposed for the non-standard case may also be applicable in hypotheses tests concerning the number of latent states, that is, when comparing models with  $C$  and  $C + 1$  states. For the power computation, we estimate the (incorrect) model under  $H_0$  for data sets generated under  $H_1$ . That is, the measurement parameters are not fixed, but estimated under this incorrect model. These parameters will be biased: the measurement error will be estimated to be larger than it actually is. But for the power computation this is not important. What matters is whether the "allowed" bias in the measurement parameter can compensate (in term of log-likelihood value) for the misspecification of the transition model. If this is not the case, the power will be large. Future research investigating whether the allowed bias compensate for the misspecification of the transition model would be interesting, as the focus of the current paper is more of on how to compute the power for hypothesis.

## ACKNOWLEDGMENTS

This work is part of research project 406-11-039 "Power analysis for simple and complex mixture models" financed by the Netherlands Organisation for Scientific Research (NWO).

## References

- Aitkin, M., Anderson, D., & Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society. Series A (General)*, *4*, 419–461.
- Bartolucci, F. (2006). Likelihood inference for a class of latent Markov models under linear hypotheses on the transition probabilities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(2), 155–178.
- Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics*, *41*(1), 164–171.
- Chung, H., Park, Y., & Lanza, S. T. (2005). Latent transition analysis with covariates: pubertal timing and substance use behaviours in adolescent females. *Statistics in Medicine*, *24*(18), 2895–2910.
- Collins, L. M., Fidler, P. L., Wugalter, S. E., & Long, J. D. (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research*, *28*(3), 375–389.
- Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). John Wiley & Sons, Inc., Hoboken, New Jersey.
- Collins, L. M., & Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research*, *27*(1), 131–157.
- Giudici, P., Ryden, T., & Vandekerckhove, P. (2000). Likelihood-ratio tests for hidden Markov models. *Biometrics*, *56*(3), 742–747.
- Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*(2), 215–231.
- Gudicha, D. W., Vermunt, J. K., & Tekle, F. B. (under review). Power and sample size determination for latent class models. *Classification Journal*.

- Langeheine, R., & Van de Pol, F. (1993). Multiple indicator markov models. In *Psychometric methodology. proceedings of the 7th european meeting of the psychometric society in trier. stuttgart: Fischer* (pp. 248–52).
- Lanza, S. T., & Collins, L. M. (2008). A new SAS procedure for latent transition analysis: transitions in dating and sexual risk behavior. *Developmental psychology*, *44*(2), 446.
- MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological methods*, *11*(1), 19.
- McLachlan, G., & Krishnan, T. (2007). *The EM algorithm and extensions* (Vol. 382). John Wiley & Sons.
- Mooijart, A., & Van der Heijden, P. G. (1992). The em algorithm for latent class analysis with equality constraints. *Psychometrika*, *57*(2), 261–269.
- Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide. seventh edition*. Los Angeles, CA: Muthén & Muthén.
- Poulsen, C. S. (1990). Mixed Markov and latent Markov modelling applied to brand choice behaviour. *International Journal of Research in Marketing*, *7*(1), 5–19.
- Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.
- Reboussin, B. A., Reboussin, D. M., Liang, K.-Y., & Anthony, J. C. (1998). Latent transition modeling of progression of health-risk behavior. *Multivariate Behavioral Research*, *33*(4), 457–478.
- Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, *50*(1), 83–90.
- Self, S. G., Mauritsen, R. H., & Ohara, J. (1992). Power calculations for likelihood ratio tests in generalized linear models. *Biometrics*, *48*(1), 31–39.
- Shapiro, A. (1988). Towards theory of inequality. *International Statistical Review*, *56*(1), 49–62.
- Tein, J.-Y., Coxe, S., & Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(4), 640–657.
- Van de Pol, F., & De Leeuw, J. (1986). A latent Markov model to correct for measurement error. *Sociological Methods & Research*, *15*(1-2), 118–141.

- Vermunt, J. K., Langeheine, R., & Bockenholt, U. (1999). Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, *24*(2), 179–207.
- Vermunt, J. K., & Magidson, J. (2013). *Lg-syntax user’s guide: Manual for latent gold 5.0 syntax module*. Statistical Innovations Inc.
- Visser, I., Raijmakers, M. E. J., & Molenaar, P. C. M. (2002). Fitting hidden Markov models to psychological data. *Scientific Programming*, *10*(3), 185–199.
- Visser, I., & Speekenbrink, M. (2010). depmixS4: an R-package for hidden Markov models. *Journal of Statistical Software*, *36*(7), 1–21.
- Wall, M. M., & Li, R. (2009). Multiple indicator hidden Markov model with an application to medical utilization data. *Statistics in medicine*, *28*(2), 293–310.
- Wiggins, L. M. (1973). *Panel analysis: Latent probability models for attitude and behavior processes*. Elsevier Scientific, San Francisco.
- Williamson, J. M., Lin, H., Lyles, R. H., & Hightower, A. W. (2007). Power calculations for zip and zinb models. *Journal of Data Science*, *5*(4), 519–534.

## APPENDIX A: Latent GOLD syntax for power computation

This appendix illustrates the application of the proposed power computation methods using the Latent GOLD program. As an example, we use a 2-state LM model with six binary response variables (y1 through y6). The  $H_1$  population model contains unequal transition probabilities, and we test the  $H_0$  model with equal transition probabilities against the  $H_1$  model.

In order to perform a power computation, one should first define a data file indicating the time structure and the variables in the model. With  $T = 4$  and  $p = 6$ , this file could be of the form

```
id time y1 y2 y3 y4 y5 y6 n100000 n300
1 1 0 0 0 0 0 0 100000 300
1 2 0 0 0 0 0 0 100000 300
1 3 0 0 0 0 0 0 100000 300
1 4 0 0 0 0 0 0 100000 300
```

This data file contains 4 records (one for each measurement occasion) which are connected by an identifier variable, arbitrary values for the response variables, and variables indicating sample sizes to be used later on.

A Latent GOLD syntax model consists of three sections: “options”, “variables”, and “equations”. The relevant LM model is defined as follows

```
// basic model
options
  output parameters=first standarderrors profile;
variables
  caseid id;
  dependent y1 nominal 2, y2 nominal 2, y3 nominal 2,
    y4 nominal 2,y5 nominal 2,y6 nominal 2;
  latent State nominal dynamic 2;
equations
  State[=0] <- 1;
  State2 <- (beta~tra) 1 | State[-1];
  y1-y6 <- 1 | State;
```

The “output” option indicates that we wish to use dummy coding for the logit parameters with the first category as the reference category. Subsequently, we define the variables which are part of the model. Note that the latent variable “State” is specified to be dynamic, which yields a latent variable which changes its value across measurement occasions.

The three equations represent the logit equations for the initial state, the transitions, and the measurement part of the model, respectively. Note that “1” indicate an intercept, and “|” that the intercept depends on the variable concerned. A special type of coding (called transition coding and indicated with “~ tra” ) is used for the logit parameters of the transition model and, moreover, a label (beta) is specified for these parameters. This label will be used below to impose restrictions.

#### a) Standard case

**Option 1.** Implementation of the Monte Carlo based power computation method ( $MC_1$ ) involves defining the  $H_0$  and  $H_1$  model in a single input file. Denoting the parts which remain the same as in basic model defined above by “...”, the  $H_1$  model may equal:

```
// H1 model for MC-based method
...
equations
```

```

...
{0.000000000
-0.54729786  -1.14729786
1.386294361  -1.386294361
1.386294361  -1.386294361
1.386294361  -1.386294361
1.386294361  -1.386294361
1.386294361  -1.386294361
1.386294361  -1.386294361}

```

The numbers shown inside the curly bracket represent the values for the logit parameters in the  $H_1$  model. The first row contains the initial state parameter(s), the second row the transition parameters, and the remaining rows the response parameters.

The  $H_0$  model in which the transitions are restricted to be equal is defined as follows:

```

// H0 model for MC-based method
options
...
montecarlo replicates=1000 power='H1' N=300 alpha=0.05;
variables
...
equations
...
beta[1,1] = beta[2,1];

```

As can be seen, the equality restriction on the transition logits is defined at the end of the equations section. What is important to note is that the  $H_0$  model should contain the “montecarlo” option indicating the number of Monte Carlo replications, the “power” command with the name of the  $H_1$  model, the sample size “N”, and the level of significance “alpha”. Running the  $H_0$  model will yield the power for LR test comparing the two models.

**Option 2.** When using the large data set method, one should first simulate a large data set from the population defined by  $H_1$  and subsequently analyze this data set using both the  $H_0$  and  $H_1$  model. Simulating the large data set is done as follows:

```

// H1 model for simulating a large data file
options
...
  outfile 'sim.dat' simulation;
variables
...
  caseweight n100000;
equations
...
{...}

```

Compared to the basic specification, we use the “outfile” option to indicate that a data file should be simulated, use the “caseweight” to indicate the size of the large data set (here 100000), and specify the parameter values of the population model.

To obtain the power we analyze the large data set with an input file containing both the  $H_0$  and  $H_1$  model. The  $H_1$  model equals:

```

// H1 model for large data based power computation method
options
...
  output LLdiff='H0' LLdiffPower=300;
...

```

That is, we indicate that a log-likelihood difference test should be performed (“LLdiff”) and that the power of this test should be computed for the specified sample size (“LdiffPower”). We also define the  $H_0$  model itself, which again is the basic LM model with the constraint “beta[1,1] = beta[2,1]”.

**Option 3.** Power computation using the exemplary data method is similar to the large data method. First an exemplary data file which is exactly in agreement with the  $H_1$  model is created, and subsequently this data file is analyzed with both the null and the alternative model. That is, first create an exemplary data file as

```

// H1 model for creating the exemplary data file
options
...
  output WriteExemplaryData='exemplary.dat';

```

```

variables
...
equations
...
{...}

```

Next compute power using the crated exemplary data file, by specifying the  $H_0$  and  $H_1$  model in the same way as the power computation using the simulated large data file method. The only difference, when compared with the simulated large data file method discussed above, is that the case weight of the exemplary data file has to be specified in both the  $H_0$  and  $H_1$  model. This requires adding the line “caseweight frequency;” to the “variables” section.

#### b) Non-standard case

We will illustrate MC-based power computation for the non-standard case using an example in which one of the transition probabilities is fixed to 0, implying that the transition logit concerned is fixed to a large negative value (say -100). Power computation in the non-standard case proceeds in two steps. First, we obtain the critical values under  $H_0$  by simulation and subsequently we obtain the power given this critical value.

To obtain the critical value, we define the  $H_0$  and  $H_1$  model in the same input file. In the  $H_1$  model, we use the “MCstudy” option and specify the number of Monte Carlo replications, the  $H_0$  model, the sample size, and level of significance “alpha”, that is,

```

// H1 model for obtaining CV by simulation
options
....
montecarlo replicates=5000 MCstudy='H0' N=300 alpha=0.05;
...

```

The H0 model contains the population values for the free parameters as well as the constraint. That is,

```

// H0 model for obtaining CV by simulation
...
equations

```

```
...  
beta[1,1] = -100;  
{...}
```

Running the  $H_1$  model gives us the critical value (CV).

In the final step, which is identical to the MC-based power computation described above, except that instead of defining the significance level, we specify the CV value that we obtained by running the  $H_1$  model; that is, define the  $H_0$  and  $H_1$  model in a single input file as

```
// H0 model for obtaining power by simulation  
options  
...  
montecarlo replicates=5000 power='H1' N=300 CV=2.2344;  
variables  
...  
equations  
...  
b[1,1]=-100;
```

The  $H_1$  model is again equal to the basic model with the population values for the parameters. Running the  $H_0$  model will give us the power for the specified sample size  $N$  and the estimated critical value  $CV$ , which we set here to  $N = 300$  and  $CV = 2.2344$ .