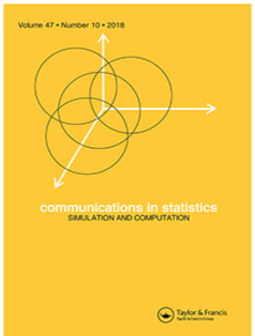
Volume 47 • Number 10 • 2018
communications in statistics
SIMULATION AND COMPUTATION
Taylor & Francis

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

The development of a goodness-of-fit test for high level binary multilevel models

Gayara Fernando & Roshini Sooriyarachchi

To cite this article: Gayara Fernando & Roshini Sooriyarachchi (2020): The development of a goodness-of-fit test for high level binary multilevel models, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2019.1700275

To link to this article: https://doi.org/10.1080/03610918.2019.1700275

Published online: 29 Jan 2020.

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2019.1700275

The development of a goodness-of-fit test for high level binary multilevel models

Gayara Fernando and Roshini Sooriyarachchi

Department of Statistics, University of Colombo, Colombo, Sri Lanka

ABSTRACT

Before making inferences about a population using a fitted model, it is necessary to determine whether the fitted model describes the data well. A poorly fitted model may lead to biased and invalid conclusions, resulting in incorrect inferences. Recent studies show the necessity of goodness-of-fit tests for high level binary multilevel models. The focus here was to develop a goodness-of-fit test to use in the model adequacy testing of high level binary multilevel models and to examine, whether the type I error and power hold for the newly developed goodness-of-fit test considering a three-level random intercept model.

ARTICLE HISTORY
Received 6 April 2017
Accepted 28 November 2019

KEYWORDS
Goodness-of-fit test; Limited-information goodness-of-fit testing; High level models; Type I error; power; Simulation

1. Introduction

1.1. Background

Before a model is trusted for drawing conclusions or in predicting future outcomes about a population of interest, it is crucial to show that the fitted model suits the scenario under consideration. This is where a goodness-of-fit test will show its importance. A goodness-of-fit test determines the model adequacy of a fitted model. Else in simple terms, a goodness-of-fit test shows how well the fitted values of a response variable under the model compared with the observed values. If a model is ill-fitted, this may lead to biased and invalid conclusions which will mislead the person fitting the model to make incorrect inferences. Therefore, it is essential to carry out a goodness-of-fit test to determine the model adequacy before making inferences from the fitted model. For single-level data structures, there are many accepted goodness-of-fit tests for binary and other response data types, as the concepts are well developed (Hosmer et al. 1997). In modeling clustered or hierarchical data, two-level models are most common, but three and higher levels are not frequently examined. Also, there are goodness-of-fit tests developed to test the adequacy of two-level multilevel models (Perera et al. 2016; Epasinghe and Sooriyarachchi 2017). However, when considering the higher level scenarios, it is seen that no satisfactory goodness-of-fit tests are available (Cool et al. 2015) and this may lead to problems, as high level scenarios can occur, though not that frequently. It should be studied whether the same techniques that are used in the case of

CONTACT Roshini Sooriyarachchi roshinis@hotmail.com Department of Statistics, University of Colombo, Colombo 3, Sri Lanka.

two-level situations can be used to assess the higher level scenarios as well as if separate concerns need to be made. Thus, the development of a goodness-of-fit test for high dimensional multilevel data may be considered as a novel development.

1.2. Objectives

The main concern under this research study is to develop a suitable goodness-of-fit test for high level binary multilevel models. The secondary objectives of this study is to determine the properties of the developed goodness-of-fit test for varying cluster sizes and intracluster correlation (ICC) values with the use of simulations and to assess the model adequacy by applying the developed goodness-of-fit test to a real-life dataset. What is new in this research study, compared to Perera et al. (2016) test for a two-level binary logistic model, is the introduction and adoption of the concepts of limited-information goodness-of-fit testing, introduced by Maydeu-Olivares and Joe (2006); Maydeu-Olivares and Garcı́A-Forero (2010) and Maydeu-Olivares et al. (2011) to the high level binary multilevel models. It should be noted that these concepts have not previously been used in the context of multilevel modeling.

1.3. Literature review

In this section, the research work required for building up the new goodness-of-fit test for high level multilevel models is discussed.

The goodness-of-fit of a statistical model describes how well the statistical model of interest fits into a set of observations (Maydeu-Olivares and Garcı́A-Forero 2010). A person may conclude that a model fits the data well if the differences between the observed and fitted values from the model are small and if there is no systematic contribution of these differences to the error structure of the model (Archer and Lemeshow 2006). When it comes to the goodness-of-fit statistics, a variety of goodness-of-fit tests are available. Most of them are full information goodness-of-fit tests. However, limited-information goodness-of-fit testing, piecewise goodness-of-fit testing and assessing model fit using goodness-of-fit indices have become much more common in the last decade (Maydeu-Olivares and Garcı́A-Forero 2010). The proposed goodness-of-fit test is developed under the consideration of the limited-information goodness-of-fit testing.

1.3.1. Full information goodness-of-fit testing

Full information goodness-of-fit tests use all the information available in coming up with the test statistic, and are defined as the goodness-of-fit tests that utilize all the information in the contingency tables (Cai et al. 2006). These are the most commonly occurring goodness-of-fit tests for assessing the model fit. As information regarding this type of testing (i.e., for single-level binary responses and for two-level hierarchical structures when the response is binary) is well documented (Perera et al. 2016) no further discussion of this will be made here.

1.3.2. Limited-information goodness-of-fit testing

Recently, limited-information goodness-of-fit testing has received increased attention in the psychometric literature. In contrast to full information test statistics like Pearson's chi-square or the likelihood ratio, instead of using the information in the full contingency table these limited-information tests utilize lower order marginal tables (Joe and Maydeu-Olivares 2010; Cai and Hansen 2013). The Theory of limited-information goodness-of-fit tests has been well developed in recent years, thanks to a theoretical breakthrough made by Maydeu-Olivares and Joe (2006).

(i) Limited-information goodness-of-fit for discrete observed data for a single-level. In the case of limited-information goodness-of-fit tests, the information contained in the high-order margins of the contingency table are ignored and only information low-order margins are used. Maydeu-Olivares and Joe (2006) recommended testing at the highest level of margins for which the model is specified disregarding the higher order margins. Joe and Maydeu-Olivares (2010) then suggest a family of goodness-of-fit tests. Their theory is based on unifying limited-information and full information goodness-of-fit statistics. This statistic is denoted by M_r and is obtained by integrating the residual proportions up to order r, their asymptotic covariance matrix, the matrix of derivatives of the marginal probabilities up to order r and the number of observations. The asymptotic distribution of any statistic of the above family is shown to be chi-square with degrees of freedom equal to the number of residuals used – the number of model parameters to be estimated. Though now the higher order margins are ignored, Joe and Maydeu-Olivares (2010) have shown that the asymptotic p-values are accurate, even for large models fitted to small samples for the developed limited-information goodness-of-fit test. They have also demonstrated that within this framework more power is obtained compared to the case of using all available information regarding the data (Maydeu-Olivares and GarcíA-Forero 2010).

(ii).Limited-information goodness-of-fit for discrete observed data for hierarchical modeling. The limited-information goodness-of-fit testing concept has been extended to hierarchical item factor models. In this case, Cai and Hansen (2013) have proposed a dimension reduction method that can take advantage of the hierarchical factor structure so that the integrals can be approximated far more efficiently. This statistic is best understood as a further reduction of the univariate and bivariate marginal tables. The residuals used in the quadratic form are linear functions of the multinomial cell residuals, but they are not marginal probabilities as these proposed by Maydeu-Olivares and Joe (2006). Maydeu-Olivares et al. (2011) have also suggested a similar statistic in the context of unidimensional graded Item Response Theory (IRT) models. However, this idea of limited-information goodness-of-fit has not been used in any other hierarchical modeling concept (for binary, ordinal, survival, etc.) yet. Under the current research, to develop goodness-of-fit test for higher level multilevel models, the ideas of limited-information goodness-of-fit testing introduced by Maydeu-Olivares and GarcíA-Forero (2010), Cai et al. (2006), and Cai and Hansen (2013) have been incorporated in coming up with the novel method. What is characteristically different among the aforementioned tests and the current test is that these previous tests depend on the complicated approximation of integrals and there is no available statistical package for the user to

try out these models. Therefore, most of these tests remain in the theoretical realm. The current test is developed within the framework of MLwiN which is a very versatile package for multilevel modeling. Moreover, the suggested test statistic is very simple and is based on the Hosmer and Lemeshow (2000) test statistic.

The rest of this article is organized as follows. In Sec. 2, the fundamental statistical theories on which this work is based are discussed, and the statistical methodologies used in the development of the goodness-of-fit test are presented. In Sec. 3 extensive simulations are carried out in order to determine the properties of the developed goodness-of-fit test. In Sec. 4, an application of the developed goodness-of-fit test to a real-life dataset is presented to assess the model adequacy. In Sec. 5, some concluding remarks are made, and the research issues involved in this article are discussed.

2. Theory and methodology

2.1. A three-level random intercept model

Similar to that of the model for the two-level scenario (Perera et al. 2016), the logistic model can be extended for the three-level multilevel model with binary response variables. Let y_{ijk} be the response variable for the i$^{\text{th}}$ individual (level 1 observation) lying in the j$^{\text{th}}$ second level cluster which in turn lies within the k$^{\text{th}}$ third level cluster. As the response of interest is binary,

$$y_{ijk} = \begin{cases} 1 \; \textit{if a success occurs for the } \mathrm{i}^{\mathrm{th}} \textit{ individual in the } \mathrm{j}^{\mathrm{th}} \textit{ cluster that is within the } \mathrm{k}^{\mathrm{th}} \textit{ cluster} \\ 0 \; \textit{if a failure occurs for the } \mathrm{i}^{\mathrm{th}} \textit{ individual in the } \mathrm{j}^{\mathrm{th}} \textit{ cluster that is within the } \mathrm{k}^{\mathrm{th}} \textit{ cluster} \end{cases}$$

For the three-level multilevel model, the random intercept-only, functional model considering a single explanatory variable measured at the lowest level of the hierarchy can be written as

$$\text{logit }(\pi_{ijk}) = \beta_{0jk} + \beta_1 x_{ijk} \tag{1}$$

where $\beta_{0jk} = \beta_0 + v_{0k} + u_{0jk}$ and $v_{0k} \sim N\left(0, \sigma^2_{v0}\right),\; u_{0jk} \sim N(0, \sigma^2_{u0})$.

As in the case of the two-level model, π_{ijk} is the probability that $y_{ijk} = 1$. In the case of the three-level model the intercept consists of three components, a fixed component β_0, and two random components for the two higher levels; that is, for the second level and third level. These random components are now independent of each other and are distributed normally with mean zero and variances as shown in Eq. (1) It should be noted that v_{0k} is the random component for the third level and u_{0jk} is the random component for the second level. Similarly, the higher dimensional multilevel models can also be defined. The model fitting and simulations are done based on a three-level random intercept-only model under the current research. As now there are two cluster levels (2$^{\text{nd}}$ level and 3$^{\text{rd}}$ level), the combined ICC will suggest how strongly the observations within a considered 2$^{\text{nd}}$ level cluster and a 3$^{\text{rd}}$ level cluster are correlated. The formula for calculating the combined ICC is described as follows.

$$\textit{Combined ICC} = \frac{\sigma^2_{u0} + \sigma^2_{v0}}{\sigma^2_{u0} + \sigma^2_{v0} + \sigma^2_{e}} \;\; \text{(Gregorich 2013).} \tag{2}$$

Table 1. Multilevel model for the kth level of the large cluster (level 3) as a contingency table.

	Second Cluster Level (X_2)					
First level (i)	1	2	....	j	...	J
1	π_{11k}	π_{12k}	...	π_{1jk}	...	π_{1Jk}
2	π_{21k}	π_{22k}	...	π_{2jk}		π_{2Jk}
...	...	...	...	...	...	...
i	π_{i1k}	π_{i2k}	...	π_{ijk}	...	π_{iJk}
...	...	...	...	...	...	...
I	π_{I1k}	π_{I2k}	...	π_{Ijk}	...	π_{IJk}

σ_ε^2 is the variance of the residuals at the 1st level. σ_ε^2 can be estimated by $\frac{\pi^2}{3} \approx 3.29$ for binary data.

2.2. The main theories used in developing the goodness-of-fit test

Perera et al. (2016) discuss theories regarding the building up of the goodness-of-fit test for the two-level multilevel case.

The main new theory incorporated in coming up with the proposed goodness-of-fit test is the limited-information goodness-of-fit testing concepts for discrete observed data developed by Maydeu-Olivares et al. (Maydeu-Olivares and Joe 2006; Maydeu-Olivares and Garcı́A-Forero 2010; Maydeu-Olivares et al. 2011) which is adapted to the proposed test. Under the current study, the three-level multilevel model can be considered as a contingency table. Let i, j, and k denote the first, second and third levels respectively. That is simply the i^{th} individual (1st level) that belongs to the j^{th} small cluster (2nd level) that in turn lies in the k^{th} big cluster (3rd level). Here, let $i = 1, 2, \ldots I$ or $i = 1\ (1)\ I$ where i corresponds to the i^{th} first level unit. Let X_1 be the value (0 or 1) which is the response given by the unit. Here $j = 1\ (1)\ J$ where J is the number of small clusters and X_2 corresponds to the j^{th} 2nd level cluster. Here $k = 1\ (1)\ K$ where K is the number of large clusters and X_3 corresponds to the k^{th} 3rd level cluster. For the k^{th} cluster at the 3rd level, ($k = 1\ (1)\ K$), the obtained probabilities (fitted values) can be represented in a two-way contingency table as given in Table 1. If the K number of contingency tables is drawn considering each 3rd level cluster, the multilevel structure will correspond to a three-way contingency table.

Here, X_3 denotes the third cluster level. The three-way contingency table that can be created using a K number of two-way contingency tables considering all the clusters of the third level can now be represented using all the cell probabilities as follows.

$$\boldsymbol{\pi} = (\pi_{111}, \pi_{112}, \ \ldots, \pi_{11k}, \ldots, \pi_{121}, \ \pi_{122}, \ \ldots, \pi_{12k}, \ \ldots, \pi_{ijk}, \ \ldots, \pi_{IJK})$$

Alternatively, it can be expressed using the univariate, bivariate, and trivariate probabilities; $\boldsymbol{\pi}_{\mathrm{i}}^{(\mathrm{l})} = \Pr(X_{\mathrm{i.}} = \mathrm{l})$ and similarly $\boldsymbol{\pi}_{\mathrm{j}}^{(\mathrm{m})}$ and $\boldsymbol{\pi}_{\mathrm{k}}^{(\mathrm{n})}$, $\boldsymbol{\pi}_{\mathrm{ij}}^{(\mathrm{l,m})} = \Pr(X_i{=}l\ and\ X_j{=}m)$ and similarly, $\boldsymbol{\pi}_{\mathrm{jk}}^{(\mathrm{m,n})}$ and $\boldsymbol{\pi}_{\mathrm{ik}}^{(\mathrm{l,n})}$, and $\boldsymbol{\pi}_{\mathrm{ijk}}^{(\mathrm{l,m,n})} = \Pr(X_i{=}l\ and\ X_j{=}m\ and\ X_k{=}n)$ respectively, as shown in Maydeu-Olivares and Garcı́A-Forero (2010). Both expressions are equivalent, and the equivalence is valid for contingency tables of any dimension. In the case of the newly proposed goodness-of-fit test statistic, instead of the M_r statistic introduced by Maydeu-Olivares and Joe (2006), the Hosmer and Lemeshow (2000) statistic has been used. The Hosmer and Lemeshow test statistic here is based on ordering the fitted values $\widehat{\pi_{ijk}}$ at the highest level margin under the fitted model. Maydeu-Olivares et al. (2011) developed a goodness-of-fit test for Item Response theory and Factor Analysis Models using univariate and bivariate covariances. Their work was based on Samejima (1969) and computing the model-implied variances and covariances under the graded logistic model. Limited-information goodness-of-fit methods disregard information contained in the high-order margins of the table. Thus, quadratic forms of univariate and bivariate residuals are used instead of using all marginal residuals up to order n (if n levels) as suggested by Maydeu-Olivares and Joe (2006) and Joe and Maydeu-Olivares (2010) in calculating fitted values. Based on Maydeu-Olivares and GarcíA-Forero (2010) and Maydeu-Olivares et al. (2011) the fitted probability values in this article are computed using the univariate (third level variance V_{0k}) and bivariate (second level variance U_{0jk}) residuals. In the Hosmer and Lemeshow statistic, applied to the 2-level binary multilevel model by Perera et al. (2016), indicator variables are given within the lowest level (1^{st} level) in ranking the sorted probabilities. This is extended for the three-level case, following Maydeu-Olivares and GarcíA-Forero (2010) where indicator variables are given at the highest margin (π_{ijk}). This corresponds to the lowest level (1^{st} level) which is within the 2^{nd} level cluster and the 3rd level cluster. The fitted values are calculated using univariate and bivariate residuals of the higher levels as suggested by Maydeu-Olivares and GarcíA-Forero (2010). This now clarifies the two phrases in the limited-information goodness-of-fit testing that states, (i) "testing is done at the higher order margins", where the π_{ijk}'s are tested and (ii) "ignoring information at the higher order margins" and using only the quadratic forms of univariate and bivariate residuals, in calculating fitted values that correspond to the two lower level marginals.

2.3. Proposed goodness-of-fit test for high dimensional binary multilevel models

Incorporating the theories discussed above, the following steps are carried out in developing the proposed goodness-of-fit statistic.

Step 1: The three-level binary random intercept-only model as specified in Section 2.1. is initially fitted and the model parameters are estimated using the 1^{st} order Marginal Quasi-Likelihood (MQL1) method. The MQL1 procedure of estimation is used here as all other methods of estimation result in non-convergence problems. The model fitted as suggested in Sec. 2.1 is as follows and the definitions are as in Sec. 3.1. (three-level random intercept model)

$$\text{logit } (\pi_{ijk}) = \beta_{0jk} + \beta_1 x_{ijk}$$

where

$$\beta_{0jk} = \beta_0 + v_{0k} + u_{0jk}$$

and $v_{0k} \sim N(0, \sigma^2_{v0})$ and $u_{0jk} \sim N\left(0, \sigma^2_{u0}\right)$

Step 2: The predicted probabilities, π_{ijk} based on the estimated parameters are calculated for the i^{th} first level units (observations) residing within the j^{th} small cluster (2^{nd} level), which in turn lies within the k^{th} large cluster (3^{rd} level) considering the quadratic forms of univariate (v_{0k}) and bivariate (u_{0jk}) residuals as suggested by Maydeu-Olivares and GarcíA-Forero (2010).

Step 3: Then the estimated predicted probabilities are sorted in ascending order, within each second level cluster which resides within each third level cluster as suggested by Maydeu-Olivares and GarcíA-Forero (2010) such that the information at the higher order margins are discarded. Sorting follows the method of Rosner et al. (2003).
Step 4: The sorted probabilities are then ranked within the first level.

In the Hosmer and Lemeshow statistic applied to multilevel data by Perera et al. (2016), for ranking the sorted probabilities, indicator variables were specified within the lowest level. Now, for the three-level binary multilevel model, this is further extended using limited-information goodness-of-fit as suggested by Maydeu-Olivares and GarcíA-Forero (2010). Then, the sorted probabilities are ranked again within the lowest level (i.e., the highest margin) by assigning indicator variables (This is repeated within levels 2 and 3).

The sorted probabilities are grouped into G groups as suggested in the Hosmer and Lemeshow test from ranking and assigning indicator variables as suggested in Lipsitz et al. (1996), for each level one, sorted unit.

As an example, in creating ten groups, they are created such that the 1st group contains observations with the smallest predicted probabilities (i.e., the observations are ranked as I_1), and the last group includes observations with the largest estimated probabilities (these observations are ranked as I_10), where Hosmer and Lemeshow (2000) suggested 10 as a suitable value for the number of groups). The 1st grouping strategy, under Section 2.2. is proposed here. The ten groups, however, may not always be of the same size (roughly of equal size).

Step 5: After ranking the probabilities, π_{ijk} (allocation of indicator variables), the dataset is re-arranged in the way it originally was (originally, as before it was sorted).
Step 6: A pooled indicator variable is now created across the third level for the whole dataset, for each of the created indicators under step 4. The observations within any third level cluster are independent of another observation within any other third level cluster. Therefore, all indicator variables that pertain to a specific group created under Hosmer and Lemeshow method, are pooled into a single indicator variable. (i.e., now a single indicator variable is used to represent all probabilities). These were ranked earlier using indicator variables within each second and third level cluster, for all units over all clusters. Likewise $G - 1 = 10 - 1 = 9$ pooled indicator variables will be created to represent all the groups holding the group that contains the smallest probabilities as the reference group. This can be justified because now, as at the third level the observations are independent of each other.

The pooled indicator variable can now be interpreted as follows.

$$I_{g_{ijk}} \begin{cases} 1; & if \ \ \pi_{ijk} \ \ is \ in \ region \ g \\ 0; & otherwise \end{cases}$$

With $g = 2, 3, \ldots, 10$

Step 7: An alternative model is now fitted by incorporating the pooled indicator variables, and the parameters are estimated using the 1st order MQL method. The fitted alternative model takes the form given in equation (3).

$$\log\left(\frac{\pi_{ijk}}{1-\ \pi_{ijk}}\right) = \beta_{0jk} + \beta_1 x_{ijk} + \sum_{g=2}^{10} \gamma_g I_{g_{ijk}} \tag{3}$$

Where $\sum_{g=2}^{10} \gamma_g I_{g_{ijk}} = \gamma_2 I_{2_{ijk}} + \gamma_3 I_{3_{ijk}} + \ldots + \gamma_{10} I_{g_{ijk}}$
With i, j, and k as in the usual notation.

Step 8: The joint Wald statistic is now calculated (Liao 2004) for the model in Eq. (3) to determine whether the coefficients of the indicator variables are simultaneously equal to zero, using MLwiN (at the α% significance level.). The hypothesis of interest can be written as

$H_0 : \gamma_2 = \gamma_3 = \ldots = \gamma_{10} = 0$ and H_1 : Not all coefficients of the model are zero.

Step 9: The calculated joint Wald statistic value is compared to the $\chi^2_{(G-1)}$ distribution at an α% significance level.

If the calculated joint Wald statistic value is less than the $\chi^2_{(G-1)}$ at α% value, then the null hypothesis is not rejected at an α% level. (all the indicator variables are simultaneously equal to zero). That is, in this case, the model 3 reduces to the model 1. Hence, it can be said that the model 1 fits the data well. If, however, the calculated joint Wald statistic value is greater than $\chi^2_{(G-1)}$ at an α% level, then the model 1 has a questionable fit. Hence, it does not fit the data well (all indicator variables are not simultaneously equal to zero).

These steps can be summarized in the following flow diagram.

These are the steps to be followed in determining the goodness-of-fit of a fitted model for 3-level binary multilevel models.

The sorting can now be generalized. As the highest possible dimension in MLwiN is five, the extension to four and five dimensions can be discussed. Based on Maydeu-Olivares and GarcíA-Forero (2010), Maydeu-Olivares and Joe (2006) for the four level model the fitted probabilities in the first level will be sorted and ranked within the 2nd, 3rd, and 4th levels. To calculate the fitted values the univariate (w_{0l}) and bivariate (v_{0kl}) variance components will be used. Similarly, for the five level model, the fitted probabilities in the first level will be sorted and ranked within the 2nd, 3rd, 4th, and 5th levels. To calculate the fitted values the univariate (z_{0m}) and bivariate (w_{0lm}) variance components will be used. Then the same procedure may be followed in coming up with a goodness-of-fit statistic.

3. A simulation study

To justify the proposed goodness-of-fit test for high level multilevel models, the stipulated type I error needs to hold and power should be reasonable for this test. Therefore, these properties need to be checked and this can be used in assessing the model fit and hence, making inferences. The simulations were done using macros for varying number of individuals (observations) at the lowest level of the hierarchy, varying cluster sizes for the higher levels and also for the different variances of random effects at the higher levels of the hierarchy. It should be noted that in this study far more extensive simulations have been carried out comparatively to the formerly developed goodness-of-fit test by Perera et al. (2016) for both type I error and power. Whereas, in the former case only a few (12 simulations each) have been carried out. However, in the current situation, as it deals with three levels, substantial amounts of simulations have been carried out in justifying the developed goodness-of-fit test. Under each of the determined properties (type I error and power), the simulations carried out are discussed in detail, and the necessary techniques to be used for differing cluster sizes and ICC values are recommended.

3.1. Considerations used under the simulation study

For the simulations of both type I error and power, 1000 datasets were generated under several specified conditions depending on both cluster sizes and standard deviation of random effects (ICC combinations) considering each level of the three-level hierarchy. The description of the combinations is provided in the following sections. As mentioned in Sec. 2.1, selection of the model (random intercept with a single explanatory variable) was due to the simplicity. The explanatory variable is defined at the lowest level of the hierarchy. The simulation of the explanatory variable is done based on a normal distribution as suggested by Perera et al. (2016), and the simulation of the random effects is also done based on the normal distribution.

(i) Incorporating sample size for simulations. Based on the suggestions in the literature and taking into account practical considerations the number of clusters and the individuals to be used within the clusters were decided. Two cluster sizes for each of the three levels were taken such that the sizes can be justified for real-life scenarios, as very large cluster sizes are unlikely to occur. Also, it should be noted that when the total sample size is considerable, MLwiN is unable to handle these heavy simulations due to non-convergence. Considering all these aspects, the cluster sizes used in the simulation study, to determine the properties of the proposed goodness-of-fit test are as follows.

- 1st level (individual observations) – 30 and 50
- 2nd level (small clusters) – 15 and 30
- 3rd level (large clusters) – 10 and 15

These three specifications result in 8 combinations of cluster sizes to be used in the case of simulations.The sample sizes that result from the above cluster sizes range from 4500 to 22,500.

(ii) Incorporating standard deviations. The variances for the random effects of the two levels were chosen according to the guidelines set out in the literature. A constraint had to be met in coming up with the selected values for the standard deviations. That is, the variance of the random effect pertaining to the third level should be less than or equal to the variance of the random effect pertaining to the second level (Gregorich 2013). The standard deviation values chosen for the random effects of the two levels are

- 1, 1.5, and 2 for the second level (Perera et al. 2016)
- 0.4, 0.5, and 1 for the third level (Fotouhi 2003).

The eight cluster combinations together with nine standard deviation combinations result in $9 \times 8 = 72$ extensive simulations to be carried out.

3.1.1. Combined ICC

As discussed under Sec. 2.1 it is clear that the interest now lies in the combined ICC and not on individual ICCs at the two higher levels. Therefore, for each identified standard deviation combination the combined ICC values can be calculated using the formula of Gregorich (2013).

The ICC values can be calculated for each of the identified cluster combinations. As an example, for the first standard deviation combination considered, which is 0.4 at the third level and one at the second level, $\sigma^2_{u0} = 1.0$ and $\sigma^2_{v0} = 0.16$ the combined ICC can be calculated as $= \frac{1.0+0.16}{1.0+0.16+\frac{\left(\frac{22}{7}\right)^2}{3}} = 0.2605$

Similarly, for each identified standard deviation combination, the combined ICC can be calculated as given in the above case.

3.2. Simulation studies to determine the type I error (TOE)

The probability of rejecting the null hypothesis given that it is true is the type I error rate or significance level, and it is denoted by α (alpha) and is also called the alpha level. Often the significance level is set at 5% (0.05). It is of interest to determine the probability interval for the type I error in the case of simulating 1000 datasets in determining the properties of the suggested goodness-of-fit test.

A 100 (1- β) % probability Interval for a significance level of size α is given by the formula $\alpha \pm z_{\beta/2}\sqrt{\frac{\alpha\ (1-\alpha\ )}{n}}$. (4)

For a 95% probability interval for an α of 5% = 0.05 substitution to the above formula gives $0.05 \pm 1.96\sqrt{\frac{0.05(1-0.05)}{1000}} = (0.036491, 0.063508) \approx (0.036, 0.064)$(5)

Hence, in order to see whether the developed goodness-of-fit test achieves the type I error, it is necessary to simulate the proposed 1000 datasets under each of the identified 72 combinations in accordance with the null hypothesis.

3.2.1. Data generation procedure under the null hypothesis

In determining the type I error, the data were generated under the null hypothesis. The simulations were done so that the resultant data are all random. Macros under the MLwiN version 2.19 were used throughout the simulations carried out. During the data generation procedure, first, the explanatory variable was generated from the normal distribution with mean 2 and standard deviation 1 ($x_{ijk} \sim N(2, 1)$) as suggested by Perera et al. (2016) and Archer et al. (2007). Then the random effects of the two higher levels were generated according to the values of standard deviations specified earlier, also from the normal distribution according to the definition of multilevel modeling $V_{0k} \sim N(0, \sigma_{v0}^2)$ and $U_{0jk} \sim N(0, \sigma_{u0}^2)$. Once the explanatory variable and the random effects have been generated, to determine the estimated probabilities (π_{ijk}), from trial and error method β_0 and β_1 were selected as in Perera et al. (2016) so that $\beta_0 = -0.125$ and $\beta_1 = 0.50$. Then from the fitted model under data generation, the probabilities were estimated such as follows.

$$\pi_{ijk} = \frac{\exp(-0.125 + V_{0k} + U_{0jk} + (0.50 \times x_{ijk}))}{1 + \exp(-0.125 + V_{0k} + U_{0jk} + (0.50 \times x_{ijk}))} \tag{6}$$

The binary response variable, y_{ijk} was then created based on the Bernoulli distribution, by considering the above-estimated probabilities as the probabilities of success. If the TOE holds for the developed goodness-of-fit test, then out of the 1000 datasets, the rejection proportion should lie within the calculated 95% probability limits of the TOE.

3.3. Simulation studies to determine the power

The power of any considered test of statistical significance is the probability that it will reject the null hypothesis if it is false. Statistical power is inversely related to β which is the probability of making a Type II error. Very high power will indicate the ability to detect the deviation specified under the alternative hypothesis, from the null hypothesis. Section 3.3 will mainly look at the power of the developed goodness-of-fit test when the ICC values change and the cluster sizes at the three levels of the hierarchy change.

3.3.1. Data generation procedure for power

In determining the power of the developed goodness-of-fit test, data were generated under the alternative hypothesis. In this, the model fitted to data are mis-specified by incorporating an incorrect form for the explanatory variable. Several suggestions have been made in the literature for the selection of an alternative model (Archer et al. 2007) and the model fitted here was one with a single explanatory variable X, and it has been transformed to generate data as LogX^2 as suggested by Perera et al. (2016). The simulations are done so that the resultant data are all random.

All x_{ijk}, V_{0k} and U_{0jk} are simulated as stated under Sec. 3.2 as was done for the null hypothesis. The model fitted in generating data can be given as follows.

$$\log\left[\frac{P_{ijk}}{1 - P_{ijk}}\right] = \beta_{0jk} + 0.50 \times log(x_{ijk} \times x_{ijk}) \tag{7}$$

Where $\beta_{0jk} = -0.125 + V_{0k} + U_{0jk}$ and $V_{0k} \sim N(0, \sigma_{v0}^2)$ and $U_{0jk} \sim N\left(0, \sigma_{u0}^2\right)$. The same estimates for β_0 and β_1 have been used here as done in the null hypothesis. From the fitted model the probabilities can now be estimated using the following.

$$\pi_{ijk} = \frac{\exp\Big(-0.125 + V_{0k} + U_{0jk} + \big(0.50 \times log(x_{ijk} \times x_{ijk})\big)\Big)}{1 + \exp\Big(-0.125 + V_{0k} + U_{0jk} + \big(0.50 \times log(x_{ijk} \times x_{ijk})\big)\Big)} \tag{8}$$

Table 2 (a). Rejection proportions for case 1.

Cluster combination	Level 3 size	Level 2 size	Level 1 size	Rejection proportion under H_o	Rejection proportion under H_1	Status under H_o
1	15	30	50	0.046	1.0	Within limits
2	15	30	30	0.049	1.0	Within limits
3	15	15	50	0.054	1.0	Within limits
4	15	15	30	0.041	0.998	Within limits
5	10	30	50	0.040	1.0	Within limits
6	10	30	30	0.049	1.0	Within limits
7	10	15	50	0.054	1.0	Within limits
8	10	15	30	0.055	0.981	Within limits

Table 2 (b). Rejection proportions for case 2.

Cluster combination	Level 3 size	Level 2 size	Level 1 size	Rejection proportion under H_o	Rejection proportion under H_1	Status under H_o
c	15	30	50	0.054	1.0	Within limits
2	15	30	30	0.054	1.0	Within limits
3	15	15	50	0.053	1.0	Within limits
4	15	15	30	0.043	0.997	Within limits
5	10	30	50	0.062	1.0	Within limits
6	10	30	30	0.045	0.999	Within limits
7	10	15	50	0.060	1.0	Within limits
8	10	15	30	0.073		Outside limits

Table 2 (c). Rejection proportions for case 3.

Cluster combination	Level 3 size	Level 2 size	Level 1 size	Rejection proportion under H_o	Rejection proportion under H_1	Status under H_o
1	15	30	50	0.059	1.0	Within limits
2	15	30	30	0.056	1.0	Within limits
3	15	15	50	0.056	1.0	Within limits
4	15	15	30	0.045	0.989	Within limits
5	10	30	50	0.053	1.0	Within limits
6	10	30	30	0.048	0.999	Within limits
7	10	15	50	0.059	0.996	Within limits
8	10	15	30	0.101	0.970	Outside limits

3.4. Simulation results

This section provides the results obtained under the simulation study for determining the type I error and power. Under each of the identified standard deviation combinations in the second and third levels, the rejection proportion for the nine cluster combinations are presented and discussed in Tables 2(a–i).

Case 1: Level 3 standard deviation = 0.4, Level 2 standard deviation = 1.0, combined ICC = 0.2605.

Case 2: Level 3 standard deviation = 0.4, Level 2 standard deviation = 1.5, combined ICC over =0.4226.

Case 3: Level 3 standard deviation = 0.4, Level 2 standard deviation = 2.0, The combined ICC = 0.5582.

Table 2 (d). Rejection proportions for case 4.

Cluster combination	Level 3 size	Level 2 size	Level 1 size	Rejection proportion under H_o	Rejection proportion under H_1	Status under H_o
1	15	30	50	0.045	1.0	Within limits
2	15	30	30	0.048	1.0	Within limits
3	15	15	50	0.052	1.0	Within limits
4	15	15	30	0.035	0.998	Just outside limits
5	10	30	50	0.051	1.0	Within limits
6	10	30	30	0.046	1.0	Within limits
7	10	15	50	0.055	1.0	Within limits
8	10	15	30	0.069	0.974	Outside limits

Table 2 (e). Rejection proportions for case 5.

Cluster combination	Level 3 size	Level 2 size	Level 1 size	Rejection proportion under H_o	Rejection proportion under H_1	Status under H_o
1	15	30	50	0.047	1.0	Within limits
2	15	30	30	0.060	1.0	Within limits
3	15	15	50	0.055	1.0	Within limits
4	15	15	30	0.038	0.997	Within limits
5	10	30	50	0.045	1.0	Within limits
6	10	30	30	0.043	0.999	Within limits
7	10	15	50	0.056	1.0	Within limits
8	10	15	30	0.081	0.970	Outside limits

Table 2 (f). Rejection proportions for case 6.

Cluster combination	Level 3 size	Level 2 size	Level 1 size	Rejection proportion under H_o	Rejection proportion under H_1	Status under H_o
1	15	30	50	0.057	1.0	Within limits
2	15	30	30	0.065	1.0	Just outside limits
3	15	15	50	0.055	1.0	Within limits
4	15	15	30	0.037	0.988	Within limits
5	10	30	50	0.054	1.0	Within limits
6	10	30	30	0.048	0.999	Within limits
7	10	15	50	0.050	0.995	Within limits
8	10	15	30	0.094	0.962	Outside limits

Table 2 (g). Rejection proportions for case 7.

Cluster combination	Level 3 size	Level 2 size	Level 1 size	Rejection proportion under H_o	Rejection proportion under H_1	Status
1	15	30	50	0.057	1.0	Within limits
2	15	30	30	0.050	1.0	Within limits
3	15	15	50	0.054	1.0	Within limits
4	15	15	30	0.038	0.999	Within limits
5	10	30	50	0.053	1.0	Within limits
6	10	30	30	0.049	0.999	Within limits
7	10	15	50	0.056	1.0	Within limits
8	10	15	30	0.078	0.971	Outside limits

Case 4: Level 3 standard deviation = 0.5, Level 2 standard deviation = 1.0, The combined ICC =0.2752.

Table 2 (h). Rejection proportions for case 8.

Cluster combination	Level 3 size	Level 2 size	Level 1 size	Rejection proportion under H_o	Rejection proportion under H_1	Status
1	15	30	50	0.057	1.0	Within limits
2	15	30	30	0.054	1.0	Within limits
3	15	15	50	0.049	0.998	Within limits
4	15	15	30	0.041	0.999	Within limits
5	10	30	50	0.046	1.0	Within limits
6	10	30	30	0.047	0.999	Within limits
7	10	15	50	0.041	0.999	Within limits
8	10	15	30	0.096	0.970	Outside limits

Table 2 (i). Rejection proportions for case 9.

Cluster combination	Level 3 size	Level 2 size	Level 1 size	Rejection proportion under H_0	Rejection proportionUnder H_1	Status
1	15	30	50	0.063	1.0	Within limits
2	15	30	30	0.057	0.996	Within limits
3	15	15	50	0.050	0.998	Within limits
4	15	15	30	0.046	0.993	Within limits
5	10	30	50	0.036	1.0	Within limits
6	10	30	30	0.046	0.999	Within limits
7	10	15	50	0.051	0.998	Within limits
8	10	15	30	0.115	0.965	Outside limits

Case 5: Level 3 standard deviation = 0.5, Level 2 standard deviation = 1.5, The combined ICC =0.4316.

Case 6: Level 3 standard deviation = 0.5, Level 2 standard deviation = 2.0, The combined ICC =0.5635.

Case 7: Level 3 standard deviation = 1.0, Level 2 standard deviation = 1.0, The combined ICC =0.3779.

Case 8: Level 3 standard deviation = 1.0, Level 2 standard deviation = 1.5, The combined ICC =0.4968.

Case 9: Level 3 standard deviation = 1.0, Level 2 standard deviation = 2.0, The combined ICC =0.6029.

3.5. Interpretation of results

The previous sub-sections extensively discussed the simulation study carried out in determining the properties of the new goodness-of-fit test. The properties studied were the power and the type I error. Under each property, extensive simulations (72 each) were carried out by varying cluster sizes at the three levels of the hierarchy and for the differing combined ICC values with different standard deviations at the two higher levels. The results were presented in tabular form. Overall, it is seen that for the developed

Table 3. The TOE and power for the unbalanced cases.

Level 3 Standard Deviation	Level 2 Standard Deviation	Level 3 size and break down		Level 2 size	Level 1 size	Rejection proportion under H_o	Rejection proportion under H_1	Status under H_o
0.4	1.0	10	1 … 3	50	50	0.049	1.000	Within limits
			4 … 6	30	50			
			7 … 10	50	30			
0.4	1.0	15	1 … 5	15	30	0.058	0.987	Within limits
			4 … 6	10	30			
			7 … 10	15	20			
0.4	1.0	10	1 … 3	15	30	0.043	0.921	Within limits
			4 … 6	15	20			
			7 … 10	20	30			
0.4	2.0	10	1 … 3	15	30	0.057	0.933	Within limits
			4 … 6	15	20			
			7 … 10	20	30			
0.4	2.0	10	1 … 3	20	20	0.064	0.646	Within limits
			4 … 6	15	20			
			7 … 10	20	10			
0.5	2.0	10	1 … 3	15	30	0.063	0.934	Within limits
			4 … 6	15	20			
			7 … 10	20	30			
0.5	2.0	10	1 … 3	20	20	0.069	0.662	Outside limits
			4 … 6	15	20			
			7 … 10	20	10			
1.0	1.5	10	1 … 3	15	30	0.043	0.945	Within limits
			4 … 6	15	20			
			7 … 10	20	30			
1.0	1.5	10	1 … 3	20	20	0.059	0.655	Within limits
			4 … 6	15	20			
			7 … 10	20	10			

goodness-of-fit test the type I error holds for moderate to large samples (total cells = 6750, 7500, 9000, 11250, 13500, 15000, 22500) and is somewhat inflated for small samples (total cells = 4500).

Except in case 1, the type I error in cluster combination eight (which is the smallest cluster size (4500)) is inflated in the other cases. In all other combinations of all other cases the type I error holds. Cai et al. (2006) have looked at limited-information goodness-of-fit testing of item response theory models for sparse 2^P tables. Though they have not used the Hosmer and Lemeshow test followed by the joint Wald statistic, they have tested many other statistics which are considered to be more modern for this small sample case. These are YBL1-YBL3 the first, second and third-moment adjustments to the quadratic form in the bivariate residuals; YCL1-YCL3 the first, second and third-moment adjustments to ours using corrected moments that take Maximum Likelihood (ML) estimation into account; Y2CL1-Y2CL3 first, second and third-moment adjustments to the quadratic form in univariate and bivariate residuals. They have also used two full information statistics, G^2 and X^2. They have performed simulations for a total sample size of 4000 which is close to our smallest sample case of 4500. These results are given in Table 3 of their article. For $\alpha = 5\% = 0.05$ YBL1-YBL3 and G^2 gives, conservative type I errors and YCL1-YCL3, Y2CL1-Y2CL3 and X^2 gives inflated type I errors. Our statistic uses univariate and binary residuals so it is somewhat similar to the Y2CL's. So it seems that this explains the failure of type I error for the small sample size.

For the alternative hypothesis considered, all combinations give very high power. The method of estimation used was Marginal Quasi-Likelihood with first-order Taylor series approximation (MQL1), as the other methods of estimation provide convergence problems. Though MQL 1 gives biased estimates in small samples our samples are assumed to be large enough for this bias to be minimal.

As the total sample size used in our simulations is rather large, the reader may speculate whether it is practically useful. We have come across the following studies which have a very large sample size. There is sure to be a lot more cases in practice.

In de Jong, Moerbeek, and van der Leeden (2010), where real multidimensional research data sets are discussed, a study total sample size of 13,112 has been used. Also, Lipsitz et al. (1996) in her book on "Changes in reading comprehension across cultures and over time" (Section 10.4) discusses two examples of a three-level study where the total sample sizes are 25,611 and 24,505 respectively. Three-level data examples also appear in "Big Data". (https://www.crowdflower.com/three-levels-of-big-data/. Retrieved on 10th June 2017). Here level 3 is up to 20,000 rows, level two is up to 200,000 rows, and level one is up to 2,000,000 rows. Therefore, very large sample sizes have been used in the past for 3-level data structures. Thus, our approach does have its advantages even though the sample sizes used are reasonably large.

3.6. The unbalanced case

Another speculation on the part of the reader would be whether our test statistic would work well in the unbalanced data scenario, though we have shown that it works well for the balanced scenario when the sample size is reasonably large. Why we initially tried out only the balanced case was because unlike in the balanced case, the unbalanced case cannot be automated for simulation. Thus, it is a very tedious task to carry out 1000 simulations at a time manually. However, for completeness purposes we selected a few combinations of some cases and performed the type I error and power simulations manually for the unbalanced case. This is given in Tables 2(a–i).

These tables show that the unbalanced case gives a better type I error, and slightly less power as the balanced case for similar sample sizes and the type I error is tolerable even for smaller sample sizes. Chatterjee, Chakraborty, and Chowdhury (2019) show in their results reported in Table 5 of their article that the properties of the unbalanced case can be better than in the balanced case when the unbalanced Bayesian D-optimal in the design exceeds that of the balanced design. Also, the relaxing of the balance constraint allows the Bayesian D-optimal in the design to improve on the "E(s^2)-optimal" value and this is also addressed in Jones, Lin, and Nachtsheim (2008).

4. Application to a practical example

4.1. Description

Under this section a real-life dataset was analyzed, using the MLwiN software version 2.19. Due to the lack of binary response datasets that are freely available on higher level models, and also as the available datasets are not given on the internet by the owners, it was not possible to obtain a dataset satisfying the conditions at hand. Thus, a dataset

Table 4. Description of variable.

Variable name	Description	Type	Reference category
year	Year in which the data was collected for a particular location	categorical	Year_1
month	Year in which the data was collected for a particular location	categorical	Month_1
Source	Source type of water denoting where the samples are obtained from (coded 1 to represent running water and 2 to represent standing water)	categorical	Source_1
Rain	Mean monthly rainfall	continuous	–
Temp	Mean monthly air temperature	continuous	–
Humidity	Mean monthly humidity	continuous	–

from a research study by an undergraduate in the year 2015 has been used in for the application of the goodness-of-fit test. It has to be noted that this dataset has been used only for illustration purposes.

The data set consisted of information regarding water quality data collected from all over Sri Lanka. The response variable in the original dataset was the chemical component of water quality, which was initially of continuous scale, and was coded under this study into a binary variable such that the values less than the median (0.66795) are coded 0 and those that are greater than the median are coded 1. Then the developed test was applied to this dataset. The three levels were the location in which the data were collected which is the 1st level, the district into which the location belongs is the 2nd level, and the province into which the district belongs is the 3rd level that composes of the hierarchy. There are four provinces, nine districts and 1448 locations from which the samples have been collected. All the explanatory variables have been measured at the lowest level of the hierarchy. Table 4 represents the explanatory variables under consideration in the dataset.

4.2. Fitting the model

For fitting the model, it is necessary to identify which of the explanatory variables affect the response of interest. Forward selection was used for variable selection, initially starting with the constant term only. A significance level of 5% is used throughout the model fitting stage. For determining the variables that significantly affect the response, Wald statistic and the corresponding p-values are used. It should be noted under this case that, Wald statistic calculated for each parameter has a chi-square distribution with a degree of freedom equal to 1. Parameter estimation was done using the 1st order MQL method using MLwiN version 2.19.

The final model fitted, considering parsimony is as follows.

$$\mathrm{logit}(\pi_{ijk}) = \beta_{0jk}const - 0.415(0.175)source_{2_{ijk}} \tag{9}$$

where $\beta_{0jk} = 0.634(0.736) + v_{0k} + u_{0jk}$

4.3. Carrying out the goodness-of-fit test

The hypothesis of interest can be denoted as H_0: The model Eq. (9) fits the data well, and the alternative hypothesis can be denoted as H_1: The model Eq. (9) does not fit the data well.

For checking whether the model fits the data well, a model is created using pooled indicator variables and this can be denoted as follows.

$$\mathrm{logit}(\pi_{ijk}) = \beta_{0jk} const - \beta_1 source2_{ijk} + \sum_{g=2}^{10} \gamma_g I_{g_{ijk}}. \tag{10}$$

For the new model with pooled indicators, the joint Wald statistic is calculated to determine the fit of the model. The joint Wald statistic obtained for the model is 10.062 on 9 degrees of freedom. This test corresponds to a p-value of 0.345. Thus the coefficients of the indicator variables are simultaneously zero, and the original model fits the data well. Hence the fitted model adequately describes the data according to the goodness-of-fit test.

4.4. A comparative study considering the same example using a a two-level model

For the same example, a two-level model was fitted for comparison purposes. Forward selection of variables was used, initially starting with the constant term only for fitting the model at a significance level of 5%. To determine the variables that significantly affect the response, Wald statistic and the corresponding p-values were used. The 1st order MQL method was used for parameter estimation.

The final model fitted, considering parsimony is as follows.

$$\mathrm{logit}(\pi_{ij}) = \beta_{0j} const - 1.335(0.147) source2_{ij} + 0.057(0.014) Humidity\ _{i}\mathrm{s} \tag{11}$$

Where $\beta_{0j} = -4.348(1.316) + U_{0j}$

Comparing Eq. (9) with Eq. (11) suggests that if the same data is considered as a two-level study, for simplicity, with the first level as location and the second level as province ignoring the district, the fitted model would be a different one.

A goodness-of-fit test was carried out for the model denoted by Eq. (11), considering a two-level model with the null hypothesis as H_0: The model Eq. (11) fits the data well, and the alternative hypothesis can be denoted as H_1: The model Eq. (11) does not fit the data well. Similar to the case in the three-level model to check whether the model fits the data well, a model was created using pooled indicator variables as follows.

$$\mathrm{logit}(\pi_{ij}) = \beta_{0j} const - \beta_1 source2_{ij} + \beta_2 Humidity_{ij} + \sum_{g=2}^{10} \gamma_g I_{g_{ij}}. \tag{12}$$

The joint Wald statistic obtained for the model was 23.845 on 9 degrees of freedom. This test corresponds to a p-value of 0.0046. Hence the coefficient of the indicator variables are not simultaneously zero, and indicates that the fitted model given in Eq. (11) should be rejected and concluded that it does not fit the data well according to the goodness-of-fit test. These results are contrast with the conclusion under Sec. 5.3. This indicates that when the intermediary levels are ignored a wrong model which does not fit the data is obtained. Further, this shows the importance of using limited-information methods instead of ignoring intermediary levels.

5. Discussion and conclusion

5.1. General discussion

The main objective of this research was to develop a new goodness-of-fit test based on the Hosmer and Lemeshow goodness-of-fit test for the single-level binary model. For achieving the stated objective, the theory behind the goodness-of-fit testing for ordinal data for a single-level by Lipsitz et al. (1996), the goodness of the fit testing method developed by Perera et al. (2016), and limited fit testing concepts for goodness-of-fit testing introduced by Maydeu-Olivares and Joe (2006), Maydeu-Olivares and GarcıA-Forero (2010) and Maydeu-Olivares et al. (2011) were incorporated and extended for high level binary multilevel data. As the secondary objective, the properties of the new goodness-of-fit test were studied using simulations. What is new in the current study is the introduction and adoption of the concepts of limited-information goodness-of-fit which were introduced by Maydeu-Olivares and Joe (2006), Maydeu-Olivares and GarcıÁ-Forero (2010) and Maydeu-Olivares et al. (2011). No previous study has incorporated the theories of limited-information goodness-of-fit testing in multilevel modeling.

5.2. Comparison of the newly proposed goodness-of-fit test and the previously developed goodness-of-fit test of *Perera et al. (2016)*

For the topic of multilevel modeling, due to unavailability of a satisfactory goodness-of-fit test for assessing the model adequacy of binary response models, Perera et al. (2016) extended the theory behind the Hosmer and Lemeshow test for the single-level data to multilevel hierarchies. However, this test has been developed considering only a two-level scenario and not thinking about higher level scenarios. Cool et al. (2015) stressed the need for goodness-of-fit tests for such high level multilevel models. Under this research study, the test developed by Perera et al. (2016) has been extended to high level binary multilevel models. For demonstrating this case, a three-level model was used. It should be noted now that the structure of the data is different from a two-level multilevel model. Hence a new approach had to be thought of, to extend the theory of Hosmer and Lemeshow test for the study at hand. It should now be taken into consideration that the highest level units (big clusters) in this case, the third level units are independent of or non-correlated to each other. However, the small clusters (second level units) which are inside the big clusters are correlated, and so are the first level units that in turn lie inside the small clusters. The main difference of the previously developed goodness-of-fittest for the two-level study and the current study for a higher level multilevel model is that the goodness-of-fit statistic suggested under the former research uses full information at both levels in coming up with the goodness-of-fit test. The current research suggests a method based on the limited-information goodness-of-fit testing according to the suggestions made by Maydeu-Olivares et al. The way of calculating the fitted probabilities, the method of sorting the fitted probabilities and ranking these by giving indicator variables are different from what has been proposed by the former research study. Several major changes have been incorporated to that of the previous research.

5.3. Computational complexity and the stability of the proposed goodness-of-fit test

Regarding the computational complexity of the proposed test for the balanced case, it should be noted that the results from simulations for a very large number of 3rd level clusters, are computationally intensive and not so stable as some non-converging problems were found. Also when the cluster combinations are unbalanced, the simulation study was computationally very tedious, and the automation of this type of simulations is very difficult. Both balanced and unbalanced case involves heavy computations; however, powerful computers will find it not so time-consuming.

5.4. Conclusions from the study

The developed test is simple and not difficult to understand. The type I error holds for the developed goodness-of-fit test under the considered cluster combinations, except for the case when the cluster sizes for all the three levels are at their minimum. The power of the developed test is very high under the alternative model considered for data generation. The sample sizes being higher may be an added advantage for the power to be very high. It is important to maintain a higher number of cluster sizes at all levels of the hierarchy to obtain precise results from the novel goodness-of-fit test. The developed test can be used for datasets with unequal cluster sizes without a problem. The explanatory variables used can be either continuous or discrete, and also these can be measured at any level of the hierarchy. Suggestions are given for applying the developed test for high level binary multilevel model.

References

Archer, K. J., and S. Lemeshow. 2006. Goodness-of-fit test for a logistic regression model fitted using survey sample data. *The Stata Journal: Promoting Communications on Statistics and Stata* 6 (1):97–105. doi:10.1177/1536867X0600600106.

Archer, K. J., S. Lemeshow, and D. Hosmer. 2007. Goodness-of-fit test for logistic regression models when data are collected using a complex sample design. *Computational Statistics & Data Analysis* 51:4450–64. doi:10.1016/j.csda.2006.07.006.

Cai, L., and M. Hansen. 2013. Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology* 66 (2):245–76. doi:10.1111/j.2044-8317.2012.02050.x.

Cai, L., A. Maydeu-Olivares, D. L. Coffman, and D. Thissen. 2006. Limited-information goodness-of-fit testing of item response theory models for sparse 2○P tables. *British Journal of Mathematical and Statistical Psychology* 59 (1):173–94. doi:10.1348/000711005X66419.

Chatterjee, T., S. Chakraborty, and R. Chowdhury. 2019. A critical review of surrogate assisted robust design optimization. *Archives of Computational Methods in Engineering* 26 (1):245–74.

Cool, G., A. Lebel, R. Sadiq, and M. J. Rodriguez. 2015. Modelling the regional variability of the probability of high trihalomethane occurence in municipal drinking water. *Environment Monitoring Assess* 187 (12):746.

de Jong, K., M. Moerbeek, and R. van der Leeden. 2010. A priori power analysis in longitudinal three-level multilevel models: An example with therapist effects. Psychotherapy Research 20 (3):273–84. doi:10.1080/10503300903376320.

Epasinghe, N, and R. Sooriyarachchi. 2017. A goodness of fit test for the multilevel proportional odds model. Communications in Statistics – Simulation and Computation. 46 (7):5610–5626. doi:10.1080/03610918.2016.1169293.

Fotouhi, A. R. 2003. Comparisons of estimation procedures for nonlinear multilevel models. *Journal of Statistical Software* 8 (9):8984.

Gregorich, S. 2013. Models of binary outcomes with 3-level data: A comparison of some options within SAS. CAPS Method Core Seminar, April 19. https://prevention.ucsf.edu/sites/prevention.ucsf.edu/files/uploads/2011/02/130419g-SLIDES.pdf.

Hosmer, D. W., and S. Lemeshow. 2000. *Applied logistic regression*. 2nd ed. New York: John Wiley & Sons, Inc.

Hosmer, D. W., T. Hosmer, S. L. Cessie, and S. Lemeshow. 1997. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine* 16 (9):965–80. doi:10.1002/(SICI)1097-0258(19970515)16:9<965::AID-SIM509>3.3.CO;2-F.

Joe, H. and A. Maydeu-Olivares. 2010. A General Family of Limited Information Goodness-of-Fit Statistics for Multinomial Data. *Psychometrika* 75 (3):393–419.

Jones, B., D. K. J. Lin, and C. J. Nachtsheim. 2008. Bayesian D-optimal supersaturated designs. *Journal of Statistical Planning and Inference* 138:86–92.

Liao, T. 2004. Comparing social groups: Wald statistics for testing equality among multiple logit. *International Journal of Comparative Sociology* 45 (1-2):3–16. doi:10.1177/0020715204048308.

Lipsitz, S. R., G. M. Fitzmaurice, and G. Molenberghs. 1996. Goodness-of-fit tests for ordinal response regression models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 45 (2):175–90. http://www.jstor.org/stable/2986153. doi:10.2307/2986153.

Maydeu-Olivares, A., and C. GarcíA-Forero. 2010. Goodness-of-fit testing. *International Encyclopedia of Education* 7:190–6.

Maydeu-Olivares, A., and H. Joe. 2006. Limited information goodness-of- fit testing in multidimensional contingency table. *Psychometrika* 71:713.

Maydeu-Olivares, A., L. Cai, and A. Hernández. 2011. Comparing the fit of item response theory and factor analysis models. *Structural Equation Modeling: A Multidisciplinary Journal* 18 (3):333–56. doi:10.1080/10705511.2011.581993.

Perera, A. A., M. R. Sooriyarachchi, and S. L. Wickramasuriya. 2016. A goodness of fit test for the multilevel logistic. *Communications in Statistics - Simulation and Computation* 45 (2):643–59. doi:10.1080/03610918.2013.868906.

Rosner, B., R. J. Glynn, and M.-L. T. Lee. 2003. Incorporation of clustering effects for the Wilcoxon rank sum test: A large-sample approach. *Biometrics* 59 (4):1089–98. doi:10.1111/j.0006-341X.2003.00125.x.

Samejima, F. 1969. *Estimation of Latent Ability Using a Response Pattern of Graded Scores (Psychometric Monograph No. 17)*. Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf.