International Journal on Advanced Science Engineering Information Technology

Vol.11 (2021) No. 5
ISSN: 2088-5334

A Latent Class Model for Multivariate Binary Data Subject to Missingness

Samah Zakaria [a,*], Mai Sherif Hafez [a], Ahmed M. Gad [a]

[a] *Department of Statistics, Cairo University, 1 University* Street, *12613, Giza, Egypt*
Corresponding author: [*]samahzkaria@feps.edu.eg

***Abstract*— When researchers are interested in measuring social phenomena that cannot be measured using a single variable, the appropriate statistical tool to be used is a latent variable model. A number of manifest variables is used to define the latent phenomenon. The manifest variables may be incomplete due to different forms of non-response that may or may not be random. In such cases, especially when the missingness is nonignorable, it is inevitable to include a missingness mechanism in the model to obtain valid estimates for parameters. In social surveys, categorical items can be considered the most common type of variable. We thus propose a latent class model where two categorical latent variables are defined; one represents the latent phenomenon of interest, and another represents a respondent's propensity to respond to survey items. All manifest items are considered to be categorical. The proposed model incorporates a missingness mechanism that accounts for forms of missingness that may not be random by allowing the latent response propensity class to depend on the latent phenomenon under consideration, given a set of covariates. The Expectation-Maximization (EM) algorithm is used for estimating the proposed model. The proposed model is used to analyze data from 2014 Egyptian Demographic and Health Survey (EDHS14). Missing data is artificially created in order to study results under the three types of missingness: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).**

***Keywords*— Binary variables; latent class model; item non-response; non-random missingness; response propensity.**

Manuscript received 9 Apr. 2021; revised 13 Jun. 2021; accepted 6 Jul. 2021. Date of publication 31 Oct. 2021.

I. INTRODUCTION

In many social science applications, the main interest is to measure constructs or concepts, such as behavior or abilities, which cannot be measured using a single observable variable. These are known as latent (unobserved) variables and can be measured through a set of manifests (observed) variables using what is known in Statistics as latent variable models. Observed variables can be of any type, and latent variables can be assumed to be either categorical or continuous, depending on the nature of the problem. This results in different classifications of latent variable models. Verbeke and Molenberghs [1] has given general overview of latent variable models and their inference. Network models provide an alternative to latent variable models [2].

Categorical latent variables are usually assumed when there is a reason to believe that a particular phenomenon is inherently categorical, justified by prior evidence or theory which leads to latent categories, or when it would be practically useful to have such categories; for example, to organize respondents into a number of relevant subgroups [3]. When the observed items are categorical, latent class analysis has been adopted by many authors to achieve this objective. Some previous studies also concern behavioral and psychometric fields that employ different versions of this type of model [4]–[10]. The covariates and direct effects within latent class models are also discussed in Janssen *et al.* [11] and Bakk and Kuha [12]. Bakk and Kuha [13] has fitted a latent class model with structural regression models for the relationships between the latent classes and observed manifest variables and covariates.

Item non-response is a common type of missing data, especially in surveys, in which a respondent may provide answers to some of the variables but not the others leading to different patterns of incomplete data. Despite trials to reduce non-response, such as probing "Don't Know" answers [14], almost all surveys still suffer from missing data due to non-response. Little and Rubin [15] has classified missing data in general into three types: MCAR, where missingness neither depends on observed data nor on data that is missing; MAR, where missingness is independent of missing data conditional on the observed data; or MNAR, where missingness may depend on the missing values, and possibly the observed data

too. When data is MCAR, the observed data may be considered as a random subset of the full data. If data is MAR, it can still be considered as a random subset defined for specific values of the observed data. In such cases, the missingness is labeled as "ignorable missingness". When data is MNAR, the missingness is related to the value that was not observed itself, reflecting a systematic difference between respondents and nonrespondents. Hence, it is said that this type of missingness is "nonignorable" or "informative." Incorporating a missingness mechanism is crucial in this case to avoid biasedness in the estimation of model parameters.

There are various approaches to incorporate a missing nonignorable mechanism, mostly developed for longitudinal data. These include the selection approach and the pattern-mixture approach. Selection models factorize the joint distribution of observed and missing responses into a marginal distribution for the full data multiplied by the conditional distribution of the missing data given the full data. On the other hand, the pattern-mixture approach specifies the marginal distribution for the missing responses and the distribution of the complete data conditional on the missing responses. Du *et al.* [16] has proposed imputing data that is MNAR using a latent variable approach and fit it within a Bayesian framework.

For multiple observed variables, latent response propensities have been employed by many authors to account for missingness in data. Rose *et al.* [17] and Cursio *et al.* [18] are among the most recent publications that use this concept. The main idea is to create a binary indicator variable corresponding to each manifest item of those measuring the latent variable of interest that indicates whether this manifest item is observed or missing for each subject. The number of those created binary variables is thus the same as the number of manifest items. A latent variable of the phenomenon of interest is measured by a number of binary manifest variables that may include some missing values. Another latent variable labeled as response propensity as measured by a set of binary items that are created to indicate whether a value is observed or missing for each of the manifest variables. Both latent variables are assumed to be continuous.

Models, where the response propensity is a categorical latent variable have received less attention. Jung [19] has used a categorical response propensity variable to gather with the joint distribution of the observed items. Harel and Schafer [20] has proposed using latent class models that deal with partially ignorable missingness by fitting a latent class model that includes binary missingness indicators as additional items. A possible criticism of this specification is that the latent class variable is a summary for both the main observed items and the response propensity, thus possibly changing the meaning of the latent variable itself. Kuha *et al.* [21] has proposed models for survey data that contain non-response, assuming the main latent variable to be continuous and the response propensity latent variable to be categorical. Bacci and Bartolucci [22] has defined similar models assuming both latent variables to be categorical. It also assumes that the two latent variables are independent conditional on a set of covariates. The non-response model may be dependent on one or both latent variables. Sterba [23] has presented a shared parameter latent transition analysis, assuming categorical latent variables, in the case of longitudinal data.

This article considers models where the latent variable of interest may affect the probability of non-response through another latent variable that summarizes response propensity. The response propensity determines the non-response probabilities is affected by the main latent variable in the structural part of the model. The non-response is nonignorable if the response propensity is associated with the main latent variable and ignorable otherwise. We propose a latent class model that considers binary manifest variables subject to non-response, assuming categorical latent variables, both for the main latent variable and used to measure response propensity. Unlike Bacci and Bartolucci [22], the two latent variables are related by allowing members of the latent class to affect the probability of response. This means that the missingness mechanism may be nonignorable. However, the model allows the two latent variables to be affected by covariates. To illustrate the proposed model, data from the EDHS14 is analyzed [24]. Missingness is artificially created under three scenarios: MCAR, MAR, and MNAR. The aim is to study how the results of the model may change according to the type of missingness.

Section (II) of this article presents the specification of the proposed latent class model (LCM), the estimation process for the LCM parameters, and the methods from the literature for model selection and fit evaluation. Discussion and results of the proposed model using a real data set appear in Section (III). Concluding remarks appear at the end of the article in Section (IV).

II. Materials and Method

A. Research Methodology

The general outline for systematic stages to conduct such research is as follows. In the initial stage of the research, the research problem is formulated. This involves an attempt to measure an unobserved phenomenon of interest using a number of observed binary variables (items). In many cases, the observed items will have some missing values. The first step is to select the most appropriate items for measuring the latent variable, depending on the suitable selection criteria given in the next section. The implementation stage then starts by creating a missingness pseudo-item corresponding to each of the original selected items to indicate whether a value is missing or not. The proposed latent class model, outlined in the next subsection, is then ready to be implemented. It assumes two categorical latent variables, one to summarize the main phenomenon of interest and the other to summarize response propensity. One of the main contributions of this model is that it allows the two latent variables to be dependent, thus allowing for missingness to be not at random. Estimation of this model is then carried out before moving to the final stages of the research, where the model is evaluated and models having different numbers of classes are compared. The best-fitting model concerning fit, and interpretability is then selected, and its results are interpreted. In our study, we create artificial missingness to evaluate the model's performance under different types of missingness. Fig. 1 is a flowchart representation of the outlined research methodology.

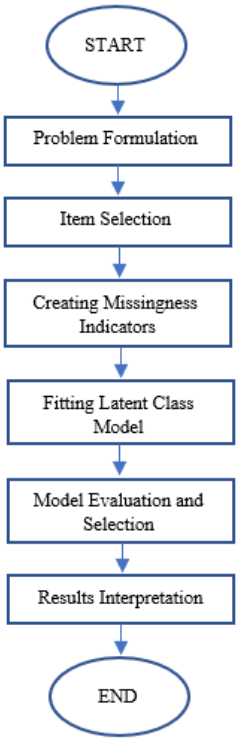

Fig. 2 Research Methodology Flowchart

B. Model Specification

Finite mixture models are models in which categorical latent variables are made of classes where class membership is inferred from the data. A special case is latent class analysis, where the latent classes explain relationships among the manifest items. The model proposed here considers the case where all manifest variables are binary. The latent variable used to summarize the manifest variables is assumed to be categorical too. A missingness mechanism to account for item non-response is incorporated. This involves another categorical latent variable to measure a respondent's propensity to respond based on binary indicators representing whether a respondent has given an answer to each observed item. The latent variable of interest is allowed to affect response propensity makes the missingness possibly non-random. It is thus assumed that an individual's probability of responding to items depends on their response propensity, observed covariates, and possibly their class membership of the main latent variable.

Fig. 2 is a path diagram representation of the proposed model. The main latent variable is denoted by z_a, while that representing response propensity is denoted by z_r. There are p binary manifest variables, each denoted by y_i. The vector $\mathbf{x}$ represents a number of observed covariates that may influence the main latent variable, while $\mathbf{w}$ another vector of covariates may be the same or different from $\mathbf{x}$, influencing the response propensity latent variable. An indicator variable r_i takes value 1 if the manifest variable y_i is observed and 0 if it is missing.

An LCM consists of two main parts known as the measurement and structural parts. In the case of our proposed model, a third part is added to incorporate the missingness mechanism.

1) Measurement model: The measurement part for an LCM is a multivariate regression model where the dependent variables are the manifest variables, and the categorical latent variables represent the independent variables. When manifest variables are binary, logistic regression equations are used to model the relationships between the observed and latent variables.

Let $\mathbf{y}$ denote a vector of p binary manifest variables and $\pi_{ai}(z_a)$ be the probability of a positive response on the variable y_i for an individual in each category of the latent class $(i = 1, 2, \dots, p)$. The latent classes for the latent variable of interest z_a are mutually exclusive and exhaustive. Each respondent belongs to one and only one latent class. Each observed binary variable follows a Bernoulli distribution. The probability of responding to variable y_i positively can thus be presented as,

$$\text{logit}\,\pi_{ai}(z_a) = \alpha_{i0} + \alpha_{ia}\, z_a, \qquad (1)$$

where

$$\pi_{ai}(z_a) = P(y_i = 1 \mid z_a).$$

2) Missingness mechanism: To include the missingness mechanism within the proposed model, a random indicator variable for the missingness is defined for each observed item. For the i^{th} observed variable of the m^{th} individual, a random indicator variable r_{mi} is defined as

$$r_{mi} = \begin{cases} 1; & y_{mi}\ observed \\ 0; & y_{mi}\ missing. \end{cases} \qquad (2)$$

Let $\mathbf{r}$ denote a vector of p indicator variables and $\pi_{ri}(z_r)$ be the probability that variable y_i is observed $(r_i = 1)$ for a respondent, given their membership to the latent class categories $(i = 1, 2, \dots, p)$. The latent classes for the latent variable of response propensity z_r are also mutually exclusive and exhaustive.

Each of the response/non-response indicator variables follows a Bernoulli distribution. The probability that a manifest variable y_i is not missing $(r_i = 1)$ can thus be modeled as,

$$\text{logit}\,\pi_{ri}(z_r) = \upsilon_{i0} + \upsilon_{ir}\, z_r, \qquad (3)$$

where

$$\pi_{ri}(z_r) = P(r_i = 1 \mid z_r).$$

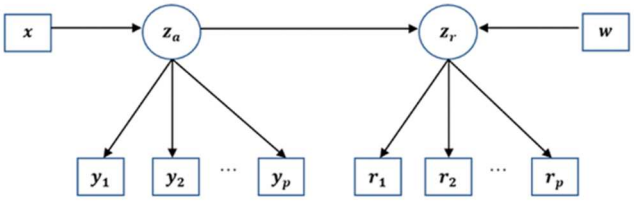

Fig. 2 Path diagram for a latent variable model with one main latent variable and another for propensity to respond

3) Structural model: Relationships among latent variables and possibly covariates too, if they exist, are outlined within the structural part of the latent variable model. Both latent variables z_a and z_r are assumed to be binary, each of them having a Bernoulli distribution. Logistic regression equations are used to model these relationships. According to the model specification in Fig. 2, the structural model will be given by

$$\text{logit}\,\pi_{z_a}(x) = \alpha_{a0} + \sum_{h=1}^{H} \beta_h x_h, \qquad (4)$$

where $\pi_{z_a}(x) = P(z_a = 1 \mid x)$ is the probability of being a member of the first class of the latent variable z_a given a set of H observed covariates $\mathbf{x}$ affecting z_a, and

$$\text{logit}\,\pi_{z_r}(z_a, w) = \alpha_{r0} + \phi z_a + \sum_{l=1}^{L} \gamma_l w_l, \quad (5)$$

where $\pi_{z_r}(z_a, w) = P(z_r = 1 \mid z_a, w)$ is the probability that an individual is a member of the first class of a latent variable for response propensity z_r given their class membership of the main latent variable z_a , and a number of L observed covariates $\mathbf{w}$ affecting z_r. If the regression coefficient ϕ is significant, this can reflect that the non-response is non-random, since the probability of non-response will be associated with certain levels of the main latent variable, and hence including a missingness mechanism is crucial.

C. Model Estimation

Estimating the LCM involves estimation of parameters determining the probability of latent class membership, represented in equations (4) and (5); in addition to parameters determining item-response probabilities, conditional on latent class membership. The latter parameters are present in equations (1) and (3).

The loglikelihood for a random sample of size n is given by

$$L = \sum_{m=1}^{n} \log \{f(\mathbf{y}_m, \mathbf{r}_m)\}. \quad (6)$$

According to the model specified by equations (1), (3), (4) and (5), the joint distribution of all observed items is given by

$$f(\mathbf{y}_m, \mathbf{r}_m) = \sum_{z_a} \sum_{z_r} g(\mathbf{y}_m \mid z_a)\; g(\mathbf{r}_m \mid z_r)\, h(z_a, z_r \mid \mathbf{x}, \mathbf{w}), \quad (7)$$

where $\mathbf{y}_m$ and $\mathbf{r}_m$ represent the $2*p$ manifest variables for the m^{th} respondent. The conditional distribution of $y_m \mid z_a$ is

$$g(\boldsymbol{y}_m \mid z_a) = \prod_{i=1}^{p} [\pi_{ai}(z_a)]^{y_{im}}\; [1 - \pi_{ai}(z_a)]^{1-y_{im}}, \quad (8)$$

while the conditional distribution of $r_m \mid z_r$ is

$$g(r_m \mid z_r) = \prod_{i=1}^{p} [\pi_{ri}(z_r)]^{r_{im}}\; [1-\pi_{ri}(z_r)]^{1-r_{im}}. \quad (9)$$

The joint distribution of z_a and z_r can be factorized as

$$h(z_a, z_r \mid x, w) = h(z_r|z_a, w)\, h(z_a|x), \quad (10)$$

where a Bernoulli distribution is assumed for the main latent variable conditional on covariates $h(z_a|\mathbf{x})$and the response latent variable conditional on the main latent variable z_a and covariates $h(z_r|z_a, \mathbf{w})$. For estimating the outlined model, a given response to an observed item is weighted by the probability of responding to this item. This probability is a direct function of response propensity and is indirectly affected by class membership of the main latent variable through the response propensity latent variable.

Collins and Lanza [25] has shown that model parameters cannot be estimated in closed form in this case. These parameters are estimated from the data (for a given number of classes) using the EM algorithm, combined with another iterative algorithm, such as Newton-Raphson. These algorithms attempt to maximize the likelihood function, thus obtaining maximum likelihood (ML) parameter estimates. In our application, estimation is carried out in M*plus* [26].

D. Model Selection and Fit

This section addresses two practical issues when fitting an LCM: selecting the relevant items to measure a latent class variable and determining the appropriate number of latent classes. Selecting items for latent class analysis is crucial to help interpretability of the latent variable and the model. In general, classification performance and precision of parameter estimates are better for more parsimonious models.

Collins and Lanza [25] has shwon that two aspects are characterized by a strong relation between each manifest item and a latent variable. The first aspect is how the item-response probabilities for manifest item y_i vary across the latent classes. The second aspect is whether the item-response probabilities corresponding to the observed variable y_i are close to 1 or 0. Real data can be checked if the item-response probabilities are close to 0 or 1, as it is not common to find item-response probabilities that are exactly 0 or 1.

Goodman [27] has discussed the identifiability of a latent class model given a specific number of classes for a given number of variables. For K classes and p observed binary variables, the following condition needs to be satisfied for the model to be identified

$$2^p > (p + 1) \times K. \quad (11)$$

This assures that there is enough information for estimation of model parameters. However, in practice, the available data may be not sufficient to estimate the model parameters. Xu [28] has resolved identifiability issues for a restricted family of latent class models with binary manifest items. Deciding on the appropriate number of classes for a model usually involves comparing models with different classes (e.g., 2, 3, and 4 latent classes) and selecting the model that gives the best fit and most interpretability. There is no common standard for the best fit criterion, and researchers often use a number of fit criteria in selecting the appropriate number of latent classes. Reference Koo and Kim [29] fits a latent class model for longitudinal data, allowing the number of latent classes to be determined from the data based on a Bayesian estimation method. Tein *et al.* [30] stated that methods for determining the number of classes are classified into three categories: likelihood ratio statistical test methods, information-theoretic methods, and entropy-based criterion. In the data analysis, we use these methods for selecting models with the best fit.

III. Results and Discussion

This paper applies the proposed LCM incorporating a missingness mechanism to analyze data from the 2014 Egyptian Demographic and Health Survey (EDHS14) about people's access to knowledge sources. The EDHS14 consisted of two questionnaires, one for the household and another for individuals. The household questionnaire included social and economic questions. Out of 29,471 households sampled for the EDHS14, 28,630 households were found, and a response rate among those was 98.4 percent.

A. Access to Knowledge Sources: Measurement Model

The Bristol definition for information (knowledge sources) deprivation, based on the ''deprivation approach'' to poverty [31], was originally developed for children between $2 - 18$

years old. It defines children without access to the following media: radio, television, telephone (landline or mobile phone), computer or newspapers at home as "information deprived". We propose a more flexible definition based on a latent variable approach that, unlike the original definition, provides a more flexible concept of deprivation. The latent class model does not merely consider an individual as "not deprived" if they have any of those devices but classifies subjects according to the measurement model that gives different weights to different items in measuring the latent variable.

The following eight items are measured on the household level, indicating whether certain sources of knowledge are available for all household members or not. We use these items to measure a latent variable that is interpreted as "Access to Knowledge Sources".

1- Does your household have a radio? (Radio)
2- Does your household have a television? (Television)
3- Does your household have a telephone (land line)? (Telephone)
4- Does your household have a mobile phone? (Mobile)
5- Does your household have a computer? (Computer)
6- Does your household have a video/ T.V player? (Video)
7- Does your household have a smart phone? (Smart phone)
8- Does your household have a satellite dish? (Satellite)

All these items are binary variables. Each of them is given the value "1" for a household having this source of knowledge and "0" for a household that does not have that source.

The analysis aims to determine the contribution of each of the eight items to the measurement of access of people to knowledge sources. First, we investigate which of the eight items has the highest contribution in measuring the latent variable. Then, we choose the best items that represent the latent variable well according to the two criteria outlined in Section II-C. Next, the missingness mechanism is incorporated within the model framework. Data analysis is implemented in M*plus* [26].

The results from the model in Equation (1) are presented in Table 1, assuming a two-class latent variable. Among 28,175 of households who were successfully interviewed 27,850 provided complete answers. The analysis is based on those who gave complete answers.

TABLE I
ITEM-RESPONSE CONDITIONAL PROBABILITIES FROM A TWO-CLASS LCM FOR "ACCESS TO KNOWLEDGE SOURCES" LATENT VARIABLE

Item	Probability of a "Yes" response	
	1st class	2nd class
Radio	0.459	0.221
Television	0.998	0.966
Telephone	0.401	0.092
Mobile	1.000	0.861
Computer	0.798	0.096
Video	0.083	0.002
Smartphone	0.570	0.033
Satellite	0.999	0.950

Note: The complement of the above probabilities indicates the probability of responding with a "No" to the corresponding item.

By applying item selection criteria, four items are selected as measures of the latent variable that we label as "Access to Knowledge Sources"; namely, access to radio, telephone (landline), computer, and smartphone. The conditional probabilities of the other four items (television, mobile phone, video, and satellite dish) do not change much whether they belong to the first or second class of the latent variable, as shown in Table 1. This can indicate that they do not have a great contribution in measuring the latent variable and are thus excluded.

According to the identifiability condition in equation (11), the latent variable of interest "Access to Knowledge Sources" model will be identifiable for either two or three classes. By comparing the results of the measurement model with two versus three classes (see Table 2), it is found that the *p*-value for both the Lo-Mendell-Rubin (LMR) likelihood ratio test and the bootstrap likelihood ratio test (BLRT) indicates a good fit in both cases. The AIC, BIC, and entropy-based criteria for model selection are not too far for the two models, although they slightly favor the three-class model. Thus, we resort to ease of interpretability and labeling of the latent classes in determining the number of classes that will be used.

TABLE II
DETERMINING THE NUMBER OF CLASSES OF "ACCESS TO KNOWLEDGE SOURCES."

Selection criterion	Two classes	Three classes
BIC	119066.100	118789.763
AIC	118991.989	118674.479
LMR $\boldsymbol{p}$-value	0.0000	0.0000
BLRT $\boldsymbol{p}$-value	0.0000	0.0000
Entropy	0.699	0.758

Table 3 presents the estimated conditional probabilities for a two-class measurement model versus those of a three-class measurement model, respectively. On examining the estimated probabilities, we decide that the suitable number of classes for our latent variable "Access to Knowledge Sources" is two, as it reflects a clear pattern of probabilities that are higher in the first-class than the second, while no specific pattern can be inferred from the three-class model.

TABLE III
ITEM-RESPONSE CONDITIONAL PROBABILITIES FROM A TWO-CLASS AND A THREE- CLASS LCM FOR "ACCESS TO KNOWLEDGE SOURCES" LATENT VARIABLE

Items	Probability of a "Yes" response				
	Two classes		Three classes		
	1st	2nd	1st	2nd	3rd
Radio	0.499	0.234	0.498	1.000	0.141
Telephone	0.470	0.102	0.463	0.287	0.086
Computer	0.887	0.145	0.960	0.000	0.142
Smart phone	0.656	0.064	0.629	0.164	0.069

Note: The complement of the above probabilities indicates the probability of responding with a "No" to the corresponding item.

B. Response Propensity: Missingness mechanism

The EDHS14 data has a negligible percentage of missingness. Therefore, in order to carry out a comparison of the proposed model under the three types of missingness MCAR, MAR and MNAR, missingness is artificially created within the selected items under the three scenarios. An indicator variable r_{mi} is created to indicate whether a manifest item y_{mi} is observed or not. It is thus given the value "1" if item y_i is observed for an individual m, and the value "0" if it is artificially missing. For MCAR, the missingness is created in a totally random manner. The probability of an individual having a missing value for one of the manifest variables is neither dependent on observed nor unobserved data. This is

achieved by randomly deleting 10% of each item resulting in 34.5% of missingness in the overall data. That is, 34.5% of individuals will have at least one of the four items missing. In case of data MAR, the probability of a missing value for one of the manifest variables is generated as a function of covariates. In this application, wealth index (x_1) and educational level of household head (x_2) are used as covariates for this matter. The probability of a missing response is thus given by

$$\mathrm{P(missing)} = \mathrm{logit}\,(\alpha_0 + \alpha_1 \mathrm{x}_1 + \alpha_2 \mathrm{x}_2). \qquad (12)$$

For each manifest item, a uniform random variable $[0, 1]$ of the same length is generated, such that an observation will be deleted and treated as missing if the corresponding $P(missing) > U_i\,[0, 1],$. The percentage of missingness in each item was approximately 10% resulting in 27.7% missingness in the overall data. This was achieved at $\alpha_0 = 0$, $\alpha_1 = -1$, and $\alpha_2 = 0.1$. It is worth noting that both the choice of covariates and values for parameters α_0, α_1 and α_2 in the equation (12) are arbitrary. In order to obtain data that is MNAR, the missingness has to be generated such that the probability of missing observations depends on the unobserved values themselves. This will lead to a systematic difference between those who respond and those who do not. So, we randomly deleted 10% each item from those who do not have that device (those whose response to the item is 0), resulting in 34% missingness in the overall data.

The four missingness indicators are used as measures of another latent variable that we label as "Response Propensity". The same steps for creating the latent variable "Response Propensity" and selecting the appropriate number of classes are repeated as we have done for the "Access to Knowledge Sources". Again, "Response Propensity" will be identifiable with either two or three classes. By comparing the two-class and the three-class models for the different types of missingness in Table 4, it is decided to go for a latent variable with two classes which also provides the best model fit.

TABLE IV
DETERMININIG THE NUMBER OF CLASSES OF "RESPONSE PROPENSITY" LATENT VARIABLE UNDER THREE TYPES OF MISSINGNESS

	MCAR		MAR		MNAR	
Selection criterion	**Two Class**	**Three Class**	**Two Class**	**Three Class**	**Two Class**	**Three Class**
BIC	72507	72555	67771	67820	72488	72535
AIC	72433	72440	67697	67705	72414	72420
LMR p-value	0.004	0.454	0.000	0.793	0.038	0.053
BLRT p-value	0.000	0.286	0.000	0.667	0.000	0.250
Entropy	0.810	0.879	0.479	0.370	0.222	0.536

C. Overall model

Having fitted each latent class variable "access to Knowledge Sources" and "Response Propensity" separately, the overall proposed model illustrated in Fig. 2 is now presented. Covariates used to affect the latent variable "Access to Knowledge Sources" are sex (male/ female), age in years, place of residence (urban/ rural), educational level, and wealth index of the household head. The wealth index is obtained in five categories: poorest, poorer, middle, richer, and richest. The highest level of education is obtained in six categories: no education, incomplete primary, complete primary, incomplete secondary, complete secondary, and higher. The covariates affecting the missingness latent variable "Response propensity" are sex (male/ female), age, and place of residence (urban/ rural) of the household head. We avoid studying the effect of wealth index and educational level of household head-on "Response Propensity" as these covariates are used in creating the MAR dataset.

TABLE V
PARAMETER ESTIMATES FOR A TWO-CLASS LCM FOR "ACCESS TO KNOWLEDGE SOURCES" AND "RESPONSE PROPENSITY" LATENT VARIABLES IN CASE OF COMPLETE DATA AND UNDER DIFFERENT TYPES OF MISSINGNESS

		Complete data with covariates	MCAR	MAR	MNAR
Measurement Model					
Radio	α_{10}	- 0.104***	-0.111***	-0.136***	0.067***
	α_{1a}	-1.143***	-1.140***	-1.123***	-1.164***
Telephone	α_{20}	- 0.219***	-0.213***	-0.279***	- 0.076***
	α_{2a}	-2.196***	-2.253***	-2.186***	-2.213***
Computer	α_{30}	1.396***	1.395***	1.173***	1.558***
	α_{3a}	-3.369***	-3.383***	-3.164***	-3.364***
Smart phone	α_{40}	0.284***	0.281***	0.164***	0.415***
	α_{4a}	-3.145***	-3.170***	-3.083***	-3.136***
Missingness Model					
r (Radio)	v_{10}		2.012*	4.503***	2.488***
	v_{1r}		0.190	-3.259***	-0.379***
r (Telephone)	v_{20}		2.303***	4.630***	2.663***
	v_{2r}		-0.108	-3.365***	-0.596****
r (Computer)	v_{30}		4.103	4.421***	5.444**
	v_{3r}		-1.929	-3.200***	-3.578**
r (Smart phone)	v_{40}		-6.921	4.571***	3.295***
	v_{4r}		9.397	-3.356***	-1.330***
Structural Model					
z_a on z_r	ϕ		-0.129	4.033***	17.134***

Note: *** denotes a p-value < $\mathbf{0.01}$, ** denotes a p-value < $\mathbf{0.05}$ and * denotes a p-value < $\mathbf{0.10}$.

Table 5 gives parameter estimates for the overall proposed model presented by Equations (1), (3), (4) and (5) and illustrated in Fig. 2 for the three types of missingness, assuming a two-latent class model for each of the latent variables "Access to Knowledge Sources" and "Response Propensity".

The measurement model is quite robust under different missingness scenarios. This indicates that the contribution of the manifest variables in measuring the latent variable of interest is not affected by the type of missing data. On the contrary, the missingness model exhibits differences under the different types of missingness. The four missingness indicators are insignificant in defining the latent variable "Response Propensity" when data is MCAR. This is a logical result as the items were created to reflect a completely random pattern of missingness in the data, and thus the created dataset is a random subset of the data. The indicators do not measure "Response Propensity" in this case.

However, they are significant in measuring the latent variable when missing data is MAR or MNAR, as in both cases, those who have missing values are different from those who respond. However, while their contribution in measuring "Response Propensity" is almost equal for the four indicators in data MAR, it is higher for computers and smartphones than radio and telephone in the case of data MNAR.

The structural parameter ϕ, reflecting the relationship between "Response Propensity" and "Access to Knowledge Sources", is given at the bottom of Table 5. As one would expect, this relationship is insignificant in case of data MCAR. On the contrary, there is a significant relationship in the case of data MAR and MNAR. The significant positive effect ϕ indicates that higher response levels are more likely to be found with higher levels of access to knowledge sources, even after controlling for covariates. The magnitude of the structural parameter is much higher in the case of data MNAR. A possible explanation of the significant effect indicating nonignorable missingness in the case of data that was originally created at random is that levels that are used in creating missingness are themselves confounded with certain levels of "Access to Knowledge Sources", and thus there still exists kind of dependence of "Response Propensity" on "Access to Knowledge Sources".

Table 6 shows the estimated conditional probabilities for the manifest items given class membership of the main latent variable "Access to Knowledge Sources", and those of the missingness indicators given class membership of the latent variable "Response Propensity," assuming that both latent variables are binary. The conditional probabilities are reported under different types of missingness. We may consider the first latent class of "Access to Knowledge Sources" to indicate "High access to knowledge sources" and the second to indicate "Low access to knowledge sources", as the conditional probabilities of having any of the devices are consistently higher for the first-class than the second. The conditional probabilities resulting from the "Response Propensity" latent variable are not reliable in the case of data MCAR since the indicators are not significant in defining the "Response Propensity" latent variable (see Table V).

However, there is a clear pattern of higher estimated probabilities for the first class than the second, in data MAR or MNAR, although the differences are sometimes not too big. The first latent class may thus be labeled as "High response propensity" and the second latent class as "Low response propensity".

TABLE VI
ITEM-RESPONSE CONDITIONAL PROBABILITIES FROM A TWO-CLASS LCM FOR "ACCESS TO KNOWLEDGE SOURCES" AND "RESPONSE PROPENSITY" LATENT VARIABLES UNDER THREE TYPES OF MISSINGNESS

	MCAR		**MAR**		**MNAR**	
"Access to Knowledge Sources"	**1st class**	**2nd class**	**1st class**	**2nd class**	**1st class**	**2nd class**
	Probability of a "Yes"					
Radio	0.472	0.223	0.466	0.221	0.517	0.250
Telephone	0.447	0.078	0.431	0.078	0.481	0.092
Computer	0.801	0.120	0.764	0.120	0.826	0.141
Smart phone	0.570	0.053	0.541	0.051	0.602	0.062
"Response Propensity"	**1st class**	**2nd class**	**1st class**	**2nd class**	**1st class**	**2nd class**
	Probability of being observed					
$r_{(Radio)}$	-	-	0.989	0.776	0.923	0.892
$r_{(Telephone)}$	-	-	0.990	0.780	0.935	0.888
$r_{(Computer)}$	-	-	0.988	0.772	0.996	0.866
$r_{(Smart\ phone)}$	-	-	0.990	0.771	0.964	0.877

Notes: The complement of the above probabilities indicates the probability of responding with a "No" to the corresponding item. The complement of the response indicator probabilities gives the probability of a "Missing" response.

From Table 7, and considering how the latent variable is defined, it can be concluded that the probability of having high access to knowledge sources is generally higher for more privileged people. That is to say; it is higher for males than females and people with a higher wealth index, of older age, and with higher levels of education. An unexpected result is that the probability of having high access to knowledge sources is higher for those living in rural areas compared to urban. However, it is not known whether the available media are used to access knowledge or mainly for entertainment and communication.

In our application, covariates effects on the missingness part of the model seem to be significant only in the case of data MAR. The sex covariate has a negative effect on "Response Propensity," which means that females are more likely to respond. Place of residence and age positively affect "Response Propensity," which means that younger people and people in urban areas are more likely to respond. For data MCAR, the "Response Propensity" latent variable is not well-defined due to randomness in creating the missingness in this case, so its relationship with "Access to Knowledge Sources" and covariates turned out to be insignificant. In the case of data MNAR, only place of residence significantly affects "Response Propensity".

A possible explanation is that the effect of the other two covariates is already indirectly carried within the latent variable of interest "Access to Knowledge Sources", since their effect on "Access to Knowledge Sources" is already highly significant.

TABLE VII
ESTIMATED COVARIATES EFFECTS FROM A TWO-CLASS LCM FOR "ACCESS TO KNOWLEDGE SOURCES" AND "RESPONSE PROPENSITY" LATENT VARIABLES IN CASE OF COMPLETE DATA AND UNDER DIFFERENT TYPES OF MISSINGNESS

		Complete data with covariates	MCAR	MAR	MNAR
Covariates effects on Z_a					
Place of residence (Rural)	β_1	-2.791***	-2.751***	-3.540***	-2.839***
Wealth index	β_2	-2.691***	-2.645***	-3.209***	-2.705***
Sex (Female)	β_3	0.376***	0.365***	0.408***	0.355***
Age	β_4	-0.043***	-0.045***	-0.047***	-0.044***
Educational level	β_5	-0.628***	-0.626***	-0.607***	-0.630***
Covariates effects on Z_r					
Place of residence (Rural)	γ_1		-0.047	2.548***	0.698**
Sex (Female)	γ_2		-0.109	-0.481***	0.305
Age	γ_3		0.001	0.018***	0.012

Note: *** denotes a p-value < $\mathbf{0.01}$ and ** denotes a p-value < $\mathbf{0.05}$.

IV. Conclusion

When multiple manifest variables are used as measures of a latent variable, it is quite often to have some missing values in the data due to item non-response. In this paper, we proposed to summarize item non-response by another latent variable that can be labeled as "Response Propensity". The missingness can thus be allowed to be non-random by allowing the "Response Propensity" latent variable to depend on the main latent variable of interest.

A model specification incorporating a missingness mechanism within a latent class model framework has been proposed to model multivariate binary data used as measures of a categorical latent variable. This model specification allows for nonignorable item non-response by letting the response propensity latent variable summarize the response indicators, depending on the latent variable of interest and covariates. Logistic regression equations are used to model relationships within the latent class model under the categorical nature of all manifest and latent variables in the model. Estimation of model parameters and goodness of fit measures use conventional methods that are usually used to fit latent variable models for multivariate data.

The proposed model has been applied to data from Egypt's Demographic and Health Survey 2014. Data missingness has been artificially created to generate three different types of missingness, MCAR, MAR and MNAR, to study the results of the model in each case. An important result of the model was that the measurement part defining the latent variable of interest, "Access to Knowledge Sources" is quite robust no matter how missingness was created. For data MAR and MNAR, the relationship between the "Response Propensity" latent variable and the "Access to Knowledge Sources" latent variable remains significant even after controlling for covariates.

Unlike other models already existing in the literature, such as Bacci and Bartolucci [22] and Beesley *et al* [32], the proposed model accounts for missingness and allows for this missingness to be non-random by depending on levels of the latent class of interest. The estimated probabilities of class membership of the "Response Propensity" latent variable are affected by class membership of the "Access to Knowledge Sources" latent variable making the missingness nonignorable. Lower levels of response were associated with lower levels of "Access to Knowledge Sources". This result confirms the importance of accommodating the missingness mechanism within the modeling of the data due to the systematic difference between respondents and nonrespondents. Covariates effects are also found to be robust on the measurement model; however, they are quite sensitive to the type of missingness in the missingness part of the model. We have used Bayesian estimation in Zakaria *et al.* [33] to fit the same model specification proposed in this article and to study the sensitivity of the results to different levels of missingness.

References

[1] G. Verbeke and G. Molenberghs, "Modeling Through Latent Variables," *Annual Review of Statistics and Its Application*, vol. 4, pp. 267–282, 2017.

[2] R. van Bork, M. Rhemtulla, L. J. Waldorp, J. Kruis, S. Rezvanifar, and D. Borsboom, "Latent Variable Models and Networks: Statistical Equivalence and Testability," *Multivariate Behavioral Research*, vol. 56, no. 2, pp. 175–198, 2021.

[3] D. J. Bartholomew, M. Knott, and I. Moustaki, *Latent Variable Models and Factor Analysis*, 3rd ed., Wiley series in probability and statistics, 2011.

[4] S. Jeon, J. Lee, J. C. Anthony, and C. H., "Latent class analysis for multiple discrete latent variables: a study on the association between violent behavior and drug-using behaviors," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 24, pp. 911–925, 2017.

[5] J. W. Lee and H. Chung, "Latent class analysis with multiple latent group variables," *Communications for Statistical Applications and Methods*, vol. 24, pp. 173–191, 2017.

[6] K. J. Petersen, P. Qualter, and N. Humphrey, "The Application of Latent Class Analysis for Investigating Population Child Mental Health: A Systematic Review," *Frontiers in Psychology*, vol. 10, p. 1214, 2019.

[7] E. Kim, H. Chung, and S. Jeon, "Joint latent class analysis for longitudinal data: an application on adolescent emotional well-being," *Communications for Statistical Applications and Methods*, vol. 27, pp. 241–254, 2020.

[8] K. J. Petersen, N. Humphrey, and P. Qualter, "Latent Class Analysis of Mental Health in Middle Childhood: Evidence for the Dual-Factor Model," *School Mental Health*, vol. 12, pp. 786–800, 2020.

[9] A. Robitzsch, "Regularized Latent Class Analysis for Polytomous Item Responses: An Application to SPM-LS Data," *Journal of Intelligence*, vol. 8, p. 30, 2020.

[10] J. W. Lee and H. Chung, "A multivariate latent class profile analysis for longitudinal data with a latent group variable," *Communications for Statistical Applications and Methods*, vol. 27, pp. 15–35, 2020.

[11] J. H. M. Janssen, S. van Laar, M. J. de Rooij, J. Kuha, and Z. Bakk, "The Detection and Modeling of Direct Effects in Latent Class Analysis," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 26, no. 2, pp. 280–290, 2019.

[12] Z. Bakk and J. Kuha, "Relating latent class membership to external variables: An overview," *British Journal of Mathematical and Statistical Psychology*, vol. 74, pp. 340–362, 2020.

[13] Z. Bakk and J. Kuha, "Two-step estimation of models between latent classes and external variables," *Psychometrika*, vol. 83, no. 4, pp. 871–892, 2018.

[14] J. Kuha, S. Butt, M. Katsikatsou, and C. J. Skinner, "The Effect of Probing 'Don't Know' Responses on Measurement Quality and Non-response in Surveys," *Journal of the American Statistical Association*, vol. 113, no. 521, pp. 26–40, 2018.

[15] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 2nd ed., John Wiley and Sons, 2002.

[16] H. Du, C. Enders, B. T. Keller, T. N. Bradbury, and B. R. Karney, "A Bayesian Latent Variable Selection Model for Nonignorable Missingness," *Multivariate Behavioral Research*, 2021, doi: 10.1080/00273171.2021.1874259.

[17] N. Rose, M. von Davier, and B. Nagengast, "Modeling Omitted and Not-Reached Items in IRT Models," *Psychometrika*, vol. 82, pp. 795–819, 2017.

[18] J. F. Cursio, R. J. Mermelstein, and D. Hedeker, "Latent trait shared-parameter mixed models for missing ecological momentary assessment data," *Statistics in Medicine*, vol. 38, no. 4, pp. 660– 673, 2019.

[19] H. Jung, J. L. Schafer, and B. Seo, "A latent class selection model for nonignorably missing data," *Computational Statistics and Data Analysis*, vol. 55, pp. 802–812, 2011.

[20] O. Harel and J. L. Schafer, "Partial and latent ignorability in missing-data problems," *Biometrika*, vol. 96, no. 1, pp. 37–50, 2009.

[21] J. Kuha, M. Katsikatsou, and I. Moustaki, "Latent variable modelling with non-ignorable item non-response: multigroup response propensity models for cross-national analysis," *Journal of the Royal Statistical Society, Series A*, vol. 181, pp. 1169–1192, 2018.

[22] S. Bacci and F. Bartolucci, "A Multidimensional Latent Class IRT Model for Non-Ignorable Missing Responses," *Structural Equation Modeling: A Multidisciplinary Journal*, 2014, [Online]. Available: https://arxiv.org/abs/1410.4856.

[23] S. K. Sterba, "A Latent Transition Analysis Model for Latent-State-Dependent Nonignorable Missingness," *Psychometrika*, vol. 81, no. 2, pp. 506–534, 2016.

[24] El-Zanaty, Associates, and I. International, "Egypt demographic and health survey 2014 (Data file and code book)," Ministry of Health and Population, Egypt, 2014. [Online]. Available: https://www.dhsprogram.com/data/available-datasets.cfm

[25] L. M. Collins and S. T. Lanza, *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. Wiley series in probability and statistics, 2010.

[26] L. K. Muthén and B. O. Muthén, *Mplus User's Guide*, 8th ed., Los Angeles, CA: Muthén & Muthén, 1998-2017.

[27] L. A. Goodman, "Exploratory latent structure analysis using both identifiable and unidentifiable models," *Biometrika*, vol. 61, pp. 215–231, 1974.

[28] G. Xu, "Identifiability of restricted latent class models with binary responses," *The Annals of Statistics*, vol. 45, no. 2, pp. 675–707, 2017.

[29] W. Koo and H. Kim, "Bayesian nonparametric latent class model for longitudinal data," *Statistical Methods in Medical Research*, vol. 29, no. 11, pp. 3381–3395, 2020.

[30] J. Tein, S. Coxe, and H. Cham, "Statistical Power to Detect the Correct Number of Classes in Latent Profile Analysis," *Structural Equation Modeling*, vol. 20, no. 4, pp. 640–657, 2013.

[31] D. Gordon, S. Nandy, C. Pantazis, S. Pemberton, and P. Townsend, "The distribution of poverty in the developing world," *University of Bristol, Centre for International Poverty Research, UK, 2003.*

[32] L. J. Beesley, J. M. G. Taylor, and R. J. A. Little, "Sequential imputation for models with latent variables assuming latent ignorability," *Aust. N. Z. J. Stat.*, vol. 61, no. 2, pp. 213–233, 2019.

[33] S. Zakaria, M. S. Hafez, and A. M. Gad, "Bayesian Estimation of Latent Class Model for Survey Data Subject to Item Nonresponse," *Pakistan Journal of Statistics and Operation Research*, vol. 15, no. 2, pp. 303–318, 2019.