THE PHILIPPINE STATISTICIAN

Volume 71, Number 1 (2022)

The Official Publication of the
Philippine Statistical Association, Inc.

Indexed in Scopus since 2015

The **Philippine Statistical Association, Inc. (PSAI)** is the Philippines' sole scientific society of professionals committed to the promotion of the proper use of statistics.

The PSAI is a private non-stock and non-profit foundation organized on December 22, 1951 and registered and incorporated on September 24, 1952 with the Securities and Exchange Commission. It was also registered with the Department of Science and Technology on August 31, 1993 as a foundation for scientific advancement.

Room 214, Philippine Social Science Center,
Commonwealth Avenue, Diliman, 1101, Quezon City, Philippines

Telephone: (632) 9-920-6513 Telefax: (632) 3-456-1928
E-mail: secretariat@psai.ph
www.psai.ph

Contents

Editorial

This first issue for the year 2022, *The Philippine Statistician* identified as Volume 71, No. 1 has six papers authored not only by local researchers but authors from other countries as well. We have articles that are theoretical in content and there are also articles that showcase the application of Statistics in the fields of Education and Health Science. Papers in sampling survey are also included in this issue.

A paper on how to handle missing data in the response variable and covariates using the stochastic EM algorithm as authored by **A. Gad** and **N. Darwish** is included in this issue. This paper is theoretical in nature which is in the same category of the paper of **S. Bhushan** and **A. Kumar** which presented some efficient classes of estimators based on higher order moments of an auxiliary attribute.

The paper of **J. Grajo**, **J. Reyes**, **L. Comia**, **L. Ebal**, **J. Mendoza** and **M. Talento** on applying the analytic hierarchy process with Rasch measurement in the construction of a composite metric of student online learning readiness scale showcases the application of Statistics in the field of Education. Likewise, there is also a paper that illustrates the application of CATANOVA and logistic regression in the field of Health Science as it was applied in the analysis of a case study involving most prevalent sexually transmitted infection disease in Nigeria. This paper was co-authored by **N. Eze**, **O. Asogwa**, **S. Ugwu**, **C. Eze**, **F. Ohanuba**, and **T. Ugah**.

The papers in sampling survey serve as a guide on conducting survey. Specifically, one paper tackles the survey operations for research and development survey while the other paper deals with a sampling technique in a clinical survey. The former authored by **R. Cambel**, **D. Maligalig**, **M. Borromeo**, **R. Roldan Jr.**, and **C. Lesmoras** addressed the implementation of an effective survey operations for research and development in the Philippines. On the other hand, the latter paper authored by **M. Bujang**, **T. Tg Abu Bakar Sidik** and **N. Sa'at** investigated if the application of consecutive sampling techniques in a clinical survey for an ordered population will also generate accurate statistics.

As a summary, the articles in this issue introduced some ideas on estimation as well as illustrated the diversity of the application of Statistics in certain field of studies.

Prof. Zita Villa Juan Albacea, PhD
2022 TPS Editor-in-Chief

Analysis of Longitudinal Data with Missing Values in the Response and Covariates Using the Stochastic EM Algorithm

Ahmed M. Gad[1]
Business Administration Department, Faculty of Business Administration, Economics and Political Science, The British University in Egypt (BUE), Cairo, Egypt

Nesma M. Darwish
Business Administration Department, Faculty of Politics, Economics and Business Administration, May University in Cairo (MUC), Egypt

ABSTRACT

Longitudinal data are not uncommon in many disciplines where repeated measurements on a response variable are collected for each subject. Missing values are unavoidable in longitudinal studies. Missing values could be in the response variable, the covariates or in both. Dropout pattern occurs when some subjects leave the study prematurely. When the probability of missingness depends on the missing value, and may be on the observed values, the missing data mechanism is termed as non-random. Ignoring the missing values in this case leads to biased inferences. In this paper we will handle missing values in covariates using multiple imputations (MI) and the selection model to fit longitudinal data in the presence of non-random dropout. The stochastic EM (Expectation-Maximization) algorithm is developed to obtain the model parameter estimates. Also, parameter estimates of the dropout model have been obtained. Standard errors of estimates have been calculated using the developed Monte Carlo method. The proposed approach performance is evaluated through a simulation study. Also, the proposed approach is applied to a real data set.

Keywords: *Interstitial Cystitis data; missing covariates; dropout missingness; multiple imputation; selection model; the SEM algorithm.*

1. Introduction

In longitudinal studies each subject is measured repeatedly, for the same response variable at different times or under different condition or both. Longitudinal data are very common in biomedical research and clinical trials where some of measurements on a subject develops over time. In these cases, one variable is the underlying characteristic or measurement. The main advantage of longitudinal studies is that it can distinguish changes over time within individuals and enabling direct study of that change.

[1] **Address correspondence to Ahmed M. Gad**: ahmed.gad@feps.edu.eg

Missing data are not uncommon in longitudinal studies. The missing values could be due to many reasons. The missing data in the response occur whenever one or more of measurement sequences are incomplete. Missing data in the response can be categorized into two different patterns: intermittent missing pattern and dropout pattern. In intermittent pattern a missing value could be followed by an observed value. Dropout pattern means a missing value is never followed by an observed value.

It is commonly assumed that the responses are missing at the time of dropout, but all covariates are completely observed. Little (1995) reviews some approaches where the covariates are completely observed. However, the response variable and the associated covariates maybe not observed at the time of dropout. Hence, the assumption of completely observed covariates is often not realistic. In this article we focus on missingness in response and covariates.

In the case of missingness in the response, Erler et al. (2016) evaluate the performance of multiple imputation chained equation using different strategies to include a longitudinal response into the imputation models and compare it with a fully Bayesian approach. Nooraee et al. (2018) investigate a hybrid approach which is a combination of maximum likelihood and multiple imputation, i.e. scales from the imputed data are eliminated if all underlying items were originally missing. Abdelwahab et al. (2019) propose a sensitivity analysis index for shared parameter models in longitudinal studies. Darwish et al. (2020) propose using multiple imputation for missing at random (MAR) cross-sectional covariates. They employ a shared parameter model to fit response variable in the presence of non-random dropout.

Assume that Y_{ij} is the longitudinal response of subject i at time point j and R_{ij} is the missing data mechanism indicator, where R_{ij} equals 1 if Y_{ij} is observed and 0 if Y_{ij} is missing. In the selection model the joint distribution of the response Y_i and R_i are factorized as product of the marginal distribution of Y_i and conditional distribution of R_i given Y_i. Thus

$$f(Y_i, R_i|\theta, \Psi) = f(Y_i|\theta)P(R_i = r_i|Y_i, \Psi), \tag{1}$$

where θ is a vector containing the model parameters, $P(R_i = r_i|Y_i, \Psi)$ is the distribution that characterizes the missing data mechanism, and Ψ is a vector of parameters that govern the missing data mechanism. According to Rubin's taxonomy, the missing data mechanism can be classified to three different mechanisms (Rubin, 1976). The first is missing completely at random (MCAR) if R_i and Y_i are independent, i.e.

$$P\big(R_i = r_i\big|Y_{i,obs}, Y_{i,mis}\Psi\big) = P(R_i = r_i|\Psi), \tag{2}$$

where $Y_{i,obs}$ and $Y_{i,mis}$ are the observed and missing parts of Y_i, respectively. The second is missing at random (MAR) if the conditional distribution of R_i given Y_i depends only on the observed, $Y_{i,obs}$, i.e.

$$P\big(R_i = r_i\big|Y_{i,obs}, Y_{i,mis}\Psi\big) = P(R_i = r_i|Y_{i,obs}, \Psi). \tag{3}$$

The third is nonrandom (informative) if it is neither MCAR nor MAR. In dropout pattern, Diggle and Kenward (1994) propose a selection model for longitudinal data with nonrandom dropout. They specified a normal linear model for the response variable, Y_i, and a logistic model for the probability of dropout. They suggest modelling the probability of dropout at time d_i as

a function of the measurement at time d_i and the observed measurements (history H_{d_i}) up to time $d_i - 1$; that is,

$$P(D_i = d_i | history) = P_{d_i}(H_{d_i}, y_{d_i}, \Psi). \quad (4)$$

Also, they suggest using the logistic model for the dropout process as

$$\text{logit}\{P_{d_i}\{H_{d_i}, y_{d_i}, \Psi)\} = \psi_0 + \sum_{j=1}^{d_i} \psi_j \, y_{d_{i-j+1}}. \quad (5)$$

The SEM algorithm has been proposed by Celuex and Diebolt (1985) as a stochastic version of the EM algorithm. The SEM algorithm overcomes the main difficulty of the EM algorithm, in some situations, by avoiding explicit calculation of the E-step. The E-step is replaced by the stochastic step (S-step) where the missing data are imputed with a single draw from the conditional distribution of the missing data given the observed data. In the M-step, the log-likelihood function of the pseudo-complete can be maximized using standard maximization procedures. So, the algorithm involves iterating two steps, the S-step and the maximization step (M-step) for sufficient number of iterations.

The estimated parameter values corresponding to each pseudo-complete data form a Markov chain. This Markov chain converges reasonably quickly to its stationary distribution, which is unique (Diebolt and Ip, 1996). The mean of the points, ignoring the early first points as a burn-in period, generated by the SEM algorithm can be considered as an estimate for the parameter β. This mean is called the SEM estimate and denoted by $\tilde{\beta}$ (Diebolt and Ip,1996). Gad and Ahmed (2006) apply the SEM algorithm to longitudinal data with dropout in the response. Different variants of SEM algorithm are also used to escape poor local maxima using the concepts of simulated annealing (Allassonnière and Chevallier, 2021). Yassen and Gad (2020) introduce different variants of the SEM algorithm to deal with mixed continuous and discrete longitudinal data.

The EM algorithm, also the SEM algorithm, does not provide the standard errors of the parameter estimates. Several methods have been proposed in literature to solve this problem. Louis' formula (Louis, 1982) relates the observed information matrix to the conditional expectation of the second derivatives of complete data log-likelihood function and the covariance of the first derivatives of complete data log-likelihood function. Evaluating the integrals in this formula, in the current setting, may not be easy. Efron (1994) suggests using simulation (the Monte Carlo method) to approximate the integrations. The missing values are simulated from their conditional distribution and then each integration is evaluated by its empirical version.

The aim of this article is to suggest a multiple imputation approach for cross-sectional covariates. The selection model is adopted for fitting linear regression model between longitudinal response and cross-sectional covariates where both the response and the covariates have missingness. The rest of the article is organized as follows. In Section 2 we present the two common multiple imputation methods that can be used to handle missingness in covariates. In Section 3 the proposed approach is described in addition to the Monte Carlo method as a way for obtaining the standard error estimates. In section 4, a simulation study is presented to validate the proposed approach. In section 5 the proposed approach is applied to a real data. Finally, Section 6 presents the conclusion and future work.

2. Multiple Imputations (MI) Methods

Grannell and Murphy (2011) discuss the application of four multiple imputations (MI) methods using the SOLAS package. Salfran and Spiess (2015) describe some of the most common imputation methods included in software packages. The most common two multiple imputations methods are described below in more details.

2.1 Regression-based imputation

In the regression method, a regression model is fitted for each variable with missing values using the complete cases. Based on the resulting model we impute the missing values because the data set has a monotone missing data pattern. The process is repeated sequentially for all variables with missing values. That is, for a variable Y_j, with missing values, a model

$$Y_j = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k \tag{6}$$

is fitted using the observed values and the corresponding covariates $(X_1, \ldots, X_2)$. The fitted model includes the regression parameter estimates $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ and the association covariance matrix $\widehat{\sigma^2} V_j$.

The following steps are used to generate the imputed values.

1- New parameter $\beta_* = (\beta_{0*}, \beta_{1*}, \ldots, \beta_{k*})$ and $\sigma^2_{*_j}$ are obtained from the posterior predictive distribution of the parameters. That is, they are simulated from $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ and the covariance matrix $\widehat{\sigma^2} V_j$. The variance $\sigma^2_{*_j}$ is obtained as

$$\sigma^2_{*j} = \frac{\sigma^2 (n_j - k - 1)}{g}, \tag{7}$$

where g is a chi-square $\chi^2_{n_j - k - 1}$ random variate and n_j is the number of non-missing observations of Y_j. The regression coefficients are obtained as

$$\beta_* = \hat{\beta} + \sigma^2_{*j} V'_{hj} Z, \tag{8}$$

where V'_{hj} is the upper triangular matrix in the Cholesky decomposition of V_j and Z is a vector of $k+1$ independent standard normal variates with 0 means and a variance of 1.

2- The missing values are then replaced by

$$y_{i*} = \beta_{0*} + \beta_{1*} x_1 + \beta_{2*} x_2 + \ldots.. + \beta_{k*} x_k + z_i \sigma_{*_j}, \tag{9}$$

where $(x_1, x_2, \ldots, x_k)$ are the values of the covariates and z_i is a simulated standard normal (random) deviate using standard statistical packages.

The regression method can be extended to deal with longitudinal data with dropout. In the sequential imputation method, assuming that the data are complete at the first time point, the regression imputation-based method described above can be used to impute the missing values at the second time point. The previously imputed values (possibly a subset of them) are used in the imputation model as predictors for future values. This process is repeated at the third time point and sequentially up to the final time point. The process is repeated M times to obtain M completed data sets. In principle, the normal-based regression model can be replaced by any appropriate model, for other types of responses, for example logistic regressions for binary data, or proportional odds models for ordinal data.

2.2 Predictive mean matching method (PMM)

The predictive mean matching method (PMM) can also be used for imputation. It is like the regression method, except that for each missing value, it imputes an observed value which is closest to the predicted value using the simulated regression model (Rubin 1987). The predictive mean matching method ensures that imputed values are plausible, and may be more appropriate than the regression method, if the normality assumption is violated. The steps of the PMM method are the same as the regression method. However, the PMM method needs generating a set of k_0 observations whose corresponding predicted values are closest to y_{i*}. The missing value is then replaced by a value drawn randomly from these k_0 observed values.

3. The Proposed Approach

First, we handle the missingness in covariates through multiple imputation using the regression method or predictive mean matching method. Second the SEM algorithm can be applied using the pseudo complete covariates using the two steps: the S-step and the M-step.

- **Imputing continuous cross-sectional covariates with monotone missingness using regression method.**

Depending on the observed part of the response and the observed cross-sectional covariates, the missing covariates are imputed using the model $x_{i,mis} = \beta_0 + \beta_1 y_{i,obs} + \beta_2 x_{i,obs}$.

The following steps are used to generate the imputed values for each imputation.

1. New parameters $\beta_* = (\beta_{0*}, \beta_{1*}, \beta_{2i*})$ and σ^2_{*i} are drawn from the posterior predictive distribution of the parameters. That is, they are simulated from $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)$ and the associated covariance matrix $\hat{\sigma}^2 V_i$. The variance is obtained as $\sigma^2_{*i} = \frac{\sigma^2(n_i - k - 1)}{g}$, where g is a chi-square $\chi^2_{n_i-k-1}$ random variate and n_i is the number of non-missing observations for X_i. The regression coefficients are drawn as $\beta_* = \hat{\beta} + \sigma^2_{*j} V'_{hj} Z$, where V'_{hj} is the upper triangular matrix in the Cholesky decomposition of V_j and Z is a vector of k+1 independent random standard normal variates.
2. The missing values are then replaced by $x_{i*} = \beta_{0*} + \beta_{1*} y_i + \beta_{2*} x_{i,obs} + z_i \sigma^2_{*i}$, where y_i is the longitudinal response and $x_{i,obs}$ is the observed covariates and z_i is a simulated standard normal deviate.

- **Imputing continuous cross-sectional covariates with monotone missingness using predictive mean matching method.**

The above steps are used to generate imputed values with extra following two steps.

1- Generate a set of k_0 observations whose corresponding predicted values are the closest to x_{i*}.
2- The missing value is then replaced by a value drawn randomly from these k_0 values.

Here, the SEM algorithm can be applied using the pseudo complete covariates using the two steps: the S-step and the M-step.

The S-Step

In this step, the missing response values are simulated from their conditional distribution, given the observed values and the current parameter estimates, $Y_{i,mis} \sim f\left(Y_{i,mis} | Y_{i,obs}, R_i; \theta\right)$. This distribution does not have a standard form; hence it is not possible to use the direct simulation. To overcome this problem, we adopt an acceptance/rejection Monte Carlo simulation method, to generate the missing values $y_{i,mis}$. This procedure mimics the dropout process assuming that the postulated dropout model is correct. A draw from $f\left(y_{i,mis} | Y_{i,obs}, \theta^{(t)}\right)$ is obtained instead of $f\left(y_{i,mis} | Y_{i,obs}, R_i, \theta^{(t)}\right)$. Then, this value can be accepted or rejected using Metropolis Hasting procedure (Gilks et al, 1996)). The steps of this procedure can be summarized as follows:

1. Generate a candidate value, $\boldsymbol{y}^*$ from the conditional distribution function $\boldsymbol{f\left(y_{i,mis} | Y_{i,obs}, \theta^{(t)}\right)}$ which is normal distribution. Only the first dropped observation is simulated and the remaining dropped values are considered missing at random (Gad and Ahmed, 2006).

2. Calculate the probability of dropout for the candidate value $\boldsymbol{y}^*$, according to the dropout model $\boldsymbol{P(D_i = d_i | H_i d_i) = \Psi_0 + \Psi_1 y_{d_i} + \sum_{j=2}^{d_i} \Psi_j y_{i,d_i+1-j}}$, where the parameters $\boldsymbol{\Psi_j}$ are fixed at the current values $\boldsymbol{\Psi_j^{(t)}}$. Let us denote this probability of dropout, $\boldsymbol{P(D_i = d_i | H_i d_i)}$, as $\boldsymbol{P_i}$.

3. Simulate a random variable U from the uniform distribution on the interval [0,1] that is, $\boldsymbol{U \sim U[0, 1]}$, then take $\boldsymbol{y_{i,mis} = y^*}$ if $\boldsymbol{U \leq P_i}$; otherwise repeat Step1.

The M-Step

It consists of two sub-steps: the logistic step (M1-step) and the normal step (M2-step).

- In the logistic step (M1-step), the MLEs for the dropout logistic model

$$\boldsymbol{\text{logit}\{P_i\} = \psi_0 + \psi_1 y_{d_i} + \sum_{j=2}^{d_i} \psi_j y_{i,d_i+1-j}} \tag{10}$$

  are obtained. The iteratively reweighted least squares method for finding the MLE of binary data models (McCullagh and Nelder,1989) can be used.

- In the normal step (M2 step), the MLE.s for the model parameters can be obtained using an appropriate optimization approach for incomplete data such as Newton- Raphson, Scoring method and Jennrich and Schluchter algorithm (Jennrich and Schluchter, 1986). Newton- Raphson method is used in this article. The obtained estimates are the average of the M imputed data sets, i.e.

$$\hat{\beta} = \frac{1}{M} \sum_{i=1}^{M} \hat{\beta}_i. \tag{11}$$

Standard errors

Louis (1982) introduces the following formula to approximate the information matrix:

$$I(\theta) = \mathrm{E}\left(-\frac{\partial^2 l(\theta|Y_{obs}, Y_{mis})}{\partial\theta\partial\theta}\middle|Y_{obs}\right) - \mathrm{cov}\left(\frac{\partial l(\theta|Y_{obs}, Y_{mis})}{\partial\theta}\middle|Y_{obs}\right)$$

$$= -E - C, \tag{12}$$

where θ is fixed at the stochastic EM estimates and $l(\theta|Y_{obs}, Y_{mis})$ is the log-likelihood function. Evaluating the integrals in this formula may not be easy. Efron (1994) suggests using simulation (the Monte Carlo method) to approximate the integrations. The main idea is to simulate *M* identically distributed samples, $q_1, q_2, \ldots, q_M$ from the conditional distribution of the missing values given the observed values and the parameters estimates, $f\left(Y_{mis}|Y_{obs}, \hat{\theta}\right)$. Then, the formula in Eq. (12) can be approximated by its empirical version, i.e.

$$E \approx \frac{1}{M}\sum_{j=1}^{M}\frac{\partial^2 l\left(\theta|Y_{obs}, q_j\right)}{\partial\theta\partial\theta} \tag{13}$$

and

$$C \approx \mathrm{cov}\left(\frac{\partial l\left(\theta|Y_{obs}, q_j\right)}{\partial\theta}\right). \tag{14}$$

The Monte Carlo method is proposed and developed to obtain the standard errors of the SEM estimates in the current setting. We simulate $q_1, q_2, \ldots, q_M$ samples from the conditional distribution $f\left(Y_{mis}|Y_{obs}, R; \hat{\theta}\right)$. Then, the information matrix in Eq. (12) can be approximated as

$$\mathrm{E} \approx \frac{1}{M}\sum_{j=1}^{M}\frac{\partial^2 l\left(\theta|Y_{obs}, R, q_j\right)}{\partial\theta\partial\theta} \tag{15}$$

and

$$C \approx \mathrm{cov}\left(\frac{\partial l\left(\theta|Y_{obs}, R, q_j\right)}{\partial\theta}\right), \tag{16}$$

where the parameters $\theta = (\beta, \alpha, \psi)$ is fixed at the SEM estimates; $\hat{\theta} = (\hat{\beta}, \hat{\alpha}, \hat{\psi})$.

Having the *M* pseudo-complete data, the first and second order derivatives of the log-likelihood function are evaluated for each sample. Then it is possible to calculate the quantities *E* and *C* and hence the information matrix. The inverse of the information matrix is the covariance matrix of the stochastic EM estimates. The standard error estimates are the square root of the main diagonal elements of this matrix.

4. Simulation Study

The aim of this simulation is to investigate the performance of the proposed approach. A complete longitudinal outcome Y_{ij}, for the subject i at the time point j, is generated from the following model $Y_{ij} = \beta_0 + \beta_1 X_{ij} + \varepsilon_{ij}$,, where $i=1,2,\ldots,n$ and $j=1,2,\ldots,t$. The continuous cross-sectional covariates X_{ij} are generated from the standard normal distribution. These covariates

are independent from the error terms ε_{ij}. The error terms ε_{ij} are assumed to follow a normal distribution with a mean of zero and $\sigma_e^2 = 0.5$. The number of subjects *n* (the sample size) is fixed at 25 and 50 subjects. The time points are restricted at *t* = 5. The parameters are fixed at $\beta_0 = 5$ and $\beta_1 = 10$. The simulation is replicated 2000 times. The missing values in the cross-sectional covariates are generated according to the logit model:

$$\text{logit}(X_i) = \eta_0 + \eta_1 X_{i-1}. \tag{17}$$

The parameters are fixed at $\eta_0 = -5$ and $\eta_1 = 0.06$. The missing values in the response are generated according to the model:

$$\text{logit}\left(r_{ij} = 1 \middle| \Psi\right) = \Psi_0 + \Psi_1 Y_{ij-1} + \Psi_2 Y_{ij}. \tag{18}$$

The parameters are assumed to be $\Psi = (\Psi_0, \Psi_1, \Psi_2) = (-17, 0.11, 0.13)$. All subjects are assumed to be observed at the first time point $j = 1$. The covariance structure, of the response, is assumed to be autoregressive of order 1, AR(1), with ρ=0.7 and σ =6.0.

We apply the proposed approach where multiple imputation to cross-sectional covariates with number of imputations *M=10*. The final parameter estimates are obtained as the average over the multiply imputed data sets, i.e.

$$\widehat{\boldsymbol{\beta}} = \frac{\mathbf{1}}{\mathbf{10}} \sum_{\boldsymbol{i=1}}^{\mathbf{10}} \widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}. \tag{19}$$

Table 1 and Table 2 present the results assuming the covariates are complete, and the sample size is 25 and 50, respectively. Table 3 and Table 4 present the results assuming that there are missing values in the covariates, and the sample size is 25 and 50, respectively.

Table 1. Parameter estimates (Est) and the relative bias (RB); sample size n= 25, complete covariates.

	β_0	β_1	ρ	σ	Ψ_0	Ψ_1	Ψ_2
True parameter	5	10	0.70	6	-17	0.11	0.13
Est.	5.13	9.61	0.64	5.8	-15.02	0.12	0.11
RB	0.03	0.04	0.08	0.03	0.12	0.09	0.15

Table 2. Parameter estimates (Est) and the relative bias (RB); sample size n= 50, complete covariates.

	β_0	β_1	ρ	σ	Ψ_0	Ψ_1	Ψ_2
True parameter	5	10	0.7	6	-17	0.11	0.13
Est.	5.34	9.67	0.63	5.81	-16.1	0.12	0.12
RB	0.07	0.03	0.10	0.03	0.05	0.09	0.08

Table 3.Parameter estimates (Est) and the relative bias (RB); sample size n= 25, MAR covariates.

	β_0	β_1	ρ	σ	Ψ_0	Ψ_1	Ψ_2
True parameter	5	10	0.7	6	-17	0.11	0.13
Predictive mean matching method							
Est.	4.54	10.09	0.65	5.75	-15.45	0.13	0.10
RB	0.09	0.01	0.07	0.04	0.09	0.18	0.23
Regression method							
Est.	4.80	9.76	0.62	5.85	-14.55	0.10	0.11
RB	0.04	0.02	0.11	0.02	0.14	0.09	0.15

Table 4. Parameter estimates (Est) and the relative bias (RB); sample size n= 50, MAR covariate.s

	β_0	β_1	ρ	σ	Ψ_0	Ψ_1	Ψ_2
True parameter	5	10	0.7	6	-17	0.11	0.13
Predictive mean matching method							
Est.	5.03	10.06	0.67	5.81	-16.01	0.11	0.12
RB	0.01	0.01	0.04	0.03	0.06	0.02	0.08
Regression method							
Est.	4.97	10.01	0.69	5.86	-15.55	0.09	0.10
RB	0.01	0.001	0.014	0.02	0.08	0.17	0.24

From Tables 1 and 2 we can see that the parameter estimates, for both sample sizes, perform well in terms of the relative bias. Tables 3 and 4 show similar results for both sample sizes. As a conclusion depending on this simulation results, the SEM estimates with multiple imputations to MAR in covariates using the regression method has the best performance in terms of relative bias. Hence, the performance of the proposed approach is acceptable even in relatively small sample sizes.

5. Application: Interstitial Cystitis Data Base (ICDB)

The Interstitial Cystitis Data Base (ICDB) have been used by Yang and Kang (2010). The ICDB characteristics are discussed in detail in Propert et al. (2000). The data include 637 patients at the baseline. Patients are followed for symptoms of pain, urgency, and urinary frequency, from January 1993 to November 1997. Yang and Kang (2010) study the joint effect of a group of covariates on the urgency and urinary frequency treating them as continuous and discrete variables, respectively. Each of these variables are measured by asking the patients to rate them in the last week on an ordinal scale ranging from 0; for the lowest severity, to 9 which is the maximum severity. In addition, the patients are required to rate the same variables in three consecutive days. The averages of the study variables over the three days are also recorded. Therefore, Yang and Kang (2010) consider only the data gathered in the first 36 months. The aim of the study is to explore the effect of continuous covariates on the continuous response (urgency).

There are missing values in the response variable (urgency). There are both dropout pattern and intermittent pattern. Because the proposed method deals with dropout pattern, so we omit the intermittently missing values. All continuous covariates are complete, and we generated

missing values in the covariates. This is result in a reduced sample of 450 patients. A brief description of the covariates is given in the Table 5.

Table 5. The definition of the continuous covariates used in ICDB data.

Variable	Definition
Age	Patient age
UROD_7	Volume at first sensation
UROD_9	Volume at maximal capacity

Figure 1 presents a histogram of the continuous outcome plotted and compared with normal density function with mean of 4.25 and variance of 2.13 s a pre-analysis step we checked the adequacy of the model to fit the data.

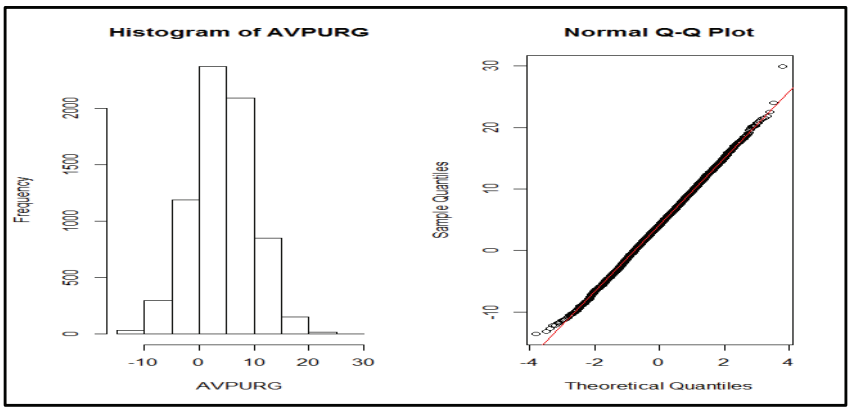

Figure 1. A histogram of the continuous outcome with normal and normal Q-Q plot.

We adopt the following model that allows each covariate to have its own effect, that is

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_{ij}. \tag{20}$$

The AR(1) covariance structure is used, the elements of the covariance matrix, $\sigma_{ij=\sigma^2\rho^{|i-j|}}$. For the missing data mechanism, we use the linear logistic regression model. To keep the model simple only the previous and the current outcomes are included, that is

$$\text{logit}(r_{ij} = 1|\Psi) = \psi_0 + \psi_1 Y_{ij-1} + \psi_2 Y_{ij}. \tag{21}$$

The proposed approach is applied to the Interstitial Cystitis Data (ICDB). The results are given in the following Tables 6 and 7.

Table 6. The SEM estimates and standard errors (SE) for the urgency response without imputation to MAR covariates.

	Est.	SE	P-value
Intercept (β_0)	0.3367	0.07	< 0.000
UROD_7(β_1)	-0.001	0.16	< 0.000
UROD_9 (β_2)	-0.003	0.17	< 0.0640
Age (β_3)	0.05	0.081	< 0.004
ρ	0.12	0.079	< 0.000
σ	4.59	0.13	0.030
Ψ_0	-1.22	0.068	< 0.000
Ψ_1	0.13	0.079	< 0.000
Ψ_2	0.18	0.098	< 0.000

Diggle and Kenward (1994) noticed that in nonrandom models, dropout tends to depend on the difference between the current and previous measurements, $Y_{ij-1} - Y_{ij}$. Using this idea the estimated model for the missing data mechanism can be viewed as:

$$\begin{aligned} \text{logit}(P) &= -1.22 + 0.13Y_{ij-1} + 0.18Y_{ij}, \\ &= -1.22 + 0.13(Y_{ij} - Y_{ij-1}) + 0.31Y_{ij}. \end{aligned} \tag{22}$$

From this model, the positive coefficient (0.13) of the difference between Y_{ij} and Y_{ij-1} also indicates that the response whose urgency increased are more likely to be missing

Table 7. The SEM estimates and standard errors (SE) for the urgency response with MAR missingness in covariate.

	Est.	SE	P-value
Predictive mean matching method			
Intercept (β_0)	0.445	0.029	< 0.000
UROD_7(β_1)	-0.005	0.01	< 0.000
UROD_9 (β_2)	-0.008	0.127	< 0.001
Age (β_3)	0.03	0.156	< 0.000
ρ	0.68	0.011	< 0.000
Ψ_0	-2.8	0.09	0.000
Ψ_1	0.07	0.07	< 0.000
Ψ_2	0.31	0.09	< 0.000
σ	5.42	0.12	< 0.000
Regression method			
Intercept (β_0)	0.544	0.023	< 0.000
UROD_7(β_1)	-0.041	0.002	< 0.000
UROD_9 (β_2)	-0.001	0.0001	< 0.000
Age (β_3)	0.022	0.001	< 0.000
ρ	0.033	0.002	< 0.000
Ψ_0	-1.023	0.022	0.000
Ψ_1	0.456	0.010	< 0.000
Ψ_2	0. 124	0.003	< 0.000
σ	4.98	0.098	< 0.000

Based in the results in Table 6 and Table 7, the positive values for the parameter Ψ_2 imply that high values of the urgency are more likely to be missing. We can conclude that the null-hypothesis that, $\Psi_2 = 0$ cannot accepted; this may be evidence for non-random dropout. Also Ψ_1 is significantly different from 0. This indicates the importance of the response at the previous time point. Also handling MAR in covariates improves results instead of ignoring

them, as ignoring the MAR in covariates at Table 6, it was an insignificant effect of UROD_9 on urgency, after handling missingness with two MI methods the effect of UROD_9 on urgency become significant as in Table 7.

6. Conclusion and Future Work

Most literature, in longitudinal studies with missing values, focus on missingness in the longitudinal response or in the covariates. However, in practice it is possible to have missingness in the longitudinal response and in the covariates at the same time. This article proposes two methods to deal with missingness in both the longitudinal response and covariates at the same time. In this paper we proposed a selection model (Diggle and Kenward, 1994) for longitudinal data with non-ignorable missing values of the response with multiple imputation to missingness on covariates. The proposed model covers the case of the dropout missingness. The obtained likelihood function is intractable and not easy to be maximized. To overcome this difficulty, we suggest using the Stochastic EM algorithm. The proposed approach is applied to a data set from (The Interstitial Cystitis Data Base (ICDB)). The approach can be easily implemented in many fields where the missingness process is suspected to be non-ignorable. The case of intermittent pattern, which has less attention in the literature compared to the dropout, is a very challenging topic for future work.

Acknowledgments

The authors would like to thank Professor Kathleen Joy Propert at University of Pennsylvania, School of Medicine for providing guidelines about how to obtain the ICDB data. The ICDB data reported here were supplied by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Central Repositories.

Literature Cited

Abdelwahab, H. A., El Kholy, R. B. and Gad, A. M. (2019) Sensitivity analysis index for shared Parameter models in longitudinal studies. *Advances and Applications in Statistics*, 57, 1-20.

Allassonnière, S. and Chevallier, J. (2021) A new class of stochastic EM algorithms. escaping local maxima and handling intractable sampling. *Computational Statistics and Data Analysis*, 159, 107159

Celeux, G. and Diebolt, J. (1985) The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem, *Computational Statistics Quarterly*, 2, 73-82.

Darwish, N. M., Gad, A. M. and Hamid, R. M. (2020) Fitting longitudinal data with missing values in the response and covariates, *Advances and Applications in Statistics,* 64(2), 127-142.

Diebolt, J. and Ip, E. H. S. (1996) Stochastic EM: method and application. In: *Markov chain Monte carlo in practice.* (eds W.R. Gilks, S. Richardson and D. J. Spiegelhalter). Chapman and Hall, London. Chapter 15, pp. 259-273.

Diggle, P. and Kenward, M. G. (1994) Informative drop-out in longitudinal data analysis, *Journal of the Royal Statistical Society C*, 43, 49 – 93.

Efron, B. (1994) Missing data, imputation, and the bootstrap, *Journal of American Statistical Association*, 89, 463–475.

Erler, N. S., Rizopoulos, D., Rasmalen, V., Jaddoe, V. W., Franco, O. H., and Leasffre, E. M. H. (2016) Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and full Bayesian approach, *Statistics in .Medicine,* 35, 2955-2974.

Gad, A. M. and Ahmed, A. S. (2006) Analysis of longitudinal data with intermittent missing values using the stochastic EM algorithm, *Computational Statistics and Data Analysis,* 50, 2702 - 2714.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996) *Markov Chain Monte Carlo in Practices*, Chapman and Hall, London.

Grannell, A. and Murphy, H. (2011) Using multiple imputation to adjust for survey nonresponse, *Shifting the Boundaries of Research Proceedings*: 123-135. University of Bristol, UK.

Jennrich, R. I. and Schluchter, M. D. (1986) Unbalanced repeated measures models with structured covariance matrices, *Biometrics*, 42,805-820.

Little, R. J. A. (1995) Modelling the dropout mechanism in repeated measures studies, *Journal of American Statistical Association*, 90, 1112-1121

Louis, T. A. (1982) Finding the observed information matrix when using the EM algorithm. *Journal of Royal Statistical Society,* B 44, 226–232.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, Chapman and Hall, London.

Nooraee, N., Molenberghs, G., Ormel, J. and Heuve, E. R. V. D. (2018) Strategies for handling missing data in longitudinal studies with questionnaires, *Journal of Statistical Computation and Simulation*, 88 (17), 3415–3436.

Propert, K. J., Schaeffer, A. J. Brensinger, C. M. Kusek, J. W. Nyberg, L. M. and Landis, J. R. (2000) A prospective study of interstitial cystitis: results of longitudinal followup of the interstitial cystitis data base cohort, *The Journal of Urology*, 163, 1434-1439.

Rubin, D. B. (1976) Inference and missing data, *Biometrika* 63, 581-592.

_____(1987) *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons Inc., New York.

Salfran, D. and Spiess, M. (2015) *A Comparison of Multiple Imputation Techniques*. Discussion Paper No. 3, Universtat Hamburg, Germany.

Yang, Y., and Kang, J. (2010) Joint analysis of mixed Poisson and continuous longitudinal data with nonignorable missing values. *Computational Statistics & Data Analysis,* 54 (1), 193–207.

Yassen, A. S. and Gad, A. M. (2020) A stochastic variant of the EM algorithm to fit mixed (discrete and continuous) longitudinal data with nonignorable missingness, *Communication in Statistics - Theory and Methods*, 49(18), 4446-4467.

Implementing an Effective Survey Operations for a Research and Development Survey in the Philippines

Ramoncito G. Cambel[1], Dalisay S. Maligalig, Maurice C. Borromeo, and Ronald R. Roldan Jr.
Institute of Statistics, University of the Philippines Los Baños

Clifford B. Lesmoras
Education Development Center

ABSTRACT

Studies have shown that research and development (R&D) is a good driver of economic growth. Policies and programs that are based on good quality data are expected to produce better results. Hence, to formulate and implement policies and programs in R&D, good quality data is vital. A good data support system is also essential in identifying critical areas that need intervention, formulating viable approaches in addressing these issues, and allocating limited resources. In the Philippines, the Department of Science and Technology (DOST) has been conducting the Survey on Research and Development Expenditures and Personnel (R&D Survey) since 2003 so that R&D data and indicators can be compiled. To ensure that good quality R&D data and indicators are achieved, the DOST granted a research fund to the Institute of Statistics (INSTAT) of the University of the Philippines Los Baños (UPLB) in 2018 to further improve the design, conduct and analysis of the R&D Survey. This paper describes the processes that were developed and implemented through this research grant in relation to the dimensions of data quality, namely, relevance, accuracy, timeliness, accessibility, coherence, and comparability. Based on the evaluation of these processes, the paper also recommends further improvement on the survey operations of future rounds of the R&D Survey.

Keywords: *Survey on Research and Development Expenditures and Personnel, data quality*

1. Introduction

Research and development (R&D) comprise creative and systematic work undertaken to increase the stock of knowledge and to devise new applications of available knowledge (OECD, 2015). It is an important driver of sustainable economic development (Khan, 2015). Research has shown that R&D expenditure is significantly correlated with economic growth (Bayarcelic and Tasel, 2012) in developing countries. Akcali and Sismanoglu (2015) also established that R&D expenditures have a positive effect on the economic growth of both developing and developed countries. Sokolov-Mladenovic et al. (2016) confirmed that R&D investment has a positive effect on the real economic growth rate. Based on the data of 28 countries of the European Union for the period 2002 to 2012, the time when a financial crisis hit Europe, a percentage increase in the gross domestic expenditure on R&D (GERD) increased the GDP growth by more than 2 percentage points. Blanco et al. (2016) stated that R&D has a

[1] **Address correspondence to Ramoncito G. Cambel**: rgcambel2@up.edu.ph

significant effect on both the states' outputs in the United States and their respective total factor productivity in the long term.

Policies and programs to promote R&D must be formulated and implemented to propel economic growth. They can be developed with better quality data. Good quality data is also needed to identify critical areas that need intervention, formulate viable approaches in addressing these issues, and allocate limited resources.

At the national level, data and indicators that are important to national issues like economic growth, inflation, poverty, and employment, are compiled and published by the Philippine Statistics Authority (PSA). Because R&D indicators are not yet part of the Philippines' official statistics, the Department of Science and Technology (DOST) has been conducting the Survey on Research and Development Expenditures and Personnel (R&D Survey) since 2003 so that R&D indicators can be compiled. This survey captures the R&D expenditure and personnel data from three different sectors of the economy, namely, government, higher education institutions (HEIs), and private non-profit institutions (PNPIs) while the Philippine Statistics Authority (PSA) gathers some R&D data from the business and industry sector through the Annual Survey of Philippine Business and Industry (ASPBI). DOST combines the data from the R&D survey and ASPBI to provide national and regional estimates of the R&D indicators.

In the interest of achieving better quality data, the DOST gave a research grant to the Institute of Statistics (INSTAT) of the University of the Philippines Los Baños (UPLB) to further improve the design, conduct and analysis of the R&D Survey in December 2018. This paper describes the processes that were developed and implemented through this research grant to achieve good quality R&D data and indicators. Based on the evaluation of these processes, the paper also recommends further improvement on the survey operations of future rounds of the R&D Survey.

2. Data Quality and Survey Operations

The quality of data collected can be assessed using the six dimensions of data quality or quality of statistical outputs (Astrologo et. al., 2019) namely: relevance, accuracy, timeliness, accessibility, coherence, and comparability. These six dimensions were applied in planning the R&D Survey. Because R&D data and indicators are already used for planning and monitoring, their relevance is already demonstrated. Usually, accuracy is measured using the mean squared error of important estimates, say $\mathrm{MSE}(\theta)$, where θ is an estimator. Since $\mathrm{MSE}(\theta) = \mathrm{Var}(\theta) + \mathrm{Bias}^2(\theta)$, where $\mathrm{Var}(\theta)$ is the variance and $\mathrm{Bias}(\theta)$ is its bias of the estimator θ, the approach that is usually taken in designing probability sample surveys is to ensure that adequate sample size is achieved at a tolerable margin of error and best practices in survey operations are implemented so that bias is kept at a minimum. While the variance of an estimator can be estimated from a probability sample survey, bias can only be measured if credible external data sources are available, or another survey is conducted specifically for this purpose. Timeliness of survey results, comparability of estimates, coherence and accessibility can also be achieved through effective survey operations.

As shown in Figure 1, conducting a sample survey involves three major stages: planning, operations, and evaluation. The objectives of the survey, its contents and procedures are developed in the planning stage and the sample design and data collection plans are implemented in the survey operation stage. The third and last stage is evaluation to assess the quality of the survey data which is important for planning the next survey rounds, if any.

Survey operations include the construction of the sampling frame, selection of the sample, data collection, processing, estimation and analysis, and dissemination of survey results. Bias is controlled and hence, accuracy could be improved, if these tasks are done well and efficiently. Timely dissemination and access to data and indicators will also be achieved in the process. Coherence is attained when the concepts and definitions and determination of the target population are uniformly applied across time and space. Thus, the development of survey instruments, the construction of sampling frame, and the training of data collectors and supervisors are important tasks that need to be done well to establish coherence. Interpretability is gained when users could easily use and properly analyze the survey data. This implies that the meta-data – the attributes of characteristics of interest to be measured like definitions of concepts, units of measures, target population, and the limitation of the data should be made available to users.

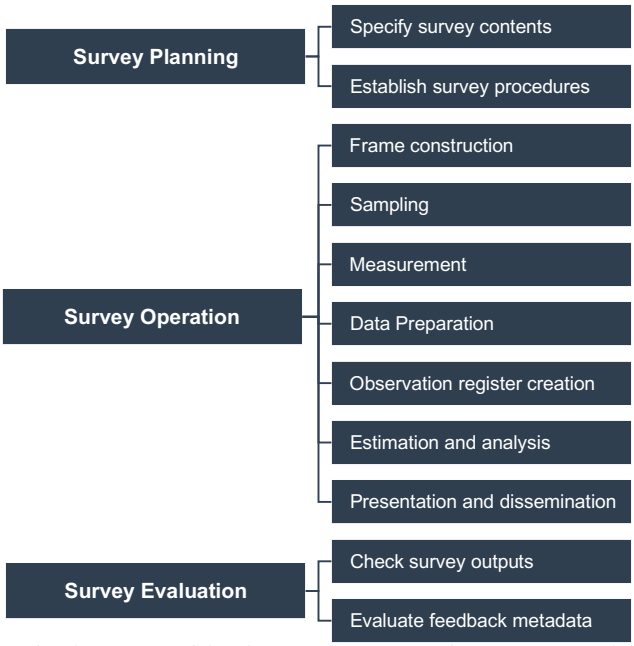

Figure 1. A visual summary of the phases and processes of a survey processing system. Source: Sundgren (1999)

2.1 Sampling Frame Construction

Survey operations begin with the construction of the sampling frame. This task is contingent upon the target population, which in the case of the R&D Survey, consists of all institutions that undertook in-house research in a given reference period, the calendar year of 2018. Since the R&D survey was designed to be a probability survey, each institution that performs R&D should have a chance of being represented in the survey. This is possible if the sampling frame that is used for selecting the sample includes all the target population units. However, unlike many surveys in which there is a readily available sampling frame, there is no comprehensive list of institutions that undertook in-house research for government, HEIs, and PNPIs. An initial sampling frame was developed from the following data sources:

For the government sector, online searches, and visits to websites of various government agencies and offices were done. An updated list of government institutions was constructed with information about their head agency, head of the office, address, telephone number, and email address.

For the HEIs, the Commission on Higher Education (CHED) provided a list of HEIs for the academic year 2017-2018. The list, however, needed updating since some contact information like the head of the institution and email address were outdated or missing for some HEIs especially for the satellite campuses of university and college systems. Moreover, new HEIs that were identified from online searches which were not found in the list were added and HEIs found to have ceased their operations were removed from the list. Updating the list was done by examining every HEI's website and contacting them through emails and telephone calls.

For the PNPIs, a list of non-stock, non-profit organizations that meet established minimum criteria for non-government organization (NGO) governance and accountability found in the Philippine Council for NGO Certification (PCNC) website was combined with the initial list of PNPIs who responded in the previous R&D surveys. Institutions known to conduct R&D but are not part of the initial list were also added to the list.

Because the web searches and other research that were done did not render a strong indication that all the institutions in the initial sampling frame have in-house R&D, it is quite likely that some of these institutions are not eligible to take part in the survey. If many ineligible institutions are selected in the sample, then the resulting survey estimates may not be as precise as planned. To improve the efficiency of the initial sampling frame, a two-phase survey was implemented, in which Phase 1 would screen out institutions that did not have any in-house research while the R&D Survey questionnaire would be administered in Phase 2 on the selected sample drawn from the improved sampling frame from Phase 1.

The Phase 1 survey was conducted online using SurveyMonkey, an online survey platform that allows the creation and sending of questionnaires in electronic format. The use of this online platform allowed respondents to complete the survey at their convenience. It resulted in a very low participation rate as shown in Table 1. It turned out that despite the preliminary activities described above, many sampled institutions did not have updated contact information. There were also changes in the management hierarchy of some institutions. Moreover, some of them also ignored the invitation to participate in the R&D Survey that was sent through email because they do not consider them as an official invitation. These cases are examples of cultural and technological practices of population units that should be considered when designing a survey so that costs and data quality are optimized (De Leeuw, Hox, and Dillman, 2008). In the context of the R&D survey, there should have been more analysis that was done on how the telephone or Internet surveys are viewed by the respondents from the various sectors – government, HEIs, and PNPIs.

Because of the low response rates in the Phase 1 of the survey, the desired improvement in the initial sampling frame did not materialize. Only 194 of the 3,378 institutions in the initial list responded, and of these, only 107 institutions have in-house R&D. To improve the sampling frame, imputations based on credible external sources were done. The 107 institutions with in-house R&D from those that responded form the initial list in the revised sampling frame. The respondents in the previous rounds of the R&D Survey were reviewed and those that did not respond to the Phase 1 survey were added in the sampling frame. Both CHED and DOST also provided lists of institutions that were given research grants. These were also added if they were not yet in the initial list. Web searches were also done to identify additional institutions that have published research or some other indications of R&D like seminars, workshops, news releases about research studies that they did.

Table 1 summarizes the resulting counts of institutions that may have R&D in the reference period in the different phases of sampling frame construction. Note that there was an increase in the number of institutions in the sampling frame at the end of Phase 2 for both government and PNPIs. The additional institutions were added when they were only identified during survey operations as having R&D. This addition to the sampling frame did not affect the sampling design since institutions in both government and PNPIs were selected with certainty.

Table 1. Number of institutions that undertake R&D by sector

Sector	Initial list	Phase 1 results		Imputations*			With R&D at the end of Phase 1	With R&D at the end of Phase 2
		Responded	With R&D	DOST R&D List	CHED/ DOST Projects	Web search		
GOV	670	73	32	88 (88)	21 (31)	14 (15)	155	263
HEI	2,354	100	66	269 (298)	43 (171)	487 (530)	865	865
PNPI	354	21	9	29 (32)	0 (0)	24 (24)	62	66

* The numbers enclosed in parentheses indicate the number of institutions found in the external data source, whereas the number above is the number of institutions in the external source that was captured in addition to the ones identified in Phase 1 or using the previously used external source, e.g., data sources from CHED and DOST listed 171 HEIs that were granted research projects, of which 43 HEIs were found to perform R&D but were not captured in Phase 1 and the DOST R&D list.

Sources: Challenges in Designing and Implementing Research and Development Surveys in the Philippines (Maligalig et al., 2019), 2018 R&D Survey Report (Maligalig et al., 2021)

The imputations did not guarantee that all institutions in the revised sampling frame have R&D and hence, the status of the sampled institutions must be identified so that adjustments for non-coverage error can be introduced after the survey. A system for monitoring the responses of the institutions and identifying their eligibility status was developed and implemented.

The approach of combining various data sources for constructing the sampling frame that was taken is similar to the study of Arora et al. (2021). The sampling frame of small and medium-sized enterprises (SMEs) was constructed by using patent, search engine, and website data. The resulting sampling frame of innovative SMEs did not have substantial coverage error. Similarly, the Italian National Statistical Institute (Istat), uses various data sources in creating a sampling frame for the Italian business R&D survey. These include previous Istat R&D surveys, other Istat business surveys with R&D-related questions, the Italian register of active enterprises, the Italian register of R&D performing institutions, data on national and European Union funding to research projects, patent databases, business reports, and data from the Italian Tax Agency (Istituto Nazionale Di Statistica, 2016).

2.2 Sampling

The planned sampling design was implemented after the revised sampling frame was finalized. For the government and PNPIs, a census of all institutions was employed since their population sizes are manageable. For the 865 HEIs in the sampling frame, proportionate stratified random sampling was employed. Since research has shown that HEIs with a higher number of graduate students are likely to have R&D, HEIs were stratified according to the graduate student size.

Those with at least 1,000 graduate students were grouped as "Large HEIs", and those with less than 1,000 were grouped in the "Small HEIs" stratum. Those that did not have data on the number of graduate students were classified as "Unknown". All the 55 HEIs in the "Large HEIs" stratum were included in the sample.

Sample sizes for the two strata (Small HEIs and Unknown HEIs) were computed using the formula:

$$n = \frac{Z_{\alpha/2}^{\ 2} PQ}{d^2 + \frac{Z_{\alpha/2}^{\ 2} PQ}{N}}$$

where $Z_{\alpha/2}$ is the abscissa of the standard normal distribution given $(1-\alpha)100\%$ confidence level; N is the population size; P is the proportion of a major characteristic of interest; $Q = 1 - P$; and d is the margin of error. Since P is usually unknown, it can be assumed based on prior information about its value from previous surveys or studies, but it can also be set to $P = 0.50$ to produce the sample size that would result to the most conservative estimate of the population variability. Considering several scenarios, the sample sizes were determined using a margin of error of 0.05 and a level of confidence of 0.95 to ensure a greater balance between the resources and the precision of estimates. Out of 355 Small HEIs, 193 were selected. Meanwhile, 218 out of 455 Unknown HEIs were selected. Regional allocations were done proportionately.

2.3 Measurement

2.3.1 Questionnaire Design

While the sampling frame was being constructed, the questionnaires were also being developed. A short questionnaire was developed for the Phase 1 survey. Questions included whether the institution performed in-house R&D during the reference year, the contact details of personnel who are most knowledgeable about their R&D activities, information on whether institutions implement centralized or decentralized reckoning systems to monitor their R&D activities, administrative units of institutions that independently conducted R&D and the incumbent heads of these units. The administrative unit is defined as the institution's constituent entity that has a certain degree of autonomy to perform its respective mandates. It is usually characterized by the appointment of an official designated to lead, supervise, and/or oversee the unit's activities. Identification of administrative units from the institutions is important so that appropriate estimates can be derived.

For the higher education sector, additional questions about the total number of graduate faculty and students were added to the Phase 1 questionnaire. As explained in the sampling part, this information was used in the stratification scheme in the sampling design of the said sector. For the government and private non-profit sectors, the Phase 1 survey collected information if the institution provided R&D funds to other institutions in 2018. In general, the information derived from Phase 1 and the records of the previous survey rounds served as the basis for identifying the target institutions in Phase 2.

In the case of the survey proper, the questionnaire for the previous survey rounds, the UNESCO Institute for Statistics (UIS) recommended template and the required data items to be collected were considered in redesigning the questionnaire. As per UIS (2014), the survey questionnaire must include a minimum number of basic questions on R&D activity. The questionnaire should

be simple and short, logically structured, and provide clear definitions and instructions including explanatory notes and hypothetical examples. Recommended guidelines indicated in the Frascati Manual (OECD, 2015) served as a reference of the definitions and classifications of R&D personnel and expenditures that were included in the questionnaires. This ensures that the indicators derived from the survey are of international standards.

A consultative workshop with key stakeholders was held in which the plan for the redesign of the questionnaire was discussed. After extensive consultations, it was decided that the questionnaire be streamlined to reduce the respondent's burden and increase the response rate. Instead of requiring the respondents to enumerate the research projects and personnel involved of institutions, summary data at the institution level became the data requirements in the survey. Although there were changes in the questions, the data collected could still measure the R&D indicators generated in the past surveys.

Definitions and key concepts were included in the questionnaires like basic concepts of research and development, the characteristics of activities that can be considered as R&D and types of activities that should be included or excluded as part of R&D, definitions of R&D personnel and personnel, types of personnel, percentage of time spent on R&D, research expenditures according to accounting categories, type of research, fields of science, and socio-economic objectives. They helped improve the accuracy of responses and minimize the risk of committing measurement error.

The revised questionnaire was pre-tested on some institutions in government, private sector, and HEIs. It was found that the questionnaire requires data from different sources within the institutions and hence, completing it requires substantial time and effort, and coordination among different units of an institution like the human resources and personnel, accounting, and planning units.

2.3.2 Data Collection

The modes of data collection of these surveys are primarily contingent on the organizational structure of sectors. Commonly, data on R&D are collected through multiple R&D surveys, which can be consolidated in coming up with national estimates. As per the UNESCO Institute of Statistics (UIS) in 2014, prevailing norms that govern information exchange, showing an understanding of the way that organizations may guard their information assets are of consideration in conducting R&D surveys.

For instance, in the annual U.S. Higher Education Research and Development (HERD) Survey, respondents may answer through a paper survey or using the web-based data collection system. For both methods, telephone and email follow-ups were employed to increase the participation rate. For the Survey of Industry Research and Development (SIRD), which is an annual sample survey that intends to include or represent all for-profit R&D-performing companies, either publicly or privately held, respondents are mailed with the survey. After a certain number of days, letters and telephone follow-ups are made. Lastly, for the annual Survey of State Government Research and Development (SGRD), respondents respond thru a self-administered questionnaire. Same with HERD and SIRD, telephone and e-mail follow-up with survey respondents are done.

In Canada, the Annual Survey of Research and Development in Canada Industry (RDCI) collects data primarily through an electronic questionnaire while providing respondents with the option of receiving a paper questionnaire, replying by telephone interview, or using other electronic filing methods. Data collected in this survey are supplemented with information that is extracted from administrative files. In addition to RDCI, the Survey on Research and Development in the Higher Education Sector (RDHES) and the Annual Survey of Research

and Development of Canadian Private Non-profit Organizations (RDNP) are also carried out to gather R&D data from higher HEIs and PNPIs, respectively, which are not covered in the RDCI. In RDNP, survey questionnaires are mailed out to target respondents and followed up with a phone call to verify receipt. Institutions who have not yet responded to the survey are followed up thru telephone up to five times, with effort devoted to organizations that are believed to perform R&D

In the 2018 Philippine R&D Survey, the questionnaires were intended to be administered online with telephone and personal follow-ups. This approach would reduce data processing and validation because online survey applications allow for self-administered questionnaire and automated validation of responses. Online probability surveys, however, require that sampled units must have updated email addresses and access to the Internet. Phase 1 of the survey was launched in the last week of May 2019. Weblinks to the questionnaires, the endorsements from the DOST Secretary and the CHED Chair for HEIs, were included in the invitation to participate in the survey that was sent to respondents through email. To increase the participation in the survey, telephone follow-ups were conducted two weeks after the launching of the online survey. These follow-ups were also used to update the contact information of the sampled institutions. Several rounds of telephone follow-ups were carried out but as shown in Table 1, the response rate was quite low, and thus, the planned two-phase sampling design had to be modified

The objective of the Phase 1 survey to eliminate the ineligible institutions in the initial sampling frame was not achieved and hence, imputations on the eligibility status of institutions were done. Because the likelihood of getting ineligible sampled units has not decreased, a monitoring system that identifies and records ineligible units was established. This procedure is presented in Figure 2. Each reply that is received is examined and its eligibility status is determined. If the respondent sent a fully filled-out questionnaire, with both expenditure and personnel data, then it is classified as eligible. If critical expenditure data is missing, a telephone follow-up by the designated supervisor or survey manager is done to check whether the responding institution has undertaken in-house R&D or not. If the institution is deemed ineligible, a letter stating its ineligibility status is requested for documentation purposes. There were also HEIs that would have several campuses in different locations that were included in the sample, but they share a common research fund with the main campus managing the R&D budget, coordinating the research activities, and outputs. In this case, the main campus is deemed eligible while the other campuses were declared ineligible. When a sampled institution conveys that it did not undertake in-house R&D, a telephone follow-up is also done to make sure that the respondent understands the definition of R&D and the reference year before the final eligibility status is determined.

The Phase 2 survey questionnaire was rolled out via SurveyMonkey in the first week of September 2019. Separate questionnaires were sent to administrative units of institutions that advised in the Phase 1 that they have independent administrative units doing R&D. Telephone follow-ups were done one month after the launch of the survey. At most three attempts were made by the telephone enumerator to each non-responding sampled institution.

Because of the lower-than-expected response rate in the last quarter of 2019, the online survey mode had to be adjusted. Based on the telephone follow-ups that were done, it turned out that many institutions do not recognize email correspondence as official, and hence, they ignored the email invitations to participate in the R&D Survey. In January 2020, formal letters enclosed with a printed copy of the questionnaire, a link to the SurveyMonkey questionnaire and, the option of using a digital copy of the questionnaire to reply were sent to the institutions through the post. Endorsement letters from the DOST Secretary, and the CHED Chairperson, in the case of HEIs, were also included. The endorsements were important in getting the trust and cooperation of the sampled institutions. Just as more responses were beginning to come in, the

COVID-19 pandemic struck which slowed down not only the planned training of data collectors that were intended to supplement the telephone follow-ups.

In addition to the telephone follow-up by INSTAT, selected DOST regional staff were also requested to help in following up and collecting the completed questionnaires from sampled institutions in their respective area. A training program was conducted to orient the data collectors about the R&D Survey, their responsibilities, and the role of INSTAT as technical support to facilitate efficient collaboration. In the training, quality practices in collecting, validating, and encoding responses in the R&D survey were discussed. Information on the follow-ups that INSTAT had already made with the institutions were forwarded to the assigned data collectors so as not to disrupt the flow of communication. Illustrations of common errors and inconsistencies in the accomplished questionnaires were also given to guide them in validating the collected responses. Because of the implementation of the community quarantine restrictions in almost all the regions beginning 16 March 2020, in-person training was conducted only for CALABARZON and the Zamboanga Peninsula, while virtual training was organized for the other regions.

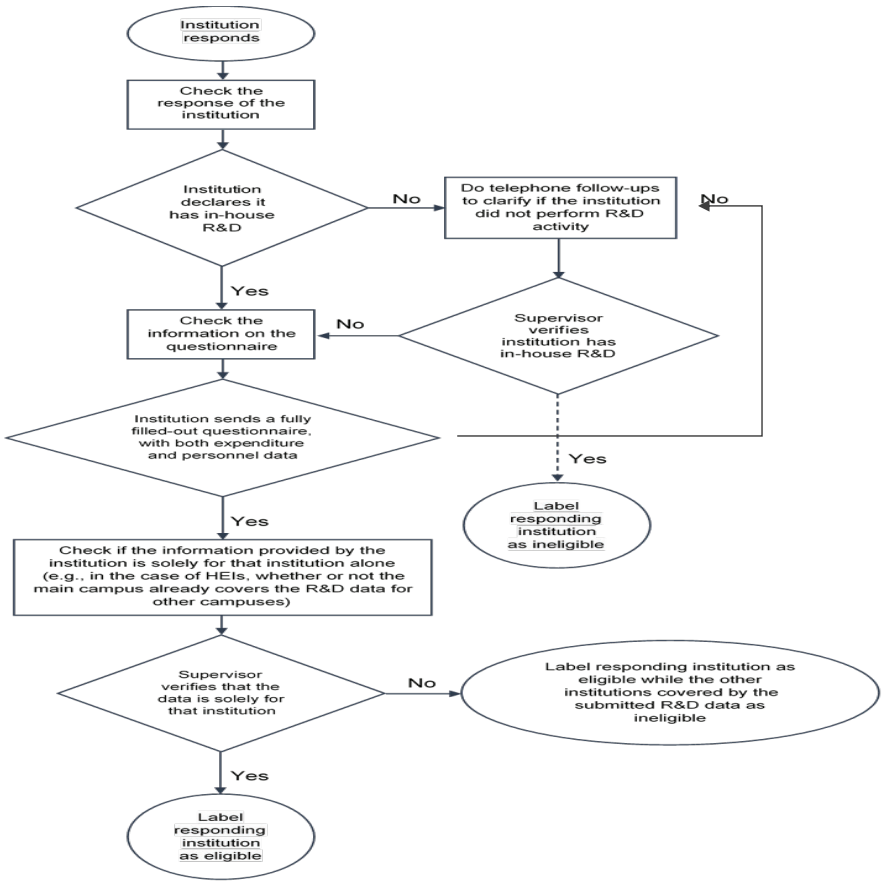

Figure 2. Process flowchart for distinguishing and validating eligibility status of responding institution.

An INSTAT survey supervisor was designated for each region to oversee the data collectors. These survey supervisors, in close collaboration with the lead data collectors for each region, served as the focal persons in providing updates to the INSTAT survey team. Regular team meetings were held to discuss and address critical issues that were encountered in the field.

Regular updates on the status of the survey were communicated to the data collectors to ensure that all responses were accounted for in the survey. The field follow-ups closed in November 2020 but a grace period on the submission of accomplished questionnaires was set until January 2021. By then, higher than targeted response rates were achieved per region and at the national level. A national-level response rate of 81% (computed as the non-weighted percentages of institutions that responded either as eligible or ineligible over the sample size) was achieved at the end of the data collection (Figure 3).

As discussed, adjustments in survey mode were adapted in the duration of the survey operation as needed. These resulted in improved response rates across the months of the implementation of Phase 2 of the survey. For instance, a very low constant series of response rates were observed from September 2019 until February 2020. An increase in the response rate for March 2020 can be attributed to the sending of questionnaires to the identified institutions through the post. For April and May, a dismal increase in the response rate is due to the adjustments of institutions due to the COVID-19 pandemic. Though most of the regional staff of DOST were already engaged as data collectors in March 2020, the effect of the various community lockdowns implemented in the country can be seen on the response rates. The response rate started to consistently increase beginning June 2020, when all the data collectors have been trained and deployed.

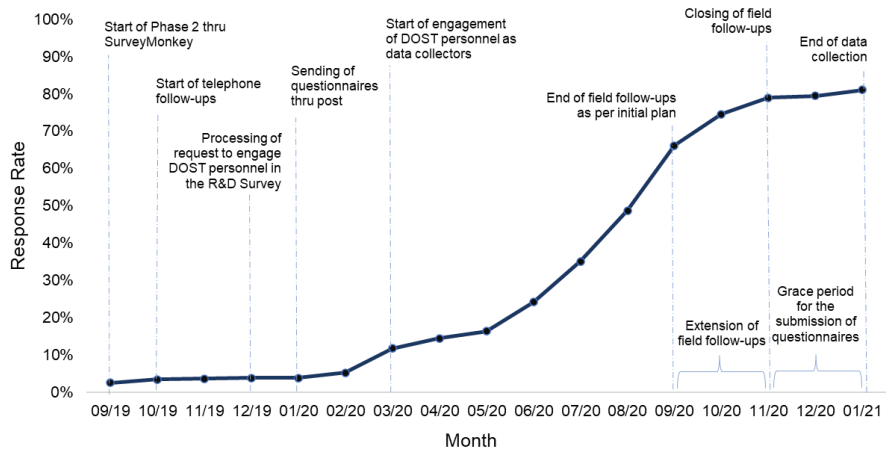

Figure 3. Response rates of Phase 2 of the 2018 R&D Survey.

2.4 Data Preparation and Observation Register Creation

To ensure that data quality is maintained, appropriate data processing and validation procedures were implemented. A coding system was developed to ensure that each institution in the sampling frame is uniquely identified, and all meta-data are properly recorded. The monitoring system that was described above was also automated so that weekly reports can be generated.

Data from the questionnaires were entered through SurveyMonkey. Scanned copies of the questionnaires were also systematically stored so that suspect data can be easily verified. Guidelines for encoding completed questionnaires were given to the data collectors to minimize coding errors. Periodic data validation was conducted with each round of data processing and preliminary data analysis, gaps between the identified eligible institutions in the monitoring form and the encoded responses in the database were flagged for the appropriate action of the designated supervisors. Institutions found to have submitted their accomplished questionnaires but were still not encoded in the database were identified. Responses recorded in the database with suspicious eligibility status in the monitoring form were clarified. Multiple responses of institutions in the database were individually cross-checked with the completed questionnaires to identify the correct entry. When all collected responses were accounted for in the database and data cleaning was completed, responses of reporting administrative units of institutions with decentralized management system were consolidated into institution-level data. Figure 4 shows the total number of eligible and ineligible institutions and nonresponding institutions.

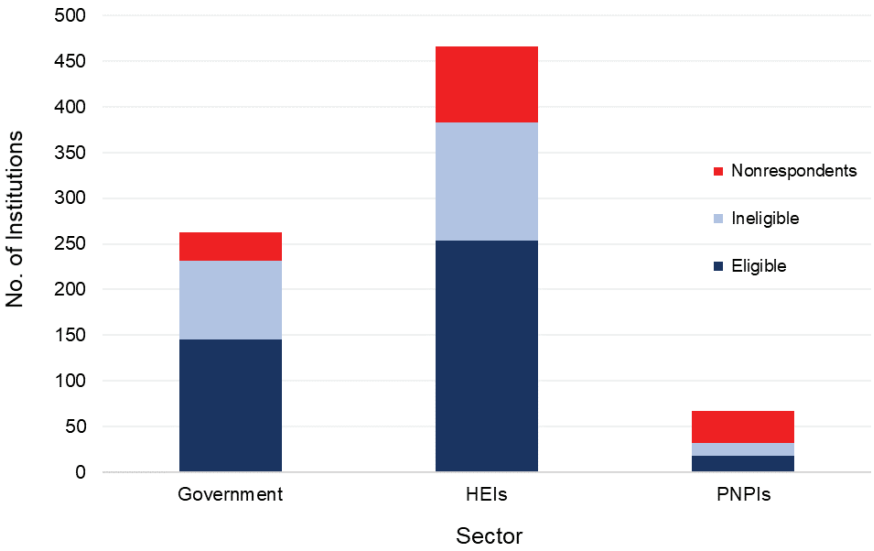

Figure 4. Distribution of the eligibility status of sampled institutions across sectors.

As shown in Figure 4, more than half of the PNPIs did not respond to the R&D survey, while less than 20% of the institutions in the government and HEIs did not respond. In terms of eligibility status, about a third of the responding institutions were deemed ineligible. These statistics were used in the adjustments for nonresponse and coverage errors.

Looking closely at the data, the management of research activities and financial reckoning systems differed widely across sampled institutions. In general, large institutions with very decentralized research management systems are likely not able to complete the R&D Survey.

At the end of the data validation process, institution-level responses were produced. Finally, design variables, like region, stratum, stratum sizes for finite population correction, and survey weights, were incorporated into the data to perform the analysis.

2.5 Estimation and Analysis

Final survey weights were determined as the product of the base weights that were computed as planned and adjustments to compensate for nonresponse and coverage errors. The R&D indicators that were computed were mostly totals and proportions, like total R&D expenditure and percent share of government R&D expenditure. To provide a measure of the precision of the estimates, the corresponding standard errors were also computed using Taylor Series Linearization (TSL), which is a variance estimation method that renders robust results even for nonlinear estimators, like subpopulations means and totals. TSL is the default variance estimation procedure in many reputable statistics software like R, Stata, and SAS. All computations in this survey report were executed using the 'survey' package in R. The complete methodological notes on the weighting adjustments and estimation implemented in the survey can be found in the 2018 R&D Survey Report that was disseminated to stakeholders (Maligalig et al., 2021).

Statistical tables were generated for all the R&D indicators at the national level, by sector, and by region. Indicators for the HEIs were also disaggregated by public and private HEIs. The corresponding standard errors of all estimates were also incorporated in the statistical tables in the survey report. Graphs were generated for the most important survey results. The estimates were also compared with those from previous R&D survey rounds. A thorough review of the estimates was undertaken before writing the survey report.

2.6 Presentation and Dissemination

The survey report was submitted to the National Research Council of the Philippines (NRCP), the monitoring agency for this research project, and DOST for review. All comments were discussed, and the survey report was finalized accordingly. The results of the survey were presented in two webinars – for the DOST staff and management and another for the respondents, UPLB constituents, and the data collectors. The survey report can also be downloaded from the INSTAT website (instat.uplb.edu.ph). Printed copies of the survey report were also sent to the responding institutions and data collectors.

A complete set of data files were turned over to NRCP and DOST. The set included an anonymized survey data file that can be turned into a public utility file that can be shared with researchers, the sampling frame, and the data dictionary.

3. Summary and Conclusions

The most challenging task in the survey operations is the construction of the sampling frame because there is no central database from which data on all the institutions that undertake R&D can be extracted. Data and information from many sources were combined to construct the sampling frame. Because there is no guarantee that all the institutions in the sampling frame are eligible, a monitoring system that could distinguish eligible and ineligible respondents was incorporated in the workflow so that appropriate weighting adjustments to compensate for coverage errors can be derived. When the response rate was low at the beginning months of the survey, a more flexible combination of survey modes was adapted. These innovations led to an effective survey operation for the R&D Survey and data of good quality was obtained.

All the statistical tables in the survey report contained the corresponding measure of the precision of the estimates being presented so that users can do their evaluation. Bias due to non-response and coverage errors were mitigated through appropriate weighting adjustments. Measurement errors were controlled by applying proven good practices. The changes

implemented in the questionnaire also helped in controlling for the non-sampling errors in the survey.

To ensure the accessibility and clarity of the information in the R&D survey, dissemination of the results through two webinars, distribution of survey report, and provision of a downloadable version of the survey report were done. Moreover, anonymized survey data that can be turned into a public utility file was turned over to DOST for possible sharing with researchers.

In terms of timeliness, there had been delays in the proposed survey activities. Though the implementation of the mixed survey modes was successful, still it has room for improvement. For instance, instead of starting the data collection approach using the online mode, it may be better to send by post the sampled institutions the questionnaire, invitation to participate and endorsement letters since most institutions have yet to consider emails and online surveys as official communication.

For the coherence and comparability dimensions, data across time and space were achieved by employing the same set of concepts and definitions. Streamlining the questionnaires based on the Frascati manual and based on the review of various R&D questionnaires ensured the coherence and comparability of the R&D survey.

4. Recommendations

The R&D questionnaire can be expanded to enable aggregates by sex and personnel characteristics such as age group, highest educational attainment, and the field of specialization. Moreover, the number of graduate students, especially the number of Ph.D. candidates can be collected to indicate the potential sources of future researchers. The size of the graduate class can also be included in the questionnaire to help sharpen the estimation strategy. The inclusions of questions on other forms of dissemination of completed research and outputs must also be explored to enable more in-depth analysis.

The construction of sampling frames for government, HEIs, and PNPIs were one of the most challenging activities that were undertaken to ensure that the 2018 R&D Survey is a probability sample survey that can render robust estimates. To reduce this difficult burden for the next survey rounds, a centralized R&D institution database that stores information on the institutions with in-house research must be maintained. Updating of this database can be done at the point of entry – when DOST or other government agencies or PNPIs provide a grant to a research project. In this regard, the collaboration with other research-granting institutions such as CHED, Department of Health (DOH), Department of Education (DepEd), Department of Agriculture (DA), Department of Environment and Natural Resources (DENR) should be strengthened. Following the endorsement of DOST, CHED, and DOH that were given to the 2018 R&D Survey, other government agencies can also encourage their respective research units and grantees to cooperate and participate in the next R&D Survey rounds.

While the 2018 R&D Survey paved for some methodological innovations in terms of data collection and analysis, there are still areas in the survey design and operations that need further improvement so that better quality data can be achieved. Of the three sectors covered by the R&D Survey, PNPIs have the lowest response rate. Despite the low response rate, both the total number of R&D personnel and expenditures have considerably increased compared to 2015 indicating that coverage errors may have been reduced. There were 12 regions that do not have PNPIs. While there could really be no PNPIs in all these regions that undertake in-house research, it is still worth exploring possible approaches that could improve the sampling frame for PNPIs. PNPIs that have R&D activities must be engaged by DOST through regular consultations so that they can also be included in the database. There must also be an

information campaign to let institutions know the importance of R&D in our economy and to encourage them to participate in the R&D Survey

The mixed survey mode that was implemented turned out to be successful. Because of this approach, and with a streamlined questionnaire and the cooperation of the DOST regional offices staff that played a critical role in implementing the mixed-mode approach, the data collection for the 2018 R&D Survey was completed at about 81% response rate. Instead of starting the data collection approach using the online mode, it may be better to send by post the sampled institutions the questionnaire, invitation to participate and endorsement letters since most institutions have yet to consider emails and online surveys as official communication. The link to the online survey can be given in the letter to offer the sampled institutions an alternative way of responding to the questionnaire.

The DOST regional offices, as well as regional development councils, must also be engaged in updating the R&D institution database. Given the help of the DOST regional staff in improving the response rates of the 2018 R&D Survey, they must be engaged at the onset of the survey to help update the database mentioned above as well as undertake field operations for the R&D Survey.

DOST can develop application software that can be used for managing research activities in government, HEIs, and PNPIs. The software should enable institutions to store and manage all the information regarding their research activities. The software can also generate reports that can be used by the institutions to manage their research agenda as well as, for completing the R&D Survey. Large institutions with decentralized research management systems can use this application software to consolidate the R&D information from different reporting administrative units and consequently, manage their research agenda more effectively. The creation of the application software will further improve the response rate and reduce measurement errors of future R&D Survey rounds.

Literature Cited

AKCALI, B.Y., and SISMANOGLU, E. 2015. Innovation and the Effect of Research and Development (R&D) Expenditure on Growth in Some Developing and Developed Countries. Procedia – Social and Behavioral Sciences, Vol. 195, pp. 768-775.

ARORA, S.K, KELLEY, S., and MADHAVAN, S. 2021. Building a Sample Frame of SMEs Using Patent, Search Engine, and Website Data. Journal of Official Statistics, Vol. 37, No. 1, 2021, pp. 1–30, http://dx.doi.org/10.2478/JOS-2021-0001

ASTROLOGO, C.J., BAÑARES, J.S., GARRAEZ, J.A.H., ROBLES, M.P., and BANTANG, J.A.O. 2019. Evaluating the Quality of Statistics Produced by the Philippine Statistics Authority Using the UN NQAF Assessment Tool. Paper presented at the 14th National Convention on Statistics, Manila, Philippines. October 1-3, 2019.

BAYARCELIK, E.B., and TASEL, F. 2012. Research and Development: Source of Economic Growth. Procedia - Social and Behavioral Sciences 58, pp. 744 – 753.

BLANCO, L.R., GU, T., and PRIEGER, J.E. 2016. The Impact of Research and Development on Economic Growth and Productivity in the U.S. States. Southern Economic Journal, Vol. 82, No. 3, pp. 914-934. Southern Economic Association.

DE LEEUW, E.D., HOX, J.J., and DILLMAN, D.A. 2008. In E.D. De Leeuw, J.J. Hox, and D.A. Dillman (Ed). International Handbook of Survey Methodology, pp. 1-17.

ISTITUTO NAZIONALE DI STATISTICA. 2016. The Behaviour of Respondents While Filling in a Web Questionnaire: The Case of the Italian Business R&D Survey. M. Masselli, A. Nuccitelli, and A.L. Palma. Istat Working Papers.

KHAN, J. 2015. The Role of Research and Development in Economic Growth: A Review. Journal of Economics Bibliography, Vol. 2, Issue 3, pp. 128-133.

MALIGALIG, D.S., BORROMEO, M.C., CAMBEL, R. G., ROLDAN, JR., R. R. and LESMORAS, C. B. 2019. Challenges in Designing and Implementing Research and Development Surveys in the Philippines. Paper presented at the 14th National Convention on Statistics, Manila, Philippines. October 1-3, 2019.

MALIGALIG, D.S., BORROMEO, M.C., CAMBEL, R. G., ROLDAN, JR., R. R. and LESMORAS, C. B. 2021. 2018 R&D Survey Report. Department of Science and Technology, and Institute of Statistics.

NATIONAL RESEARCH COUNCIL. 2015. Measuring Research and Development Expenditures in the U.S. Nonprofit Sector: Conceptual and Design Issues, Summary of a Workshop. C. House, H. Rhodes, and E. Sinha, Rapporteurs. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

NATIONAL SCIENCE FOUNDATION. Survey of Industrial Research and Development. Retrieved from https://www.nsf.gov/statistics/srvyberd/prior-descriptions/overview-sird.cfm

OECD. 2015. Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development, The Measurement of Scientific, Technological and Innovation Activities, OECD Publishing, Paris. DOI: http://dx.doi.org/10.1787/9789264239012-en

SOKOLOV-MLADENOVIĆ, S., CVETANOVIĆ, S., and MLADENOVIĆ, I. 2016. R&D expenditure and economic growth: EU28 evidence for the period 2002–2012, Economic Research-Ekonomska Istraživanja, 29:1, 1005-1020, DOI: https://doi.org/10.1080/1331677X.2016.1211948

SUNDGREN, B. 1999. Information Systems Architecture for National and International Statistics Offices - Guidelines and Recommendations, United Nations Statistical Commission and Economic Commission for Europe, Geneva.

THE AMERICAN ASSOCIATION FOR PUBLIC OPINION RESEARCH. 2015. Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 8th edition. AAPOR.

UNESCO INSTITUTE FOR STATISTICS. 2014. Guide to Conducting an R&D Survey: For countries starting to measure research and experimental development. Downloaded from: http://dx.doi.org/10.15220/978-92-9189-151-1-en

UNIVERSITY OF READING. 2009. International Household Survey Network Survey Quality Assessment Framework SQAF.

Analytic Hierarchy Process with Rasch Measurement in the Construction of a Composite Metric of Student Online Learning Readiness Scale

Joyce DL. Grajo, James Roldan S. Reyes[1], Liza N. Comia, Lara Paul A. Ebal, Jared Jorim O. Mendoza, and Mara Sherlin DP. Talento
Institute of Statistics, University of the Philippines Los Baños

ABSTRACT

This paper developed the Online Learning Readiness Composite Scale (OLRCS), a composite measure of student online learning readiness based on five dimensions, namely (1) computer/internet self-efficacy; (2) self-directed learning; (3) learner control; (4) motivation for learning; and (5) online communication self-efficacy. A single metric of online learning readiness has its advantage over its disaggregated dimensions. For one, it allows a summative description of each student which school administrators can use for an effective student targeting toward flexible learning. Rasch Analysis (RA) was performed to come up with an objective measure for each dimension while Analytic Hierarchy Process (AHP) was applied to aggregate the computed Rasch scores of the five dimensions. Three OLRCS have been constructed using weights generated by (1) teacher participants, (2) student participants, and (3) combined student and teacher participants. Results showed that motivation for learning consistently received the highest weight while online communication self-efficacy and computer/internet self-efficacy got low weights among the three OLRCS. Research findings also showed that student participants gave more importance to learner control than self-directed learning, unlike the teacher participants. The difference in the teacher and student perspectives merits detailed attention to optimize the online learning environment and enable individual support. Nevertheless, using cluster analysis, the distribution of students who are ready, undecided, or not ready for online learning is similar to the three constructed OLRCS.

Keywords: *multidimensional latent variable; multi-criteria decision analysis; linear aggregation*

1. Introduction

The benefits of online learning in formal education are highlighted due to the current worldwide situation. Horton (2006) defined online learning as the use of information and computer technologies and systems to build and design learning experiences. It covers a range of technologies such as the world wide web, email, chat, new groups and texts, and audio/video conferencing delivered over computer networks to impart education. Thus, it relies on the learner's own pace, according to their own convenience (Dhull and Sakshi, 2017).

A wide variety of literature states that online learning has various critical success factors (CSF). Identification of these CSF plays a huge role to boost the successful implementation of

[1] **Address correspondence to James Roldan S. Reyes:** jsreyes7@up.edu.ph

an online learning system. Some researchers identified student characteristics as a major success factor of online learning, above and beyond the teacher and the institution (Selim, 2007; Rohayani and Kurniabudi, 2015; Alhabeeb and Rowley, 2018). A recent study by Reyes et al. (2021) focuses on assessing students' online learning readiness, which is an important student characteristic to consider in online learning. The student's readiness for online learning must be assessed to improve content quality (Cigdem and Yildirim, 2014) and to predict the student's success in online learning (Doe et al., 2017).

Extensive literature reviews agreed that students' online learning readiness is a multidimensional metric. Kirmizi (2015) considered five dimensions which include computer/internet self-efficacy, self-directed learning, learner control, motivation for learning, and online communication self-efficacy. Motivation for learning is defined as the student's learning attitude while learner control refers to a student's control over learning efforts to direct his/her learning. Moreover, computer/internet self-efficacy refers to a student's ability to demonstrate proper computer and internet skills while self-directed learning refers to the student's responsibility for learning context to reach learning outcomes. Lastly, online communication self-efficacy defines the student's adaptability to ask questions, respond, give insights, and participate in discussions online. Using the Online Learning Readiness Scale (OLRS), which was developed and validated by Hung et al. (2010), it was found that motivation for learning is the most important predictor of success, followed by learner control. On the other hand, six dimensions were used by Yilmaz (2017) to determine the online learning readiness of students. His research applied the OLRS developed by Demir and Yurdugül (2015) which consists of six dimensions, similar to Hung et al. (2010) but separated online self-efficacy from computer self-efficacy. Moreover, the study by Watkins et al. (2004) assessed the readiness for online learning by using their developed self-assessment instrument. This instrument has six dimensions namely, technology access, online skills and relationships, motivation, online audio/video, internet discussions, and importance to success. Although the developed instrument was deemed reliable based on Cronbach's alpha coefficient, there was no validity measure used to assess the readiness for online learning. Dray et al. (2011) also developed a tool to assess student readiness for online learning. Unlike the previously mentioned research, student readiness was defined by only two dimensions namely learner characteristics and technology capabilities. There are also two dimensions, such as self-management of online learning and comfort with online learning, which were considered by Smith (2005) to develop a tool that was both reliable and valid by using Cronbach's alpha and factor analysis (FA), respectively, for online learning readiness. Moreover, in the study of Ascari et al. (2005), online learning readiness has two dimensions, namely perceived ease of use and perceived usefulness.

Since dimensions related to online learning readiness vary across various types of research, Demir and Yurdugül (2015) explored a reference model by desk reviewing various studies which developed tools for assessing student readiness for online learning. The created model includes six dimensions, namely competency of technology usage, self-directed learning, access to technology, confidence in prerequisite skills, motivation, and time management. This research adopted the OLRS by Hung et al. (2010) in measuring student readiness for online learning since their instrument wholly captured Demir and Yurdugül's (2015) reference model. Some studies which applied OLRS were those by Kaymak and Horzum (2013), Cigdem and Yildirim (2014), Kirmizi (2015), Buzdar et al. (2016), Kayaoglu and Akbas (2016), Elnakeeb and Khalifa (2016), Engin (2017), Cavusoglu (2019), Mutambik et al. (2018), Fearnley and Malay (2021), and Reyes et al. (2021).

Although student online learning readiness is a multidimensional metric, aggregation of its dimensions to come up with a composite measure has an intuitive appeal, especially for policy or decision-makers (Phelps et al., 2018). A composite measure provides a summary picture of the multiple facets or dimensions of complex, multidimensional phenomena in a way

that facilitates evaluation and comparison (Becker et al., 2017). This composite measure serves as an overall measure or assessment that has not been fully established in online learning readiness yet. In the case of an overall measure of online readiness, the capabilities of the higher education system were assessed to introduce and implement online learning readiness programs. The low value of the overall online readiness of a university suggested a serious deficit of some e-readiness factors (Darab and Montazer, 2011). This paper, however, aims to construct a composite measure of student readiness for online learning which takes into account all five dimensions in OLRS (Hung et al., 2010), namely (1) computer/internet self-efficacy; (2) self-directed learning; (3) learner control; (4) motivation for learning; and (5) online communication self-efficacy.

In constructing a composite measure, a multi-criteria decision analysis (MCDA) such as the Analytic Hierarchy Process (AHP) is used so that the relative importance of each dimension of a characteristic/variable can be established (Kil et al., 2016; Blagojevic et al., 2019). A composite measure provides decision-makers with an explicit view of the dimensions that should be prioritized. Moreover, knowing the degree of student readiness through a single metric provides school administrators valuable insights for effective student targeting. Several kinds of literature related to education applied the AHP technique. Sael et al. (2019) implemented AHP in profiling students. Their objective was to detect and classify the most important factors that increase Moroccan students' dropout and failure. Further, Anis and Islam (2015) reviewed the AHP application in higher learning institutions. Their systematic analysis found that AHP was often applied to measure quality education, evaluate faculty members, measure performance, strategically plan, and choose a university and select university students. The AHPs in their study were applied together with other methods according to the objectives of the study. One example was measuring the quality of education by Yeşim and Ortaburun (2011). They applied AHP together with the Spearman Rank Correlation test to redesign the undergraduate curriculum of one of the faculties of Marmara University. On the other hand, in the case of using AHP in strategic planning, Begičevič et al. (2007) applied AHP together with factor analysis in modeling the systematic implementation of e-learning and online education distance education at a university in Croatia. Their study concluded that organizational readiness, which includes the university framework and faculty strategy for development and financial readiness served as the most influential criterion in implementing online learning in this institution. In this paper, AHP was used along with Rasch analysis.

Rasch measurement was performed to come up with an objective measure for each of the five dimensions wherein a linear transformation of the ordinal raw score was obtained and expressed in logits. These scores can be used for parametric analyses given that distributional assumptions are met. Hence, this research is an attempt to combine the five dimensions of Hung et al.'s (2010) OLRS using AHP with Rasch scores. The resulting composite metric can be used to assess the factors that contribute to students' readiness for e-learning. Moreover, the procedure demonstrated in this paper can be replicated to apply to future investigations related to the construction of composite metrics for other multidimensional latent variables.

2. Methodology

2.1 Data source

This study made use of the data collected by Reyes et al. (2021) using Hung et al.'s (2010) OLRS. This scale has five dimensions, namely (1) internet/computer self-efficacy; (2) self-directed learning; (3) learner control; (4) motivation for learning; and (5) online communication self-efficacy. Online learning readiness was measured through a series of statements rated on a Likert scale with a score of 1 for a strong disagreement and 5 for a strong

agreement. In the study by Reyes et al. (2021), the results showed that the overall Cronbach's alpha coefficient is 0.89. This suggests that the OLRS has a high level of interaction between Filipino higher education students and items. Further, Fearnley and Malay (2021) confirmed the OLRS' reliability in the local context wherein Cronbach's alpha coefficient in each dimension ranged from 0.727 to 0.871 with an overall internal consistency of 0.889. Hence, OLRS is highly reliable and can be adopted in studies involving Filipino higher education students.

The subjects involved a stratified random sample of 290 students from the University of the Philippines Los Baños (UPLB). The sample size was determined using a margin of error of ± 0.04, a confidence level of 0.95, and a design effect of 0.50. Proportional allocation was implemented wherein the sample students were stratified according to program classification with 80% undergraduate and 20% graduate students.

The weights for the composite measure of online learning readiness were obtained by asking the AHP participants, who were four faculty members and seven students, to rank the dimensions in terms of what they perceived as relatively more important to the scale. This was done by pairwise comparisons of the five dimensions with the more important dimension given a higher rating than the others. At the end of the process, two sets of ratings were collected, one from the teachers, and another from the students. These AHP participants were chosen based on their length of experience in online education. The teacher and student participants were from different institutions and fields of study, with an average of three years of experience in online teaching and learning, respectively. Since inputs in AHP are obtained from experts, it does not need a statistically significant sample size to generate robust results (Dias and Ioannou, 1996; Doloi, 2008). Aside from simplicity and a high level of consistency, one of the well-known reasons for using AHP is its applicability for small sample sizes even a judgment from a single expert is already a representative (Darko et al., 2019). In this research, from the list of identified participants, the final number of experts, which is 11 participants, was determined based on their willingness to participate in the study.

2.2 Construction of a composite metric for OLRS

A. Rasch analysis

The Rasch model measures a latent trait based on the functional trade-off between a person's trait and item difficulty in a series of questions. This study specifically used the following Rasch-Andrich (RA) Rating Scale model:

$$B_n - T_j - F_k = ln\left(\frac{P_{njk}}{P_{nj(k-1)}}\right) \tag{1}$$

where it specifies the probability, P_{njk}, that student n of online readiness B_n is observed in category k of a rating scale applied to item j of difficulty T_j as opposed to the probability $P_{nj(k-1)}$ of being observed in category (k–1). Moreover, F_k is the Rasch-Andrich threshold. Since Rasch analysis is capable of transforming the ordinal responses into interval measures (expressed in logits), it can then be used in the parametric analysis. In this paper, the online learning readiness score D_i for each dimension (where i = 1, 2, 3, 4, 5) was generated by the RA model. The higher the logit score D_i, the more ready the student is for the i^{th} dimension. One logit or one natural log of odds ratio refers to the distance along the line of the online readiness scale that increases the odds of being ready that is specified in the measurement model by a factor of 2.718. For example, an item with a logit score of +1.0 means that this item increases the odds of a student being in the k^{th} level of readiness as compared to being in the

(k-1)th level of readiness multiplicatively by $e^{+1.0} = 2.178$. A logit unit of zero means a neutral response. Data processing was done using WINSTEPS 4.5 and AHP Software (Goepel, 2018).

B. Analytic Hierarchy Process

An important part of the construction of composite metrics for OLRS is the assignment of weights, W_i to be used in aggregating the five dimensions. This research recruited inputs of participants from the field of online education, four teachers and seven students, to generate these dimension weights using the Analytic Hierarchy Process (AHP). The process starts with participants establishing a pairwise priority among the five dimensions. Using these pairwise priorities, a matrix of pairwise comparisons of the five dimensions was constructed, with whole odd numbers from 1 to 9 given to those with higher priorities and its inverse to its less prioritized partner. The even numbers in between were assigned as a compromise. The scale of comparisons is presented in Table 1.

Table 1. The scale used for the comparison of each dimension (Saaty and Vargas, 1991)

Scale	Degree of importance
1	Equal importance
3	Moderate importance of one dimension over another
5	Strong or essential importance
7	Very strong importance
9	Extreme importance
2, 4, 6, 8	Intermediate values
Reciprocals	Values for inverse comparison

The resulting matrix was normalized by dividing each cell by its corresponding column total. The average scores in the rows are the resulting weights. These series of comparisons produced the normalized principal eigenvector (or priority vector) from which weights of each dimension were obtained. Higher weight, W_i, means greater importance of the dimension as perceived by the participants. To check if the assumption of matrix consistency is satisfied, the consistency ratio (CR) was computed and checked. A CR of at most 0.10 is considered acceptable; otherwise, the judgments often need reexamination, that is, the ratings should be revisited by transforming them (Saaty, 1987; Teknomo, 2006; Franek and Kresta, 2014). Reducing the square root of the 1 to 9 scale gives higher consistency.

The compatibility of the weighing method is considered in deciding the suitability of the aggregation method. According to Nardo et al. (2008), AHP as a weighing method is compatible with either linear or geometric aggregations. The simplest and most widely used linear aggregation was used to combine the five dimensions since it preserves the relative importance of the dimension as reflected by the generated weights. The constructed composite metric satisfies the preferential independence condition of a linear aggregation, i.e., the composite metric permits the assessment of the marginal contribution of each dimension separately (Munda, 2012). Further, since linear aggregation possesses a full compensability property, this means that a student with a low readiness score in some dimensions can still be compensated by sufficiently high scores in other dimensions to tweak his/her composite score. The formula to compute the composite metric of OLRS (OLRCS*)* is:

$$OLRCS = \sum_{i=1}^{5} W_i D_i \tag{2}$$

where W_i and D_i are the weight and Rasch scores of the i^{th} dimension, respectively.

Finally, cluster analysis was done on the OLRCS to classify students into three groups namely, (1) not ready for online learning; (2) undecided; (3) ready for online learning.

2.3 Evaluation of the constructed composite metric for OLRS (OLRCS)

A perceived overall student readiness was also obtained during the survey done by Reyes et al. (2021). To assess the relationship between the OLRCS and the perceived overall student readiness, the Spearman correlation coefficient was obtained. A positive correlation between the two scores indicates the same direction, i.e., as the perceived overall student readiness rating increases, OLRCS also increases. A low correlation, however, suggests that overall online readiness cannot be measured through personal perception but requires an objective measure like the OLRCS.

The robustness of the OLRCS is assessed by sensitivity analysis, deleting one dimension at a time and recomputing the OLRCS for each deletion (Dating, 2019). The result of each OLRCS re-computation was compared with the original OLRCS. A low percentage of matched results in the resulting confusion matrix will indicate sensitivity to that particular dimension.

3. Results and Discussion

3.1 Reliability and validity of OLRS

The OLRS by Hung et al. (2010), which this study adopted, was tested for reliability and validity as administered to higher education students in the Philippines. Rasch principal components analysis of residuals was used to check the construct validity of the instrument. This was done by weighing the index of raw variance explained by the instrument over the total raw variance of the sampled responses. Using all the 18 questionnaire items, Table 2 showed that 50% of the index of raw variance is more than 40%, the standard set by Fisher (2007) and Adams et al. (2018). However, when all 18 questionnaire items were pooled together, the unidimensionality property of the construct was not achieved since the eigenvalue in the first contrast is equal to 2.52, exceeding the threshold of less than 2 (Raîche, 2005). On the other hand, the index of raw variance in each dimension ranges from 56% to 68% and their eigenvalues in the first construct are all within the threshold value. This means that each dimension exhibited a good unidimensional scale to effectively measure the online readiness of the students.

Table 2. Standardized Residual Variance in Eigenvalue Units

Dimension	Eigenvalue	Observed
All		
Raw variance explained by measures	17.99	50.00%
Unexplained variance in 1st contrast	2.52	7.00%
Computer/Internet Self-Efficacy		
Raw variance explained by measures	5.21	63.50%
Unexplained variance in 1st contrast	1.67	20.30%
Self-directed Learning		
Raw variance explained by measures	6.20	67.40%
Unexplained variance in 1st contrast	1.62	17.60%
Learner Control		
Raw variance explained by measures	6.39	56.10%
Unexplained variance in 1st contrast	1.52	13.30%
Motivation for Learning		
Raw variance explained by measures	5.33	57.10%
Unexplained variance in 1st contrast	1.50	16.10%
Online Communication Self-Efficacy		
Raw variance explained by measures	4.47	59.90%
Unexplained variance in 1st contrast	1.54	20.50%

Further, Table 3 shows the item reliability of each OLRS dimension. With all item reliabilities having a value of at least 0.90 (or separation index of at least 3.00), the sample is large enough to confirm the construct validity of OLRS (Linacre, n. d.). This means that the set of items in each dimension is well-targeted (Wright and Stone, 1999; Linacre, 2012). These results supported the results of the studies made by Kirmizi (2015), Hung et al. (2010), and Yilmaz (2017) that student readiness for online learning is a multidimensional measure. Table 3 also summarized the person reliability of all dimensions which are all below 0.80 (or separation index below 2.00). This implies that OLRS is not sensitive enough to distinguish between more or less-ready students. Though the person reliability of the four dimensions ranged from 0.71 to 0.76, which is somewhat close to the standard, additional survey items can be included to improve the OLRS (Linacre, n. d.), particularly in the learner control dimension (person reliability = 0.56).

Table 3. Person and Item Reliability Measures

Dimension		Student	Item
Computer/ Internet Self-Efficacy (CS)	Separation	1.55	7.33
	Reliability	0.71	0.98
Self-directed Learning (SL)	Separation	1.70	7.15
	Reliability	0.74	0.98
Learner Control (LC)	Separation	1.14	15.00
	Reliability	0.56	1.00
Motivation for Learning (ML)	Separation	1.79	7.48
	Reliability	0.76	0.98
Online Communication Self-Efficacy (OS)	Separation	1.57	5.60
	Reliability	0.71	0.97

The results of the rating scale analysis presented in Table 4 revealed that the 5-rating scale, from strongly disagree to strongly agree, was used approximately as modeled, based on the fit statistics having values close to 1 (Fisher, 2007; Adams et al., 2018). However, some categories were overly improbable, particularly those in dimensions of computer/internet self-efficacy (category: strongly disagree) and learner control (category: strongly agree). A few students expressed poor levels of readiness on items in the computer/internet self-efficacy dimension that they were expected to be ready for. Similarly, some students expressed high levels of readiness on items in the learner control dimension that they were expected to be less ready for. Both the Rasch-Andrich thresholds and the average measures exhibited the desired monotonically ascending pattern across categories indicating that the categories represent advancing levels of online learning readiness (Linacre, n. d.).

Overall, the OLRS showed a considerable ability to measure student readiness for online learning. In particular, the instrument may be improved by adding items that can measure student readiness in the learner control dimension.

Table 4. Results of the Rating Scale Analysis

Dimension/Category Label	Average Measure	Infit MNSQ	Outfit MNSQ	Andrich Threshold
Computer/Internet Self-Efficacy (CS)				
1	-2.42	1.53	1.44	NONE
2	-1.44	0.78	0.65	-3.77
3	0.42	1.14	1.13	-0.69
4	2.26	0.90	0.86	-0.39
5	4.72	0.99	0.89	4.85
Self-directed Learning (SL)				
1	-2.02	0.99	0.98	NONE
2	-0.75	0.99	1.05	-2.49
3	0.14	1.01	1.23	-0.51
4	1.17	0.84	0.84	0.01
5	2.36	1.10	1.06	2.99
Learner Control (LC)				
1	-2.53	0.97	0.98	NONE
2	-1.48	0.78	0.94	-2.51
3	-0.12	0.78	0.94	-0.71
4	1.26	0.82	1.00	0.01
5	1.94	1.55	1.35	3.21
Motivation for Learning (ML)				
1	-2.46	1.25	1.12	NONE
2	-0.80	0.94	1.00	-3.14
3	0.34	0.86	0.78	-0.99
4	2.10	0.84	0.94	-0.06
5	4.03	1.16	1.10	4.19
Online Communication Self-Efficacy (OS)				
1	-2.56	1.18	1.11	NONE
2	-1.25	0.99	1.08	-3.34
3	0.11	0.78	0.75	-0.39
4	1.25	0.97	0.96	0.33
5	2.40	1.12	1.08	3.40

3.2 Five dimensions of OLRS using Rasch Score

From the paper by Reyes et al. (2021), Table 5 shows that students were found to be quite ready in the computer/internet self-efficacy (mean = +2.3 logit) and motivation for learning (mean= +2.06) dimensions, while students were not ready in terms of controlling their learning program and the environment under the learner control (mean= -0.08 logit) dimension. For the self-directed learning and online communication self-efficacy dimensions, a neutral response was obtained. For each dimension, the following are the items where students felt least ready based on item difficulty: (1) not being distracted by other online activities while learning online (LC); (2) having confidence in knowledge and skills of how to manage online learning platforms (CS); (3) managing time well (SL); (4) having the motivation to learn (ML); and (5) feeling confident in posting questions in online discussions (OS). On the other hand, the following are the items where students were most ready in each dimension: (1) repeating the online instructional materials based on their needs (LC); (2) being open to new ideas (ML); (3) having confidence in performing the basic functions of Microsoft Office programs or their

counterparts (CS); (4) seeking assistance when facing learning problems (SL); and (5) improving from their mistakes.

Table 5. Student Readiness for online learning in Rasch scores in each dimension with their corresponding minimum and maximum values of item difficulty

Dimension	Mean	Standard Deviation (SD)	Item Difficulty	
			Minimum	Maximum
Computer/ Internet Self-Efficacy (CS)	2.30	2.67	-1.11	1.10
Self-directed Learning (SL)	0.65	1.52	-0.75	1.09
Learner Control (LC)	-0.08	1.53	-1.42	1.72
Motivation for Learning (ML)	2.06	2.44	-1.13	1.03
Online Communication Self-Efficacy (OS)	0.08	2.01	-0.49	0.68

3.3 A composite metric for OLRS using Analytic Hierarchy Process

Table 6 shows that all identified participants in the field of online learning gave ratings with acceptable consistency ratios (CR) of less than 0.10. The lowest CR are 0.031 and 0.017 for teacher participants and student participants, respectively.

Table 6. Consistency Ratio (CR) of each participant

Number	Participants	CR
1	Teacher	0.071
2	Teacher	0.076
3	Teacher	0.031
4	Teacher	0.087
5	Student	0.090
6	Student	0.017
7	Student	0.073
8	Student	0.037
9	Student	0.022
10	Student	0.062
11	Student	0.088

Table 7 shows the weights of each dimension for teacher participants, student participants, and combined student and teacher participants (overall). In the case of combined participants, the dimension of motivation for learning has the highest weight (0.317) followed by self-directed learning (0.228), and learner control (0.200). On the other hand, online communication self-efficacy and computer/internet self-efficacy have low weights of 0.147 and 0.108, respectively. The weights of teacher participants were consistent with that of combined participants. Meanwhile, the weights for dimensions of learner control and self-directed learning have different orders of magnitude for student participants. The CRs for teacher participants, student participants, and combined participants are acceptable.

Table 7. AHP weights and its Consistency Ratios

Dimension	Weights Teacher	Weights Student	Weights Combined
Computer/Internet Self-Efficacy (CS)	0.086	0.122	0.108
Self-directed Learning (SL)	0.282	0.200	0.228
Learner Control (LC)	0.171	0.217	0.200
Motivation for Learning (ML)	0.311	0.317	0.317
Online Communication Self-Efficacy (OS)	0.150	0.144	0.147
Consistency Ratio	**0.028**	**0.004**	**0.005**

The perceived online readiness rating is moderately and positively correlated with Online Learning Readiness Composite Scale (OLRCS) as shown in Table 8, implying that as the perceived readiness score increases, OLRCS also increases. Moreover, there is an almost perfect direct correlation among OLRCS for teacher participants, student participants, and combined participants.

Table 8. Spearman's Rank Order Correlation Coefficient of perceived online readiness rating with OLRCS

	Perceived	Teacher	Student	Combined
Perceived	1.0000			
Teacher	0.5217*	1.0000		
Student	0.5243*	0.9958*	1.0000	
Combined	0.5223*	0.9982*	0.9993*	1.0000

*Significant at 5% level

Table 9 shows cluster analysis results to classify students according to their online learning readiness. Using OLRCS, the majority of the students (68%) are undecided on their online learning readiness having cluster centroids ranging from 0.66 to 0.70. Only one-fourth of the students are ready for online learning. The predetermined three clusters were justified since they gave the lowest variation within clusters.

Table 9. Cluster Analysis with Cluster Centroids and Cluster Sizes

	Classification		
OLRCS	**Not Ready**	**Undecided**	**Ready**
Teacher	-2.6136	0.6571	2.9146
	(19)	(197)	(74)
Student	-2.6520	0.7000	2.9883
	(19)	(198)	(73)
Combined	-2.6397	0.6818	2.9545
	(19)	(197)	(74)

Table 10 shows that the students who are similarly classified by their perception and OLRCS based on combined/teacher participants are 10, 75, and 46 for not ready, undecided, and ready, respectively. These comprise 45% of the students. However, 117 students (40%) perceive themselves as not ready for online learning. This is contrary to the results of OLRCS where only 19 students (7%) are classified as not ready.

Table 10. Distribution of respondents based on their perceived online readiness rating and OLRCS for teacher participants and combined participants

	OLRCS Teacher Participants/ Combined Participants			
Perceived	**Not Ready**	**Undecided**	**Ready**	**Total**
Not Ready	10	93	14	117
Undecided	6	75	14	95
Ready	3	29	46	78
Total	19	197	74	290

Similar distributions were obtained using OLRCS based on student participants except for one student who shifted from ready to undecided. This is shown in Table 11.

Table 11. Distribution of respondents based on their perceived online readiness rating and using OLRCS for student participants

	OLRCS Student Participants			
Perceived	**Not Ready**	**Undecided**	**Ready**	**Total**
Not Ready	10	93	14	117
Undecided	6	75	14	95
Ready	3	30	45	78
Total	19	198	73	290

Figure 1 shows the results of the sensitivity analysis. If the dimension of online communication self-efficacy is removed from the OLRCS model, at least 96% of the students are consistently classified as ready, undecided, and not ready by both models. This high percentage indicates that the model with four dimensions still provides almost the same predictions compared to the model where all five dimensions are present. On the other hand, if the dimension of motivation for learning is removed from the OLRCS model, at least 85% of the students are consistently classified as ready, undecided, and not ready by both models. This relatively low percentage implies that the model without the dimension motivation for learning provides around 15% prediction inconsistency against the model where all five dimensions are present.

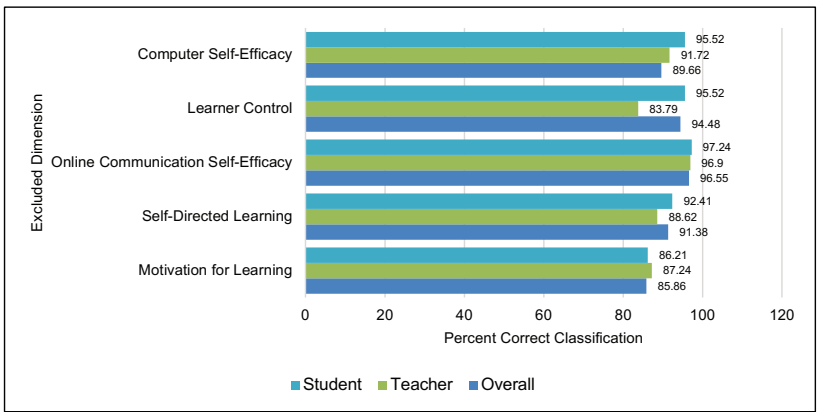

Figure 1. Distribution of students who are similarly classified using the original OLRCS model and with one dimension excluded

4. Conclusions and Recommendations

The research findings prove that personal perception cannot effectively quantify a student's overall online learning readiness. It requires an objective measurement through the use of the OLRS, whose reliability and validity have been supported by numerous literature. Most students perceive themselves as not ready while OLRCS classified the majority of them as undecided. The results differ because OLRCS considered the multidimensionality of the construct, unlike the perceived online readiness rating. Further, the classification of a student using OLRCS also accounted for the readiness of other students in the group.

The Rasch analysis (RA) model was used to generate scores per dimension while the Analytic Hierarchy Process (AHP) was utilized to assign weights to aggregate the five dimensions. Three composite metrics of OLRS (OLRCS) have been constructed using weights generated by (1) teacher participants, (2) student participants, and (3) combined student and teacher participants. The weights generated by the teacher participants ranked the dimensions in the following order: (1) motivation for learning; (2) self-directed learning; (3) learner control; (4) online communication self-efficacy; and (5) computer/internet self-efficacy. The student participants ranked the five dimensions similarly except for learner control and self-directed learning. Student participants gave more weight to learner control than self-directed learning. Students feel that they need to be more prepared in learner control than in self-directed learning. This resulting difference in the teacher and student perspectives merits detailed attention to optimize the online learning environment and enable individual support. Nevertheless, combining the two gave rankings similar to that of the teacher participants, with weights across different dimensions not varying too much from each other. Generally, the results of classifying a student as whether s/he is ready, undecided, or not for online learning are similar regardless of the OLRCS used.

Sensitivity analysis validated the results of AHP on the weights of each dimension on the OLRCS. The dimension motivation for learning gave the largest alteration to the result once removed from the computation of OLRCS. This result validates that this dimension has the highest weight in the OLRCS model. Interestingly, the dimension computer/Internet self-efficacy displayed the second largest change in the percentage of similar classifications when removed from the OLRCS despite having been ranked last by the participants. In addition, the dimension of online communication self-efficacy remains to be the least important among the dimensions. In the face of increasing familiarity with computers and the Internet, self-efficacy in its usage continues to be a requisite for online learning. As a result, students can develop the confidence to effectively communicate online.

The OLRCS is useful in flexible learning specifically for planning and policy-making post-pandemic, such that school administrators can perform initial assessments of students to identify those who are ready or not for online learning. They may treat students who were classified as not ready and undecided to be both unprepared and find ways to make them ready for online learning. However, considering resource constraints such as time, budget, and labor, those who were not ready should be given priority. Moreover, a high proportion of undecided students requires further investigation like holding dialogues/consultations with such students. School administrators should provide a support structure for students' specific needs (Brooks, 2003), especially for those who are not ready and undecided. This may include devices and technological support and connectivity, financial assistance, and psychological support and guidance (Zincirli, 2021). Also, they should offer students with forthright information and proper advising (Brooks, 2003). For instance, schools may organize seminars and lectures to help these groups of students overcome the challenges they face during online learning and encourage them to seek assistance once they experience difficulty in online learning. In addition, school administrators should train and equip their teachers with the necessary skills to support and handle students in online learning (Murray, 2021). The training should include

how to communicate more openly with students who have such limitations and consider individual differences between students and their capacities to adapt to the online learning environment.

Moreover, the resulting weights of individual dimensions of online learning readiness can help teachers design and manage classes effectively. For instance, a module might be designed to increase a student's motivation for learning to stimulate and sustain student interest. Ease and confidence in computer use also play a vital role in deciding whether a student is ready for online learning. School administrators should create programs or activities that would support students with their ability to direct and control their learning. For instance, the school can invest in programs and activities for training teachers on various approaches to create interesting and engaging activities, which target lifelong learning and empower students to choose alternatives, and then assess the student's progress.

Finding associations and modeling relationships among critical factors can be constructed using the OLRCS instead of using individual dimensions. Such a model can help identify the extent of influence of key factors on online learning readiness and subsequently, its outcomes such as student performance, drop-out rate, etc. Combining Rasch analysis with AHP may lead to constructing composite metrics for other multidimensional latent variables like happiness, anxiety, and stress.

5. Limitations and Future Directions

This paper focused only on constructing a composite measure of student readiness for online learning which takes into account all five dimensions of the OLRS (Hung et al., 2010). It must be noted that this study adopted the OLRS without the intention of changing it. In addition, the study focused on higher education students; thus, similar research can be done to assess the online learning readiness of primary and secondary education students.

This study used the Rasch logit scores to come up with a measure of online learning readiness per dimension. The possibility of converting the OLRCS from Rasch logit score to positive interval scores can also be done. This will allow geometric aggregation which can be compared to the results of this study. Given the OLRCS, conducting concrete validation is recommended to further verify the scores.

Literature Cited

ADAMS, D., SUMINTONO, B., MOHAMED, A., and NOOR, N. S. M., 2018, E-learning Readiness among Students of Diverse Backgrounds in a Leading Malaysian Higher Education Institution, *Malaysian Journal of Learning and Instruction*, 15(2), 227-256.

ALHABEEB, A. and ROWLEY, J., 2018, E-learning Critical Success Factors: Comparing Perspectives from Academic Staff and Students, *Computers and Education*, 127, 1-12.

ANIS, A. and ISLAM, R, 2015, The application of analytic hierarchy process in higher-learning institutions: a literature review, *Journal of International Business and Entrepreneurship Development*, 8(2), 166-182.

ASAARI, A. H., HASMI, M., and KARIA, N., 2005, Adult Learners and E-learning readiness: A case study, 2005 European College Teaching & Learning Conference, Athens, Greece.

BECKER, W., SAISANA, M., PARUOLO, P. and VANDECASTEELE, I., 2017, Weights and Importance in Composite Indicators: Closing the Gap, *Ecological Indicators*, 80, 12-22.

BEGIČEVIČ, N., DIVJAK, B., and HUNJAK, T., 2007, Prioritisation of e-learning forms: a multi-criteria methodology, *Journal of Operations Research*, 15, 405–419.

BLAGOJEVIC, B., ATHANASSIADIS, D., SPINELLI, R., RAITILA, J., and VOS, J., 2019, Determining the relative importance of factors affecting the success of innovations in forest technology using AHP, *Journal of Multi-Criteria Decision Analysis,* 27(1-2), 129-140.

BROOKS, L., 2003, How the Attitudes of Instructors, Students, Course Administrators, and Course Designers Affects the Quality of an Online Learning Environment. *Online Journal of Distance learning Administration*, 6(4), 1-6.

BUZDAR, M. A.., ALI, A., and TARIQ, R. U. H., 2016, Emotional Intelligence as a Determinant of Readiness for Online Learning, *International Review of Research in Open and Distributed Learning*, 17(1), 148-158.

CAVUSOGLU, M., 2019, Online and Self-Directed Learning Readiness Among Hospitality and Tourism College Students and Industry Professionals, Dissertation, College of Education, University of South Florida.

CIGDEM, H. and YILDIRIM, O. G., 2014, Effects of Students' Characteristics on Online Learning Readiness: A Vocational College Example, *Turkish Online Journal of Distance Education*, 15(3), 80-91.

DARAB, B. and MONTAZER, G. A., 2011, An eclectic model for assessing e-learning readiness in the Iranian universities, *Computers & Education*, 56(3), 900-910.

DARKO, A., CHAN, A. P. C., AMEYAW, E. E., OWUSU, E. K. PÄRN, E., and EDWARDS, D. J., 2018, Review of Application of Analytic Hierarchy Process (AHP) in Construction, International Journal of Construction Management, 19(5), 436-452.

DATING, M. J. P., 2019, Mapping of vector-borne disease hotspots in the Philippines using official statistics and analytic hierarchy process, Unpublished Master's Thesis, University of the Philippines Manila.

DEMIR, O. and YURDUGÜL, H., 2015, The Exploration of Models Regarding E-learning Readiness: Reference Model Suggestions, *International Journal of Progressive Education*, 11(1), 173-194.

DHULL, I. and SAKSHI M. S., 2017, Online Learning. *International Education and Research Journal*, 3(8), 32-34.

DIAS JR., A., and IOANNOU, P. G., 1996, Company and Project Evaluation Model for Privately Promoted Infrastructure Projects, *Journal of Construction Engineering and Management*, 122(1), 71-82.

DOLOI, H., 2008, Application of AHP in Improving Construction Productivity from a Management Perspective, *Construction Management and Economics*, 26(8), 841-854.

DOE R., CASTILLO M. S. and MUSYOKA, M. M., 2017, Assessing Online Readiness of Students, *Online Journal of Distance Learning Administration* 20(1).

DRAY, B. J., LOWENTHAL, P. R., MISKIEWICZ, M., and MARCZYNSKI, Z., 2011, Developing a Tool for Assessing Student Readiness for Online Learning: A Validation Study, *Distance Education*, *32*(1), 29-47.

ELNAKEEB, M. and KHALIFA, S. M. A., 2016, The Relationship between Online Learning Readiness and Social Interaction Anxiety among Nursing Students in Alexandria University, *World Journal of Nursing Sciences*, 2(3), 140-152.

ENGIN, M., 2017, Analysis of Student's Online Learning Readiness on Their Emotional Intelligence Level, *Universal Journal of Educational Research*, 5(12A), 32-40.

FEARNLEY, M. R. and MALAY, C. A., 2021, Assessing Students' Online Learning Readiness: Are College Freshmen Ready?, *Asia-Pacific Social Science Review*, 21(3), 249-259.

FISHER, W. P. J., 2007, Rating scale instrument quality criteria, Rasch measurement transactions 21(1): 1095.

FRANEK, J. and KRESTA, A., 2014, Judgment scales and consistency measure in AHP, Procedia Econ. 12, 164-173.

GOEPEL, K. D., 2018, Implementation of an Online Software Tool for Analytic Hierarchy Process (AHP-OS), *International Journal of Analytic Hierarchy Process*, 10(3).

HUNG, M., CHOU, C., CHEN, C., and OWN, Z., 2010, Learner Readiness for Online Learning: Scale development and Student Perceptions, *Computers & Education*, 55, 1080–1090.

HORTON, W., 2006, E-Learning by Design, Pfeiffer: San Francisco, CA, USA, ISBN -13.

KAYAOGLU, M. C. and AKBAS, R. D., 2016, Online Learning Readiness: A Case Study in the Field of English for Medical Purposes, *Participatory Educational Research,* 4, 212-220.

KAYMAK, Z. D. and HORZUM, M. B., 2013, Relationship between Online Learning Readiness and Structure and Interaction of Online Learning Students, *Educational Sciences: Theory & Practice*, 13(3), 1792-1797.

KIL S. H., LEE, D. K., KIM, J. H., LI, M. H., and NEWMAN, G., 2016, Utilizing Analytic Hierarchy Process to Establish Weighted Values for Evaluating the Stability of Slope Revegetation based on Hydroseeding Applications in Korea, *Sustainability*, 8(58), 1-17.

KIRMIZI, O., 2015, The Influence of Learner Readiness on Student Satisfaction and Academic Achievement in an Online Program at Higher Education. *Turkish Online Journal of Education Technology*, 14(1), 133-142.

LINACRE, J. M., n. d., Andrich Thresholds: Disordered Rating or Partial Credit Structures, https://www.winsteps.com/winman/disorder.htm. Accessed 03 April 2020.

LINACRE, J. M., 2012, A User Guide to Winsteps Ministep Rasch model Computer Programs: Program manual 3.75.0. http://www.winsteps.com/a/winstepsmanual.pdf. Accessed 03 April 2020.

MUNDA, G., 2012, Choosing Aggregation Rules for Composite Indicators, *Social Indicators Research*, 109(3), 337-354.

MURRAY, B., 2001, What Makes Students Stay. eLearn Magazine. Retrieved from https://elearnmag.acm.org/featured.cfm?aid=566901

MUTAMBIK, I., LEE, J., and FOLEY, Y., 2018, Identifying the Underlying Factors of Student's Readiness for E-learning in Studying English as a Foreign Language in Saudi Arabia: Students' and Teachers' Perspectives, *Proceedings of the 2018 Computing Conference*, 267-280.

NARDO, M., SAISANA, M., SALTELLI, A., and TARANTOLA S., 2008, Handbook on Constructing Composite Indicators: Methodology and Userguide, OECD.

PHELPS, C. E., LAKDAWALLA, D. N., BASU, A., DRUMMOND, M. F., TOWSE, A., and DANZON, P. M., 2018, Approaches to Aggregation and Decision Making—A Health Economics Approach: An ISPOR Special Task Force Report, Value in Health, 21(2), 146-154.

RAÎCHE, G., 2005, Critical eigenvalue sizes (variances) in standardized residual Principal Components Analysis (PCA). Retrieved from www.rasch.org/rmt/rmt191h.htm on 29 July 2020.

ROHAYANI, H. and KURNIABUDI, S., 2015, A Literature Review: Readiness Factors to Measure E-learning Readiness in Higher Education, *Procedia Computer Science*, 59, 230-234.

REYES, J. R. S., GRAJO, J. DL., COMIA, L. N., TALENTO, M. S. DP., EBAL, L. P. A. and MENDOZA, J. J. O., 2021, Assessment of Filipino Higher Education Students' Readiness for e-Learning During a Pandemic: A Rasch Technique Application, *Philippine Journal of Science*, 150(3).

SAATY, R. W., 1987, The Analytic Hierarchy Process - What is it and how it is used, *Mathematical Modelling*, 9 (3-5), 161-176.

SAATY, T. L. and VARGAS, L. G., 1991, Prediction, Projection, and Forecasting. Boston: Kluwer Academic Publishers.

SAEL, N., HAMIM, T., and BENABBOU, F., 2019, Implementation of the Analytic Hierarchy Process for Student Profile Analysis, *International Journal of Emerging Technologies in Learning*, 14(15), 78-93.

SELIM, H. M., 2007, E-Learning Critical Success Factors: An Exploratory Investigation of Student Perceptions, *International Journal of Technology Marketing*, 2(2), 157-182. DOI:10.1504/IJTMKT.2007.014791

SMITH, P. J., 2005, Learning Preferences and Readiness for Online Learning, *Educational Psychology*, *25*(1), 3-12.

TEKNOMO, K., 2006, Analytic Hierarchy Process (AHP) Tutorial. https://Revoledu.com. Accessed 23 April 2021.

WATKINS, R., LEIGH, D., and TRINER, D., 2004, Assessing Readiness for E-learning, *Performance Improvement Quarterly*, *17*(4), 66-79.

WRIGHT, B. and STONE, M., 1979, Best Test Design, Chicago, IL: Mesa Press.

YEŞIM, Y.A. and ORTABURUN, Y., 2011, Redesigning curriculum in higher education by using analytical hierarchy process and spearman rank correlation test, *European Journal of Scientific Research*, 53(2), 271–279.

YILMAZ, R., 2017, Exploring the Role of E-learning Readiness on Student Satisfaction and Motivation in Flipped Classroom, *Computers in Human Behavior*, 251-260.

ZINCIRLI, M, 2021, School Administrators' Views on Distance Education During the Covid-19 Pandemic Process. Malaysian Online Journal of Educational Technology, 9(2), 52-66.

APPENDIX: List of items in OLRS per dimension

Dimension
Computer/ Internet Self-Efficacy (CS)
CS1. I feel confident in performing the basic functions of Microsoft Office programs or their counterparts.
CS2. I feel confident in my knowledge and skills of how to manage online learning platforms
CS3. I feel confident in using the Internet to find or gather information for online learning.
Self-directed Learning (SL)
SL1. I carry out my own study plan.
SL2. I seek assistance when facing learning problems.
SL3. I manage time well.
SL4. I set up my learning goals.
SL5. I have higher expectations for my learning performance.
Learner Control (LC)
LC1. I can direct my own learning progress.
LC2. I am not distracted by other online activities while learning online.
LC3. I repeat the online instructional materials on the basis of my needs.
Motivation for Learning (ML)
ML1. I am open to new ideas.
ML2. I have motivation to learn.
ML3. I improve from my mistakes.
ML4. I like to share my ideas with others.
Online Communication Self-Efficacy (OS)
OS1. I feel confident in using online tools (e.g., email, discussion) to effectively communicate with others.
OS2. I feel confident in expressing myself (e.g., emotions and humor) through text.
OS3. I feel confident in posting questions in online discussions.

An Application of CATANOVA and Logistic Regression on the Most Prevalent Sexually Transmitted Infection (A Case Study of the University of Nigeria Teaching Hospital)

Nnaemeka Martin Eze[1]
Department of Statistics, University of Nigeria, Nsukka, Nigeria

Oluchukwu Chukwuemeka Asogwa
Department of Mathematics, Computer Science, Statistics and Informatics, Alex Ekwueme Federal University Ndufu-Alike Ikwo, Nigeria

Samson Offorma Ugwu, Chinonso Michael Eze
Felix Obi Ohanuba, Tobias Ejiofor Ugah
Department of Statistics, University of Nigeria, Nsukka, Nigeria

ABSTRACT

This research focused on the application of CATANOVA and logistic regression on the most prevalent Sexually Transmitted Infection (STI) reported in the University of Nigeria Teaching Hospital from 2010-2020. A population of 20,704 patients was recorded to have contracted eight(8) selected STIs. Prevalence analysis was computed to determine the most prevalent STI. Two-way CATANOVA cross-classification was computed to ascertain the age group and gender that suffer more from the most prevalent STI. Three-way CATANOVA was computed to ascertain the association among drug prescription, age, and gender of the Gonorrhea patients. A logistic regression model was fitted to predict infertility as an effect of the most prevalent STI. The prevalence analysis showed Gonorrhea infection as the most prevalent STI at 33.08%. A population of 6,850 patients recorded to have contracted Gonorrhea infection from 2010-2020 was employed for the analysis. Two-way CATANOVA cross-classification showed that gender, age, and interaction effects were statistically significant at a 5% significance level. Male (3,752; 54.8%) suffers Gonorrhea infection more than female (3,098;45.2%) and aged 30-39 years (1,946; 28.4%) suffers it more than any other age interval. The interaction effect shows that the rate of contracting Gonorrhea infection by gender differs from one age interval to another. Three-way CATANOVA results showed that drugs prescribed for the treatment of Gonorrhea infection depend on gender and age. Logistic regression results showed that an increase in age, body mass index, blood pressure, blood sugar, bacteria quantity, and Gonorrhea history were associated with an increased likelihood of the Gonorrhea patient being infertile.

Keywords: *Chi-square test, Prediction, Prevalence*

[1] **Address correspondence to Nnaemeka Martin Eze:** nnaeneka.eze@unn.edu.ng

1. Introduction

This study focuses on the occurrence of different kinds of Sexually Transmitted Infections (STIs) in our societies. Scientists have proved that several infections have their origin and some can be cured while some cannot be cured. The U.S. Department of Health and Human Services reported that there are several ways in which one can contract these infections and this can be through sexual practices (Scatterwhite et al., 2013). Sexually Transmitted Infections (STIs) also known as Sexually Transmitted Diseases (STDs) are harmful micro-organisms that are very hard to control their growth in the body of their host. These infections are easily contracted through sex. Most STIs initially do not show symptoms. According to medical experts, infections can be called diseases only when they show symptoms and this is the reason STDs are known as STIs. Medical experts had said that the infections can easily be spread when there is no presence of symptoms of these infections. Some of the symptoms of STIs are vaginal discharge, penile discharge, ulcers on or around the genitals, and pelvic pain. Some STIs may cause infertility in both males and females and also poor development of a baby if contracted before or during pregnancy. Different bacteria, viruses, fungi, and parasites pathogenic are the major causes of STIs. Some of the bacterial STIs are chlamydia infections, gonorrhea or gonococci infection, cancroids, granuloma inguinal, and syphilis. Some of the viral STIs are genital herpes, HIV/AIDS, Viral hepatitis (Hepatitis B virus), and genital warts. Some of the fungal STIs are candidiasis and Parasitic STIs include crab louse, scabies, and Trichomoniasis (Scatterwhite et al., 2013). Despite the contamination of some STIs through sex, one can contact them through blood and tissues, breastfeeding or during child delivery.

The contamination of STIs from one, and another or from surrounding objects can be prevented (Center for Diseases Prevention and Control, 2013). Azmi et al., (2008) presented their prevalence analysis from child-bearing-age women and the result showed that the prevalence of C. trachomatis infection was 0.6% and 0.5%, among symptomatic and asymptomatic women respectively, N. gonorrhoeae was 0.9% and 2.2%, T. pallidum 0.0% and 0.0%, and Tr. vaginalis was 0.7% and 0.5%. It was noted from the result that there was no significant difference in the prevalence rate between symptomatic and asymptomatic women. Kesah et al., (2013) stated that improvement in hand washing, clean toilets, abstaining from sex, condom usage, rational employment of examination methods, medical diagnostics testing for both men and women, attitude change, and prevention education should be consistently highlighted. Otaru and Ogbonda (2020) studied the application of categorical data-nested design of knowledge and control practices of Hepatitis B Virus (HBV) infection using the two-way CATANOVA technique. They considered frequency data from university students in three universities involving response rate of student's knowledge and control practices of HBV infection using a scale of good, fair, and poor. It was noted from the result that there was no significant difference in the knowledge and control practices of HBV infection of the students in the three considered universities at a 5% level of significance. Deyhoul et al., (2017) studied infertility rate risk factors and the result showed that infertility in men and women could be caused by sexually transmitted infections and hormonal disorders. Some lifestyle factors can also cause infertility such as obesity, nutrition, smoking/alcohol consumption, mobile phone use, sexual violence, and anxiety.

It has been known that Sexually Transmitted Infections (STIs) have sporadically increased over the years and of course have caused more harm than good in our societies. These infections could lead to various dangerous ailments such as infertility, pelvic inflammatory

disease in women, ectopic pregnancy, and serious effects on pregnancy which might lead to miscarriage, failure of development of a new baby, blindness, congenital defects, and so on. This study aims to know the most prevalent Sexually Transmitted Infection (STI) among the reported cases of STIs in the University of Nigeria Teaching Hospital; the gender and ages that always suffered from the most prevalent STI; examine if the prescribed drugs depend on patient's gender and age; and examine the reproductive status of the patients suffering the most prevalent STI, that is, to know if the carrier of the infection is fertile or infertile.

This study focuses only on eight (8) major sexually transmitted infections (Chlamydia, Gonorrhea, Syphilis, Herpes, Hepatitis B, Trichomoniasis, Human Immuno Deficiency Virus (HIV), Human Papilloma Virus (HPV)) contracted by both males and females which has attained sexual age as reported at University of Nigeria Teaching Hospital (UNTH) from 2010 to 2020. The significance of this study tends to educate Nigerians and the world at large about the existence of sexually transmitted infections in our societies and their risk factors. It will also notify people about the most prevalent STI, the gender and age interval that is more likely to be at risk of it, and more precisely, educate them on how to take precautionary measures.

2. Materials and Methods

2.1 Data and sampling design

The data used in this study were secondary data collected from eight (8) types of Sexually Transmitted Infections (STIs) reported in the Department of Micro Biology, University of Nigeria Teaching Hospital (UNTH). To determine the most prevalent STI, a population of 20,704 patients that reported to have contracted eight (8) selected STIs (Chlamydia (4,855), Gonorrhea (6,850), Syphilis (1,680), Trichomoniasis (1,770), Herpes (483), Hepatitis-B (602), Human Papilloma Virus (619) and Human Immune-deficiency Virus (3,845)) in the years 2010 through 2020 were collected and the prevalence method of analysis was used to ascertain the most prevalent STI among them. Furthermore, the record also showed that there were 6,850 reported cases of the most prevalent STI, and the data were presented using a randomized complete block design in which a K-dimensional vector $\left[n_{ijk}\right]$ of nominal responses are observed in frequencies in the ij^{th} plot (see Table 1). These most prevalent STI data were analyzed using categorical analysis of variance (CATANOVA) and logistic regression.

2.2 Ethical approval

The ethical issues in this study were addressed by making sure that anonymity and confidentiality are highly maintained when the need arises either from the data collection or any sources of information, and the consent of patients was respected. Therefore, all procedures performed in this research that involved patients and healthcare workers were in accordance with the ethical standards of the University of Nigeria Teaching Hospital (UNTH).

2.3 Models

2.3.1 Prevalence Rate: Prevalence is an epidemiology characteristic that is easily measured using survey data or medical records. To establish prevalence, researchers randomly select a sample (smaller group) from the entire population they want to describe. Using random selection methods increases the chances that the characteristics of the sample will be representative of (similar to) the characteristics of the population. For a representative sample, prevalence is the number of people in the sample with the characteristics of interest divided by the total number of people in the sample.

$$\left(\text{i. e., Prevalence formula} = \frac{\text{number of people in the sample with the characteristics of interest}}{\text{total number of people in the sample}}\right).$$

2.3.2 CATANOVA: The categorical analysis of variance (CATANOVA) is a technique designed to help the researcher identify the variation between treatments of interest**.** This CATANOVA is used to solve the problem in the analysis of variance when the observations are nominal without any underlying metric and it was also formulated to solve the erroneous analysis of nominal data by using the chi-square test (Onukogu, 1985; Otaru and Ogbonda, 2020). In addition, there are several methods for analyzing categorical data in which some of these methods use data transformation before proceeding to analyze the data. The transformation method to be used may depend on the classification of categorical data (Fienberg, 1973; Florian, 2008; Onukogu, 2014; Singh, 2004). In this research, two-way and three-way CATANOVAs are adopted and there is no loss in generality using the method for unequal levels of factors that do not differ significantly.

Table 1 shows the data layout for two-way cross classification or a randomized complete block design in which a K-dimensional vector $[n_{ijk}]$ of nominal responses are observed in frequencies in the ij^{th} plot. In this Table 1, the main factor A ranging from 1 to I and main factor B ranging from 1 to J have from 1 to K quanta responses per unit (D'Ambra et al., 2005; Anderson and Landis, 1980, 1982; Light and Margolin, 1971; Margolin and Light, 1974). Table 2 depicted the CATANOVA table that contains the source of variation, degrees of freedom (df), the sum of squares (SS) which is the trace of its variance-covariance matrix, test ratio from chi-square calculated, a critical value from chi-square tabulated and hypotheses for the study.

Furthermore, this study assumed that the data follows:

- Multi-nominal distribution

  $$P(\{n_{ijk}\};\{\pi_{ijk}\}) = \binom{n_{ij}}{n_{ij1},\ldots,n_{ijK}}\prod_{k=1}^{K}(\pi_{ijk})^{n_{ijk}}$$

  $n_{ijk} = 0, 1, \cdots, n_{ij}$ and $\pi_{ijk} = \frac{n_{ijk}}{n_{ij}};\ 0 \le \pi_{ijk} \le 1$
- Independence: The levels and blocks each act independently. That is, n_{ijk} and $n_{i'j'k}$ are statistically independent $\forall i \neq i'$ and $\forall j \neq j'$.
- Constant variance: $\text{var}(n_{ijk}) = n\pi_{ijk}(1-\pi_{ijk})$. The variance is not constant because it depends on i, j and k.
  $\pi_{ijk} > 0,\ \sum_{k=1}^{K}\pi_{ijk} = 1,\ \sum_{k}^{K} n_{ijk}$ is held fixed (i.e., grand total over k for j)

Table 1: The data layout for two-way CATANOVA cross-classification or randomized complete block design.

	B(j)												
	b1				b2				. . .	bJ			
A(i)	1	2	….	K	1	2	….	K	. . .	1	2	….	K
1	n_{111}	n_{112}	….	n_{11K}	n_{121}	n_{122}	….	n_{12K}	….	n_{1J1}	n_{1J2}	….	n_{1JK}
2	n_{211}	n_{212}	….	n_{21K}	n_{221}	n_{222}	….	n_{22K}	….	n_{2J1}	n_{2J2}	….	n_{2JK}
.	.	.		.	.	.		.		.	.		.
.	.	.	….	.	.	.	….	.	….	.	.	….	.
.	.	.		.	.	.		.		.	.		.
I	n_{i11}	n_{i12}	….	n_{i1K}	n_{i21}	n_{i22}	….	n_{i2K}	….	n_{iJ1}	n_{iJ2}	….	n_{iJK}

Table 2: Summary of two-way CATANOVA cross-classification of nominal data.

Source	df	SS	Test Ratio	Critical Value	Hypothesis
Row(Ai)	I-1	RSS	χ^2_{RT}	$\chi^2_{(I-1)(K-1)}$	$HO_R: \pi_{ijk} = \pi_{jk} \forall_i$
Column(Bj)	J-1	CSS	χ^2_{CT}	$\chi^2_{(J-1)(K-1)}$	$HO_C: \pi_{ijk} = \pi_{ik} \forall_j$
Interaction(AB)	(I-1)(J-1)	NSS	χ^2_{NT}	$\chi^2_{(I-1)(J-1)(K-1)}$	$HO_{RC}: \pi_{ijk} = \pi_k \forall_{ij}$
Weight Units	n-IJ	WUSS	-		-
Total	n-1	TSS	-		-

Computation of Sum of Squares

Total Sum of Square (TSS) = $n - \frac{\sum_k n^2_{..k}}{n}$; where $n_{..k} = \sum_{ij} n_{ijk}$ (1)

Within Unit Sum of Square (WUSS) = $n - \sum_{ij} \frac{\sum_k n^2_{ijk}}{n_{ij}}$ (2)

Between Row Sum of Square (BRSS) = $n - \sum_i \frac{\sum_k n^2_{i.k}}{n_i}$; where $n_{i.k} = \sum_j n_{ijk}$ (3)

Between Column Sum of Square (BCSS) = $n - \sum_j \frac{\sum_k n^2_{.jk}}{n_j}$; where $n_{.jk} = \sum_i n_{ijk}$ (4)

Row Sum of Square (RSS) = TSS – BRSS (5)

Column Sum of Square (CSS) = TSS – BCSS (6)

Interaction Sum of Square (NSS) = BCSS + BRSS – TSS – WUSS (7)

Two-way CATANOVA cross classification model

$$E(\hat{\pi}_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij} \quad (8)$$

where $\hat{\pi}_{ijk}$ is the probability that k^{th} observation occurs in the i^{th} level of factor A and j^{th} level of factor B, i.e., $\hat{\pi}_{ijk} = P_{ijk} = \frac{n_{ijk}}{n_{ij}}$, ($n_{ijk}$ is the k^{th} observation in the ij^{th} cell, n_{ij} is the sum of k^{th} observation in the ij^{th} cells, i.e., $n_{ij} = \sum_k n_{ijk}$), μ is a constant for k^{th} observation, α_i ($i = 1, 2, \ldots, I$) is the effect of the i^{th} level of factor A, β_j ($j = 1, 2, \ldots, J$) is the effect of the j^{th} level of factor B, γ_{ij}(($i = 1, 2, \ldots, I$) and ($j = 1, 2, \ldots, J$)) is the interaction between the i^{th} level of factor A and j^{th} level of factor B. In nominal data, the sum of squares is the trace of its variance-covariance matrix and the parameter π_{ijk} may be considered fixed or random with probability density $h(\pi_{ijk})$ ranging from 0 to 1 depending on whether I and J are random or fixed (Anderson, 1958; Onukogu, 1985; Onukogu, 2014; Scheffe, 1959).

Hypotheses

$H_{0R}: \pi_{ijk} = \pi_{jk}$, i. e., $\alpha_i = 0\ \forall_i$ (There is no row effect)
$H_{1R}: \pi_{ijk} \neq \pi_{jk}$, i. e., $\alpha_i \neq 0\ for\ at\ least\ one\ (i)$ (There is row effect)

$H_{0C}: \pi_{ijk} = \pi_{ik}$, i. e., $\beta_j = 0\ \forall_j$ (There is no column effect)
$H_{1C}: \pi_{ijk} \neq \pi_{ik}$, i. e., $\beta_j \neq 0\ for\ at\ least\ one\ (j)$ (There is column effect)

$H_{0RC}: \pi_{ijk} = \pi_k$, i. e., $\gamma_{ij} = 0\ \forall_{ij}$ (There is no interaction effect)
$H_{1RC}: \pi_{ijk} \neq \pi_k$, i. e., $\gamma_{ij} \neq 0\ for\ at\ least\ one\ pair\ (ij)$ (There is an interaction effect)

Test Statistic

$$\chi^2_{RT} = \frac{(K-1)(n-1)RSS}{TSS} \sim \chi^2_{(I-1)(K-1)};\ \alpha$$

$$\chi^2_{CT} = \frac{(K-1)(n-1)CSS}{TSS} \sim \chi^2_{(J-1)(K-1)};\ \alpha$$

$$\chi^2_{NT} = \frac{(K-1)(n-1)NSS}{TSS} \sim \chi^2_{(I-1)(J-1)(K-1)};\ \alpha$$

Decision rule

Reject H_{0R} if $\chi^2_{RT} \geq \chi^2_{(I-1)(K-1)}$,

Reject H_{0C} if $\chi^2_{CT} \geq \chi^2_{(J-1)(K-1)}$, and

Reject H_{0RC} if $\chi^2_{NT} \geq \chi^2_{(I-1)(J-1)(K-1)}$, at specified level of significance (5%). Fail to reject if otherwise.

Table 3: The data layout for the 3-way contingency table.

	Y_1				Y_2					
	Z_1	Z_2	n_{+j+}	π_{+j+}	Z_1	Z_2	n_{+j+}	π_{+j+}	n_{i++}	π_{i++}
X_1	$n_{111}(\hat{f}_{111})$	$n_{112}(\hat{f}_{112})$	n_{11+}	π_{11+}	$n_{121}(\hat{f}_{121})$	$n_{122}(\hat{f}_{122})$	n_{12+}	π_{12+}	n_{1++}	π_{1++}
X_2	$n_{211}(\hat{f}_{211})$	$n_{212}(\hat{f}_{212})$	n_{21+}	π_{21+}	$n_{221}(\hat{f}_{221})$	$n_{222}(\hat{f}_{222})$	n_{22+}	π_{22+}	n_{2++}	π_{2++}
n_{++k}	n_{+11}	n_{+12}	n_{+1+}	-	n_{+21}	n_{+22}	n_{+2+}	-	n	-
π_{++k}	π_{+11}	π_{+12}		π_{+1+}	π_{+21}	π_{+22}		π_{+2+}		$\sum_{ijk} \pi_{ijk} = 1$

where; $X_i (i = 1, 2, \cdots, \mathrm{I})$, $Y_j (j = 1, 2, \cdots, \mathrm{J})$, $Z_k (k = 1, 2, \cdots, \mathrm{K})$, n_{ijk} is the observed frequency in ijk cell, $n = \sum_i \sum_j \sum_k n_{ijk}$ is the total observation, $n_{i++} = \sum_j \sum_k n_{ijk}$ is the marginal row total, $n_{+j+} = \sum_i \sum_k n_{ijk}$ is the marginal column total, and $n_{++k} = \sum_i \sum_j n_{ijk}$ is the marginal k^{th} observation total, $\hat{f}_{ijk} = n\left(\frac{n_{i++}}{n}\right)\left(\frac{n_{+j+}}{n}\right)\left(\frac{n_{++k}}{n}\right)$ is the estimated expected frequency in ijk cell, π_{ijk} is the probability value in ijk cell, $\pi_{i++} = \left(\frac{n_{i++}}{n}\right)$ is the row marginal probability, $\pi_{+j+} = \left(\frac{n_{+j+}}{n}\right)$ is the column marginal probability, $\pi_{++k} = \left(\frac{n_{++k}}{n}\right)$ is the k^{th} marginal probability, $\pi_{+jk} = \left(\pi_{+j+} \cap \pi_{++k}\right) = \left(\frac{n_{+jk}}{n}\right)$ is the intersection of column marginal probability and k^{th} marginal probability, and $\sum_{ijk} \pi_{ijk} = \sum_i \pi_{i++} = \sum_j \pi_{+j+} = 1$.
Note: $\cap$ is an intersection symbol.

Hypothesis for conditional independency test in the 3-way contingency table

$H_0: \pi_{ijk} = \pi_{i++} \times \left(\pi_{+j+} \cap \pi_{++k}\right)$ (X variable is independent of Y and Z variables)

$H_1: \pi_{ijk} \neq \pi_{i++} \times \left(\pi_{+j+} \cap \pi_{++k}\right)$ (X variable depends on Y and Z variables)

Test Statistic

$$\chi^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} \frac{\left(n_{ijk} - \hat{f}_{ijk}\right)^2}{\hat{f}_{ijk}} \sim \chi^2_{ijk-(i+j+k)+2}$$

Decision Rule: Reject H_0 if $\chi^2{}_{cal} \geq \chi^2{}_{tab}$. Fail to reject if otherwise.

2.3.3 Logistic Regression: This is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analysis, logistic regression is a predictive analysis used to describe data and to explain the relationship between one dependent binary response variable, which takes values 1 and 0, and one or more nominal, ordinal, interval, or ratio level independent variable(s). The logistic regression gives each predictor a coefficient that measures its independent contribution to variation in the dependent variable. The dependent variable Y takes the value 1 if the response is "yes'' and takes a value 0 if the response is "no''. Logistic regression calculates the probability of success over the probability of failure. The results of the analysis are in the form of an odds ratio (Boateng and Abaye, 2019).

The model form for predicted probabilities is expressed as a natural logarithm (ln) of the odds ratio:

$$\mathrm{Ln}(ODDS) = \ln\left(\frac{\mathrm{P(Y)}}{1-\mathrm{P(Y)}}\right) = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \cdots + \beta_\mathrm{m} \mathrm{X}_\mathrm{m} \quad (9)$$

$$\frac{\mathrm{P(Y)}}{1-\mathrm{P(Y)}} = e^{\beta_0+\beta_1 \mathrm{X}_1+\beta_2 \mathrm{X}_2+\cdots+\beta_\mathrm{m} \mathrm{X}_\mathrm{m}} \quad (10)$$

$$\mathrm{P(Y)} = e^{\beta_0+\beta_1 \mathrm{X}_1+\beta_2 \mathrm{X}_2+\cdots+\beta_\mathrm{m} \mathrm{X}_\mathrm{m}} - \mathrm{P(Y)} e^{\beta_0+\beta_1 \mathrm{X}_1+\beta_2 \mathrm{X}_2+\cdots+\beta_\mathrm{m} \mathrm{X}_\mathrm{m}} \quad (11)$$

$$= \frac{e^{\beta_0+\beta_1 \mathrm{X}_1+\beta_2 \mathrm{X}_2+\cdots+\beta_\mathrm{m} \mathrm{X}_\mathrm{m}}}{1+e^{\beta_0+\beta_1 \mathrm{X}_1+\beta_2 \mathrm{X}_2+\cdots+\beta_\mathrm{m} \mathrm{X}_\mathrm{m}}} \quad (12)$$

where; $\frac{P(Y)}{1-P(Y)}$ is the odds ratio, $\ln\left(\frac{\mathrm{P(Y)}}{1-\mathrm{P(Y)}}\right)$ is the log odds or "logit'' of the outcomes, Y is the dichotomous outcome, P(Y = 1) is the probability of an event, $\times_i$ ($\mathrm{i} = 1,2, \ldots, \mathrm{m}$) are the predictors, β_i($\mathrm{i} = 1, 2, \ldots, \mathrm{m}$) are unknown regression parameters to be estimated and β_0 is the intercept (i.e., constant).

2.3.3.1 Goodness of Fit Test. It is also known as the Hosmer-Lemeshow test which represents a chi-square test used for testing the adequacy of the model for fitting the data. The null hypothesis is that the model is adequate to fit the data and we will only reject this null hypothesis if the p-value is less than 0.05 (Abdulqader, 2017). It is given as

$$H = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i} \tag{13}$$

where O_i and E_i denote the observed and expected frequencies , respectively.

Table 4: Values of the logistic regression model when the independent variable is dichotomous.

Outcome variable (Y)	Independent variable (X)	
	X = 1	X = 0
Y = 1	$P(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$P(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
Y = 0	$1 - P(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$	$1 - P(0) = \frac{1}{1+e^{\beta_0}}$

The odds ratio is then computed as:

$$Odds\ ratio\ (OR) = \frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} \Big/ \frac{1}{1+e^{\beta_0+\beta_1}}}{\frac{e^{\beta_0}}{1+e^{\beta_0}} \Big/ \frac{1}{1+e^{\beta_0}}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{(\beta_0+\beta_1)-\beta_0} = e^{\beta_1} \tag{14}$$

Hence, for logistic regression with a dichotomous independent variable coded 1 and 0, the relationship between the odds ratio and the regression coefficient is $Odds\ ratio(OR) = e^{\beta_1}$.

3. Results and Discussions

Figure 1 is a pie chart representation of the yearly percentage of reported cases of Sexually Transmitted Infections (STIs) as depicted in Table 5. From this figure, the years 2019(12%) and 2017(12%) had the highest reported cases of eight (8) types of STIs and were followed by the years 2016(11%), 2012(11%), 2013(10%), 2011(10%), 2018(8%), 2015(8%), 2010(8%), 2020(5%) and 2014(5%). Figure 1 also shows that there is a difference in yearly reports of STIs. This may be due to a lack of knowledge about the harmfulness of STIs in society.

Table 5: Reported cases of selected STIs from 2010-2020.

YEARS	CHL.	GON.	SYPH.	TRICO.	HERPES	HEPA.B	HPV	HIV	TOTAL
2010	201	603	137	159	85	61	18	414	1678
2011	167	941	226	173	11	28	71	455	2072
2012	437	987	148	158	35	73	15	441	2294
2013	698	434	76	178	43	24	88	397	1938
2014	116	392	64	147	33	35	78	243	1108
2015	576	503	115	131	15	84	31	260	1715
2016	705	741	123	188	35	95	23	272	2182
2017	461	982	326	118	58	159	54	262	2420
2018	516	316	178	174	16	19	98	337	1654
2019	826	751	148	181	89	17	119	418	2549
2020	152	200	139	163	63	7	24	346	1094
Total	4855	6850	1680	1770	483	602	619	3845	**20,704**

CHL.= Chlamydia, GON.= Gonorrhea, SYPH.= Syphilis, Herpes, HPV = Human Papilloma Virus, TRICO.=Trichomoniasis, HEPA B.= Hepatitis B Virus, HIV= Human Imuno Deficiency Virus

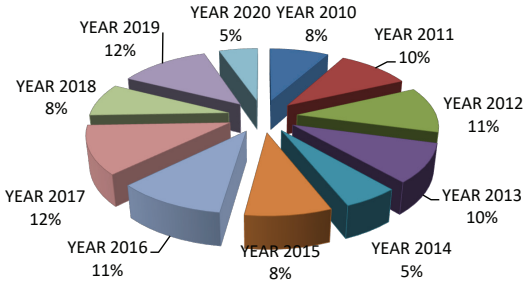

Figure 1: Yearly percentage of reported cases of Sexually Transmitted Infections.

Table 6: Prevalence rate of the eight (8) selected sexually transmitted infections (2010-2020)

Sexually Transmitted Infection (STI)	Prevalence Rate	Percentage of Prevalence Rate
Chlamydia	0.2344	23.44
Gonorrhea	0.3308	33.08
Syphilis	0.0812	8.12
Triconomiasis	0.0854	8.54
Herpes	0.0233	2.33
Hepatitis B Virus	0.0290	2.90
Human Papilloma Virus	0.0298	2.98
Human Immuno Deficiency Virus	0.1856	18.56

$$\text{Prevalence formula} = \frac{number\ of\ each\ STI}{Total\ number\ of\ STI} \times 100 \qquad (15)$$

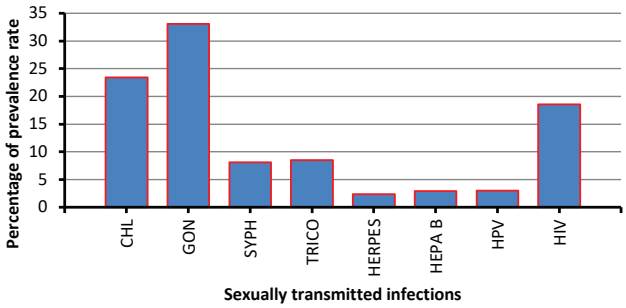

Figure 2: Bar chart for the prevalence rate of reported cases of sexually transmitted infections.

Figure 2 is a bar chart representation of the results of the prevalence rate percentage as depicted in Table 6. As can be seen from this Figure 2, Gonorrhea infection with a 33.08% rate appears to be the most prevalent among the eight selected sexually transmitted infections reported in the University of Nigeria Teaching Hospital from 2010-2020 when compared with Chlamydia, Syphilis, Trichomoniasis, Herpes, Human Papilloma Virus (HPV), Hepatitis B, Human Immuno-deficiency Virus (HIV) with 23.44%, 8.12%, 8.54%, 2.33%, 2.90%, 2.98%, and 18.56% respectively. From the data in Table 7, the percentage (54.8%) of males that suffered from Gonorrhea infection is more than the percentage (45.2%) of females. This shows that the male suffers from Gonorrhea infection more than female. Also, note that 28.4% of Gonorrhea patients are at the age interval of 30-39 years, 23.4% are at the age of 20-29 years, 18.5% are at the age of 40-49 years, 15% are the age of 50 years and above while14.7% are at the age of fewer than 20 years. These show that the age interval of 30-39 years suffers Gonorrhea infection more than any other age interval.

Table 7: Two-way contingency table depicting the response of gender and ages of gonorrhea patients reported in UNTH from 2010-2020.

Gender (i)	Age (j)															Total $n_{i.k}$		Total $n_{i..}$
	< 20 (j_1)			20-29 (j_2)			30-39 (j_3)			40-49 (j_4)			50+ (j_5)					
	History response			History response			History response			History response			History response					
	YES	NO	Total $n_{i1.}$	YES	NO	Total $n_{i2.}$	YES	NO	Total $n_{i3.}$	YES	NO	Total $n_{i4.}$	YES	NO	Total $n_{i5.}$	YES	NO	
Male	128	383	511	395	437	832	463	545	1008	388	415	803	347	251	598	1721	2031	**3752**
Female	154	340	494	272	498	770	385	553	938	283	181	464	234	198	432	1328	1770	**3098**
Total $n_{.jk}$	282	723	**1005**	667	935	**1602**	848	1098	**1946**	671	596	**1267**	581	449	**1030**	**3049**	**3801**	**6850**

Table 8: Table for the significance of gender, age, and interaction between gender and age effects.

Source	DF	Sum of Squares	Test Ratio	Critical Value	Decision
Gender (Row)	1	3.06	6.19	3.84	significant, (reject H_{0R})
Age (Column)	4	104.63	211.78	9.49	significant, (reject H_{0C})
Gender*Age	4	23.15	46.86	9.49	significant, (reject H_{0RC})
Within unit	6840	3252.88	-	-	
Total	6849	3383.72	-	-	

From the results in Table 8, we noticed a statistically significant difference in gender $\left(\chi^2_{RT(Cal)} = 6.19 > \chi^2_{RT(tab)} = 3.84\right)$, and a statistically significant difference in age $\left(\chi^2_{CT(Cal)} = 211.79 > \chi^2_{CT(tab)} = 9.49\right)$. The significant difference in gender means that a particular gender suffers more from Gonorrhea infection than another gender. The significant difference in age means that a particular age group is the most likely age group that suffers from Gonorrhea infection. It was noticed also that there is a statistically significant difference in the interaction between gender and age at a 5% significance level $\left(\chi^2_{NT(Cal)} = 46.86 > \chi^2_{NT(tab)} = 9.49\right)$. The significant difference in the interaction between gender and age intervals of Gonorrhea patients means that the rate of contracting Gonorrhea infection by males and females differs from one age interval to another. Moreover, the data in Table 7 showed that 3801(55.5%) of Gonorrhea patients do not have a Gonorrhea infection history, while 3049(44.5%) have it. It shows that there is a spread of Gonorrhea infection between the genders. (see Appendix A for the computation of results in Table 8)

Table 9.1: Three-way contingency table depicting gender, ages, and drug prescription for gonorrhea infection (2010 - 2020).

GENDER	**Male**						**Female**						
AGE	**Response in Male Ages**					$\mathbf{n_{+j_1+}}$	**Response in Female Ages**						
DRUG	**< 20**	**20-29**	**30-39**	**40-49**	**50 +**		**< 20**	**20-29**	**30-39**	**40-49**	**50+**	$\mathbf{n_{+j_2+}}$	$\mathbf{n_{i++}}$
CEFTR.	110	188	206	171	118	793	129	196	229	80	63	697	1490
STREPT.	107	163	202	164	112	748	111	135	216	55	75	592	1340
DOXY.	109	123	178	115	116	641	61	102	110	103	96	472	1113
GENTA.	130	192	248	205	159	934	104	211	184	128	132	759	1693
OFLO.	55	166	174	148	93	636	89	126	199	98	66	578	1214
$\mathbf{n_{++k}}$	511	832	1008	803	598	**3752**	494	770	938	464	432	**3098**	**6850**

CEFT=Ceftriaxone, STREPT.= Streptomycin, DOXY.= Doxycycline, GENTA.=Gentamicin , OFLO.= Ofloxacin

Table 9.2: $n_{++k}-$ table computed from table 9.1 for three-way contingency table.

Age	**Total responses in ages**		$\mathbf{n_{++k}}$
	Male	**Female**	
< 20	511	494	1005
20-29	832	770	1602
30-39	1008	938	1946
40-49	803	464	1267
50+	598	432	1030

Hypotheses:

$H_0: \pi_{ijk} = \pi_{i++} \times (\pi_{+j+} \cap \pi_{++k})$ (The drugs used to treat Gonorrhea infection are independent of gender and age)

$H_1: \pi_{ijk} \neq \pi_{i++} \times (\pi_{+j+} \cap \pi_{++k})$ (The drugs used to treat Gonorrhea infection are dependent on gender and age)

Computed Test Statistic:

$\chi^2_{cal} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\frac{(n_{ijk}-f_{ijk})^2}{f_{ijk}} = 221.30$ (see appendix Table 1 in Appendix B)

$\chi^2_{tab} = \chi^2_{40} = 55.75$ (see Appendix B)

The result of the analysis for Tables 9.1 and 9.2 showed that the prescribed drugs for patients suffering from Gonorrhea infection depend on the age and gender of the patient, $\chi^2_{cal} = 221.30$ is greater than $\chi^2_{tab} = 55.75$, at a 5% significance level.

Table 10: Logistic regression analysis code sheet for dependent and independent variables data.

Variable	Description	Codes/ Values	Name	Data type
$\times_1$	Age	Years	Age	Numerical
$\times_2$	History of Gonorrhea	0 = No 1 = Yes	History	Nominal
$\times_3$	Body Mass Index	kg/m^2	BMI	Numerical
$\times_4$	Blood Pressure	mm Hg	BP	Numerical
$\times_5$	Blood Sugar	mg/dl	BS	Numerical
$\times_6$	Bacteria Quantity	(cfu/ml)*10^8	BQ	Numerical
Y	Reproductive Status (Dependent variable)	0 = fertile 1 = infertile	Reproductive Status	Nominal

The way a particular data is presented goes a long way in determining its analytical case. In order to prevent some problems usually encountered in the poor presentation of data, extra care is taken, in Table 10, to present the independent variables and their data type and values.

Table 11: Test statistics for test on multi-collinearity.

Model	Unstandardized Coefficients		Standardized Coefficients	t	P-value	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	-.995	.046		-21.478	.001		
AGE	.013	.001	.322	18.895	.001	.380	2.631
History of Gonorrhea	.220	.086	.172	2.543	.012	.774	1.292
BMI (Kg/m^2)	.013	.002	.121	11.029	.001	.918	1.089
BP (mmHg)	.002	.001	.060	5.420	.001	.914	1.094
BS (mg/dl)	.002	.001	.169	10.619	.001	.437	2.287
BQ (cfu/ml)*10^8	.141	.013	.114	10.596	.001	.958	1.043

Multi-collinearity occurs when independent variables in a model are correlated. In logistic regression, this kind of correlation is a problem because independent variables should be have weak or no relationship at all among themselves. If there is a presence of multicollinearity, logistic regression estimates will be unstable and have high standard errors. A researcher can use the tolerance method or Variance Inflation Factor (VIF) method to check

presence of multi-collinearity. The high value of tolerance is an indication that there is no multi-collinearity in the model while the low value of tolerance is known to affect adversely the results associated with the model. The minimum tolerance value should be < 0.25. Variance Inflation Factor (VIF) is the reciprocal of tolerance. It identifies the correlation between independent variables and the strength of that correlation. The minimum value of VIF is 1 and has no upper limit. The value between 1 and 4 indicates that there is no correlation between this independent variable and any other independent variable and it suggests an absence of multi-collinearity. A VIF value between 5 and 9 indicated that there is a moderate correlation but it is not severe enough to cause a problem. A VIF value of more than 10 is said to be highly collinear and it indicates critical levels and causes a problem (Eze et al., 2021; Warner, 2013). From Table 11, the independent variables had no multi-collinearity since the tolerance values for the variables were greater than 0.25. Also, to confirm our claim the VIF values were between 1 and 4.

Omnibus Tests of Model Coefficients are used to assess the fitness of the overall logistic regression model. The overall model contains all the considered independent variables, unlike the null model which contains no independent variables. From Table 12, the Omnibus Tests of Model Coefficients tested the model fit to predict the reproductive status (i.e., fertility or infertility) of Gonorrhea patients. It tested the significance of the independent variables coded as age, history, blood sugar, bacteria quantity, body mass index, and blood pressure as predictors of the model with reproductive status as a dependent variable (fertile = 0 and infertile = 1). Also, the resuls show in Table 12, a chi-square value of 1678.063 with 6 degrees of freedom (df) and P-value less than 0.05 (i.e., $\chi^2_{(6)} = 1678.063$, P-value < 0.05). It means that the overall model is statistically significant, that is, the model as a whole fits significantly to predict the reproductive status of Gonorrhea patients better than a model with no predictors at a 5% significance level.

Table 12: Omnibus Tests of Model Coefficients

		Chi-square	df	P-value
Step 1	Step	1678.063	6	.000
	Block	1678.063	6	.000
	Model	1678.063	6	.000

The Cox & Snell R^2 and Nagelkerke R^2 seen in Table 13 are similar to R^2 which is in linear regression that gives us an idea of how much variance in the dependent variable is explained by the independent variables. The R^2 ranges from 0 to 1, with 1 being a perfect fit. These Cox & Snell R^2 and Nagelkerke R^2 values are sometimes called pseudo R^2 and have lower values than R^2 in linear regression (Laerd Statistics, 2018; Cox and Snell, 1989; Nagelkerke, 1991). The Cox & Snell R^2, both corrected and uncorrected, was discussed earlier by Maddala (1983) and Cragg and Uhler (1970). From the results in Table 13, we noticed that Cox & Snell R^2 is 0.215(21.7%) and Nagelkerke R^2 is 0.334(33.4%); this is to say that R^2 ranges between 21.7% to 33.4%. It is preferable to report the Nagelkerke R^2 because it is a modification of Cox & Snell R^2 that cannot achieve a value of 1but Nagelkerke R^2 can reach a maximum of 1 (Laerd Statistics, 2018). It can be seen from Nagelkerke's R^2 result that 33.4% of the variance in the outcome variable is affected by predictor variable and it can be said that there is evidence to say that the logistic model is adequate or a good fit for the data.

Table 13: Model summary statistics

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	5520.271	.217	.334

The Hosmer and Lemeshow Test in Table 14 tests the null hypothesis that predictions made by the logistic model fit perfectly with observed group memberships. The statistical test makes use of a chi-square statistic computed to compare the observed frequencies with those expected under the linear model. A nonsignificant chi-square statistic indicates that the model fits well with the data. This Hosmer and Lemeshow Test has several problems in which one of which is that it relies on a test of significance. The implication of this is that with large sample sizes, the test may be significant even when the fit is good, and with small sample sizes, it may not be significant even with a poor fit (Hosmer and Lemeshow, 2000; Wuensch, 2021). The result from the Hosmer and Lemeshow Test in Table 14 showed a chi-square value of 102.127 with 8 degrees of freedom (df) and a P-value greater than 0.05 (i.e., $\chi^2_{(8)} = 102.127$, P-value > 0.05) and this means that the model adequately fits the data perfectly well. Hence, there is no difference between the observed frequencies and the predicted model at a 5% significance level.

Table 14: Hosmer and Lemeshow test

Step	Chi-square	df	P-value
1	102.127	8	.0901

Table 15 shows the classification table for the reproductive status (fertility or infertility) of Gonorrhea patients. The logistic regression model estimates the probability of an event occurring using the values of the independent variables on a certain cut-off point, usually 0.5. If the estimated probability of the event occurring is greater than or equal to 0.5, the event is classified as occurring but if the probability is less than 0.5, the event is classified as not occurring. The classification table compares the actual and predicted groups to asses how many would be correctly classified. This method of classification of individuals into one of the outcome groups (YES or NO) is a way to assess the model's reliability for prediction. Therefore, it becomes necessary to have a method to examine the effectiveness of the predicted classification against the actual classification. In Table 15, the Gonorrhea patients with predicted probabilities of fertility greater than or equal to 0.5 are classified into the fertile group while those with predicted probabilities of infertility greater than or equal to 0.5 are classified into the infertile group. The model correctly classified 1121(74.8%) Gonorrhea patients into the infertile group; this is known as the sensitivity of prediction, that is, the percentage of occurrences correctly predicted. The model also correctly classified 5085(95%) Gonorrhea patients into the fertile group and this is known as a specificity of prediction, that is, the percentage of nonoccurrences correctly predicted. The overall correct prediction was 6206 out of 6850 Gonorrhea patients with an overall success rate of 90.6%. The model predicted the total number of infertility as 1387 against the actual observation of 1499 Gonorrhea patients. It predicted 266(19.2%) infertile Gonorrhea patients that were wrongly classified into the fertile group at the time of actual observation recording and this is known as a false positive prediction. Also, the model predicted the total number of fertility as 5463 against the actual observation of 5351 Gonorrhea patients. It predicted 378(7.1%) fertile Gonorrhea patients that were wrongly classified into the infertile group and this is known as a false negative of prediction.

Table 15: Classification Table[a] for reproductive status (fertility or infertility) of a Gonorrhea patient

Observed			Predicted		
			Reproductive Status		Percentage
			Fertile	Infertile	Correct
Step 1	Reproductive Status	Fertile	5085	266	95.0
		Infertile	378	1121	74.8
	Overall Percentage				90.6

a. The cut value used is 0.500

From Table 16, it was noticed that column 2 shows results for logistic regression coefficients, column 4 shows the Wald Chi-Square statistic that tests the unique contribution of each predictor to the model, and column 6 shows probability values (P-values). The unique contribution of each predictor is significant if P-value is less than the 5% level of significance. Since the P-values in this Table 16 are less than 0.05 (i.e., P-value < 0.05), we conclude that the variables coded as age, history, blood sugar, bacteria quantity, body mass index, and blood pressure and used as the predictors of the model are statistically significant at a 5% significance level. Column 7 in Table 16 shows the result for the odds ratio related to variables coded age, history, blood sugar, bacteria quantity, body mass index, and blood pressure, and column 8 shows a 95% confidence interval for the odds ratio. An odds ratio is used to predict the probability of an event occurring based on a one-unit change in a predictor when all other predictors are kept constant. The odds ratio (OR) can be less than 1 (< 1), greater than 1 (>1), or equal to 1 (= 1). There is no change in odds if the odds ratio is 1. The odd decreases for every unit change in the predictor variable if it is less than 1. The odd increases for every unit change in the predictor variable if it is greater than 1. Thus, the higher the odds ratio is above 1, the more likely a patient is to be infertile. The result from Table 16 shows that for every unit increase in the variables coded as age, blood sugar, bacteria quantity, body mass index, and blood pressure of a Gonorrhea patient, the odds ratio of being infertile are 1.086, 1.104, 1.013, 1.014, and 2.314 when other predictors are constant respectively. The odds ratio of a Gonorrhea patient with a Gonorrhea history is 3.718 times more likely to be infertile than a Gonorrhea patient without a Gonorrhea history when variables coded as age, blood sugar, bacteria quantity body mass index, and blood pressure are held constant. We noticed that the odds ratio for the constant is less than 0.001, that is, the odds ratio for the model without the variables coded as age, history, blood sugar, bacteria quantity, body mass index, and blood pressure as predictors is less than 0.001.

Table 16: The logistic regression model table to predict the reproductive status (fertility or infertility) of Gonorrhea patients

	B	**S.E.**	**Wald**	**Df**	**P-value**	**Exp(B)**	**95% C.I.for EXP(B)**	
							Lower	**Upper**
AGE	.082	.005	270.821	1	.000	1.086	1.075	1.096
HISTORY(1)	1.320	.475	7.645	1	.006	3.718	1.466	9.431
BMI (Kg/m^2)	.099	.010	104.038	1	.000	1.104	1.083	1.125
BP (mmHg)	.013	.002	29.562	1	.000	1.013	1.008	1.018
BS (mg/dl)	.014	.001	99.015	1	.000	1.014	1.011	1.016
BQ (cfu/ml)*10^8	.839	.099	71.873	1	.000	2.314	1.906	2.809
Constant	-9.921	.398	622.020	1	.000	.000		

Obtained Model: The model form for predicted probabilities is expressed as a natural logarithm (ln) of the odds ratio:

$$\ln\left(\frac{\mathrm{P(Y)}}{1-\mathrm{P(Y)}}\right) = -9.921 + 0.082(\mathrm{Age}) + 1.320(\mathrm{History}) + 0.099(\mathrm{BMI}) + 0.013(\mathrm{BP}) \quad + 0.014(\mathrm{BS}) + 0.839(\mathrm{BQ}) \tag{16}$$

3.1 Predictions

The odds ratio prediction that is formed from the model in *Equation* (16) is given as

$$\frac{\mathrm{P(Y)}}{1-\mathrm{P(Y)}} = \mathrm{e}^{-9.921+0.082(\mathrm{Age})+1.320(\mathrm{History})+0.099(\mathrm{BMI})+0.013(\mathrm{BP})+0.014(\mathrm{BS})+0.839(\mathrm{BQ})} \tag{17}$$

The conversion of the odds ratio in *Equation* (17) to general probability form for the prediction of Gonorrhea patients that are infertile is given as

$$P(Y) = \frac{e^{-9.921+0.082(\mathrm{Age})+1.320(\mathrm{History})+0.099(\mathrm{BMI})+0.013(\mathrm{BP})+0.014(\mathrm{BS})+0.839(\mathrm{BQ})}}{1+e^{-9.921+0.082(\mathrm{Age})+1.320(\mathrm{History})+0.099(\mathrm{BMI})+0.013(\mathrm{BP})+0.014(\mathrm{BS})+0.839(\mathrm{BQ})}} \tag{18}$$

The results in column 7 of Table 17 were obtained using *Equation* (18)

Table 17: Probability computations for classification of reproductive status of a gonorrhea patient.

Age	History of Gonorrhea	Body Mass Index (BMI)	Blood Pressure (BP)	Blood Sugar (BS)	Bacteria Quantity (BQ)	Probability (Y)	Reproductive status of a Gonorrhea patient
52	1	21.333	130	279	0.103	0.97***	Infertile
25	0	26.9	120	114	0.302	0.14**	Fertile
34	1	26.439	100	168	0.206	0.65***	Infertile
64	1	20.08	130	116	0.502	0.91***	Infertile
40	0	22.676	120	127	0.168	0.28**	Fertile
49	1	26.8	139	132	0.123	0.86***	Infertile
16	0	24.9	120	104	0.033	0.04**	Fertile
38	1	21.4	135	164	0.092	0.68**	Infertile

***P(Y) greater than 0.5 = Infertile; **P(Y) less than 0.5 = Fertile
*P(Y) equal to 0.5 = Equal chances of being Infertile or Fertile

4. Conclusions

In this study, we used data on the eight (8) types of sexually transmitted infections (STIs) recorded from 2010 through 2020 in the Department of Micro Biology, University of Nigeria Teaching Hospital to obtain the most prevalent sexually transmitted infection. Firstly, the prevalence analysis method was used to determine the most prevalent sexually transmitted infection among eight (8) selected infections (Chlamydia, Gonorrhea, Syphilis, Trichomoniasis, Hepatitis B, Herpes, Human papilloma Virus (HPV), and Human Immuno-deficiency Virus (HIV)). The results showed that Gonorrhea is the most prevalent STI at 33.08%. Secondly, two-way CATANOVA cross-classification was used to ascertain the gender and age of those who always suffer from Gonorrhea infection and the results showed that gender, age, and its interaction effect were statistically significant at a 5% level. This implies that a particular gender and age interval always suffer from Gonorrhea infection. The data showed that the percentage of males is more than the percentage of females that suffer from Gonorrhea infection. The percentage of 30-39 years old suffer Gonorrhea infection more than any other age interval. The data also showed that 55.5% out of 6850 Gonorrhea patients, do not have a Gonorrhea infection history.

The results also showed that there is a spread of Gonorrhea infection between the genders. The significance of the interaction effect showed that the rate of contracting Gonorrhea infection by gender differs from one age group to another. The three-way CATANOVA result showed that the drug prescription for the treatments of Gonorrhea infection depends on gender and age at a 5% significance level. A logistic regression was performed to ascertain the effects of the variables coded as age, history, blood sugar, bacteria quantity, body mass index, and blood pressure on the likelihood that a Gonorrhea patient is infertile. The logistic regression model was statistically significant, $\chi^2_{(6)} = 1678.063$, P-value < 0.001. The model explained 33.4% (Nagelkerke R^2) of the variance in reproductive status (fertility or infertility) of Gonorrhea patients and correctly classified 90.6% of cases into the fertile and infertile groups. A Gonorrhea patient with a Gonorrhea history is 3.718 times more likely to be infertile than a Gonorrhea patient without a Gonorrhea history. An increase in age, body mass index, blood pressure, blood sugar, and bacteria quantity of a Gonorrhea patient were associated with an increased likelihood of being infertile.

The findings of this study also showed drugs used in treating Gonorrhea infection depend on the patient's gender and age which means that some drugs are not for the treatment of a Gonorrhea patient because of the patient's gender or age. Gonorrhea patients are advised not to lie about their age as it helps the physicians who prescribe these medicines and the female gender should know their pregnancy status to avoid health complications. Moreover, we noticed a spread of Gonorrhea infection between the genders since the number of Gonorrhea patients that do not have an infection history is more than those with an infection history.

The previous studies showed that these infections, especially Gonorrhea, caused infertility if poorly treated or left untreated over a long period. In this study, we use the fitted logistic regression model to make some predictions on the fertility of Gonorrhea patients. Our findings (see Table 17) showed that a Gonorrhea patient with a certain age, gonorrhea history, body mass index, blood pressure, and bacterial quantities can be infertile.

This study is an eye-opener to different types of sexually transmitted infections for Nigerians. The findings in this study showed that significant steps are to be used to create awareness and motivate adults about the need for regular health check-ups for proper termination or cure of these infections. More precisely, the concerned authorities need to make efforts to educate people on STIs and this may be through mass media, social media, schools, and any other means of communication. The authorities should also provide appropriate healthcare facilities in both urban and rural areas with government intervention for the benefit of the poor ones. These measures against STIs, especially Gonorrhea infection, with their risk reduce STIs drastically in Nigeria.

Acknowledgements

The authors would like to acknowledge the department of Micro Biology, University of Nigeria Teaching Hospital (UNTH) who made their data available for free for this research.

Data availability statement

We used secondary data from the Department of Micro Biology, University of Nigeria Teaching Hospital. The data had been presented in this research work and any other information needed on the data used in this work will be made available.

Conflict of Interest

We (authors) declare that there are no conflicts of interest.

Literature Cited

ANDERSON, T. W. (1958). *An introduction to multivariate analysis*. John Wiley: New York

ANDERSON, R. J., LANDIS, J. R. (1980). CATANOVA for multidimensional contingency tables: Nominal-scale response. *Commun. Statist. Theor. Meth.* 9:1191–1206.

ANDERSON, R. J., LANDIS, J. R. (1982). CATANOVA for multidimensional contingency tables: Ordinal-scale response. *Commun. Statist. Theor. Meth.* 11:257–270.

AZMI M.,MUATAZ A.,ALI M.A. and MOHAMMAD S.E (2008). Prevalence of Sexually Transmitted Infections Among Sexually Active Jordanian Females. *Sex Transm. Dis.*, 35 (6):607-710

BOATENG, E.Y. and ABAYE, D. A. (2019). A Review of the Logistic Regression Model with Emphasis on Medical Research. *Journal of Data Analysis and Information Processing*, 7: 190-207

CENTERS FOR DISEASES CONTROL AND PREVENTION (2013). Incidence, prevalence, and cost of living in the United States.

COX, D. R and SNELL, E. J. (1989). *Analysis of Binary Data*. Second Edition. Chpman & Hall.

CRAGG, J. G. and UHLER, R. S. (1970). The demand for automobiles. *The Canadian Journal of Economics* 3: 386-406.

D'AMBRA, L., BEH, J. E. and AMENTA, P. (2005). Analysis of contingency tables: Catanova for two-way contingency tables with ordinal variables using orthogonal polynomials. *Communications in Statistics—Theory and Methods*, 34: 1755–1769 DOI:10.1081/STA-200066325

DEYHOUL N., MOHAMADDOOST T. and HOSSINI M. (2017). Infertility related risk factors:A systematic review. *International Journal of women health and reproduction science*, 5(1):24-29.

EZE, N. M., ASOGWA, O. C. and EZE, C. M. (2021). Principal component factor analysis of some development factors in southern Nigeria and its extension to regression analysis. *Journal of Advances in Mathematics and Computer Science, 36*(3): 132-160. DOI:10.9734/JAMCS/2021/v36i330351

FIENBERG, S. E. (1973). Analysis of incomplete multiway contingency tables. Biometrics, 28:177-202. DOI: https://doi.org/10.2307/2528967

FISHER, L. D (1998). Self-designing clinical trials. Statistics in Medicine; 17:1551-1562

FLORIAN, T. J.(2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *J Mem Lang, 59*(4):434-446.

HOSMER, D. and LEMESHOW, S. (2000). *Applied Logistic Regression (Second Edition)*. New York: John Wiley & Sons, Inc.

KESAH F.N.C., VINCENT K.P and AUGUSTINE A.(2013). Prevalence and etiology of Sexually Transmitted Infections I gynecologic unit of a developing country. *Annals of tropical medicine and public health, 6*(5):526.

LAERD STATISTICS (2018). Binomial Logistic Regression using SPSS Statistics. Accessed 14 August, 2021. Available: https://statistics.laerd.com/spss-tutorials/binomial-logistic-regression-using-spss-statistics.php

LIGHT, R. J., MARGOLIN, B. H. (1971). An analysis of variance for categorical data. *J. Amer. Statist. Assoc.* 66:534–544.

MADDALA, G. S. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press.

MARGOLIN, B. H., LIGHT, R. J. (1974). An analysis of variance for categorical data II. Small samples comparisons with chi-square and other competitors. *J. Amer. Statist. Assoc.* 69:755–544

NAGELKERKE, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika* 78: 691-692.

ONUKOGU, I. B. (1985). Reasoning by analogy from ANOVA to CATANOVA, *Biom J,* 27:839-849.

_____ (2014). Analysis of variance of categorical data-nested designs. *Journal of Statistics: Advances in Theory and Applications*, 12: 109-116

OTARU O. P. and OGBONDA N. P (2020). CATANOVA analysis of knowledge and control practices of hepatitis B virus infection amongst tertiary university students. Galician Medical Journal, 27(1).

SCATTERWHITE, C.L.,TORRONE, E., MEITES, E., DUNNE, E.F., MAHAJAN, R., OCFEMIA, C.SU, J., XU,F. and WEINSTOCK, H. (2013). Sexually transmitted infections among U.S. women and men: Prevalence and incidence estimates, 2008 *Sexually Transmitted Diseases*;40 (3):187-193.

SCHEFFE, H. (1959). *The Analysis of Variance*. Wiley: New York

SINGH, B. (2004). CATANOVA for analysis of nominal data from repeated measures design. *J Ind Soc Agril Statist, 58*(3):257-268

WARNER, R.M. (2013). *Applied Statistics (2nd. Edition)*. Thousand Oaks, CA: SAGE.

WUENSCH, K. L. (2021). Binary Logistic Regression with SPSS. Accessed 14 September, 2021. https://core.ecu.edu/wuenschk/MV/Multreg/Logistic-SPSS.PDF

Appendix A

Computation of Sum of Squares for Two-way contingency table depicting the response of gender and ages of gonorrhea patients reported in UNTH from 2010-2020 (see Table 7).

$$TSS = 6850 - \frac{3049^2 + 3801^2}{6850} = 3383.72$$

$$WUSS = 6850 - \frac{128^2 + 383^2}{511} + \frac{395^2 + 437^2}{832} + \cdots + \frac{234^2 + 198^2}{432} = 3252.88$$

$$BRSS = 6850 - \frac{1721^2 + 2031^2}{3752} + \frac{1328^2 + 1770^2}{3098} = 3380.66$$

$$BCSS = 6850 - \frac{282^2 + 723^2}{1005} + \frac{667^2 + 935^2}{1602} + \cdots + \frac{581^2 + 449^2}{1030} = 3279.09$$

$$RSS = 3.06$$
$$CSS = 104.63$$
$$NSS = 23.15$$

Chi-square calculated

$$\chi^2{}_{RT} = \frac{(2-1) \times (6850-1) \times 3.06}{3383.72} = 6.19$$

$$\chi^2{}_{CT} = \frac{(2-1) \times (6850-1) \times 104.63}{3383.72} = 211.78$$

$$\chi^2{}_{NT} = \frac{(2-1) \times (6850-1) \times 23.15}{3383.72} = 46.86$$

Chi-square tabulated

$$\chi^2{}_{RT} = \chi^2{}_{(2-1)(2-1)} = \chi^2{}_{(1)} (at\ 5\%\ from\ chi - square\ table = 3.841)$$
$$\chi^2{}_{CT} = \chi^2{}_{(5-1)(2-1)} = \chi^2{}_{(4)} (at\ 5\%\ from\ chi - square\ table = 9.49)$$
$$\chi^2{}_{NT} = \chi^2{}_{(2-1)(5-1)(2-1)} = \chi^2{}_{(4)} (at\ 5\%\ from\ chi - square\ table = 9.49)$$

Appendix B

Computation for Three-way contingency table depicting gender, ages, and drug prescription for gonorrhea infection through 2010 – 2020 (see Table 9.1 and 9.2).

Test Statistic:

$$\chi^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} \frac{(n_{ijk} - f_{ijk})^2}{f_{ijk}} \sim \chi^2_{ijk-(i+j+k)+2}$$

Where i = number of drugs, j = number of age intervals in age, k = number of genders

$$f_{ijk} = n\left(\frac{n_{i++}}{n}\right)\left(\frac{n_{+j+}}{n}\right)\left(\frac{n_{++k}}{n}\right)$$

Appendix Table 1: Summary table for calculated observed and estimated expected frequencies.

Cell	Observed Frequencies (n_{ijk})	Estimated Expected Frequencies $(\hat{f}_{ijk})$	$\frac{(n_{ijk}-\hat{f}_{ijk})^2}{\hat{f}_{ijk}}$
f_{111}	110	119.74	0.79
f_{112}	188	190.87	0.04
f_{113}	206	231.85	2.88
f_{114}	171	150.95	2.66
f_{115}	118	122.72	0.18
f_{121}	129	98.87	9.18
f_{122}	196	157.60	9.36
f_{123}	229	191.44	7.37
f_{124}	80	124.64	15.99
f_{125}	63	101.33	14.50
f_{211}	107	107.68	0.01
f_{212}	163	171.65	0.44
f_{213}	202	208.51	0.20
f_{214}	164	135.76	5.88
f_{215}	112	110.36	0.02
f_{221}	111	88.91	5.49
f_{222}	135	141.73	0.32
f_{223}	216	172.17	11.16
f_{224}	55	112.09	29.08
f_{225}	75	91.13	2.85
f_{311}	109	89.44	4.28
f_{312}	123	142.57	2.69
f_{313}	178	173.19	0.13
f_{314}	115	112.76	0.04
f_{315}	116	91.67	6.46
f_{321}	61	73.85	2.24
f_{322}	102	117.72	2.10
f_{323}	110	143.00	7.62
f_{324}	103	93.10	1.05
f_{325}	96	75.69	5.45
f_{411}	130	136.05	0.27
f_{412}	192	216.87	2.85
f_{413}	248	263.44	0.90
f_{414}	205	171.52	6.54
f_{415}	159	139.44	2.74
f_{421}	104	112.34	0.62
f_{422}	211	179.07	5.69
f_{423}	184	217.52	5.17
f_{424}	128	141.62	1.31
f_{425}	132	115.13	2.47
f_{511}	55	97.56	18.57
f_{512}	166	155.51	0.71
f_{513}	174	188.90	1.18
f_{514}	148	122.99	5.08
f_{515}	93	99.99	0.49
f_{521}	89	80.55	0.89
f_{522}	126	128.40	0.05
f_{523}	199	155.98	11.87
f_{524}	98	101.55	0.12
f_{525}	66	82.56	3.32

$$\chi^2_{cal} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\frac{(n_{ijk}-f_{ijk})^2}{f_{ijk}} = 221.30$$

$\chi^2_{tab} = \chi^2_{ijk-(i+j+k)+2}$

Where i = 5, j = 2, k = 5(from the contingency Table 9.1)

$\chi^2_{tab} = \chi^2_{(5\times2\times5)-(5+2+5)+2} = \chi^2_{40} = 55.75$ (From the chi-square table, size 40 under 5% level of significance)

On Some Efficient Classes of Estimators Based on Higher Order Moments of an Auxiliary Attribute

Shashi Bhushan
Department of Statistics, University of Lucknow, Lucknow, India, 226007

Anoop Kumar[1]
Department of Statistics, Amity University, Lucknow, U.P., India, 226028

Abstract

This paper discusses the problem of estimating the population mean utilizing information on the mean and variance of qualitative characteristics. We introduce some efficient classes of estimators based on higher order moments such as the variance of an auxiliary attribute. The conventional mean estimator, Bhushan and Gupta (2016) estimator, and the traditional regression and ratio estimators proposed by Naik and Gupta (1996) are shown to be the sub-class of the proposed estimators for properly chosen valuations of the described scalars. The effective performance of the suggested estimators has been assessed empirically and theoretically with respect to the contemporary estimators.

Keywords: *mean square error, efficiency, qualitative characteristics*

1. Introduction

In survey research, it has been found that the consideration of auxiliary (supplementary) information assists to improve the efficiency of the estimator provided that there arises a high degree of correlation between the variable of interest and the supplementary variable. Several classes of improved and modified estimators have been suggested till date. Bhushan et al. (2020a, b, c, 2021a, b), Bhushan and Kumar (2020, 2022), and the references listed therein are a few recent noteworthy contributions in this regard. In real life scenarios, many times the study variable may be associated with some easily available qualitative auxiliary elements. For example: height of persons (y) and sex (ϕ); quantity of production of gram crop (y) and a specific variety of gram crop (ϕ); quantity of milk production (y) and a specific breed of cow (ϕ). Moreover, if measuring a quantitative variable is costly, then such an auxiliary attribute may be considered, which can be constructed from the auxiliary variable and is highly associated with the study variable. For example: (1) the tax paid by a company (y) may depend on its turnover (ϕ) which can be converted into a large/small companies; (2) the family expenditure (y) may depend on the household size (ϕ) which can be classified as large/small household; (3) the yield of crop (y) may depend on large/small land holdings (ϕ).

Various renowned authors suggested a wide range of classes of modified and improved population mean estimators consisting of auxiliary attributes in simple random sampling (SRS). The classical product, regression and ratio estimators were suggested by Naik and Gupta (1996) for the population mean estimation utilizing auxiliary attribute. Jhajj et al.

[1] **Address correspondence to Anoop Kumar:** Department of Statistics, Amity University, Lucknow, U.P., India; E-mail: anoop.asy@gmail.com

(2006) investigated a general class of population mean estimator based on auxiliary attribute. Influenced by the work of Ray and Singh (1981), Singh et al. (2008) developed a class of estimators utilizing auxiliary attribute. Adapting the procedure of Kadilar and Cingi (2006), Abd-Elfattah et al. (2010) developed a class of population mean estimators by combining different ratio estimators. Grover and Kaur (2011) extended their own work and investigated an improved population mean exponential estimator utilizing auxiliary attribute. Singh and Solanki (2012) proposed a more effective auxiliary attribute-based estimation method for population mean. Koyuncu (2012) envisaged an efficient population mean estimator utilizing auxiliary attribute. Bhushan and Gupta (2016) introduced a ratio type estimators based on higher order moments such as the variance of an attribute ϕ. A family of ratio exponential estimators consisting of an auxiliary attribute were explored by Zaman and Kadilar (2019). Zaman (2020) developed an exponential kind of estimators utilizing attribute. An improved auxiliary attribute-based log type estimator was suggested by Bhushan and Gupta (2020). Motivated by Bhushan et al. (2021b), Bhushan et al. (2022a) suggested a few attribute-based enhanced classes of estimators for population mean. In order to compute the population mean and variance utilizing auxiliary attribute, Bhushan et al. (2022b) presented certain linear combination type estimators.

This paper discusses some efficient auxiliary attribute-based classes of regression and ratio type estimators using higher order moments such as the variance. The following sections make up the schema of this paper. *Section* 2 devotes to the existing population mean estimators based on auxiliary attribute. In *Section* 3, a few efficient classes of regression and ratio type estimators have been proposed using higher order moments such as variance of auxiliary attribute along with their characteristics. The comparative study between the suggested and existing estimators has been performed in *Section* 4. Results of an empirical application are presented in *Section* 5 for the verification of the efficiency conditions followed by the discussion of the numerical results tabulated in Section 6. Finally, we reach to the conclusion in *Section* 7.

2. Review of Existing Estimators

To compute the population mean $\bar{Y}$ of the study variable y, let a simple random sample s of size n be drawn without replacement from a finite population $\aleph = (\aleph_1, \aleph_2, \ldots, \aleph_N)$ of size N. Let ϕ_i and y_i be the observations on the auxiliary attribute ϕ and the variable of choice y for unit i of the population $\aleph$. Note that if the unit $i \in \phi$ then ϕ_i=1 and if $i \notin \phi$ then ϕ_i=0. Suppose $A = \Sigma_{i=1}^{N}\phi_i$ and $a = \Sigma_{i=1}^{n}\phi_i$ represent the total number of units in the population $\aleph$ and sample s, respectively, with attribute ϕ, whereas $P = (A/N)$ and $p = (a/n)$, respectively, represent the population and sample proportion with attribute ϕ. Let $\bar{Y} = N^{-1}\Sigma_{i=1}^{N}y_i$ and $\bar{y} = n^{-1}\Sigma_{i=1}^{n}y_i$ be, respectively the population and sample means of study variable, y with the expressions $S_y = \sqrt{(N-1)^{-1}\Sigma_{i=1}^{N}(y_i - \bar{Y})^2}$ and $S_\phi = \sqrt{(N-1)^{-1}\Sigma_{i=1}^{N}(\phi_i - P)^2}$ be, respectively, the population standard deviation of the study variable y and the auxiliary attribute ϕ. Furthermore, let us define $S_{y\phi} = (N-1)^{-1}(\Sigma_{i=1}^{N}y_i\phi_i - N\bar{Y}P)$ and $\rho = S_{y\phi}/S_yS_\phi$ be the population covariance and population coefficient of correlation between the study variable y and the auxiliary attribute ϕ. Likewise, $C_y = S_y/\bar{Y}$ and $C_\phi = S_\phi/P$ be, respectively, the coefficient of variation of study variable y and auxiliary attribute ϕ.

We take into consideration the following expressions to define the mean square error (MSE) and bias of the estimators: $\bar{y} = \bar{Y}(1 + e_0)$, $p = P(1 + e_1)$ and $s_\phi^2 = S_\phi^2(1 + e_2)$ given $E(e_0) = E(e_1) = E(e_2) = 0$ and $E(e_0^2) = \gamma C_y^2$, $E(e_1^2) = \gamma C_\phi^2$, $E(e_2^2) = \gamma(\lambda_{04} - 1), E(e_0 e_1) = \gamma\rho C_y C_\phi$ $E(e_0 e_2) = \gamma C_y \lambda_{12}$ and $E(e_1 e_2) = \gamma\lambda_{03} C_\phi$ where $\gamma = (N - n)/Nn$, $\lambda_{ab} = \mu_{ab}/\mu_{20}^{a/2}\mu_{02}^{b/2}$, and $\mu_{ab} = (N-1)^{-1}\Sigma_{i=1}^{N}(y_i - \bar{Y})^a(\phi_i - P)^b$.

The classical mean estimator is stated as $T_m = \bar{y}$ while the variance of the estimator T_m is expressed as

$$V(T_m) = \gamma \bar{Y}^2 C_y^2 . \tag{1}$$

An attribute-based classical ratio estimator was presented by Naik and Gupta (1996) $T_r = \bar{y}\left(\frac{P}{p}\right)$ while the MSE of the estimator T_r is given by

$$MSE(T_r) = \gamma \bar{Y}^2 \left(C_y^2 + C_\phi^2 - 2\rho C_y C_\phi\right). \tag{2}$$

An attribute-based classical regression estimator was also presented by Naik and Gupta (1996) as

$$T_{lr} = \bar{y} + \beta_\phi (P - p),$$

where β_ϕ is the regression coefficient of y on ϕ. The MSE of the estimator T_{lr} is provided as

$$MSE(T_{lr}) = \gamma \bar{Y}^2 \left(C_y^2 + \beta_\phi^2 \frac{P^2}{\bar{Y}^2} C_\phi^2 - 2\beta_\phi \frac{P}{\bar{Y}} \rho C_y C_\phi\right).$$

By minimizing the $MSE(T_{lr})$ with respect to (w.r.t.) β_ϕ, we obtain $\beta_{\phi(opt)} = \rho\left(\frac{\bar{Y}C_y}{PC_\phi}\right)$. By replacing the value of β_ϕ with $\beta_{\phi(opt)}$ in the $MSE(T_{lr})$, we obtain

$$minMSE(T_{lr}) = \bar{Y}^2 \gamma C_y^2 (1 - \rho^2). \tag{3}$$

Motivated by the work of Ray and Singh (1981), Singh et al. (2008) investigated the undermentioned estimators utilizing information on auxiliary attribute as

$$T_s = \left\{\bar{y} + \beta_\phi (P - p)\right\}\left(\frac{m_1 P + m_2}{m_1 p + m_2}\right),$$

where β_ϕ is the same as defined earlier, $m_1 (\neq 0)$ and m_2 are either real numbers or functions of the known parameters of the attribute, namely, the population coefficient of kurtosis $\beta_2(\phi)$, the population coefficient of variation C_ϕ of the attribute ϕ and the population correlation coefficient ρ between the variable y and the attribute ϕ. Furthermore, some members of the estimator T_s are given hereunder for the suitably chosen values of constants m_1 and m_2 as

$$T_{s_1} = \left\{\bar{y} + \beta_\phi (P - p)\right\}\left(\frac{P}{p}\right),$$

$$T_{s_2} = \left\{\bar{y} + \beta_\phi (P - p)\right\}\left(\frac{P + \beta_2(\phi)}{p + \beta_2(\phi)}\right),$$

$$T_{s_3} = \{\bar{y} + \beta_\phi(P-p)\}\left(\frac{P+C_\phi}{p+C_\phi}\right),$$

$$T_{s_4} = \{\bar{y} + \beta_\phi(P-p)\}\left(\frac{P\beta_2(\phi)+C_\phi}{p\beta_2(\phi)+C_\phi}\right),$$

$$T_{s_5} = \{\bar{y} + \beta_\phi(P-p)\}\left(\frac{PC_\phi+\beta_2(\phi)}{pC_\phi+\beta_2(\phi)}\right),$$

$$T_{s_6} = \{\bar{y} + \beta_\phi(P-p)\}\left(\frac{P+\rho}{p+\rho}\right),$$

$$T_{s_7} = \{\bar{y} + \beta_\phi(P-p)\}\left(\frac{PC_\phi+\rho}{pC_\phi+\rho}\right),$$

$$T_{s_8} = \{\bar{y} + \beta_\phi(P-p)\}\left(\frac{P\rho+C_\phi}{p\rho+C_\phi}\right),$$

$$T_{s_9} = \{\bar{y} + \beta_\phi(P-p)\}\left(\frac{P\beta_2(\phi)+\rho}{p\beta_2(\phi)+\rho}\right),$$

$$T_{s_{10}} = \{\bar{y} + \beta_\phi(P-p)\}\left(\frac{P\rho+\beta_2(\phi)}{p\rho+\beta_2(\phi)}\right),$$

where $\beta_2(\phi) = \mu_{04}/\mu_{02}^2$ is the population coefficient of kurtosis of attribute ϕ. The MSE of the estimator $T_{s_i}, i = 1, 2, \ldots, 10$ is provided by

$$MSE(T_{s_i}) = \gamma\bar{Y}^2\{R_i^2 C_\phi^2 + C_y^2(1-\rho^2)\}, \tag{4}$$

Where $R_1 = 1$, $R_2 = P/(P+\beta_2(\phi))$, $R_3 = P/(P+C_\phi)$, $R_4 = P\beta_2(\phi)/(P\beta_2(\phi)+C_\phi)$, $R_5 = PC_\phi/(PC_\phi+\beta_2(\phi))$, $R_6 = P/(P+\rho)$, $R_7 = PC_\phi/(PC_\phi+\rho)$, $R_8 = P\rho/(P\rho+C_\phi)$, $R_9 = P\beta_2(\phi)/(P\beta_2(\phi)+\rho)$, and $R_{10} = P\rho/(P\rho+\beta_2(\phi))$.

Bhushan et al. (2015) suggested the auxiliary attribute-based logarithmic estimator under SRS as $T_{bg} = \bar{y}\left\{1 + \alpha\log\left(\frac{p^*}{P^*}\right)\right\}$, where α is a duly selected scalar. Additionally, $p^* = \eta p + \lambda$ and $P^* = \eta P + \lambda$ such that η and λ are either real numbers or functions of the known population parameters of the attribute, namely, the population coefficient of kurtosis $\beta_2(\phi)$, the population coefficient of variation C_ϕ of auxiliary attribute and the population correlation coefficient ρ between y and ϕ. The MSE of the estimator T_{bg} is given by

$$MSE(T_{bg}) = \gamma\bar{Y}^2\left(C_y^2 + \alpha^2\nu^2 C_\phi^2 + 2\alpha\nu\rho C_y C_\phi\right),$$

where $\nu = \eta P/(\eta P + \lambda)$. By minimizing the $MSE(T_{bg})$ w.r.t. α, we get $\alpha_{(opt)} = -\rho\left(\frac{C_y}{\nu C_\phi}\right)$

and by replacing the value of α with $\alpha_{(opt)}$ in the $MSE(T_{bg})$, we get

$$minMSE(T_{bg}) = \bar{Y}^2\gamma C_y^2(1-\rho^2). \tag{5}$$

This resembles the minimum MSE of the estimator for classical regression.

Bhushan and Gupta (2016) introduced the following class of estimator utilizing higher order moments such as the variance of an auxiliary attribute as $T_{bg1} = \bar{y}\left(\frac{P}{p}\right)^{\beta}\left(\frac{S_\phi^2}{s_\phi^2}\right)^{\theta}$, where β and θ are scalars. The MSE of the estimator T_{bg1} is given by

$$MSE(T_{bg1}) = \gamma\bar{Y}^2\{C_y^2 + \beta^2 C_\phi^2 + \theta^2(\lambda_{04} - 1) - 2\beta\rho C_y C_\phi + 2\beta\theta C_\phi\lambda_{03} - 2\theta C_y\lambda_{12}\}.$$

By minimizing the $MSE(T_{bg1})$ with respect to β and θ, we get $\beta_{(opt)} = \left(\frac{S_y}{S_\phi}\right)\left\{\frac{\rho(\lambda_{04}-1)-\lambda_{12}\lambda_{03}}{\lambda_{04}-1-\lambda_{03}^2}\right\}$, and $\theta_{(opt)} = \left(\frac{S_y}{S_\phi^2}\right)\frac{\rho(\lambda_{12}-\rho\lambda_{03})}{(\lambda_{04}-1-\lambda_{03}^2)}$. By replacing the value of β and θ with $\beta_{(opt)}$, and $\theta_{(opt)}$, respectively, in the $MSE(T_{bg1})$, we get

$$minMSE(T_{bg1}) = \gamma\bar{Y}^2C_y^2\left\{1 - \rho^2 - \frac{(\lambda_{12} - \rho\lambda_{03})^2}{(\lambda_{04} - 1 - \lambda_{03}^2)}\right\}. \quad (6)$$

We would like to mention that following Bhushan and Gupta (2016), one may also propose regression and Walsh (1970) type estimators using higher order moments such as the variance of auxiliary attribute but both will furnish similar results under the optimal conditions.

Following Singh et al. (2007), Zaman and Kadilar (2019) suggested the following estimator as $T_{zk} = \bar{y}\exp\left\{\frac{(\eta P+\lambda)-(\eta p+\lambda)}{(\eta P+\lambda)+(\eta p+\lambda)}\right\}$, where η and λ are same as previously specified. The MSE of the estimator T_{zk} is expressed as

$$MSE(T_{zk}) = \gamma\bar{Y}^2\left(\zeta^2 C_\phi^2 + C_y^2 - 2\zeta\rho C_y C_\phi\right), \quad (7)$$

where $\zeta = \eta P/2(\eta P + \lambda)$. A few members of the estimator T_{zk} are also included in Table 1 for quick reference. Following Ozel (2016), Zaman (2020) introduced an exponential ratio kind of estimator as $T_z = \bar{y}\left(\frac{p}{P}\right)^{\varsigma}\exp\left\{\frac{(\eta P+\lambda)-(\eta p+\lambda)}{(\eta P+\lambda)+(\eta p+\lambda)}\right\}$, where ς is a suitably chosen scalar and η and λ are same as previously specified. The MSE of the estimator T_z is expressed by

$$MSE(T_z) = \gamma\bar{Y}^2\left(C_y^2 + \varsigma^2 C_\phi^2 + \nu^2 C_\phi^2 - 2\varsigma\nu C_\phi^2 + 2\varsigma\rho C_y C_\phi - 2\nu\rho C_y C_\phi\right),$$

where $\nu = \eta P/(\eta P + \lambda)$. By minimizing the $MSE(T_z)$ w.r.t. ς, we get $\varsigma_{(opt)} = \nu - \frac{\rho C_y}{C_\phi}$. By replacing the value of ς with $\varsigma_{(opt)}$ in the $MSE(T_z)$, we get

$$minMSE(T_z) = \bar{Y}^2\gamma C_y^2(1 - \rho^2), \quad (8)$$

which is similar to the minimum MSE of the classical regression estimator T_{lr}.

Bhushan and Gupta (2020) suggested the log type family of estimator utilizing auxiliary attribute as $T_{bg2} = \left\{w_1\bar{y} + w_2\left(\frac{p}{P}\right)\right\}\left\{1 + \alpha\log\left(\frac{p^*}{P^*}\right)\right\}$, where w_1, w_2 and α are duly chosen scalars. Moreover, the optimum value of α can be determined from the log kind of estimators T_{bg} envisaged by Bhushan et al. (2015). Furthermore, p^* and P^* are the same

as defined earlier. A few members of the estimator T_{bg2} are also included in Table 1 for quick reference.

Table 1: Some members of Zaman and Kadilar (2019) and Bhushan and Gupta (2020) estimators.

Values of η	Values of λ	Members of Zaman and Kadilar (2019) estimators $T_{zk(j)}, j = 1,2,\dots,10$	Members of Bhushan and Gupta (2020) estimator $T_{bg2(j)}, j = 1,2,\dots,10$
1	$\beta_2(\phi)$	$T_{zk(1)}$	$T_{bg2(1)}$
1	C_ϕ	$T_{zk(2)}$	$T_{bg2(2)}$
$\boldsymbol{\beta_2(\phi)}$	C_ϕ	$T_{zk(3)}$	$T_{bg2(3)}$
$\boldsymbol{C_\phi}$	$\beta_2(\phi)$	$T_{zk(4)}$	$T_{bg2(4)}$
1	ρ	$T_{zk(5)}$	$T_{bg2(5)}$
$\boldsymbol{C_\phi}$	ρ	$T_{zk(6)}$	$T_{bg2(6)}$
$\boldsymbol{\rho}$	C_ϕ	$T_{zk(7)}$	$T_{bg2(7)}$
$\boldsymbol{\beta_2(\phi)}$	ρ	$T_{zk(8)}$	$T_{bg2(8)}$
$\boldsymbol{\rho}$	$\beta_2(\phi)$	$T_{zk(9)}$	$T_{bg2(9)}$
$\boldsymbol{S_\phi}$	$\beta_2(\phi)$	$T_{zk(10)}$	$T_{bg2(10)}$

The MSE of the estimator T_{bg2} is given by

$$MSE\left(T_{bg2}\right) = (\bar{Y}^2 w_1^2 A + w_2^2 B + \bar{Y}^2 w_1 D + \bar{Y} w_2 G + \bar{Y} w_1 w_2 F + \bar{Y}^2), \tag{9}$$

where $A = 1 + \gamma(C_y^2 + \alpha^2 v^2 C_\phi^2 + 4\alpha v \rho C_y C_\phi - \alpha v C_\phi^2)$, $B = 1 + \gamma(C_\phi^2 + \alpha^2 v^2 C_\phi^2 - \alpha v \rho C_\phi^2 + 4\alpha v C_\phi^2)$, $D = \gamma\left(\alpha v^2 C_\phi^2 - 2\alpha v \rho C_y C_\phi\right) - 2$, $G = \gamma\left(\alpha v^2 C_\phi^2 - 2\alpha v C_\phi^2\right) - 2$ and $F = 2 + 2\gamma\left(2\alpha v C_\phi^2 + 2\alpha v \rho C_y C_\phi + \rho C_y C_\phi - \alpha v^2 C_\phi^2 + \alpha^2 v^2 C_\phi^2\right)$.

By minimizing the $MSE(T_{bg2})$ w.r.t. w_1 and w_2, we obtain $w_{1(opt)} = \frac{(GF-2BD)}{(4AB-F^2)}$, and $w_{2(opt)} = \frac{\bar{Y}(DF-2GA)}{(4AB-F^2)}$. By replacing the values of w_1 and w_2 with $w_{1(opt)}$ and $w_{2(opt)}$, respectively, in the $MSE\left(T_{bg2}\right)$, we obtain

$$minMSE\left(T_{bg2}\right) = \bar{Y}^2 \left\{1 - \frac{(AG^2 + BD^2 - 2DFG)}{(4AB - F^2)}\right\}. \tag{10}$$

3. Proposed Classes of Estimators

The main goal of this article is to use data on the auxiliary characteristics to develop some efficient classes of estimators as an alternative to the estimators reviewed in the preceding section. Motivated by the work of Searls (1964) and Bhushan and Gupta (2016), we suggest some regression and ratio type estimators using higher order moments such as the variance of an auxiliary attribute ϕ as

$$T_{a1} = \alpha_1 \bar{y} + \beta_1 (P - p) + \theta_1 \left(S_\phi^2 - s_\phi^2\right),$$

$$T_{a2} = \alpha_2 \bar{y} \left(\frac{P}{p}\right)^{\beta_2} \left(\frac{S_\phi^2}{s_\phi^2}\right)^{\theta_2}, \text{ and}$$

$$T_{a3} = \alpha_3 \bar{y} \left\{\frac{P}{P+\beta_3(p-P)}\right\} \left\{\frac{S_\phi^2}{S_\phi^2+\theta_3\left(s_\phi^2-S_\phi^2\right)}\right\},$$

where $\alpha_i,\ \beta_i$ and $\theta_i,\ i = 1, 2, 3$ are suitably opted scalars. The classes of estimators $T_{ai},\ i = 1, 2, 3$ are reduced into:

- Usual mean estimator T_m for $(\alpha_i, \beta_i, \theta_i) = (1,0,0)$,
- Classical regression estimator T_{lr} for $(\alpha_1, \beta_1, \theta_1) = \left(1, \beta_\phi, 0\right)$,
- Classical ratio estimator T_r for $(\alpha_2, \beta_2, \theta_2) = (1, 1, 0)$
- Bhushan and Gupta (2016) estimator T_{bg1} for $(\alpha_2, \beta_2, \theta_2) = (1, \beta, \theta)$.

Theorem 3.1. The suggested class of estimator T_{a1} has the following bias and minimum MSE:

$$Bias(T_{a1}) = \bar{Y}(\alpha_1 - 1), \tag{11}$$

$$minMSE(T_{a1}) = \bar{Y}^2\left(1 - \alpha_{1(opt)}\right). \tag{12}$$

Proof. Consider the first estimator

$$T_{a1} = \alpha_1 \bar{y} + \beta_1(P - p) + \theta_1\left(S_\phi^2 - s_\phi^2\right).$$

Rewrite the estimator T_{a1} by utilizing the notations defined in Section 2, we get

$$T_{a1} = \alpha_1 \bar{Y} + \alpha_1 \bar{Y} e_0 - \beta_1 P e_1 - \theta_1 S_\phi^2 e_2.$$

By subtracting $\bar{Y}$ on both sides of the above expression, we get

$$T_{a1} - \bar{Y} = \alpha_1 \bar{Y} + \alpha_1 \bar{Y} e_0 - \beta_1 P e_1 - \theta_1 S_\phi^2 e_2 - \bar{Y}. \tag{13}$$

The bias of the estimator T_{a1} is obtained by taking the expectation on both sides of *Equation* 13 which results to $Bias(T_{a1}) = \bar{Y}(\alpha_1 - 1)$. On the other hand, the MSE of the estimator T_{a1} is obtained by squaring and taking the expectation on both sides of *Equation* 13 as

$$MSE(T_{a1}) = \{(\alpha_1 - 1)^2 \bar{Y}^2 + \alpha_1^2 \gamma S_y^2 + \beta_1^2 \gamma S_\phi^2 + \theta_1^2 \gamma S_\phi^4 (\lambda_{04} - 1) - 2\alpha_1\beta_1\gamma\rho S_y S_\phi - 2\alpha_1\theta_1\gamma S_y S_\phi^2 \lambda_{12} + 2\beta_1\theta_1\gamma S_\phi^3 \lambda_{03}\}.$$

By minimizing the $MSE(T_{a1})$ w.r.t. α_1, β_1 and θ_1, we get

$$\alpha_{1(opt)} = \frac{\bar{Y}^2}{\bar{Y}^2 + \gamma S_y^2\left\{1 - \rho^2 - \frac{(\lambda_{12} - \rho\lambda_{03})^2}{(\lambda_{04} - 1 - \lambda_{03}^2)}\right\}},$$

$$\beta_{1(opt)} = \alpha_{1(opt)}\left(\frac{S_y}{S_\phi}\right)\left\{\frac{\rho(\lambda_{04}-1)-\lambda_{12}\lambda_{03}}{\lambda_{04}-1-\lambda_{03}^2}\right\},$$

$$\theta_{1(opt)} = \alpha_{1(opt)}\left(\frac{S_y}{S_\phi^2}\right)\frac{(\lambda_{12}-\rho\lambda_{03})}{(\lambda_{04}-1-\lambda_{03}^2)}.$$

By substituting the optimum values of $\alpha_1,\ \beta_1$ and θ_1 in the $MSE(T_{a1})$, we get

$$\begin{aligned} minMSE(T_{a1}) = \Bigg[& \left(\alpha_{1(opt)} - 1\right)^2\bar{Y}^2 + \alpha_{1(opt)}^2\gamma S_y^2 + \alpha_{1(opt)}^2\left(\frac{S_y}{S_\phi}\right)^2\left\{\frac{\rho(\lambda_{04}-1)-\lambda_{12}\lambda_{03}}{\lambda_{04}-1-\lambda_{03}^2}\right\}^2\gamma S_\phi^2 + \\ & \alpha_{1(opt)}^2\left(\frac{S_y}{S_\phi^2}\right)^2\left\{\frac{\lambda_{12}-\rho\lambda_{03}}{\lambda_{04}-1-\lambda_{03}^2}\right\}^2\gamma S_\phi^4(\lambda_{04}-1) - \\ & 2\alpha_{1(opt)}^2\left(\frac{S_y}{S_\phi}\right)\left\{\frac{\rho(\lambda_{04}-1)-\lambda_{12}\lambda_{03}}{\lambda_{04}-1-\lambda_{03}^2}\right\}\gamma\rho S_y S_\phi - \\ & 2\alpha_{1(opt)}^2\left(\frac{S_y}{S_\phi^2}\right)\frac{(\lambda_{12}-\rho\lambda_{03})}{(\lambda_{04}-1-\lambda_{03}^2)}\gamma S_y S_\phi^2\lambda_{12} + \\ & 2\alpha_{1(opt)}^2\left(\frac{S_y^2}{S_\phi^3}\right)\left\{\frac{\rho(\lambda_{04}-1)-\lambda_{12}\lambda_{03}}{\lambda_{04}-1-\lambda_{03}^2}\right\}\left\{\frac{\lambda_{12}-\rho\lambda_{03}}{\lambda_{04}-1-\lambda_{03}^2}\right\}\gamma S_\phi^3\lambda_{03}\Bigg]. \end{aligned}$$

After simplification of the above equation, we get

$$\begin{aligned} minMSE(T_{a1}) &= \left[\left(\alpha_{1(opt)}^2 + 1 - 2\alpha_{1(opt)}\right)\bar{Y}^2 + \alpha_{1(opt)}^2\gamma S_y^2\left\{1 - \rho^2 - \frac{(\lambda_{12}-\rho\lambda_{03})^2}{(\lambda_{04}-1-\lambda_{03}^2)}\right\}\right], \\ &= \left(1 - 2\alpha_{1(opt)}\right)\bar{Y}^2 + \alpha_{1(opt)}^2\left[\bar{Y}^2 + \gamma S_y^2\left\{1 - \rho^2 - \frac{(\lambda_{12}-\rho\lambda_{03})^2}{(\lambda_{04}-1-\lambda_{03}^2)}\right\}\right], \\ &= \left(1 - 2\alpha_{1(opt)}\right)\bar{Y}^2 + \alpha_{1(opt)}^2\left(\frac{\bar{Y}^2}{\alpha_{1(opt)}}\right), \\ &= \bar{Y}^2\left(1 - \alpha_{1(opt)}\right). \end{aligned}$$

Theorem 3.2. The suggested classes of estimator $T_{ai}, i = 2,\ 3$ have the following bias and minimum MSE:

$$Bias(T_{ai}) = \bar{Y}(\alpha_i Q_i - 1), \tag{14}$$

$$\min MSE(T_{ai}) = \bar{Y}^2\left(1 - \frac{Q_i^2}{P_i}\right). \tag{15}$$

Proof. Consider the estimator T_{a2} as

$$T_{a2} = \alpha_2 \bar{y}\left(\frac{P}{p}\right)^{\beta_2}\left(\frac{S_\phi^2}{s_\phi^2}\right)^{\theta_2}.$$

Rewriting the above estimator by utilizing the notation defined in *Section* 2, we get

$$\begin{aligned} T_{a2} &= \alpha_2 \bar{Y}(1+e_0)(1+e_1)^{-\beta_2}(1+e_2)^{-\theta_2}, \\ &= \alpha_2 \bar{Y}(1+e_0)\left\{1-\beta_2 e_1 + \frac{\beta_2(\beta_2+1)}{2}e_1^2 + \cdots\right\}\left\{1-\theta_2 e_2 + \frac{\theta_2(\theta_2+1)}{2}e_2^2 + \cdots\right\}. \end{aligned}$$

Using Taylor series expansion, we expand the right-hand side of the above expression, multiplying out and neglecting the terms of e's having power greater than two, we get

$$T_{a2} = \alpha_2 \bar{Y}\left\{1 + e_0 - \beta_2 e_1 - \beta_2 e_0 e_1 + \frac{\beta_2(\beta_2+1)}{2}e_1^2 - \theta_2 e_2 - \theta_2 e_0 e_2 + \frac{\theta_2(\theta_2+1)}{2}e_2^2\right\}.$$

By subtracting $\bar{Y}$ on both sides of the above expression, we get

$$T_{a2} - \bar{Y} = \bar{Y}\left[\alpha_2\left\{1 + e_0 - \beta_2 e_1 - \beta_2 e_0 e_1 + \frac{\beta_2(\beta_2+1)}{2}e_1^2 - \theta_2 e_2 - \theta_2 e_0 e_2 + \frac{\theta_2(\theta_2+1)}{2}e_2^2\right\} - 1\right].$$

Letting the above equation as *Equation* 16 and taking the expectation on both sides of the equation, we get the bias of the estimator T_{a2} up to the first order of approximation as

$$\begin{aligned} Bias(T_{a2}) &= \bar{Y}\left[\alpha_2\left\{1 + \frac{\beta_2(\beta_2+1)}{2}\gamma C_\phi^2 + \frac{\theta_2(\theta_2+1)}{2}\gamma(\lambda_{04}-1) - \beta_2\gamma\rho C_y C_\phi - \theta_2\gamma C_y \lambda_{12}\right\} - 1\right] \\ &= \bar{Y}(\alpha_2 Q_2 - 1). \end{aligned}$$

The bias of the estimator T_{a3} can be obtained in the same manner.

Now, squaring and applying the expectation on both sides of *Equation* 16, we get the MSE up to the first order of approximation as

$$\begin{aligned} MSE(T_{a2}) = \bar{Y}^2\Big[1 + \alpha_2^2\{1 + \gamma C_y^2 + (2\beta_2^2+\beta_2)\gamma C_\phi^2 + (2\theta_2^2+\theta_2)\gamma(\lambda_{04}-1) - 4\beta_2\gamma\rho C_y C_\phi - \\ 4\theta_2\gamma C_y\lambda_{12} + 4\beta_2\theta_2\gamma C_\phi\lambda_{03}\} - 2\alpha_2\left\{1 + \frac{\beta_2(\beta_2+1)}{2}\gamma C_\phi^2 + \frac{\theta_2(\theta_2+1)}{2}\gamma(\lambda_{04}-1) - \right. \\ \left.\beta_2\gamma\rho C_y C_\phi - \theta_2\gamma C_y\lambda_{12} + \beta_2\theta_2\gamma C_\phi\lambda_{03}\right\}\Big], \end{aligned}$$

which can further be written as

$$MSE(T_{a2}) = \bar{Y}^2(1 + \alpha_2^2 P_2 - 2\alpha_2 Q_2). \qquad (17)$$

By minimizing the $MSE(T_{a2})$ w.r.t. α_2, we get $\alpha_{2(opt)} = \frac{Q_2}{P_2}$ and by replacing the value of α_2 with $\alpha_{2(opt)}$ in *Equation* 17, we get $minMSE(T_{a2}) = \bar{Y}^2\left(1 - \frac{Q_2^2}{P_2}\right)$. The MSE of the estimator T_{a3} can be determined on the same manner. In general, we can write

$$\begin{aligned} Bias(T_{ai}) &= \bar{Y}(\alpha_i Q_i - 1); i = 2, 3, \\ MSE(T_{ai}) &= \bar{Y}^2(1 + \alpha_i^2 P_i - 2\alpha_i Q_i). \qquad (18) \end{aligned}$$

By minimizing the $MSE(T_{ai})$ w.r.t. α_i, we get $\alpha_{i(opt)} = \frac{Q_i}{P_i}$ and by replacing the value of α_i with $\alpha_{i(opt)}$ in *Equation* 18 we get $\min MSE(T_{ai}) = \bar{Y}^2\left(1 - \frac{Q_i^2}{P_i}\right)$ where

$P_2 = 1 + \gamma\{C_y^2 + (2\beta_2^2 + \beta_2)C_\phi^2 + (2\theta_2^2 + \theta_2)(\lambda_{04} - 1) - 4\beta_2\rho C_y C_\phi - 4\theta_2 C_y\lambda_{12} + 4\beta_2\theta_2 C_\phi\lambda_{03}\}$,

$Q_2 = 1 + \gamma\left\{\frac{\beta_2(\beta_2+1)}{2}C_\phi^2 + \frac{\theta_2(\theta_2+1)}{2}(\lambda_{04} - 1) - \beta_2\rho C_y C_\phi - \theta_2 C_y\lambda_{12} + \beta_2\theta_2 C_\phi\lambda_{03}\right\}$,

$P_3 = 1 + \gamma\{C_y^2 + 3\beta_3^2 C_\phi^2 + 3\theta_3^2(\lambda_{04} - 1) - 4\beta_3\rho C_\phi C_y - 4\theta_3 C_y\lambda_{12} + 4\beta_3\theta_3 C_\phi\lambda_{03}\}$, and

$Q_3 = 1 + \gamma\{\beta_3^2 C_\phi^2 + \theta_3^2(\lambda_{04} - 1) - \beta_3\rho C_\phi C_y - \theta_3 C_y\lambda_{12} + \beta_3\theta_3 C_\phi\lambda_{03}\}$.

Furthermore, $\beta_{i(opt)} = \frac{S_y}{S_\phi}\left\{\frac{\rho(\lambda_{04}-1)-\lambda_{12}\lambda_{03}}{\lambda_{04}-1-\lambda_{03}^2}\right\}$ and $\theta_{i(opt)} = \frac{S_y}{S_\phi^2}\left(\frac{\lambda_{12}-\rho\lambda_{03}}{\lambda_{04}-1-\lambda_{03}^2}\right), i = 2, 3$ are utilized as the optimum values of scalars when $\alpha_i = 1$ is put in the corresponding estimators.

Corollary 3.1. The proposed classes of ratio type estimators $T_{ai}, i = 2, 3$ are superior than the proposed regression type of estimator T_{a1} if and only if $\frac{Q_i^2}{P_i} > \alpha_{1(opt)}$.

Proof: This can be shown by the comparison of *Equations* 12 and 15.

We would also like to note that *Theorem* 3.1 and *Theorem* 3.2 are important to obtain the efficiency conditions discussed in *Section* 4.

4. Comparative Study

We deduce the efficiency conditions by comparing the minimum MSEs of the suggested classes of estimators and existing estimators.

- From *Equation* 12 and *Equation* 1, we get $MSE((T_m) > MSE(T_{a1})$ when $\alpha_{1(opt)} > 1 - \gamma C_y^2$.
- From *Equation* 15 and *Equation* 1, we get $MSE(T_m) > MSE(T_{ai})$ when $\frac{Q_i^2}{P_i} > 1 - \gamma C_y^2, i = 2, 3$.
- From *Equation* 12 and *Equation* 2, we get $MSE(T_r) > MSE(T_{a1})$ when $\alpha_{1(opt)} > 1 - \gamma\left(C_y^2 + C_\phi^2 - 2\rho C_y C_\phi\right)$.
- From *Equation* 15 and *Equation* 2, we get $MSE(T_r) > MSE(T_{ai})$ when $\frac{Q_i^2}{P_i} > 1 - \gamma\left(C_y^2 + C_\phi^2 - 2\rho C_y C_{phi}\right), i = 2, 3$.
- From *Equation* 12 and *Equation* 3, we get $MSE(T_{lr}) > MSE(T_{a1})$ when $\alpha_{1(opt)} > 1 - \gamma C_y^2(1 - \rho^2)$.
- From *Equation* 15 and *Equation* 3, we get $MSE(T_{lr}) > MSE(T_{ai})$ when $\frac{Q_i^2}{P_i} > 1 - \gamma C_y^2(1 - \rho^2), i = 2, 3$.
- From *Equation* 12 and *Equation* 4, we get $MSE\left(T_{sj}\right) > MSE(T_{a1})$ when $\alpha_{1(opt)} > 1 - \gamma\{R_j^2 C_\phi^2 + C_y^2(1 - \rho^2)\}, j = 1, 2, \ldots, 10$.
- From *Equation* 15 and *Equation* 4, we get $MSE\left(T_{sj}\right) > MSE(T_{ai})$ when $\frac{Q_i^2}{P_i} > 1 - \gamma\{R_j^2 C_\phi^2 + C_y^2(1 - \rho^2)\}, i = 2, 3$.
- From *Equation* 12 and *Equation* 6, we get $MSE\left(T_{bg1}\right) > MSE(T_{a1})$ when $\alpha_{1(opt)} > 1 - \gamma C_y^2\left\{1 - \rho^2 - \frac{(\lambda_{12}-\rho\lambda_{03})^2}{\lambda_{04}-1-\lambda_{03}}\right\}$.
- From *Equation* 15 and *Equation* 6 , we get $MSE\left(T_{bg1}\right) > MSE(T_{ai})$ when $\frac{Q_i^2}{P_i} > 1 - \gamma C_y^2\left\{1 - \rho^2 - \frac{(\lambda_{12}-\rho\lambda_{03})^2}{\lambda_{04}-1-\lambda_{03}}\right\}, i = 2, 3$.
- From *Equation* 12 and *Equation* 7, we get $MSE(T_{zk}) > MSE(T_{a1})$ when $\alpha_{1(opt)} > 1 - \gamma\left(\zeta^2 C_\phi^2 + C_y^2 - 2\zeta\rho C_y C_\phi\right), i = 2, 3$.

- From *Equation* 15 and *Equation* 7, we get
  $MSE(T_{zk}) > MSE(T_{ai})$ when $\frac{Q_i^2}{P_i} > 1 - \gamma\left(\zeta^2 C_\phi^2 + C_y^2 - 2\zeta\rho C_y C_\phi\right), i = 2, 3.$
- From *Equation* 12 and *Equation* 8, we get
  $MSE(T_z) > MSE(T_{a1})$ when $\alpha_{1(opt)} > 1 - \gamma C_y^2(1 - \rho^2)$.
- From *Equation* 15 and *Equation* 8, we get
  $MSE(T_z) > MSE(T_{ai})$ when $\frac{Q_i^2}{P_i} > 1 - \gamma C_y^2(1 - \rho^2), i = 2, 3.$
- From *Equation* 12 and *Equation* 9, we get
  $MSE\left(T_{bg2}\right) > MSE(T_{a1})$ when $\alpha_{1(opt)} > \frac{AG^2 + BD^2 - 2DFG}{(4AB - F^2)}$.
- From *Equation* 15 and *Equation* 9, we get
  $MSE\left(T_{bg2}\right) > MSE(T_{ai})$ when $\frac{Q_i^2}{P_i} > \frac{AG^2 + BD^2 - 2DFG}{(4AB - F^2)}, i = 2, 3.$

Under the conditions mentioned above, the proposed estimators $T_{ai},\ i = 1, 2, 3$ outperform the traditional mean estimator, traditional ratio and regression estimators, Singh et al. (2008) estimators, Bhushan et al. (2015) estimator, Bhushan and Gupta (2016) estimator, Zaman and Kadilar (2019) estimators, Zaman (2020) estimator and Bhushan and Gupta (2020) estimators. Successively, an empirical assessment was conducted using two different real populations to verify the above efficiency conditions.

5. Empirical Assessment

This empirical application used two real populations which are discussed below.

Population 1. (Origin: Sukhatme and Sukhatme (1970), pp. 256)

y: area (in acres) under the wheat crop inside the circles,

ϕ: a circle based on more than five villages,

$N = 89,\ n = 23,\ \overline{Y} = 1102,\ P = 0.124, C_y = 0.65,\ C_\phi = 2.678,$ and $\rho = 0.624.$

Population 2. (Origin: Singh and Chaudhary (1986), pp. 141)

y: area under lime (in acres),

ϕ: number of bearing lime trees (> 500).

N =22, n =12, $\overline{Y}$ =22.62091, P =0.5, C_y =1.4609, C_ϕ =1.0235, and ρ =0.6292.

We now compute the MSE and percent relative efficiency (PRE) for the above two populations. The PRE is calculated for the existing and proposed estimators T w.r.t. the traditional mean estimator T_m utilizing the expression: $PRE = \frac{MSE(T_m)}{MSE(T)} \times 100$. The outcomes of this empirical asessement for both populations are presented in Table 2 by MSE and PRE, demonstrating how the suggested classes of estimators outperform the existing estimators presented in *Section* 2.

Table 2: MSE and PRE of different estimators

Estimators	Population 1		Population 2	
	MSE	PRE	MSE	PRE
T_m	16559.39	100.00	41.36	100.00
T_r	212259.00	7.80	25.21	164.07
T_{lr}	10111.56	163.76	25.00	165.46
T_{s1}	290912.50	5.69	45.30	91.30
T_{s2}	10220.13	162.02	25.11	164.71
T_{s3}	10658.08	155.36	27.18	152.15
T_{s4}	23878.64	69.34	36.44	113.51
T_{s5}	10841.23	152.74	25.12	164.67
T_{s6}	17786.42	93.10	28.98	142.72
T_{s7}	43841.90	37.77	29.08	142.21
T_{s8}	10331.60	160.27	26.12	158.35
T_{s9}	94946.42	17.44	39.00	106.05
T_{s10}	10154.47	163.07	25.04	165.14
T_{bg1}	10342.32	160.11	12.11	341.42
$T_{zk(1)}$	15731.42	105.15	40.02	103.34
$T_{zk(2)}$	14798.41	111.78	35.92	115.13
$T_{zk(3)}$	10561.38	156.63	30.53	135.45
$T_{zk(4)}$	14551.86	113.68	39.99	103.42
$T_{zk(5)}$	11421.86	144.83	34.28	120.64
$T_{zk(6)}$	10240.09	161.55	34.20	120.92
$T_{zk(7)}$	15404.10	107.39	37.36	110.72
$T_{zk(8)}$	14403.59	114.85	29.72	139.16
$T_{zk(9)}$	16026.43	103.22	40.49	102.15
$T_{zk(10)}$	16263.84	101.71	40.64	101.76
T_z	10111.56	163.76	25.00	165.46
$T_{bg2(1)}$	14142.35	116.97	17.01	243.05
$T_{bg2(2)}$	15209.51	108.76	16.45	251.35
$T_{bg2(3)}$	20479.82	80.77	11.95	345.93
$T_{bg2(4)}$	15492.64	106.78	16.96	243.76
$T_{bg2(5)}$	19245.03	85.96	12.76	323.98
$T_{bg2(6)}$	21995.06	75.21	15.15	272.92
$T_{bg2(7)}$	14516.17	113.96	12.39	333.81
$T_{bg2(8)}$	21362.40	77.44	13.28	311.45
$T_{bg2(9)}$	13805.71	119.82	17.83	231.91
$T_{bg2(10)}$	13534.92	122.22	18.10	228.43
T_{a1}	**10028.06**	**165.13**	**11.85**	**348.88**
T_{a2}	**9883.88**	**167.53**	**11.83**	**349.51**
T_{a3}	**10030.91**	**165.08**	**11.88**	**348.06**

6. Results and Discussion

From the reported findings in *Table* 2, the proposed estimators $\boldsymbol{T_{ai}}$, i=1,2,3 outperform the following:

- The traditional mean per unit estimator T_m, classical regression and ratio estimators T_{lr} and T_r, Bhushan et al. (2015) estimator T_{bg}, Bhushan and Gupta (2016) estimators T_{bg1}, Zaman (2020) estimator T_z.
- The members $T_{s_i}, i = 1,2, \dots, 10$ of Singh et al. (2008) estimator T_s.
- The member $T_{zk(j)}, j = 1, 2, \dots, 10$ of Zaman and Kadilar (2019) estimators T_{zk}.
- The members $T_{bg2(j)}, j = 1, 2, \dots, 10$ of Bhushan and Gupta (2020) estimators T_{bg2}.

Furthermore, the proposed class of estimator T_{a2} is found to be the most efficient among the proposed classes of estimators $T_{ai}, i = 1, 2, 3$ for both populations by having the minimum MSE and maximum PRE.

7. Conclusion

This manuscript suggests some efficient classes of regression and ratio kind of estimators for the computation of population mean $\bar{Y}$ of the variable of interest y utilizing a higher order moments such as the variance of an auxiliary attribute ϕ. The conventional mean estimator T_m, ratio estimator T_r, regression estimator T_{lr} and Bhushan and Gupta (2016) estimator T_{bg_1} are found to be the members of the envisaged classes of estimators for duly opted valuations of the characterizing scalars. The bias and MSE expressions have been obtained and the efficiency conditions have been derived by comparing the MSE of the proposed classes of estimators with the MSE of the contemporary estimators.

Furthermore, an empirical assessment utilized two real populations to verify the credibility of the efficiency conditions. The numerical results have been found to be highly rewarding with minimum MSE and maximum PRE exhibiting superiority over the traditional mean estimator T_m, classical regression and ratio estimators T_{lr} and T_r, members $T_{s_i}, i = 1,2, \dots, 10$ of Singh et al. (2008) estimator T_s, Bhushan et al. (2015) estimator T_{bg}, Bhushan and Gupta (2016) estimator T_{bg_1}, members $T_{zk(j)}$; j =1,2,...,10 of Zaman and Kadilar (2019) estimators T_{zk}, Zaman (2020) estimator T_z and members $T_{bg_2(j)}$; j =1,2,...,10 of Bhushan and Gupta (2020) estimators T_{bg_2}. Moreover, it has been also seen from the numerical results that the ratio type estimator T_{a_2} performs superior among the proposed estimators. Thus, due to their uncontested performance, the proposed estimators are highly recommended to the surveyors for the estimation of population mean $\bar{Y}$ of variable of interest y such that the information is present as auxiliary attribute ϕ.

Literature Cited:

ABD-ELFATTAH, A.M., EL-SHERPIENY, E.A., MOHAMED, S.M. and ABDOU, O.F. 2010. "Improvement in Estimating the Population Mean in Simple Random Sampling Using Information on Auxiliary Attribute." *Applied Mathematics and Computation* 215: 4198-4202.

BHUSHAN, S., GUPTA, R. and PANDEY, S.K. 2015. "Improved Searl's type Logarithmic Estimators Using Auxiliary Information." *Statistical and Mathematical Sciences & their Applications,* Narosa Publishing House, New Delhi.

_____ and GUPTA, R. 2016. "Efficient Class of Estimators for Population Mean Using Attribute." *Recent Advances in Applied Statistics and its Applications* 225-229.

_____ and GUPTA, R. 2020. "An Improved Log-Type Family of Estimators Using Attribute." *Journal of Statistics and Management Systems 23*(3): 593-602.

_____ and KUMAR, A. 2020. "Log Type Estimators of Population Mean Under Ranked Set Sampling." *Predictive Analytics using Statistics and Big Data: Concepts and Modelling* 28: 47-74.

_____ and KUMAR, A. 2022. "Novel Log Type Class of Estimators Under Ranked Set Sampling." *Sankhya B*, 84: 421-447.

_____, GUPTA, R., SINGH, S. and KUMAR, A. 2020a. "A Modified Class of Log Type Estimators for Population Mean Using Auxiliary Information on Variable." *International Journal of Applied Engineering Research* 15(6): 612-627.

_____, GUPTA, R., SINGH, S. and KUMAR, A. 2020b. "Some Improved Classes of Estimators Using Auxiliary Information." *International Journal for Research in Applied Science & Engineering Technology* 8(VI): 1088-1098.

_____, GUPTA, R., SINGH, S. and KUMAR, A. 2020c. "A New Efficient Log-Type Class of Estimators Using Auxiliary Variable." *International Journal of Statistics and Systems* 15(1): 19-28.

_____, KUMAR, A., KUMAR, S. and SINGH, S. 2021a. "An Efficient Class of Estimators of Population Mean Under Simple Random Sampling." *The Philippine Statistician* 70(1): 33-47.

_____, KUMAR, A. and SINGH, S. 2021b. "Some Efficient Classes of Estimators Under Stratified Sampling." *Communications in Statistics - Theory and Methods* 1-30. [Online] DOI:10.1080/03610926.2021.1939052.

_____, KUMAR, A., TYAGI, D. and SINGH, S. 2022a. "On Some Improved Classes of Estimators Under Stratified Sampling Using Attribute." *Journal of Reliability and Statistical Studies* 15(1): 187-210.

_____, KUMAR, A., KUMAR, S. and SINGH, S. 2022b. "Some Modified Classes of Estimators for Population Variance Using Auxiliary Attribute." *Pakistan Journal of Statistics* 38(2): 235-252.

GROVER, L.K. and KAUR, P. 2011. "An Improved Exponential Estimator of Finite Population Mean in Simple Random Sampling Using an Auxiliary Attribute." *Applied Mathematics and Computation* 218(7): 3093-3099.

JHAJJ, H.S., SHARMA, M.K., GROVER, L.K. 2006. "A Family of Estimators of Population Mean Using Information on Auxiliary Attribute." *Pakistan Journal of Statistics* 22(1): 43-50.

KADILAR, C. and CINGI, H. 2006. "Improvement in Estimating the Population Mean in Simple Random Sampling." *Applied Mathematics Letters* 19: 75-79.

KOYUNCU N. 2012. "Efficient Estimators of Population Mean Using Auxiliary Attributes." *Applied Mathematics and Computation* 218(22): 10900-10905.

NAIK, V.D. and GUPTA, P.C 1996. "A Note on Estimation of Mean With Known Population Proportion of An Auxiliary Character." *Journal of Indian Society of Agricultural Statistics* 48(2): 151-158.

OZEL, K.G. 2016. "A New Exponential Type Estimator for the Population Mean in Simple Random Sampling." *Journal of Modern Applied Statistical Methods* 15(2): 207-214.

RAY, S.K. and SINGH, R.K. 1981. "Difference-Cum-Ratio Type Estimators." *Journal of Indian Statistical Association* 19: 147-151.

SEARLS, D.T. 1964. "The Utilization of A Known Coefficient of Variation in the Estimation Procedure." *Journal of American Statistical Association* 59: 1225-1226.

SINGH, D. and CHAUDHARY, F.S. 1986. "*Theory and Analysis of Sample Survey Designs."* New Age International Limited, New Delhi.

SINGH, R., CHAUHAN, P., SAWAN, N. and SMARANDACHE, F. 2008. "Ratio Estimators in Simple Random Sampling Using Information on Auxiliary Attribute." *Pakistan Journal of Statistics and Operation Research* IV(1): 47-53.

SINGH, H.P. and SOLANKI, R.S. 2012. "Improved Estimation of Population Mean in Simple Random Sampling Using Information on Auxiliary Attribute." *Applied Mathematics and Computation* 218(15): 7798-7812.

SUKHATME, P.V. and SUKHATME, B.V 1970. "*Sampling Theory of Surveys with Applications."* Iowa State University Press, Ames, U.S.A.

ZAMAN, T. and KADILAR, C. 2019. "Novel Family of Exponential Estimators Using Information of Auxiliary Attribute." *Journal of Statistics and Management Systems 22*(8): 1499-1509.

_____. 2020. "Generalized Exponential Estimators for the Finite Population Mean." *Statistics in Transition New Series* 21(1): 159-168.

Application of Consecutive Sampling Technique in a Clinical Survey for an Ordered Population: Does it Generate Accurate Statistics?

Mohamad Adam Bujang[1]
Clinical Research Centre, Sarawak General Hospital, Ministry of Health, Malaysia

Tg Mohd Ikhwan Tg Abu Bakar Sidik
Department of Pharmacology, Faculty of Medicine, Universiti Kebangsaan Malaysia

Nadiah Sa'at
Institute for Clinical Research, Ministry of Health, Malaysia

ABSTRACT

This study aims to compare the statistical generalizations which are inferred from samples obtained by both systematic sampling and consecutive sampling and then compare both their results with the true population parameters of the target population. This study was conducted using two approaches. The first approach was a comparison between sample statistics and population parameters based on a simulation analysis to estimate the population parameters from three types of statistical distributions (i.e. Normal, Exponential, and Poisson) by using seven sub-samples and 1000 iterations. The second approach was a comparison between sample statistics and population parameters based on real-life data sets which comprise six sub-samples and four parameters. Based on results from the simulation analysis, systematic sampling offers a greater advantage by having a smaller value of mean square error (MSE) in 40 out of 70 comparisons (57.1%) while consecutive sampling has a smaller value of MSE in 29 out of 70 comparisons (41.4%). There is only one MSE comparison that was identical between systematic sampling and consecutive sampling. Based on a validation approach, systematic sampling produced more accurate statistics than consecutive sampling with six out of eight comparisons. In summary, systematic sampling offers a better advantage in terms of accuracy. However, consecutive sampling is still able to generate valid and accurate statistics despite the fact that it is a type of non-probability sampling, especially if a sufficiently large sample size has been obtained for statistical analysis. Therefore, it is recommended that in any situation when it can be difficult to apply a systematic sampling technique for a particular clinical setting, researchers may opt to apply the consecutive technique in the recruitment process as an alternative, with a limitation on making generalizations about the target population.

Keywords: *population parameters; sample statistics; systematic sampling.*

[1] **Address correspondence to Mohamad Adam Bujang:** adam@crc.gov.my

1. Introduction

In the medical field, patients are usually arranged in an ordered sequence such as a consecutive list of patients obtaining their appointments for seeking medical treatment. There is usually no sampling frame available since both the total population size and the entire population list are sometimes unknown and unspecified. Therefore, in an observational study, patient recruitment is mostly conducted by using consecutive sampling (Flamaing et al. 2015; Jensen et al. 2015; Weigner et al., 2015). It has already been known that consecutive sampling is typically far better than convenience sampling in controlling and minimizing the risk of sampling bias (Polit and Beck, 2010). Convenience sampling involves a recruitment procedure in which all respondents will be selected for inclusion in the sample merely by virtue of being conveniently available to the researcher. On the other hand, consecutive sampling shall necessitate a researcher to recruit all respondents on a first-come and first-served basis so long as the subjects fulfill all the eligibility requirements stipulated by both the inclusion and exclusion criteria (Bowers et al., 2011; Bujang, 2017).

Consecutive sampling is a type of non-probability sampling method which is commonly adopted as the sampling technique for an ordered population, particularly in a clinical survey. Consecutive sampling is defined as a sampling method that obtains the sample in a consecutive manner (i.e. on a first-come, first-served basis) after having established that the sample has fulfilled all the stipulated eligibility criteria and the recruitment process shall continue until the desired sample size has been achieved (Bowers et al., 2011; Bujang, 2017). In addition, consecutive sampling is also designed to recruit all the eligible subjects that can be ranked in a specified order, which is usually based on the earliest date and time. For example, only those patients who have fulfilled all eligibility criteria and are being notified for an earlier appointment at a hospital or in the clinic can be invited into the study within a stipulated timeframe (Chew et al., 2013; Bujang et al., 2015).

It is already well-known that the recommended probability sampling technique for an ordered population is systematic sampling (Fraenkal and Wallen, 2006). However, a major obstacle to adopting the systematic sampling technique in a clinical survey is the difficulty of prespecifying or setting the interval k since in some cases the total population number is not known. Therefore, the researcher may opt to apply a consecutive sampling technique even though it is a type of non-probability sampling. Therefore, this study aims to assess the viability of adopting consecutive sampling in lieu of systematic sampling by determining the accuracy of the sample estimates obtained by consecutive sampling. This can be achieved by comparing the sample statistics produced by consecutive sampling versus those produced by systematic sampling, and finally, a comparison is then made of all the sample estimates with the true value of the parameter in the target population.

Figure 1 is shown to visualize the recruitment procedure for both consecutive sampling and systematic sampling. Say, a population size (N) is designated as 8 and sample size (n) is designated as 4. For consecutive sampling, a researcher will recruit subject 1, subject 2, subject 3, and subject 4 to comply with the requirements on a ‘first-come and first-served basis. On the other hand, for systematic sampling, a researcher shall first and foremost have to calculate the value of interval (k), such as $N/n = 2$, and then shall perform the first

subject selection by random. Say, if the first chosen subject by random is 2, this means that the researcher shall recruit a series of subjects as subject 2, subject 4, subject 6, and subject 8.

Figure 1: Sample selection based on consecutive sampling and systematic sampling

As consecutive sampling is nonetheless still a type of non-probability sampling, it is therefore always necessary to determine to what extent it is able to generate accurate sample estimates for the target population. In such cases, it is, therefore, necessary to compare the sample statistics with the true population parameters which have been obtained between both consecutive sampling and systematic sampling techniques, through an initial simulation process and subsequently a validation process by using real-life data sets. Such findings are important for determining to what extent, within the remit of a sufficiently large pre-specified sample size, the estimates derived from the consecutive sampling are also close approximations of the statistics derived from the systematic sampling, and hence they shall be regarded as equally accurate and valid for representing the true values of the parameters of the target population as those of systematic sampling do. However, it should be emphasized that there is still a limitation on the generalizations about the population that can be made if consecutive sampling is used as this is a non-probability sample.

2. Study Methods

This study was conducted using two approaches. The first approach involved a simulation analysis and the second approach involved a validation technique based on a comparison of the estimates obtained by using real-life data sets.

2.1 First Approach: Simulation Analysis (This simulation analysis was conducted based on three distribution data sets; namely, normally distributed, exponentially distributed, and distributed as Poisson)

Step 1: Generate five sets of population data based on three different statistical distributions with a selected parameter as presented in Table 2. The total population size (N) is set at 3000.

Step 2: Design a sampling strategy by using systematic sampling initially and then by using consecutive sampling. The systematic sampling procedure for recruiting subjects is based on prespecifying the interval, k^{th} of five, and the first patient is selected at random from a single random number between one and k. For systematic sampling, the sample size is set at 600, and thus, therefore $k = 3000/600 = 5$. For consecutive sampling, the selection of subjects is performed in a consecutive manner (i.e on a first-come, first-served basis).

Step 3: Calculate the statistics for the mean and standard deviation for each sampling procedure. In each population, there were seven sub-samples as *n*, are 30, 50, 100, 200, 300, 500, and 1000.

Step 4: Iterate this procedure 1000 times. For each of the 1000 iterations, the mean square error (MSE) will be calculated for seven sub-samples within each population based on the two different sampling techniques. The comparisons will be made based on MSE from the two sampling techniques. Altogether, there are a total of 70 comparisons (five statistical distributions × two parameters × seven sub-samples).

2.2 Second Approach: Comparison of sample statistics and parameters between a sample obtained from consecutive sampling and another sample obtained from systematic sampling via a validation process

A validation process was conducted to determine the degree of proximity between population parameters and the sample estimates derived from samples obtained by using the two different sampling methods, namely: consecutive sampling and systematic sampling. This validation was performed in two different populations, by conducting four statistical analyses on six to eight samples. Both the definitions and explanations of the three key component data (namely: population data, statistical analysis, and sample size) are provided in Table 1. Data for this validation step have been obtained from "An Audit of Diabetes Control and Management (ADCM) 2009", which involved the collection of data from all patients with diabetes mellitus within all government health clinics in Malaysia, during the year 2009 at a national level (Ismail et al., 2009).

Table 1: Definition and explanation of the population, statistical analysis and sample size for the validation.

Component	Definition and explanation
Population	
P1 = All patients with diabetes mellitus notified in Health Clinic A, from January 1st until December 31st 2009 (N=1688)	
P2 = All patients with diabetes mellitus notified in Health Clinic B, from January 1st until December 31st 2009 (N=1986)	
Statistical analysis	
Descriptive	Mean (μ) of HbA1c for overall
Descriptive	Mean (μ) of HbA1c among male
Correlation (Spearman)	Correlation (*r*) between age and HbA1c
Multivariate (ANCOVA)	Marginal mean (μ) of HbA1c among male after controlled for age
Sample size	
Sample of 30, 50, 100, 150, 200, 300, 500 and 1000	

This study selected two tertiary health clinics with a relatively large number of patients (i.e. >1000 patients) so a series of analyses for the validation was conducted for a few sub-samples which displayed a wide range of different sizes from small to big. The consecutive sampling procedure for recruiting patients was based on a first-come, first-served basis, which depended on the notification date within the registry database. On the other hand, the systematic sampling procedure for recruiting patients was based on prespecifying the interval k^{th} of five with assumption $N = 1500$ and $n = 300$ ($N/n = 1500/300 = 5$). Then

randomly selecting the first patient who can belong to any rank number which is a single random number between one and *k*.

Four parameters were selected as a basis for comparison, namely: mean (μ) HbA1c, mean (μ) HbA1c among the male gender, correlation (*r*) between age and HbA1c, and adjusted mean calculated based on Analysis of Covariance (ANCOVA). The glycated haemoglobin A1c (HbA1c) test is a clinical measure that describes the patient's average level of blood sugar control over the past 2 to 3 months. A total number of 48 comparisons were performed (six sub-samples × four parameters × two populations). The sample estimates and reference parameters obtained from both sampling techniques were then compared.

The comparisons were made based on the selection of a sampling technique that can produce statistics with the minimum bias (between parameter and statistics) on average from the six sub-samples. The analysis involves descriptive analysis such as the calculation of mean, a univariate analysis via the correlation test, and a multivariate analysis by the Analysis of Covariance. All the statistical analyses for this study were performed using R software (R Core Team, 2014) and SPSS (IBM Corp. Released 2011. IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp.)

3. Results

Both Table 2 and Table 3 illustrate the findings obtained from a simulation analysis. For a comparison of means, systematic sampling offers a smaller value of MSE in 16 out of 35 comparisons while consecutive sampling offers a smaller value of MSE in 19 out of 35 comparisons. For a comparison of standard deviations, systematic sampling has reported smaller values of MSE in 24 out of 35 comparisons while consecutive sampling has reported smaller values of MSE in 10 out of 35 comparisons. There are a total of 35 comparisons of MSE on SD. In one of the comparisons, the MSE for systematic sampling and consecutive sampling are the same. Overall, systematic sampling has provided a slightly greater advantage by reporting smaller MSEs in 40 out of 70 (or 57.1%) of the comparisons while consecutive sampling has reported smaller MSEs in only 29 out of 70 (or 41.4%) of the comparisons.

Table 2: The comparison of statistics between statistics derived from systematic sampling and consecutive sampling, results from a simulation analysis based on normal, exponential and Poisson distributions.

Parameter	Sample size	Systematic Sampling		Consecutive Sampling	
		Mean (95% CI)	SD (95% CI)	Mean (95% CI)	SD (95% CI)
Normal Distribution μ = 80 σ = 5	30	80.014 (78.233 - 81.746)	4.934 (3.636 - 6.165)	79.985 (78.215 - 81.647)	4.946 (3.656 - 6.302)
	50	80.002 (78.632 - 81.352)	4.978 (4.012 - 6.051)	80.001 (78.729 - 81.238)	4.955 (3.895 - 5.940)
	100	80.017 (79.073 - 81.021)	4.971 (4.249 - 5.649)	80.007 (79.051 - 80.933)	4.984 (4.301 - 5.717)
	200	80.023 (79.371 - 80.730)	4.990 (4.485 - 5.515)	80.005 (79.325 - 80.684)	5.001 (4.519 - 5.498)
	300	79.998 (79.426 - 80.542)	4.993 (4.595 - 5.400)	80.012 (79.489 - 80.539)	5.006 (4.617 - 5.420)
	500	79.993 (79.598 - 80.412)	4.996 (4.706 - 5.292)	80.003 (79.580 - 80.462)	5.014 (4.695 - 5.293)
	1000	79.997 (79.697 - 80.304)	5.006 (4.800 - 5.228)	80.006 (79.706 - 80.301)	5.005 (4.813 - 5.214)
Normal Distribution μ = 80 σ = 15	30	80.042 (74.698 - 85.237)	14.803 (10.909 - 18.496)	79.955 (74.645 - 84.941)	14.838 (10.968 - 18.905)
	50	80.006 (75.895 - 84.057)	14.933 (12.037 - 18.153)	80.004 (76.188 - 83.713)	14.864 (11.684 - 17.819)
	100	80.052 (77.218 - 83.064)	14.913 (12.748 - 16.948)	80.021 (77.153 - 82.798)	14.951 (12.902 - 17.150)
	200	80.070 (78.113 - 82.189)	14.969 (13.456 - 16.546)	80.015 (77.974 - 82.052)	15.003 (13.557 - 16.495)
	300	79.993 (78.279 - 81.627)	14.979 (13.784 - 16.200)	80.035 (78.467 - 81.616)	15.018 (13.852 - 16.259)
	500	79.979 (78.793 - 81.236)	14.988 (14.118 - 15.876)	80.008 (78.739 - 81.386)	15.041 (14.085 - 15.880)
	1000	79.991 (79.092 - 80.912)	15.018 (14.401 - 15.684)	80.018 (79.118 - 80.903)	15.014 (14.438 - 15.641)
Normal Distribution μ = 80 σ = 25	30	80.070 (71.163 - 88.728)	24.671 (18.181 - 30.826)	79.925 (71.074 - 88.236)	24.729 (18.281 - 31.508)
	50	80.010 (73.159 - 86.761)	24.888 (20.062 - 30.255)	80.007 (73.647 - 86.188)	24.773 (19.473 - 29.698)
	100	80.086 (75.363 - 85.107)	24.856 (21.247 - 28.247)	80.035 (75.255 - 84.664)	24.918 (21.504 - 28.583)
	200	80.116 (76.855 - 83.648)	24.948 (22.427 - 27.577)	80.025 (76.623 - 83.421)	25.005 (22.596 - 27.492)
	300	79.989 (77.131 - 82.711)	24.965 (22.973 - 27.000)	80.058 (77.445 - 82.694)	25.029 (23.086 - 27.098)
	500	79.965 (77.988 - 82.060)	24.980 (23.530 - 26.460)	80.013 (77.898 - 82.310)	25.068 (23.476 - 26.466)
	1000	79.984 (78.487 - 81.519)	25.030 (24.002 - 26.140)	80.030 (78.529 - 81.505)	25.023 (24.063 - 26.068)
Exponential Distribution λ = 0.0125	30	79.156 (54.180 - 110.300)	76.836 (45.789 - 124.763)	80.379 (53.278 - 109.967)	77.765 (45.269 - 117.010)
	50	80.012 (60.163 - 102.489)	78.554 (53.473 - 109.301)	79.939 (59.505 - 102.181)	78.297 (51.871 - 111.098)
	100	80.352 (66.199 - 95.804)	79.783 (59.519 - 102.411)	80.081 (65.625 - 95.991)	79.245 (58.140 - 103.558)
	200	80.158 (70.035 - 91.483)	79.690 (66.141 - 96.548)	80.243 (69.416 - 92.091)	79.700 (65.563 - 96.726)
	300	80.125 (71.531 - 88.944)	79.876 (68.756 - 93.617)	80.141 (71.617 - 89.699)	79.810 (67.705 - 93.785)
	500	80.113 (73.317 - 87.362)	79.987 (70.997 - 90.226)	80.063 (73.359 - 87.045)	79.859 (70.176 - 90.819)
	1000	79.904 (74.870 - 84.991)	79.844 (73.508 - 86.659)	79.971 (75.237 - 84.665)	79.988 (73.449 - 86.807)
Poisson Distribution λ = 4	30	4.026 (3.367 - 4.768)	1.991 (1.476 - 2.548)	3.992 (3.267 - 4.734)	1.972 (1.460 - 2.527)
	50	3.987 (3.480 - 4.561)	1.981 (1.600 - 2.404)	3.986 (3.440 - 4.540)	1.980 (1.596 - 2.419)
	100	3.999 (3.630 - 4.380)	1.996 (1.713 - 2.292)	3.993 (3.590 - 4.410)	1.989 (1.705 - 2.305)
	200	3.991 (3.725 - 4.310)	1.994 (1.791 - 2.198)	4.001 (3.715 - 4.280)	1.999 (1.791 - 2.222)
	300	4.000 (3.787 - 4.213)	1.998 (1.830 - 2.175)	4.002 (3.770 - 4.237)	2.001 (1.823 - 2.187)
	500	4.000 (3.826 - 4.170)	1.998 (1.873 - 2.132)	4.001 (3.826 - 4.180)	2.001 (1.867 - 2.136)
	1000	4.000 (3.890 - 4.117)	2.001 (1.909 - 2.095)	4.002 (3.876 - 4.123)	2.002 (1.910 - 2.096)

Table 3: Comparison of mean square error between systematic sampling and consecutive sampling based on normal, exponential and Poisson distributions on different sample sizes.

Parameter	Sample size	Systematic Sampling		Consecutive Sampling	
		MSE for Mean	MSE for SD	MSE for Mean	MSE for SD
Normal Distribution μ = 80 σ = 5	30	0.7911	**0.3995**	**0.7632**	0.4715
	50	0.4779	**0.2606**	**0.4483**	0.2612
	100	0.2492	0.1259	**0.2248**	**0.1234**
	200	0.1133	0.0621	**0.1115**	**0.0587**
	300	0.0751	0.0386	**0.0656**	**0.0373**
	500	**0.0391**	**0.0191**	0.0393	0.0200
	1000	0.0167	**0.0080**	**0.0155**	0.0081
Normal Distribution μ = 80 σ = 15	30	7.1197	**3.5951**	**6.8691**	4.2435
	50	**4.3012**	**2.3451**	4.0351	2.3509
	100	2.2427	1.1330	**2.0236**	**1.1110**
	200	1.0195	0.5588	**1.0034**	**0.5281**
	300	0.6755	**0.3475**	**0.5905**	0.3356
	500	**0.3520**	**0.1717**	0.3533	0.1798
	1000	0.1503	**0.0723**	**0.1397**	0.0726
Normal Distribution μ = 80 σ = 25	30	19.7770	**9.9863**	**19.0808**	11.7874
	50	11.9477	**6.5142**	**11.2085**	6.5304
	100	6.2297	3.1473	**5.6211**	**3.0861**
	200	2.8319	1.5521	**2.7871**	**1.4670**
	300	1.8764	0.9654	**1.6403**	**0.9323**
	500	**0.9776**	**0.4770**	0.9813	0.4993
	1000	0.4174	**0.2009**	**0.3880**	0.2016
Exponential Distribution λ = 0.0125	30	**203.8285**	380.6237	210.3496	**359.6566**
	50	**117.9314**	**226.2870**	121.7321	227.4735
	100	**58.3774**	**110.5930**	60.1264	121.8113
	200	**28.4590**	**57.6523**	30.2410	61.7133
	300	**18.3875**	**36.8262**	18.8156	39.6324
	500	**10.3341**	**21.8541**	10.5601	22.3768
	1000	**4.0119**	8.2338	4.1569	**7.9792**
Poisson Distribution λ = 4	30	0.1332	**0.0770**	**0.1326**	0.0777
	50	**0.0774**	**0.0430**	0.0803	0.0455
	100	**0.0371**	**0.0216**	0.0406	0.0225
	200	0.0201	**0.0104**	**0.0193**	0.0111
	300	**0.0119**	**0.0067**	0.0122	0.0073
	500	**0.0066**	**0.0036**	0.0068	0.0039
	1000	**0.0025**	0.0015	0.0028	0.0015

Note: MSE with bold refers as smaller MSE, a comparison between systematic sampling versus consecutive sampling
For means comparisons, systematic sampling is smaller in terms of MSE in 16 out of 35 comparisons while consecutive sampling is smaller in 25 out of 35 comparisons.
For standard deviations, systematic sampling is smaller in terms of MSE in 19 out of 35 comparisons while consecutive sampling is smaller in 10 out of 35 comparisons.
For overall comparisons, systematic sampling is smaller in terms of MSE in 35 out of 70 comparisons while consecutive sampling is smaller in 34 out of 70 comparisons. One comparison of MSE was identical.

Table 4 presents the results of the validation process which involves a comparison of the sample estimates of population parameters with the true population parameters, which have been obtained by both consecutive sampling and systematic sampling. These sample estimates of population parameters obtained by both sampling techniques were also more closely approximating to the true population parameters when the sample size was at least 300, which is at least 15.1% (i.e. [300/1986] × 100%) to 17.8% (i.e. [300/1688] × 100%) of the population size. Table 5 shows the comparisons that were made between the population parameters and the sample estimates, which were obtained by both consecutive sampling and systematic sampling.

Table 4: Comparison of statistics and parameter between results from consecutive sampling (CS) and systematic sampling (SS) from various types of analysis

Parameter to be measured	Population	Sample	Parameter	*n*=30	*n*=50	*n*=100	*n*=150	*n*=200	*n*=300	*n*=500	*n*=1000
Mean (μ) HbA1c	A	CS	8.23	7.67	7.49	7.38	7.58	7.54	7.58	7.70	7.99
		SS	8.23	7.60	7.67	7.40	7.67	7.96	8.10		
	B	CS	8.11	7.16	7.90	7.75	7.80	7.92	7.93	8.08	8.14
		SS	8.11	7.95	7.97	8.41	8.37	8.33	8.29		
Mean (μ) HbA1c among male	A	CS	8.22	8.22	7.83	7.68	7.62	7.59	7.65	7.77	7.93
		SS	8.22	7.70	7.71	7.68	8.05	7.99	8.12		
	B	CS	8.00	6.80	7.89	8.02	7.99	8.05	7.89	7.93	8.01
		SS	8.00	8.00	8.13	8.35	8.22	8.26	8.18		
Correlation (*r*)[a] between age and HbA1c	A	CS	-0.25	-0.22	-0.33	-0.28	-0.21	-0.21	-0.16	-0.22	-0.26
		SS	-0.25	-0.14	-0.15	-0.19	-0.19	-0.28	-0.25		
	B	CS	-0.22	-0.09	-0.08	-0.07	-0.14	-0.23	-0.25	-0.25	-0.20
		SS	-0.22	-0.02	-0.30	-0.38	-0.24	-0.23	-0.18		
ANCOVA (marginal mean, μ)	A	CS	8.24	8.22	7.87	7.72	7.64	7.61	7.67	7.80	7.98
		SS	8.24	8.15	7.82	7.66	8.00	7.98	8.11		
	B	CS	8.02	6.76	7.89	8.01	7.99	8.04	7.92	7.96	8.04
		SS	8.02	7.99	8.25	8.38	8.22	8.24	8.17		

Note: Statistics for systematic sampling were calculated until sample size of 300 due to limited sample size.
[a]Spearman's correlation test was applied instead of Pearson's correlation test since parametric assumption was violated.

Table 5: Comparison of differences between the estimates and parameter between sample derived from consecutive sampling (CS) and systematic sampling (SS)

Parameter to be measured	Population	Sample	*n*=30	*n*=50	*n*=100	*n*=150	*n*=200	*n*=300	Mean[a]
Mean (μ) HbA1c	A	CS	0.56	0.74	0.85	0.65	0.69	0.65	0.69
		SS	0.63	0.56	0.83	0.56	0.27	0.13	**0.50**
	B	CS	0.95	0.21	0.36	0.31	0.19	0.18	0.37
		SS	0.16	0.14	-0.3	-0.26	-0.22	-0.18	**0.21**
Mean (μ) HbA1c among male	A	CS	0	0.39	0.54	0.6	0.63	0.57	0.46
		SS	0.52	0.51	0.54	0.17	0.23	0.1	**0.35**
	B	CS	1.2	0.11	-0.02	0.01	-0.05	0.11	0.25
		SS	0.00	-0.13	-0.35	-0.22	-0.26	-0.18	**0.19**
Correlation (*r*) between age and HbA1c	A	CS	-0.03	0.08	0.03	-0.04	-0.04	-0.09	**0.05**
		SS	-0.11	-0.10	-0.06	-0.06	0.03	0.00	0.06
	B	CS	-0.13	-0.14	-0.15	-0.08	0.01	0.03	0.09
		SS	-0.20	0.08	0.16	0.02	0.01	-0.04	0.09
ANCOVA (marginal mean, μ)	A	CS	0.02	0.37	0.52	0.60	0.63	0.57	0.45
		SS	0.09	0.42	0.58	0.24	0.26	0.13	**0.29**
	B	CS	1.26	0.13	0.01	0.03	-0.02	0.10	0.26
		SS	0.03	-0.23	-0.36	-0.20	-0.22	-0.15	**0.20**

Note: [a] Reported absolute value of mean where values in bold refers to smaller value
Systematic sampling has smaller bias (difference between parameter and statistics in average) in six out of eight comparisons

Results obtained from these comparisons revealed that systematic sampling was able to provide more accurate estimates of the population parameters than consecutive sampling because it had shown six out of eight comparisons with the minimum differences and also one of the comparisons had ended up in a tie. These findings illustrated an important observation whereby systematic sampling is shown to yield smaller differences between the

sample estimates and the population parameters than consecutive sampling does, indicating that it can provide a higher level of precision in these sample estimates.

4. Discussions

Due to both cost and time constraints, consecutive sampling is often deployed for recruiting a sample in a survey, especially within the medical discipline (Bujang, 2017). However, as consecutive sampling is a form of non-probability sampling which can likely be introducing bias when collecting a sample; it is, therefore, necessary to assess and evaluate the accuracy of sample estimates which have been derived from consecutive sampling. Various studies were conducted to determine whether or not the different sampling methods can potentially affect the generalisability of these findings. A study conducted by Howes (1985) had previously stated that the sampling technique would not have much influence on the study outcome so long as the researchers were able to control those variables that could potentially introduce bias (Howes et al., 1985).

Results obtained from the simulation analysis in this study further substantiated this statement, which revealed that the study findings remain unaffected regardless of whether consecutive sampling or systematic sampling technique was deployed for sample recruitment, as long as the subjects have been randomly arranged in an ordered sequence (in other words, there is no pre-existing systematic pattern that appears in the order). On the other hand, various other research studies also found that the choice of sampling technique could potentially influence the accuracy of the sample estimates, in that the use of a probability sampling method (such as systematic sampling) can potentially improve the accuracy and precision of the sample estimates obtained (Yeager et al., 2011; Erens et al., 2014).

However, the results based on the simulation analysis of this study did not provide conclusive evidence that consecutive sampling is equivalent to systematic sampling. It is well-known that systematic sampling is proven to be an unbiased and efficient sampling technique. Consecutive sampling is regarded as a type of non-probability sampling and thus it will not be possible to defy its inherent bias. Hence, the ultimate aim of this paper is not to demonstrate equivalence between consecutive sampling and systematic sampling. Nonetheless, this paper hopes to illustrate by using real examples that although consecutive sampling is a form of non-probability sampling, the statistics derived from this sampling technique can still be likely to produce highly accurate statistics especially when its sample size is sufficiently large.

To this end, this paper has included an additional validation approach to determine to what extent the statistics derived from samples are close approximations to their respective parameters of the target population. In inferential statistics, the p-value is always regarded as an indicator that provides evidence for making inferences from raw observations. However, no one will know for sure whether the inference is valid or not unless a population or census study is conducted to validate the accuracy of these sample estimates. Again, this explains why this study also incorporates a validation process to validate these sample statistics apart from conducting an initial simulation analysis.

Since the true values of these population parameters are already known, the purpose of this validation process is to validate the sample statistics against the true population parameters. Hence, the purpose of validation, in this case, is not for making inferences, which means that the indicators such as the absolute differences between sample estimates and population parameters (i.e. sample statistics minus population parameters) were being calculated, instead of p-values and their 95% confidence intervals.

From the results, it can be deduced that even though consecutive sampling is a type of non-probability sampling technique, the sample estimates derived from this sampling technique can be close approximations to the population parameters especially if the sample size is adequately large. This is an important finding obtained from this study, which conducted a validation procedure by using four different types of statistical analysis, including both univariate and multivariate analyses. This finding was also found to be consistent with those of previous studies whereby the sample statistics derived from consecutive sampling were found to be almost equivalent to the true population parameters especially when the sample size was sufficiently large (Bujang et al., 2012; Bujang et al., 2015).

Based on the theory of probability sampling, the sample estimate derived from the systematic sampling method is both unbiased and efficient because it is a type of probability sampling (Taro, 1967). Therefore, based on the validation results of this study, our findings have shown that sample statistics derived from systematic sampling provided better estimates when compared to those obtained from consecutive sampling. Hence, by basing on real-life data sets, this study has successfully contended that apart from providing an unbiased estimate, a probability sampling technique such as systematic sampling can also produce better and more accurate estimates than a non-probability sampling technique such as consecutive sampling.

In conclusion, it is already well-known that sample statistics derived from the systematic sampling method will yield better and more accurate estimates than those from the consecutive sampling method. However, consecutive sampling will still be able to elicit a sample that can provide estimates for an ordered population that consists of randomly-arranged units, especially if a relatively large sample is used. This statement will hold provided there is no pre-existing consecutive pattern for the arrangement of population data. If necessary, a runs test can always be applied to ensure there is no consecutive pattern that pre-exists in a string of subjects before deploying the consecutive sampling technique to recruit subjects (Bujang and Sapri, 2018). However, it should be emphasized that there is limitation in the use of consecutive sample in making statistical inferences about the population.

One of the major limitations of this study is that the simulation was conducted only on a few selected distributions with specific population parameters. Future studies may explore the possibility of investigating other findings which are based on various other simulation techniques and models in order to compare findings derived from both sampling techniques. Secondly, the validation was conducted by using only two real-life data sets from the medical field. However, it is beyond the scope of this study to determine the applicability of this observation within all the various fields of study by performing an audit of the various real-

life data sets from among all the various fields. Despite the above, cumulative evidence has supported the contention that when the sample size is adequately large, it becomes likely for those estimates that were derived from the samples to more closely mimic the true parameters of the intended populations (Bujang et al., 2012; Bujang et al., 2015; Stockwell and Peterson, 2001; Hernandez et al., 2006).

Acknowledgement

The authors wish to thank the Director General of Health, Ministry of Health Malaysia for permission to publish this paper. A special thanks to Mr Hon Yoon Khee for his effort to proofread the article. We also would like to thank Madam Nur Khairul Bariyyah Mohd Hatta and Ms Lim Chien Joo for their help in processing the manuscript for publication.

Conflict of Interest Statement

All the authors declared we have no conflict of interest.

Funding Statement

This study receives no funding

Literature Cited

BOWERS, D., HOUSE, A. and OWENS, D., 2011, Carrying out a systematic search: Getting started in health research, United Kingdom: John Wiley and Sons Ltd.

BUJANG, M.A., 2017, Enhancement of systematic sampling for clinical survey: Systematic sampling with consecutive approach, Selangor: Institute of Graduate Studies, Universiti Teknologi Mara.

_____, GHANI, P.A., ZOLKEPALI, N.A., SELVARAJAH, S. and HANIFF, J., 2012, A comparison between convenience sampling versus systematic sampling in getting the true parameter in a population: Explore from a clinical database: The Audit Diabetes Control Management (ADCM) registry in 2009. In: ICSSBE 2012. Proceedings, 2012 International Conference on Statistics in Science, Business and Engineering Empowering Decision Making with Statistical Sciences, 10–12 September 2012, Kedah, Malaysia: 499-503. doi:10.1109/ICSSBE.2012.6396615.

_____, MUSA, R., LIU, W.J., CHEW T.F., LIM, C.T.S. and MORAD, Z., 2015, Depression, anxiety and stress among patients with dialysis and the association with quality of life, Asian J Psychiatr, 18: 49-52.

_____, SA'AT, N., JOYS, A.R. and ALI, M.M., 2015, An audit of the statistics and the comparison with the parameter in the population. In: AIP Conference Proceedings, Shah Alam, Malaysia, 1682(1): 050019. doi:10.1063/1.4932510.

_____and SAPRI, F.E., 2018, An application of the runs test to test for randomness of observations obtained from a clinical survey in an ordered population, Malays J Med Sci, 25(4): 146-151. doi:10.21315/mjms2018.25.4.15

CHEW, B.H., SHARIFF-GHAZALI, S., MASTURA, I., HANIFF, J. and BUJANG, M.A., 2013, Age ≥ 60 years was an independent risk factor for diabetes-related complications despite good control of cardiovascular risk factors in patients with type 2 diabetes mellitus, Exp Gerontol, 48(5):485–491. doi: 10.1016/j.exger.2013.02.017.

ERENS, B., BURKILL, S., COUPER, M.P., CONRAD, F., CLIFTON, S., and TANTON, C., 2014, Nonprobability Web Surveys to Measure Sexual Behaviors and Attitudes in the General Population: A Comparison with a Probability Sample Interview Survey. J Med Internet Res, 16:e276-1–276-14. doi:10.2196/jmir.3382

FLAMAING, J., DE BACKER, W., VAN LAETHEM, Y., HEIJMANS, S., and MIGNON, A., 2015, Pneumococcal lower respiratory tract infections in adults: an observational case-control study in primary care in Belgium. BMC Fam. Pract. 16. doi:10.1186/s12875-015-0282-1.

FRAENKAL, J.R. and WALLEN, N.E., 2006, How to design and evaluate research in education, New York: McGraw-Hill.

HERNANDEZ, P.A., GRAHAM, C.H., MASTER, L.L. and ALBERT, D.L., 2006, The effect of sample size and species characteristics on performance of different species distribution modelling methods. Ecography, 29(5): 773-785. doi:10.1111/j.0906-7590.2006.04700.x

HOWES, B.L., DACEY, J.W.H. and WAKEHAM, S.G., 1985, Effects of sampling technique on measurements of porewater constituents in salt marsh sediments, Limnology and Oceanograph, 30(1): 221-227. doi:10.4319/lo.1985.30.1.0221

ISMAIL, M., CHEW, B.H., LEE, P.Y., AI, T.C., SHARIFF, G.S., and HANIFF, J. 2011, Control and treatment profiles of 70,889 adult type 2 diabetes mellitus patients in Malaysia - A cross sectional survey in 2009, Int J Collab Res Intern Med Public Health, 3(1): 98-113.

JENSEN, L.F., HVIDBERG, L., PEDERSEN, A.F., and VEDSTED, P., 2015, Symptom attributions in patients with colorectal cancer. BMC Fam Pract, 16:1–10.

POLIT, D.F. and BECK, C.T., 2010, Essentials of Nursing Research: Appraising Evidence for Nursing Practice. Lippincott Williams & Wilkins. pp. 311–312.

STOCKWELL, D.R.B. and PETERSON, A.T., 2001, Effects of sample size on accuracy of species distribution models, Ecol. Model, 148(1): 1–13. doi:10.1016/S0304-3800(01)00388-X

TARO, Y., 1967, Elementary sampling theory, 1st ed., Englewood Cliffs (NJ): Prentice-Hall.

WIEGNER, L., HANGE, D., BJORKELUND, C., JR and AHLBORG, G., 2015, Prevalence of perceived stress and associations to symptoms of exhaustion, depression and anxiety in a working age population seeking primary care - an observational study. BMC Fam Pract **16**, 38 doi:10.1186/s12875-015-0252-7

YEAGER, D.S., KROSNICK, J.A., CHANG, L.C., JAVITZ, H.S., LEVENDUSKY, M.S. and SIMPSER, A., 2011, Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. Public Opin Quart, 75: 709-747. doi:10.1093/poq/nfr020

PSAI Officers and Board of Directors

2022 Executive Committee Members

President	Rosalinda P. Bautista
Vice President	Wilhelmina C. Mañalac
Secretary	Wilma A. Perante
Treasurer	Teresita B. Deveza
Ex-Officio, Execom *(Immediate Past President)*	Dennis S. Mapa

PSAI Board of Directors

A. Individual Members

- **Joselito R. Basilio**
  Deputy Director, Department of Economic Research, Bangko Sentral ng Pilipinas
- **Rosalinda P. Bautista**
  Former Assistant Secretary and Deputy National Statistician, PhilSys Registry Office, Philippine Statistics Authority
- **Teresita B. Deveza**
  Former Deputy Director, Bangko Sentral ng Pilipinas (Retired)
- **Minerva Eloisa P. Esquivias**
  OIC Deputy National Statistician for Censuses and Technical Coordination, Philippine Statistics Authority
- **Maqtahar L. Manulon**
  Regional Director, Philippine Statistics Authority (RSSO XII)
- **Wilhelmina C. Mañalac**
  Former Assistant Governor, Bangko Sentral ng Pilipinas
- **Dennis S. Mapa**
  Undersecretary, National Statistician and Civil Registrar General, Philippine Statistics Authority
- **Genelyn Ma. F. Sarte**
  Associate Professor, UPD School of Statistics

B. Institutional Members - Government Sector

Philippine Statistics Authority	Represented by **Benjamin Arsenio Y. Navarro** *Director III, International Cooperation Unit*
Bangko Sentral ng Pilipinas	Represented by **Redentor Paolo M. Alegre, Jr.** *Senior Director, Department of Economic Statistics*
University of the Philippines School of Statistics	Represented by **Joseph Ryan G. Lansangan** *Dean*

C. Institutional Members - Private Sector

Ateneo de Zamboanga University	Represented by **Jocelyn Partosa** *Dean, College of Science and Information Technology*
Far Eastern University	Represented by **Frederick S. Gella** *Head, Mathematics Department*
Social Weather Stations	Represented by **Gerardo A. Sandoval** *Assistant Vice-President and Director for Sampling and Data Processing Group*

D. PSAI Regional Chapter

PSAI Region 8 Chapter	Represented by **Wilma A. Perante** *Regional Director Philippine Statistics Authority Region 8*

2022 Working Committees

Advocacy Committee

- Chair: **Rosalinda P. Bautista**, *Philippine Statistics Authority*
- Co-Chairs: **Dennis S. Mapa**, *Philippine Statistics Authority*; **Maqtahar L. Manulon**, *Philippine Statistics Authority*

Annual Conference Committee

- Chair: **Carmelita N. Ericta**, *Novo Trends PH*
- Co-Chair: **Minerva Eloisa P. Esquivias**, *Philippine Statistics Authority*

Sub-Committee:

Scientific Program Committee:

- Chair: **Joseph Ryan G. Lansangan**, *UPD School of Statistics*

Annual Meeting and Christmas Party Committee

- Chair: **Mi-Auree L. Bautista**, *UPLB Institute of Statistics*
- Co-Chair: **Gian Louisse A. Roy**, *UPD School of Statistics*

Finance Committee

- Chair: **Teresita B. Deveza**, *Bangko Sentral ng Pilipinas (Retired)*
- Co-Chair: **Jade Eric T. Redoblado**, *Bangko Sentral ng Pilipinas*

Institutional Development Committee

- Chair: **Dennis S. Mapa**, *Philippine Statistics Authority*
- Co-Chairs: **Romeo S. Recide**, *Philippine Statistics Authority (Retired);* **Wilhelmina C. Mañalac**, *Bangko Sentral ng Pilipinas (Retired)*

Institutional Training and Statistics Committee

- Chair: **Josefina V. Almeda**, *Philippine Statistical Research and Training Institute*
- Co-Chairs: **Joseph Ryan G. Lansangan**, *UPD School of Statistics*; **Maria Praxedes R. Peña**, *Philippine Statistical Research and Training Institute*

Membership Committee

- Chair: **Ferdinand S. Co**, *Bangko Sentral ng Pilipinas*
- Co-Chair: **Wilma A. Perante**, *Philippine Statistics Authority*

Nominations and Election Committee

- Chair: **Estela T. de Guzman**, *Philippine Statistics Authority (Retired)*
- Co-Chair: **Ruthie A. Floresta**, *Inner Sense Consulting and Research Corporation*

Publications Committee

- Chair: **Zita VJ Albacea**, *UPLB Institute of Statistics*
- Co-Chair: **Jose Ramon G. Albert**, *Philippine Institute for Development Studies*

The Philippine Statistician (TPS)

- Editor-in-Chief: **Zita VJ Albacea**, *UPLB Institute of Statistics*
- Managing Editor: **Nancy A. Tandang**, *UPLB Institute of Statistics*
- Associate Editors: **Rechel G. Arcilla**, *De La Salle University*; **Anna Maria Lourdes S. Latonio**, *Central Luzon State University*

PSAI Newsletter (PSAIEdition)

- Editor: **Genelyn Ma. F. Sarte**, **UPD School of Statistics**

Search and Awards Committee

- Chair: **Luisito Asuncion**, *Bangko Sentral ng Pilipinas*
- Co-Chair: **Maria Praxedes R. Peña**, *Philippine Statistical Research and Training Institute*

Social Media, Information and Communications Committee

- Chair: **Benjamin Arsenio Y. Navarro**, *Philippine Statistics Authority*
- Co-Chair: **Gerardo A. Sandoval**, *Social Weather Stations*

Ad Hoc Committee on Regional Affairs

- Chair: **Maqtahar L. Manulon**, *Philippine Statistics Authority (RSSO XII)*
- Co-Chair: **Jocelyn D. Partosa**, *Ateneo de Zamboanga University*

Web Development Team

- Lead Web Developer/Administrator: **Ferdinand S. Co**, *Bangko Sentral ng Pilipinas*

Guidelines for Authors for 2023 TPS

The Philippine Statistician (TPS) is the official scientific journal of the Philippine Statistical Association. Inc. (PSA). It considers papers resulting from original research in statistics and its applications. Papers will be sent for review on the assumption that this has not been published elsewhere nor is submitted in another journal.

Aims and Scope

The Journal aims to provide a media for the dissemination of research by statisticians and researchers using statistical method in resolving their research problems. While a broad spectrum of topics will be entertained, those with original contribution to the statistical science or those that illustrates novel applications of statistics in solving real-life problems will be prioritized. The scope includes, but is not limited to the following topics:

- Official Statistics
- Computational Statistics
- Simulation Studies
- Mathematical Statistics
- Survey Sampling
- Statistics Education
- Time Series Analysis
- Biostatistics
- Nonparametric Methods
- Experimental Designs and Analysis
- Econometric Theory and Applications
- Other Applications

In addition to research articles, the Journal will have the following sections that may appear in some of its issues (but not necessarily in all):

> ***Letters to the Editor.*** This section will provide a forum for the airing of opinions on issues pertinent to the statistical community or offers commentaries on articles that have appeared in the journal.
>
> ***Notes*** section will include notices and announcements of upcoming events, conferences, calls for papers.
>
> ***Review*** section will present reviews on statistics books and software.

Articles submitted for the three special sections above will be reviewed only by the Editor and/or Associate Editors.

Submission of Manuscript

Only unpublished manuscripts will be considered. They will be refereed and evaluated on content, language and presentation. The article in MS word format, without author's identification should be sent as email attachment to **Zita VJ Albacea**, Editor-in-chief, The Philippine Statistician: **tps@psai.ph** (cc: psai.tps2023@gmail.com).

A separate file containing the title of the paper, authors(s) (corresponding author identified) and their affiliations and complete address should be included in a separate file to be emailed along with the main article. Likewise, a report (in pdf format) from a software application checking on the similarity of the article with other publications should also be included in the attachment. To avoid delays and difficulties in submission, authors should follow instructions on style prescribed below.

A printed page in the Journal will have a maximum amount of space of 4.5" by 8.5".

Organization of Manuscript

The manuscript should be written in 8.5"x11" page using Times New Roman font size 12 with 1" margin in all sides.

The manuscript should be no more than 25 pages inclusive of the abstract, tables, references, figure captions, footnotes, endnotes. Submissions that exceed the prescribed page limit are unlikely to be accepted for publication and may be rejected immediately.

Further, the manuscript must be organized in the following manner:

- Title
- Abstract and Key Words
- Article Text
- Acknowledgments
- Literature Cited
- Appendices

Title: This should be brief and concise.

Abstract and Key Words: An abstract of at most 250 words must be submitted with the manuscript. It precedes the article text. The abstract should summarize objectives, results, and main conclusions, but it should not contain any graph or complex mathematical notation and no references. Three to six keywords should be identified. These keywords should not include words that are already in the title of the paper.

Article Text: Sections should be concise and numbered in a decimal system. Tables, Figures, and Artwork may be used within the body of the article, should be numbered consecutively. Tables, Figure Titles and Legends, Figure Artwork must be strategically

placed in the article text. The original files for the tables, figures, and artwork should also be submitted to facilitate typesetting. Authors must obtain written permission to reproduce or adapt all or part of a figure from a copyrighted source. Mathematical equations cited in the text should be numbered consecutively. Numbers should be placed at the rightmost margin of the equation line in parenthesis. Matrices should appear in bold and vectors in italics. All other symbols should appear in italics. The preferred software for equations are ***Mathtype*** and ***MS Equation Editor***, which are add-ins of MS Word.

Acknowledgments: An acknowledgment section may be included at the end of the article. This section should acknowledge financial assistance in the form of grants or university funding, assistance by individual colleagues, and any other pertinent information. This section will be inserted by the author only upon acceptance of the paper.

Literature Cited. All references included in the list at the end of an article must be cited in the text. References are cited in the text in the following format: (Author, Year). Up to two authors can be cited in the text. If there are three or more authors, only the first author will be cited in the text, e.g. (Author et al, Year). The following format will be followed in the listing references:

Journal Article

LANDAGAN, O. and BARRIOS, E., 2007, An Estimation Procedure for a Spatial-Temporal Model, *Statistics and Probability Letter,* 77(4):401-406.

Book

KOTTAK, C., 1991, Anthropology: The Exploration of Human Diversity, 5th ed., New York: McGraw-Hill, Inc.

Book Chapter

FINK, E. and PRATT, K., 2008, Indexing Compressed Time Series, in Last, M., Kandel, A. and Bunke, H., eds., Data Mining in Time Series Databases, Singapore: World Scientific, pp. 43-66.

Internet Document

MUNDLAK, Y., LARSON, D. and BUTZER, R., 2002, Determinants of Agricultural Growth in Indonesia, the Philippines, and Thailand, V. 1, *World Bank Working Paper 2803*, The World Bank. Available at: http://econ.worldbank.org/external/default/main?pagePK=64165259&piPK=64165421&theSitePK=469372&menuPK=64216926&entityID=000094946_02032604542948

Some Notes:

1. Sample papers may be downloaded from:
   https://www.psai.ph/tps.php?page=1&max=10
2. For particulars about the style, please download the Philippine Statistician Style Guide at https://www.psai.ph.

Appendices: A single appendix is headed, "APPENDIX: FOLLOWED BY A DESCRIPTIVE TITLE". If there are two or more appendices, they should be labeled, "APPENDIX A", "APPENDIX B", and so on.

Editorial Style: In addition to content, manuscripts are evaluated on their conciseness and clarity. Thus, the Journal gives premium to well-written and well-structured papers that will be of interest to a wide segment of the readership. Manuscripts and reviews that have been accepted for publication will be copy edited in accordance to accepted rules of correct grammar, usage, spelling, and punctuation. To avoid common problems of style, for guidelines on style, usage, and the preparation of technical manuscripts for publication, the following reference may be consulted:

The Chicago Manual of Style (14th ed.) (1993), Chicago: University of Chicago Press.

Editorial Notes

Use quotation marks only when a standard term is used in a nonstandard way and to indicate the beginning and ending of a direct quotation.

1. Hyphens are used when two or more adjectives or an adjective and a noun together modify another noun; for example, *goodness-of-fit test* is the equivalent of *test for goodness of fit*. Most words with prefixes such as sub and non are not hyphenated, for example, *subtable, nonnormal.*
2. Italics are used to introduce important terms, when appropriate; they are to be used sparingly to indicate emphasis.
3. Abbreviations and acronyms should be minimized; those that are used are spelled out on their first appearances in the manuscript with the shortened form given in parentheses, for example, *best linear unbiased estimate (BLUE)*.
4. Numbers under 10 are spelled out when they are not part of an equation or an expression containing symbols.
5. The sign % is always used when giving a specific percentage, for example, 23%, not 23 percent. Otherwise use the word *percent.*

Copyright Transfer Form

Authors of accepted papers will be required to submit an author copyright transfer form before the final release of the journal.

Philippine Statistical Association, Inc.
Room 214, Philippine Social Science Center, Commonwealth Avenue, Diliman, 1101, Quezon City, Philippines
Telephone: (632) 9-920-6513 | Telefax: (632) 3-456-1928 | E-mail: secretariat@psai.ph
www.psai.ph