JOURNAL OF STATISTICAL MODELING & ANALYTICS (JOSMA)
(ISSN: 2180-3102)
UNIVERSITI MALAYA

# Evaluating Panel Data Estimators Under Unbalanced Conditions Across Sample Sizes

O. P. Balogun[1*] and W. B. Yahya[2]

[1]*Department of Statistics, The Federal Polytechnic, Bida, Nigeria*
[2]*Department of Statistics, University of Ilorin, Nigeria*

*Corresponding e-mail:omoshadebalogun@yahoo..com*

Received 7 Aug 2024
Accepted 10 Dec 2025
Published 9 May 2026

RESEARCH ARTICLE

**Abstract**

This paper investigates the performance of panel data under the unbalanced panel dataset. The degree of missingness is varied in increments of 5% (5%, 10%, 15%, and 20%) and across different sample sizes (N = 25, 75, 100, 150, 200, 250, and 300). Monte-Carlo simulations of panel data for different N sample sizes and different degrees of missingness investigate the behaviors of the estimators through Mean Squared Error and Mean Absolute Error criteria. The rule of thumb is choosing the estimator with the lowest value of MSE and MAE as the best estimator that performs better than others at different levels of degree and different sample sizes. For the balanced panel data, the Between estimator, ranked 1st, the Within estimator, Random estimator, and Pooling estimators have no specific rank as they are not consistent in ranking, while the First Difference estimator, which has the highest MSE, MAE, and RMSE, ranked last for the panel data set. Similar results patterns were observed for the Unbalanced panel data set, which shares the same ranking as the balanced panel dataset. The result underscores that the estimator consistently outperforms its counterparts in managing unbalanced panels data under varying degrees of missingness. This estimator was therefore considered fit for the data used.

**Keywords:** Missingness, sample size, Mean Squared Error, Mean Absolute Error, Root Mean Squared Error, Unbalanced Panel Dataset

## 1. Introduction

Panel data is cross-sectional, time-invariant data. It is Longitudinal data that includes both individual cross-sectional and time series observations. An individual is chosen from a population and tracked for an extended period of time. Panel data observations are not expected to be equally distributed; hence, several panel data models and approaches have been developed to account for variations in data collected for individuals' observations across time (Wooldridge, 2012). This study considers five common panel data estimators.

Unbalanced panel data, a case where certain panels have incomplete observations in some time period (Bates et al., 2024), is a serious issue in the behavioral sciences, especially when data acquisition is expensive or entails destruction. The main technique commonly used by researchers is to reject situations with missing observations. Listwise and pairwise deletions are the most often used techniques (Peugh & Enders, 2004).

These techniques assume that data is missing completely at random (MCAR), which means that the likelihood of dependent missing values is uncorrelated with both the independent variables and the dependent variable itself (Marsh, 1998). Other forms of missingness are Missing at Random (MAR), where the missingness is due to known factors and results in a random impact that is easily estimated; and Not Missing at Random (NMAR), when the missing variable is driven by the level of the variable (Allison, 2002; Little et al., 2013; Little & Rubin, 2019).

Missing values in data are due to attrition (dropout), non-response, faulty machines, incomplete data entry, lost files, poorly designed research protocol, and so many other reasons (Popovich, 2024). The question is, what happens to data that has 50% values missing in research? Removing missing values results in bias and loss of power. How do we analyze such data without loss of power and bias? Little (2021) and Ren et al. (2023) explained in detail the three missing values or missing data types.

Among the five estimators considered by Balogun et al. (2022) on the performance of panel data estimators under the unbalanced panel data show that the Between estimator was found to be more efficient than the other estimators, vise-a-vice, the Within estimator, First Difference estimator, Random estimator, and Pooling estimator. They injected 5% missingness into a panel data sample size, $N = 25$, using mean square error (MSE) and mean absolute error (MAE) to investigate efficiency. This study is to further examine the performance of panel data estimators with small sample sizes ($N = 25, 50$), moderate sample sizes ($N = 75, 100$), and large sample sizes (150, 200,250, and 300) at different levels of missingness (5%, 10%, 15%, 20%, and 25%).

## 2. Materials and Methods

This work studies the performance of some common panel data estimators under the unbalanced panel data at different levels of sample sizes N across various periods T. In literature, the prevalence of missingness varies; 50% missingness is recommended as an acceptance level of missingness; however, data with 15% -20% missingness are common in most cases (Enders,2003). There is no rule for the rate of missingness to be analyzed.

Missing values result in bias and lower power, but it is not recommended to remove data in variables when values are missing. Removing missing values reduces the sample size, which leads to bias. Sampling bias resulting from missingness means the data may not represent a general population outside the study (Little, 2021). Unacceptance of missing values greater than 50% is considered a good rule of thumb (Salgado et al., 2016). Balogun et al. (2022) considered missingness at a 5%-20% level to investigate the performance of some estimators in panel data under an unbalanced condition for a small sample size. N=25. This work is to further examine the performances of these estimators considered by the trio at different levels of missingness, and as the data size increases.

### 2.1 Five-Panel Data Estimators Under Study

Balogun et al. (2022) used the typical panel data estimators reported in the literature. This study considers five theoretically reviewed estimators by Garba et al. (2013), which include:

i. Pooled Estimator: This Estimator stacks the data over $i$ and $t$ into one long regression with $nT$ (observations) and estimates of the parameters are obtained by OLS using the model (Greene, 2008; Garba et al., 2013).
ii. Between Estimator (BTW): This regresses the group means of Y on the group means of Xs in a regression of n observations. It uses cross-sectional variation by averaging the observations over period $t$ (Greene, 2008; Amemiya, 1971). Explicitly, it converts all the observations into individual-specific averages and performs OLS on the transformed data.

iii. Within Estimator: This regresses on the deviations from the individual or/and time mean. This is equivalent to "amemiya (Amemiya, 1971; Matyas & Sevestre, 1992).
iv. First Difference Estimator (FD): This is the ordinary least squares estimation of the difference between the original model and its one-period-lagged model (Arellano, 2003; Baltagi, 2005).
v. Random Estimator: This is equivalent to “swar” models (Baltagi, 2005; Arellano, 2003).

### 2.2 Simulation

This research follows the scheme adopted by Balogun et al. (2022) with some minor modifications. The simulation settings were as follows: For a total of $k = 5$ subjects (for instance) were studied and simulated over T = 5 times. Thus, a total of n = 25 ($k \times T$) size of observations was generated. Other sample sizes, n = 75, 100, 150, 200, 250, and 300, were also investigated.

The panel data model considered is of the form:

$$y_{it} = \beta_0 + \beta_1 X_{it} + e_{it}\,, \tag{1}$$

where:
$t = 1, \ldots, T\,;\, i = 1, \ldots, n_k$ and $k = 1, \ldots, 5$.

#### 2.2.1 Assumptions

The response variable, explanatory variable, and error simulation follow a normal distribution and assume normality.
The parameters were simulated, thus:

- The $X_{it}$ was simulated from the Gaussian population with the following: $X_{it} \sim iid\ N(20, 1)$.
- The error term $e_{it}$, was simulated from $e_{it} \sim iid\ N(0, 1)$.
- The $y_{it}$ is expected to return as normal.
- The parameters $\beta_0$ and $\beta_1$ in model (1) were set at: $\beta_0 = 20$ and $\beta_1 = 3$.
- The vectors $y_{it}$ and $X_{it}$ values are then used to obtain estimates of $\beta_x$ for each of the estimators under study.
- Unbalanced time intervals were infused into the data by randomly removing 5% of the total sample from the data. Population parameter values for the DGPs in the Monte Carlo experiments for different sample units $k$ = 5 gives various degrees of missingness with increments of 5% (i.e., % of missingness = 5%,10%,15%, and 20%).

### 2.3 Measurement Criteria

This study extends the investigation made by Balogun et al. (2022). It focuses on the two measurement criteria adopted by the trio, in addition to Root Mean Square Error, when the sample size increases. Estimates of these criteria were calculated from the resultant values generated from the simulation process, and the behavior of each estimate was observed for both Balanced and Unbalanced panel data to determine how they weigh on the error.

### 2.3.1 Mean Square Error (MSE)

The MSE equation is given as:

$$MSE = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it} - \hat{y}_{it})^2 \quad (2)$$

where *N* is the number of samples we are testing against. (Swarmy and Arora, 1972).

$$\text{If } MSE(\hat{\theta}_1) < MSE(\hat{\theta}_2) \quad (3)$$

then $\hat{\theta}_1$ *is* said to be more efficient than $\hat{\theta}_2$ (Stewart, 2024).

### 2.3.2 Mean Absolute Error (MAE)

The MAE, like the MSE, will never be negative since in this case, we are always taking the absolute value of the errors (Veroniki & Salanti, 2013). The MAE equation referred to Robeson and Willmott (2023) is given as:

$$MAE = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}|y_{it} - \hat{y}_{it}| \quad (4)$$

$$MAE(\hat{\theta}_1) < MAE(\hat{\theta}_2) \quad (5)$$

### 2.3.3 Root Mean Square Error (RMSE):

The formula for RMSE is as follows:

$$\left[\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it}-\hat{y}_{it})^2\right]^{1/2} \quad (6)$$

This follows the research of Arslanoglu (2016) and Robeson and Willmott (2023).

## 3. Results and Discussion

The Monte-Carlo study for balanced panel models under Unbalanced conditions across different sample sizes under consideration for this study is discussed in this section. The Mean Square Error and the Mean Absolute Error estimates were ranked from 1st, 2nd, 3rd, 4th, and 5th, with the 1st rank attributed to the most efficient estimator that has the lowest value of the mean square error and the absolute mean square error, while the 2nd rank was assigned to the second-best performing best and so on.

Tables 1, 2, and 3 represent the estimates of Mean Squared Error; each panel is the estimate of the five estimators across N samples for different degrees of missingness. That is, as an increment of 5% missingness was injected, the sample size reduces from left to right hand in each panel.

**Table 1.** Mean Squared Error for small sample sizes ($N = 25, 50$)

| **MSE** | **n** | **N** | **N** | **N** | **N** | **N** |
|---|---|---|---|---|---|---|
| **5%,10%, 15%,20%** | **5** | **25** | **21** | **18** | **14** | **12** |
| Pooling | | 1.1287 | 1.3040 | 1.3492 | 1.2077 | 1.4077 |
| Within | | 0.8354 | 0.9582 | 1.0302 | 1.0340 | 1.1993 |
| Random | | 0.9687 | 1.1493 | 1.2911 | 1.2077 | 1.4077 |
| First Difference | | 1.3986 | 1.7083 | 2.1913 | 2.8086 | 3.6398 |
| Between | | 0.2904 | 0.3134 | 0.3066 | 0.1549 | 0.1732 |
| $\boldsymbol{N = 50}$ | **5** | **50** | **40** | **32** | **24** | **20** |
| Pooling | | 0.8369 | 0.8022 | 0.8753 | 0.8534 | 0.5870 |
| Within | | 0.8088 | 0.7603 | 0.7892 | 0.6025 | 0.3909 |
| Random | | 0.8369 | 0.8022 | 0.8753 | 0.7230 | 0.5385 |
| First Difference | | 1.5655 | 1.3096 | 1.4266 | 1.1624 | 0.7865 |
| Between | | 0.0047 | 0.0252 | 0.0373 | 0.3636 | 0.3458 |

**Table 2.** Mean Squared Error for moderate sample sizes ($N = 75, 100$)

| **MSE** | **n** | **N** | **N** | **N** | **N** | **N** |
|---|---|---|---|---|---|---|
| **5%,10%, 15%,20%** | **5** | **75** | **61** | **49** | **37** | **28** |
| Pooling | | 0.8956 | 1.0239 | 1.0532 | 1.1682 | 1.4064 |
| Within | | 0.7922 | 0.9121 | 0.8726 | 0.8828 | 1.0802 |
| Random | | 0.8351 | 0.9673 | 0.9338 | 0.9888 | 1.2510 |
| First Difference | | 1.3905 | 1.5562 | 1.5458 | 1.9417 | 2.4479 |
| Between | | 0.0838 | 0.0972 | 0.1855 | 0.3400 | 0.4660 |
| $\boldsymbol{N = 100}$ | **5** | **100** | **81** | **67** | **52** | **40** |
| Pooling | | 0.8964 | 0.9390 | 0.8683 | 0.8483 | 0.8964 |
| Within | | 0.8737 | 0.9240 | 0.8204 | 0.7741 | 0.8737 |
| Random | | 0.8964 | 0.9390 | 0.8638 | 0.8258 | 0.8964 |
| First Difference | | 1.5406 | 1.8424 | 1.7944 | 1.8771 | 1.5406 |
| Between | | 0.0002 | 0.0106 | 0.0467 | 0.0786 | 0.0002 |

**Table 3.** Mean Squared Error for large sample sizes (150, 200,250, and 300)

| **MSE** | **n** | **N** | **N** | **N** | **N** | **N** |
|---|---|---|---|---|---|---|
| **5%,10%, 15%,20%** | **5** | **150** | **124** | **103** | **86** | **67** |
| Pooling | | 0.8578 | 0.8763 | 0.9402 | 0.9916 | 0.8542 |
| Within | | 0.8279 | 0.8390 | 0.9005 | 0.9554 | 0.8117 |
| Random | | 0.8578 | 0.8763 | 0.9402 | 0.9916 | 0.8542 |
| First Difference | | 1.5792 | 1.7220 | 2.0022 | 2.3198 | 2.0066 |
| Between | | 0.0137 | 0.0079 | 0.0130 | 0.0205 | 0.0223 |
| $\boldsymbol{N = 200}$ | **5** | **200** | **164** | **134** | **110** | **90** |
| Pooling | | 0.9890 | 0.9873 | 0.9205 | 0.9445 | 1.0651 |
| Within | | 0.9520 | 0.9477 | 0.8717 | 0.9138 | 1.0364 |
| Random | | 0.9686 | 0.9859 | 0.9205 | 0.9445 | 1.0651 |
| First Difference | | 1.9482 | 1.9112 | 1.7184 | 1.8827 | 2.0682 |
| Between | | 0.0327 | 0.0183 | 0.0135 | 0.0045 | 0.0062 |
| $\boldsymbol{N = 250}$ | **5** | **250** | **201** | **163** | **131** | **100** |
| Pooling | | 0.9572 | 0.9685 | 1.0263 | 1.0791 | 1.1265 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Within | | 0.9210 | 0.9274 | 0.9882 | 1.0363 | 1.0896 |
| Random | | 0.9327 | 0.9417 | 1.0074 | 1.0617 | 1.1264 |
| First Difference | | 1.9065 | 2.0545 | 2.2413 | 2.4238 | 2.6241 |
| Between | | 0.0351 | 0.0408 | 0.0364 | 0.0397 | 0.0332 |
| **$N = 300$** | **5** | **300** | **244** | **197** | **152** | **117** |
| Pooling | | 1.0513 | 1.0503 | 1.0713 | 0.9866 | 0.9548 |
| Within | | 1.0384 | 1.0206 | 1.0390 | 0.9748 | 0.9408 |
| Random | | 1.0491 | 1.0338 | 1.0612 | 0.9866 | 0.9548 |
| First Difference | | 2.0812 | 2.0759 | 2.1748 | 2.1953 | 1.9851 |
| Between | | 0.0128 | 0.0290 | 0.0238 | 0.0008 | 0.0026 |

Tables 4, 5, and 6 show the order of MSE ranking for the five estimators for N across T; the Between model ranked first, having the lowest values of MSE at different levels of missingness; the Pooling estimator ranked last, having the highest values of MSE at different levels of missingness. while other models have no consistent pattern of ranking.

**Table 4.** Mean Squared Error Result in Order of Ranking for small sample sizes ($N = 25, 50$)

| **MSE** | **n** | **N** | **N** | **N** | **N** | **N** |
|---|---|---|---|---|---|---|
| **5%,10%,15%,20%** | **5** | **25** | **21** | **18** | **14** | **12** |
| Pooling | | 4th | 4th | 4th | 3rd | 3rd |
| Within | | 2nd | 2nd | 2nd | 2nd | 2nd |
| Random | | 3rd | 3rd | 3rd | 3rd | 3rd |
| First Difference | | 5th | 5th | 5th | 4th | 4th |
| Between | | 1st | 1st | 1st | 1st | 1st |
| **$N = 50$** | **5** | **50** | **40** | **32** | **24** | **20** |
| Pooling | | 3rd | 3rd | 3rd | 4th | 4th |
| Within | | 2nd | 2nd | 2nd | 2nd | 2nd |
| Random | | 3rd | 3rd | 3rd | 3rd | 3rd |
| First Difference | | 4th | 4th | 4th | 5th | 5th |
| Between | | 1st | 1st | 1st | 1st | 1st |

**Table 5.** Mean Squared Error Result in Order of Ranking for moderate sample sizes ($N = 75, 100$)

| **MSE** | **n** | **N** | **N** | **N** | **N** | **N** |
|---|---|---|---|---|---|---|
| **5%,10%,15%,20%** | **5** | **75** | **61** | **49** | **37** | **28** |
| Pooling | | 4th | 4th | 4th | 4th | 4th |
| Within | | 2nd | 2nd | 2nd | 2nd | 2nd |
| Random | | 3rd | 3rd | 3rd | 3rd | 3rd |
| First Difference | | 5th | 5th | 5th | 5th | 5th |
| Between | | 1st | 1st | 1st | 1st | 1st |
| **$N = 100$** | **5** | **100** | **81** | **67** | **52** | **40** |
| Pooling | | 3rd | 3rd | 4th | 4th | 3rd |
| Within | | 2nd | 2nd | 2nd | 2nd | 2nd |
| Random | | 3rd | 3rd | 3rd | 3rd | 3rd |
| First Difference | | 4th | 4th | 5th | 5th | 4th |
| Between | | 1st | 1st | 1st | 1st | 1st |

**Table 6.** Mean Squared Error Result in Order of Ranking for large sample sizes ($N = 150, 200, 250, 300$)

| MSE | n | N | N | N | N | N |
|---|---|---|---|---|---|---|
| **5%,10%,15%,20%** | **5** | **150** | **124** | **103** | **86** | **67** |
| Pooling | | 3rd | 3rd | 3rd | 3rd | 3rd |
| Within | | 2nd | 2nd | 2nd | 2nd | 2nd |
| Random | | 3rd | 3rd | 3rd | 3rd | 3rd |
| First Difference | | 4th | 4th | 4th | 4th | 4th |
| Between | | 1st | 1st | 1st | 1st | 1st |
| $\boldsymbol{N = 200}$ | **5** | **200** | **164** | **134** | **110** | **90** |
| Pooling | | 4th | 4th | 3rd | 3rd | 3rd |
| Within | | 2nd | 2nd | 2nd | 2nd | 2nd |
| Random | | 3rd | 3rd | 3rd | 3rd | 3rd |
| First Difference | | 5th | 5th | 4th | 4th | 4th |
| Between | | 1st | 1st | 1st | 1st | 1st |
| $\boldsymbol{N = 250}$ | **5** | **250** | **201** | **163** | **131** | **100** |
| Pooling | | 4th | 4th | 4th | 4th | 4th |
| Within | | 2nd | 2nd | 2nd | 2nd | 2nd |
| Random | | 3rd | 3rd | 3rd | 3rd | 3rd |
| First Difference | | 5th | 5th | 5th | 5th | 5th |
| Between | | 1st | 1st | 1st | 1st | 1st |
| $\boldsymbol{N = 300}$ | **5** | **300** | **244** | **197** | **152** | **117** |
| Pooling | | 4th | 4th | 4th | 3rd | 3rd |
| Within | | 2nd | 2nd | 2nd | 2nd | 2nd |
| Random | | 3rd | 3rd | 3rd | 3rd | 3rd |
| First Difference | | 5th | 5th | 5th | 4th | 4th |
| Between | | 1st | 1st | 1st | 1st | 1st |

The outcomes of the Tables 4,5 and 6 show the estimators employed for a 5% missingness attributable to unbalanced data, which follow the same trend as when it was for the balanced panel data.

The results for the Balance and Unbalanced data models show that the Between estimator performs best with the measurement error (MSE) in Table 1 because it assumes an average of $y_i$ which regressed on the average of $X_i$ in an n observation. ignoring the within-time-variant observation, which solved the inaccurate random deviation from the long-run, that is, the inaccuracy has a mean of zero over time. It also measures the variation between the individual MSEs is always positive (and not zero) because of randomness.

The MSE is the second moment (about the origin) of the error and thus incorporates both the variance of the estimator (how widely spread the estimates are) and its bias (how far off the average estimated value is from the true value). For an unbiased estimator, the MSE is the variance of the estimator. Like the variance, MSE has the same units of measurement as the square of the quantity being estimated in an analogy to standard deviation. If MSE is greater than zero, which means that if $\lambda \neq 0$ therefore, the Random effect is Fixed Effect; however, this is less efficient than Within and Between (Robeson and Willmott, 2023). This assumption makes the Random estimator less efficient in this study.

The pooling (OLS) Estimator ignores time and individual characteristics and focuses only on dependencies between individuals. It is characterized by no correlation between the unobserved, independent variable(s) and the independent variables (i.e., exogeneity) for the same individual. This assumption on the error terms is very strong or unrealistic. Thus, this accounts for its high estimate

compared to Between estimators as seen in Tables 1, 2, and 3; and Tables 7, 8, 9 for both MSE and MAE, respectively.

The first difference estimator performs poorly among the five Estimators considered, as it has the highest estimate. Correlation between $X_{it}$ and $w_{j.t-2}, it$ assumption on exogeneity makes it less demanding for FD than the Within estimator and other Estimators. It is also less efficient than other Estimators because $W_{it}$ is serially correlated, and even if $W_{it}{}'$s is uncorrelated. Therefore, it does not violate this assumption in this project.

Similar to the MSE outcomes, Tables 7, 8, and 9 show the estimates of the Mean Absolute Error (MAE) for the estimators considered for balanced panel models and their corresponding estimates for unbalanced panel data models. The results were comparable to the MSE observed. Between Estimator rank 1st with the lowest MAE value, Pooling estimator, Within estimator, and Random estimator converge with almost the same values as missingness increases, and they show no consistent ranking pattern at that point of convergence, while First Difference has the highest value and is considered to rank last.

**Table 7.** Mean Absolute Error (MAE) for small sample sizes ($N = 25, 50$)

| **MAE** | **n** | **N** | **N** | **N** | **N** | **N** |
|---|---|---|---|---|---|---|
| **5%, 10%, 15%,20%** | **5** | **25** | **21** | **18** | **14** | **12** |
| Pooling | | 0.8023 | 0.8760 | 1.0171 | 0.8915 | 0.4765 |
| Within | | 0.7148 | 0.8084 | 0.7995 | 0.5667 | 0.3409 |
| Random | | 0.7541 | 0.8311 | 0.9395 | 0.8836 | 0.4765 |
| First Difference | | 0.9706 | 1.1069 | 1.1301 | 1.2277 | 0.8071 |
| Between | | 0.4724 | 0.4626 | 0.5223 | 0.4430 | 0.2754 |
| $\boldsymbol{N = 50}$ | **5** | **50** | **40** | **32** | **24** | **20** |
| Pooling | | 0.7880 | 0.8066 | 0.8015 | 0.7825 | 0.8419 |
| Within | | 0.7471 | 0.7704 | 0.7229 | 0.7362 | 0.6739 |
| Random | | 0.7664 | 0.7855 | 0.7871 | 0.7588 | 0.7680 |
| First Difference | | 1.1006 | 1.1248 | 1.0518 | 1.1986 | 1.3097 |
| Between | | 0.3355 | 0.3428 | 0.2911 | 0.3776 | 0.6184 |

**Table 8.** Mean Absolute Error (MAE) for moderate sample sizes ($N = 75, 100$)

| **MAE** | **n** | **N** | **N** | **N** | **N** | **N** |
|---|---|---|---|---|---|---|
| **5%, 10%, 15%,20%** | **5** | **75** | **61** | **49** | **37** | **28** |
| Pooling | | 0.8017 | 0.8389 | 0.8124 | 0.7764 | 0.9626 |
| Within | | 0.7608 | 0.8056 | 0.7665 | 0.7178 | 0.9328 |
| Random | | 0.7817 | 0.8389 | 0.7846 | 0.7764 | 0.9626 |
| First Difference | | 1.2507 | 1.3192 | 1.2659 | 1.2123 | 1.6186 |
| Between | | 0.1911 | 0.1567 | 0.2909 | 0.2298 | 0.1145 |
| $\boldsymbol{N = 100}$ | **5** | **100** | **81** | **67** | **52** | **40** |
| Pooling | | 0.8118 | 0.7866 | 0.8100 | 0.7746 | 0.7374 |
| Within | | 0.8041 | 0.7647 | 0.7969 | 0.7567 | 0.7234 |
| Random | | 0.8118 | 0.7866 | 0.8100 | 0.7746 | 0.7374 |
| First Difference | | 1.1497 | 1.0973 | 1.1169 | 1.1814 | 1.2052 |
| Between | | 0.1239 | 0.1502 | 0.1570 | 0.2021 | 0.1546 |

**Table 9.** Mean Absolute Error (MAE) for large sample sizes ($N = 150, 200, 250, 300$)

| **MAE** | **n** | **N** | **N** | **N** | **N** | **N** |
|---|---|---|---|---|---|---|
| **5%, 10%, 15%,20%** | **5** | **150** | **124** | **103** | **86** | **67** |
| Pooling | | 0.8132 | 0.8600 | 0.7908 | 0.7598 | 0.7756 |
| Within | | 0.8036 | 0.8394 | 0.7761 | 0.7401 | 0.7593 |
| Random | | 0.8132 | 0.8579 | 0.7725 | 0.7428 | 0.7756 |
| First Difference | | 1.1604 | 1.2211 | 1.0376 | 1.0575 | 1.3080 |
| Between | | 0.0893 | 0.1721 | 0.2174 | 0.2347 | 0.1903 |
| $\boldsymbol{N = 200}$ | **5** | **200** | **164** | **134** | **110** | **90** |
| Pooling | | 0.7800 | 0.7745 | 0.7378 | 0.7608 | 0.7717 |
| Within | | 0.7754 | 0.7564 | 0.7390 | 0.7594 | 0.7424 |
| Random | | 0.7796 | 0.7653 | 0.7373 | 0.7570 | 0.7577 |
| First Difference | | 1.0865 | 1.0594 | 1.1005 | 1.1242 | 1.2029 |
| Between | | 0.1141 | 0.1742 | 0.1348 | 0.1953 | 0.2457 |
| $\boldsymbol{N = 250}$ | **5** | **250** | **201** | **163** | **131** | **100** |
| Pooling | | 0.8212 | 0.7946 | 0.7734 | 0.8389 | 0.8313 |
| Within | | 0.8147 | 0.7943 | 0.7636 | 0.8429 | 0.8110 |
| Random | | 0.8199 | 0.7946 | 0.7682 | 0.8389 | 0.8313 |
| First Difference | | 1.1850 | 1.1138 | 1.1233 | 1.2766 | 1.2756 |
| Between | | 0.1236 | 0.0720 | 0.1717 | 0.0731 | 0.0898 |
| $\boldsymbol{N = 300}$ | **5** | **300** | **244** | **197** | **152** | **117** |
| Pooling | | 0.7875 | 0.7810 | 0.8098 | 0.8498 | 0.7826 |
| Within | | 0.7796 | 0.7640 | 0.8036 | 0.8260 | 0.7501 |
| Random | | 0.7803 | 0.7691 | 0.8098 | 0.8340 | 0.7645 |
| First Difference | | 1.0817 | 1.0717 | 1.1932 | 1.1875 | 1.0150 |
| Between | | 0.1487 | 0.1617 | 0.1102 | 0.2199 | 0.1895 |

**Table 10.** Mean Absolute Error Result in Order of Ranking for small sample sizes ($N = 25, 50$)

| **MAE** | **n** | **N** | **N** | **N** | **N** | **N** |
|---|---|---|---|---|---|---|
| **5%,10%,15%,20%** | **5** | **25** | **21** | **18** | **14** | **12** |
| Pooling | | 4th | 4th | 4th | 4th | 3rd |
| Within | | 2nd | 2nd | 2nd | 2nd | 2nd |
| Random | | 3rd | 3rd | 3rd | 3rd | 3rd |
| First Difference | | 5th | 5th | 5th | 5th | 5th |
| Between | | 1st | 1st | 1st | 1st | 1st |
| **N=50** | **5** | **50** | **40** | **32** | **24** | **20** |
| Pooling | | 4th | 4th | 4th | 4th | 4th |
| Within | | 2nd | 2nd | 2nd | 2nd | 2nd |
| Random | | 3rd | 3rd | 3rd | 3rd | 3rd |
| First Difference | | 4th | 5th | 5th | 5th | 5th |
| Between | | 1st | 1st | 1st | 1st | 1st |

**Table 11.** Mean Absolute Error Result in Order of Ranking for moderate sample sizes ($N = 75, 100$)

| MAE | n | N | N | N | N | N |
|---|---|---|---|---|---|---|
| **5%,10%,15%,20%** | **5** | **75** | **61** | **49** | **37** | **28** |
| Pooling | | 4th | 3rd | 4th | 3rd | 3rd |
| Within | | 2nd | 2nd | 2nd | 2nd | 2nd |
| Random | | 3rd | 3rd | 3rd | 3rd | 3rd |
| First Difference | | 5th | 5th | 5th | 5th | 5th |
| Between | | 1st | 1st | 1st | 1st | 1st |
| **N=100** | **5** | **100** | **81** | **67** | **52** | **40** |
| Pooling | | 3rd | 3rd | 3rd | 3rd | 3rd |
| Within | | 2nd | 2nd | 2nd | 2nd | 2nd |
| Random | | 3rd | 3rd | 3rd | 3rd | 3rd |
| First Difference | | 5th | 5th | 5th | 5th | 5th |
| Between | | 1st | 1st | 1st | 1st | 1st |

**Table 12.** Mean Absolute Error Result in Order of Ranking for large sample sizes ($N = 150, 200, 250, 300$)

| MAE | n | N | N | N | N | N |
|---|---|---|---|---|---|---|
| **5%,10%,15%,20%** | **5** | **150** | **124** | **103** | **86** | **67** |
| Pooling | | 3rd | 4th | 4th | 4th | 3rd |
| Within | | 2nd | 2nd | 3rd | 2nd | 2nd |
| Random | | 3rd | 3rd | 2nd | 3rd | 3rd |
| First Difference | | 5th | 4th | 5th | 5th | 5th |
| Between | | 1st | 1st | 1st | 1st | 1st |
| **N=200** | **5** | **200** | **164** | **134** | **110** | **90** |
| Pooling | | 4th | 4th | 3rd | 4th | 3rd |
| Within | | 2nd | 2nd | 4th | 3rd | 2nd |
| Random | | 3rd | 3rd | 2nd | 2nd | 3rd |
| First Difference | | 5th | 5th | 5th | 5th | 5th |
| Between | | 1st | 1st | 1st | 1st | 1st |
| **N=250** | **5** | **250** | **201** | **163** | **131** | **100** |
| Pooling | | 4th | 3rd | 4th | 2nd | 3rd |
| Within | | 2nd | 2nd | 2nd | 4th | 2nd |
| Random | | 3rd | 3rd | 3rd | 2nd | 3rd |
| First Difference | | 5th | 5th | 5th | 5th | 5th |
| Between | | 1st | 1st | 1st | 1st | 1st |
| **N=300** | **5** | **300** | **244** | **197** | **152** | **117** |
| Pooling | | 4th | 4th | 3rd | 4th | 4th |
| Within | | 2nd | 2nd | 2nd | 2nd | 2nd |
| Random | | 3rd | 3rd | 3rd | 3rd | 3rd |
| First Difference | | 5th | 5th | 5th | 5th | 4th |
| Between | | 1st | 1st | 1st | 1st | 1st |

Tables 10,11, and 12 show a MAE ranking pattern similar to the one observed in Tables 4, 5, and 6. The Between model ranks first, having the lowest estimate of MAE across N and a percentage

missingness of 5%. The Pooling model ranks last, having the highest values of MAE, while other models have no consistent pattern of ranking.

### 3.1 Graphical Representation of the Assessments

This section shows plots to describe patterns in the assessments. Figure 1 represents the plot of all the estimators considered when $N = 25$ across the degree of missingness. As shown on the graph, the Between estimator performance exceeds other estimators for Mean Squared Error and Mean Absolute Error criteria used in this study. These behaviors were repeated in Figures 2, 3, and 4. The plots demonstrate the accuracy of the Between estimator, which is narrowly dispersed around the central mean. It is therefore considered fit for the data samples

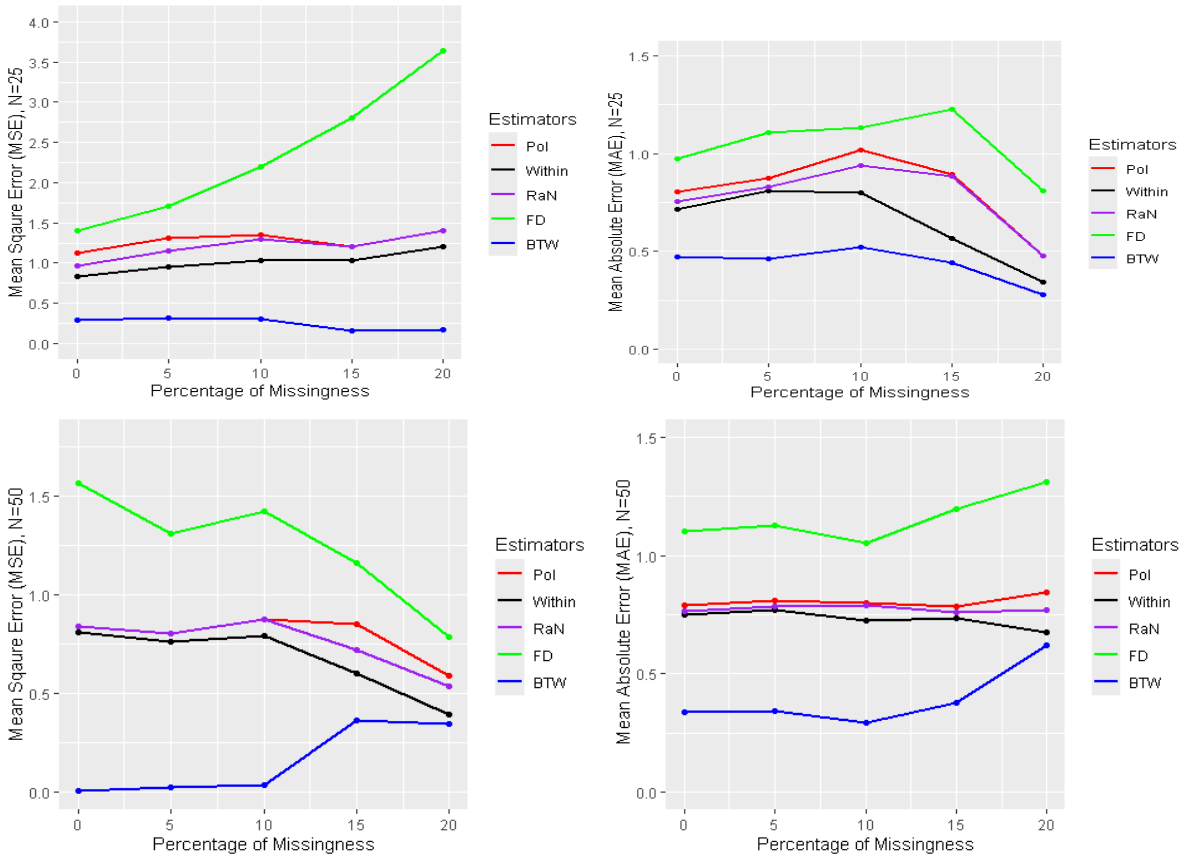


**Figure 1.** Mean Square Error (MSE) and Absolute Mean Square Error (MAE) for small Data sizes, $N\ (25, 50)$, at 0%- 20% Degree of Missingness

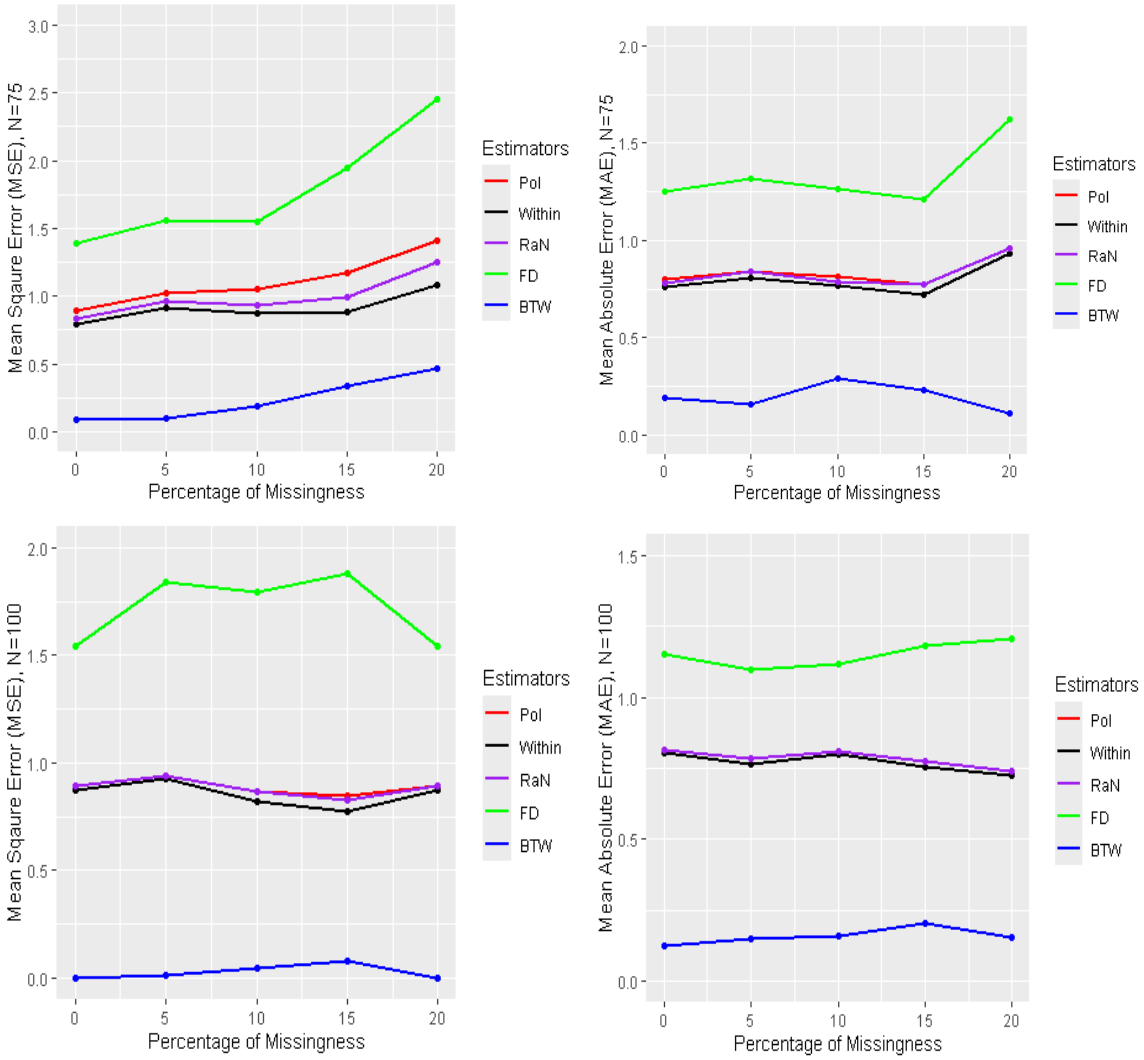


**Figure 2.** Mean Square Error (MSE) and Absolute Mean Square Error (MAE) for Moderate Data sizes, $N\ (75, 100)$, at 0%- 20% Degree of Missingness

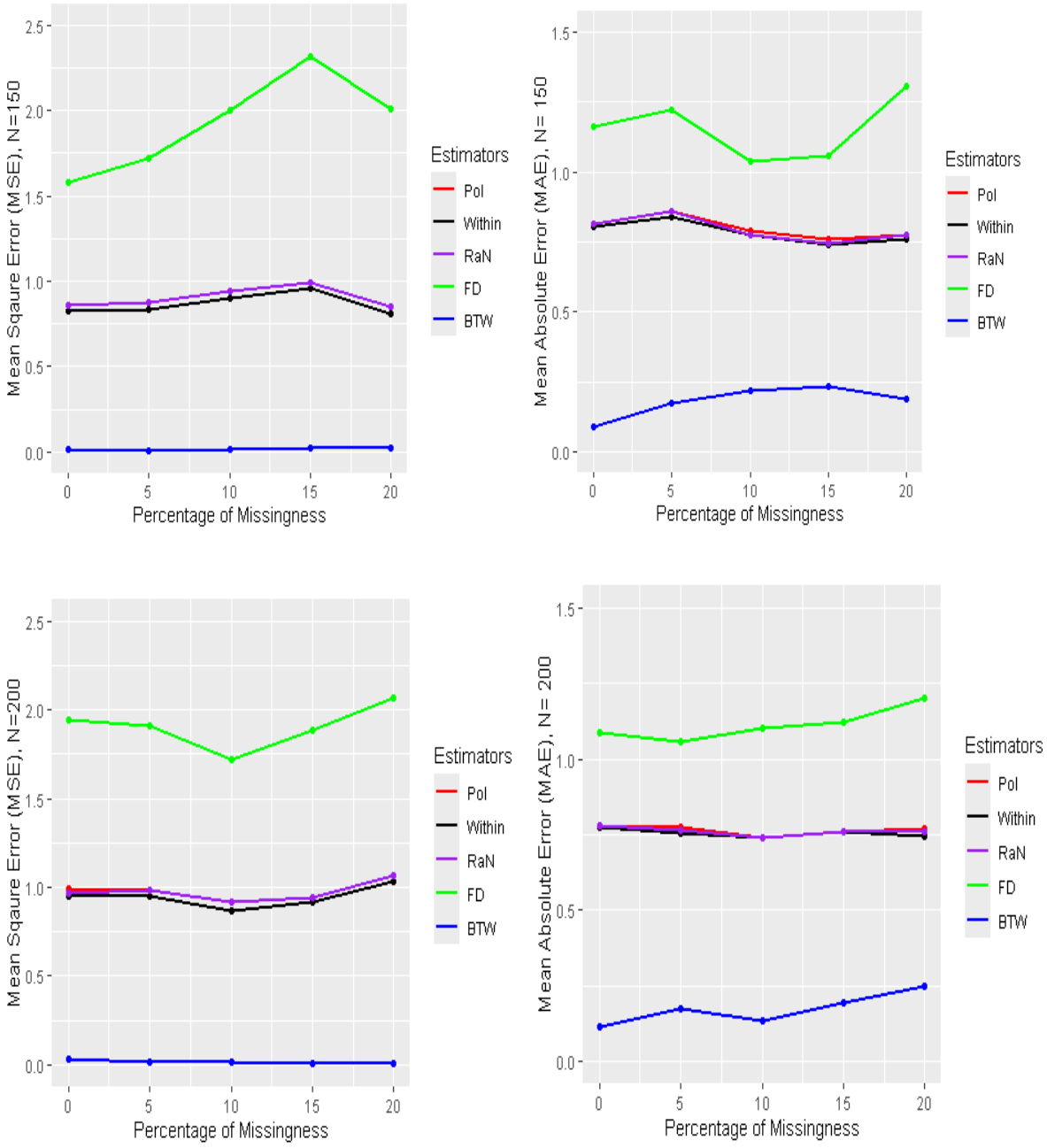


**Figure 3.** Mean Square Error (MSE) and Absolute Mean Square Error (MAE) for big data size, $N\ (150, 200)$, at 0%- 20% Degree of Missingness

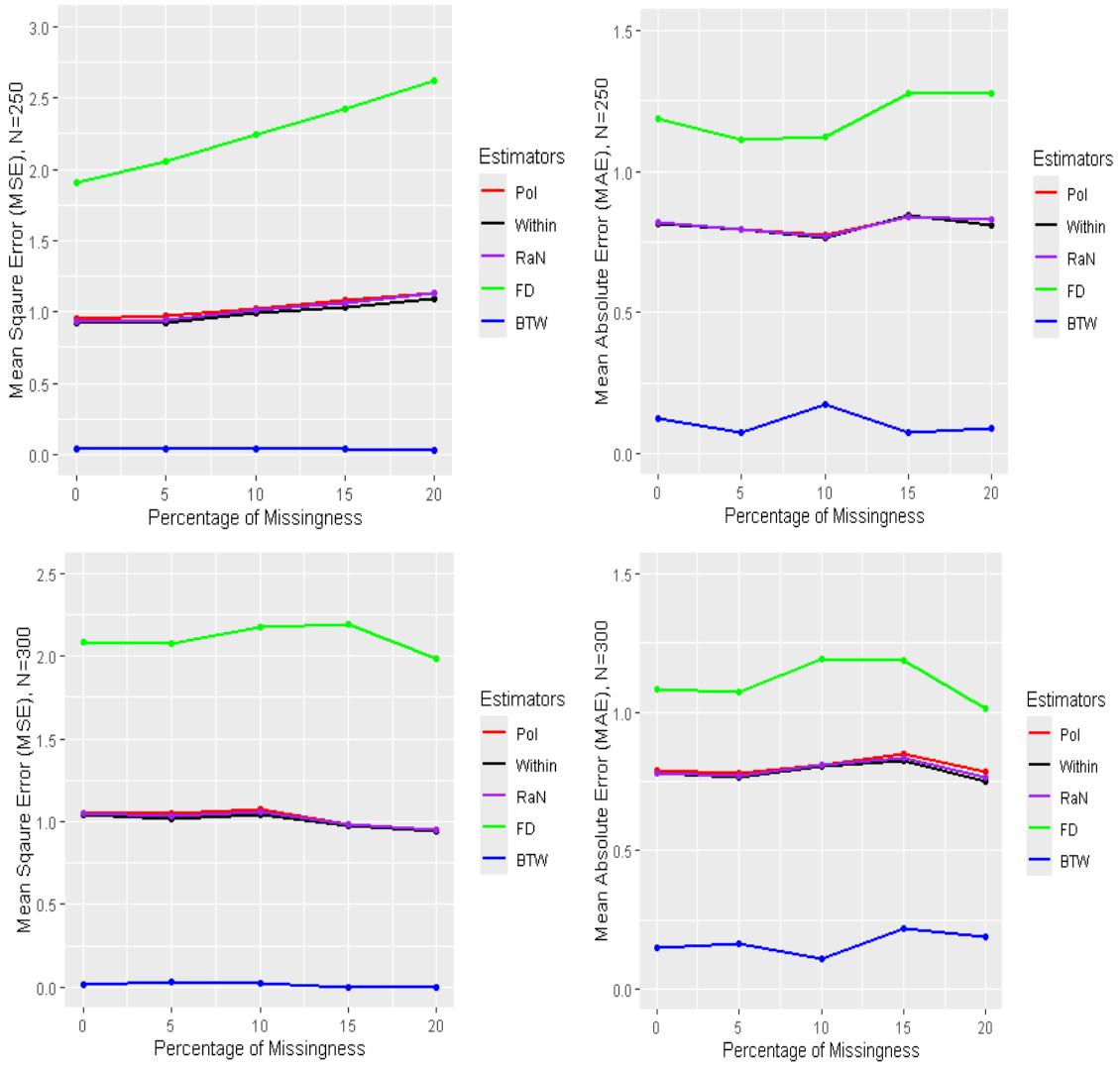


**Figure 4.** Mean Square Error (MSE) and Absolute Mean Square Error (MAE) for big data size, $N\ (250, 300)$, at 0%- 20% Degree of Missingness

### 3.2 Box Plot

A box plot for the Root Mean Square Error (RMSE) for different N sizes chosen from the categories was plotted and shown in Figure 5. To demonstrate the consistency of the performances of the Between estimator, various sample sizes, $N = 25, 75, \text{and } 300$, were painstakingly considered for small, medium, and large sample sizes, respectively.

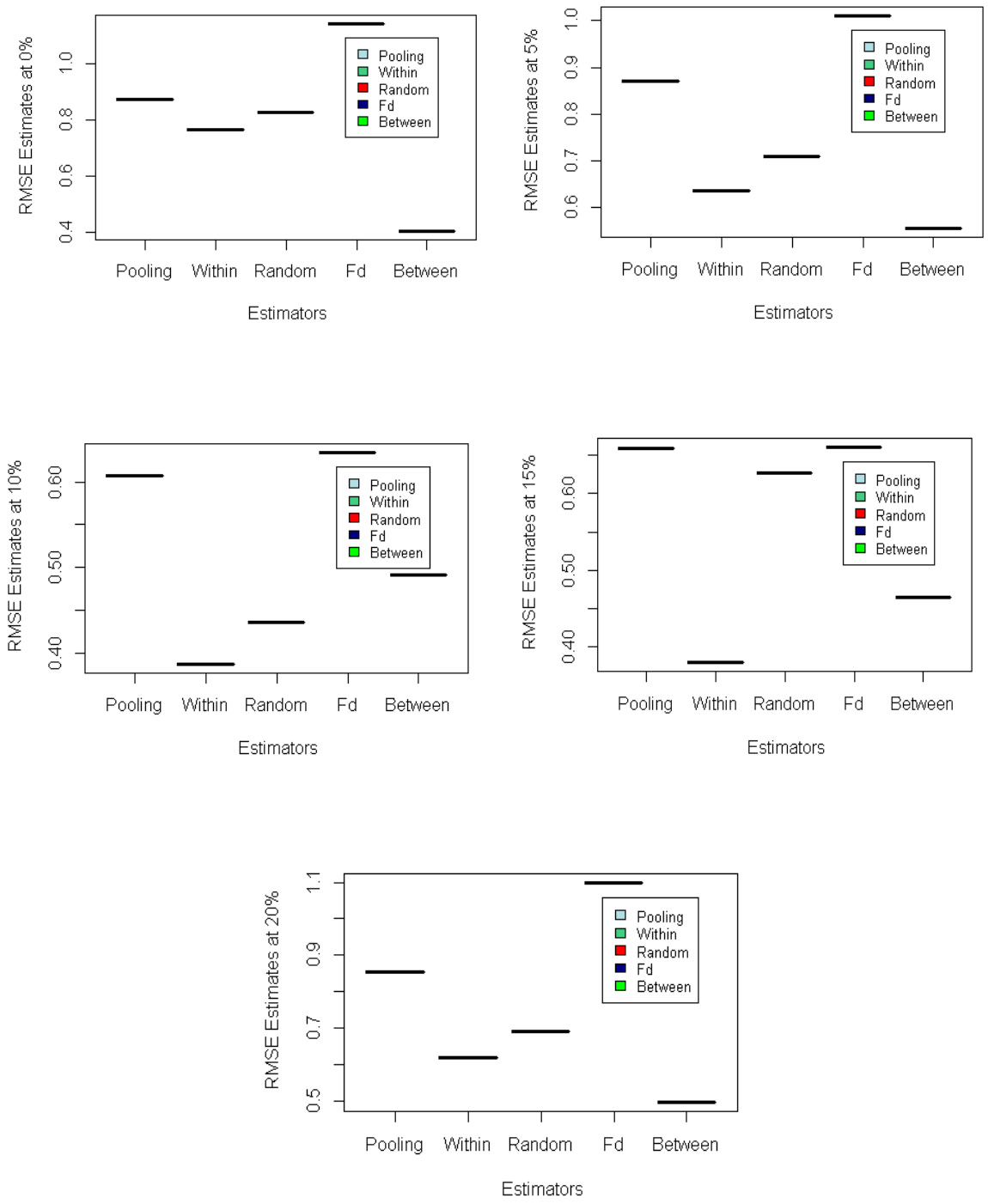


**Figure 5.** Boxplots for $N = 25$ at different degrees of missingness (that is, 0%-20%)

The results for several sample categories show that, across the five common panel estimators used in this study, the between estimator yields the lowest RMSE, indicating accurate prediction of the observed outcomes and the highest model performance. It is evidence that the Between estimator outperforms other estimators considered.

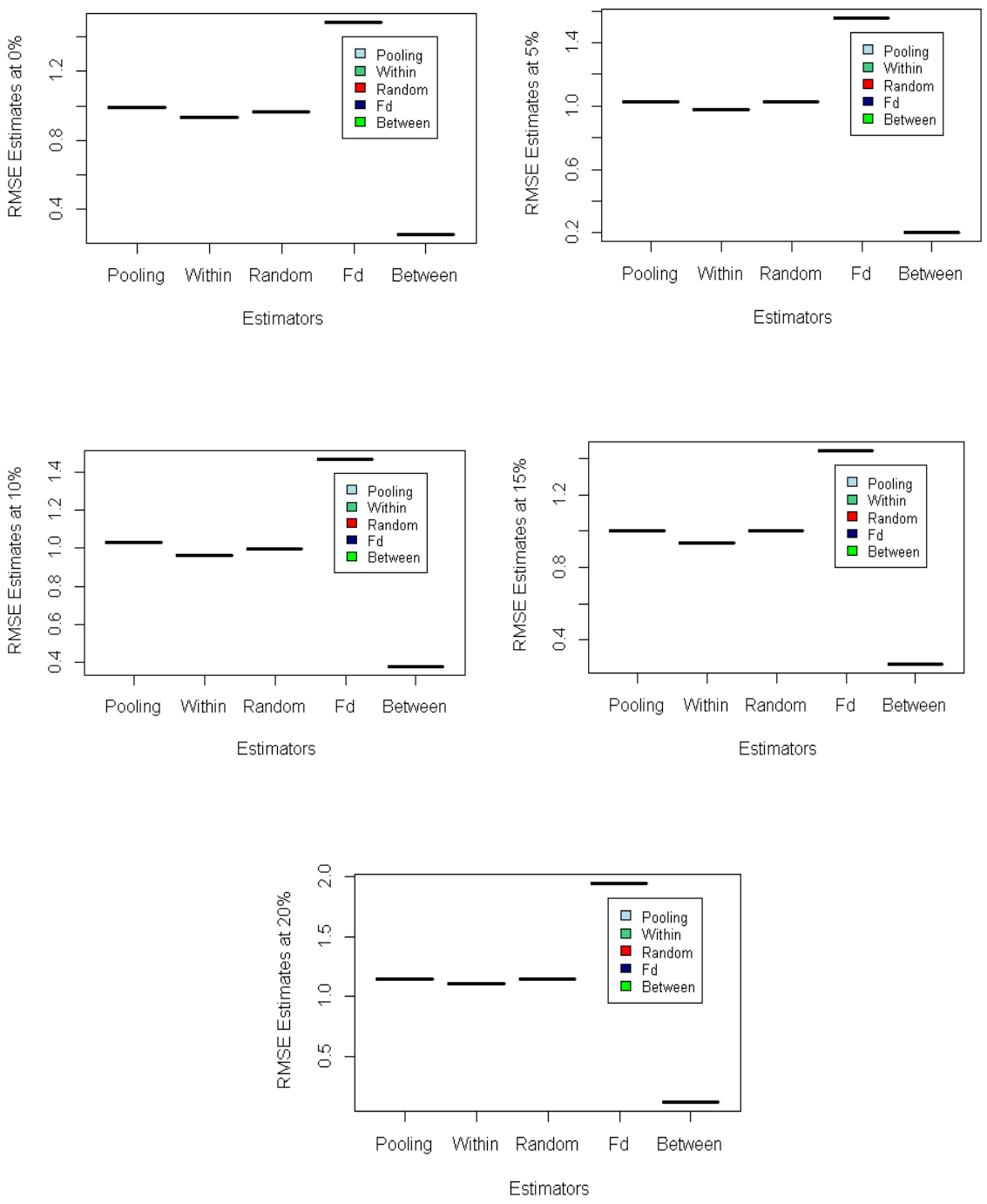


**Figure 6.** Box plot ($N = 75$, Missingness (0%-20%))

Figure 6 represents boxplots for RMSE when $N = 75$ at different degrees of missingness (that is, 0%-20%). Similar to Figure 5, the Between estimator performs better than the other estimators considered. Figure 7 shows boxplots for $N = 300$ at different degrees of missingness (that is, 0%-20).

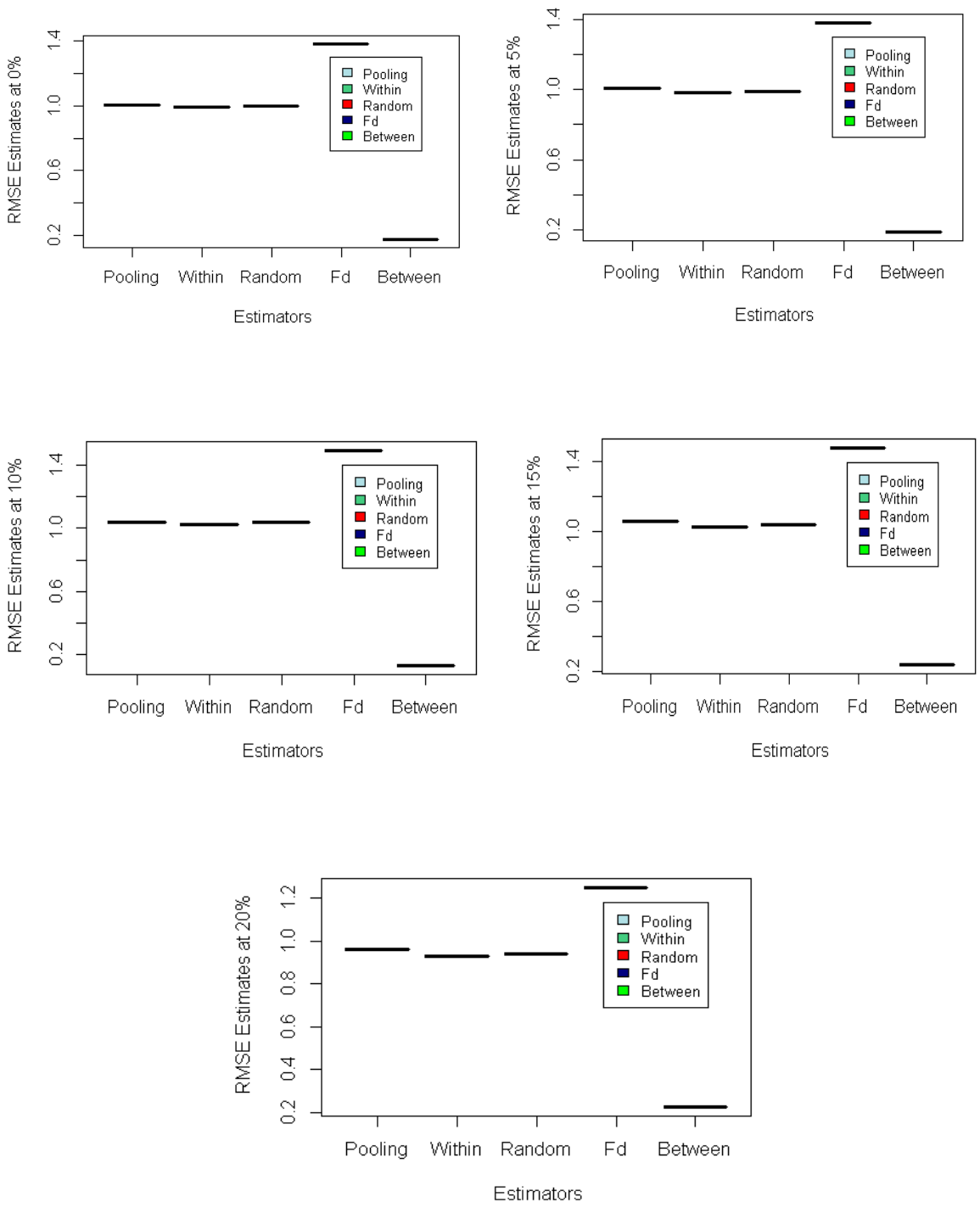


**Figure 7.** Box plot ($N = 300$, Missingness (0%-20%))

### 3.3 Data Distribution

This study focuses on the Monte Carlo study of a panel data set from (1). It follows that the resultant data values $X_{it}$ clustered around the mean. That is, Fig. 8 shows the mean of 19.89 and the standard deviation of 0.9067; thus, 68% of the values generated were 19.89 plus or minus 0.91(1 standard deviation away from the mean), and 95% of these values were 19.89 plus or minus 1.82 (2 standard

deviations away from the mean). Consequently, 99% of the values could also be represented as 19.89 plus or minus 2.73 (3 standard deviations away from the mean).

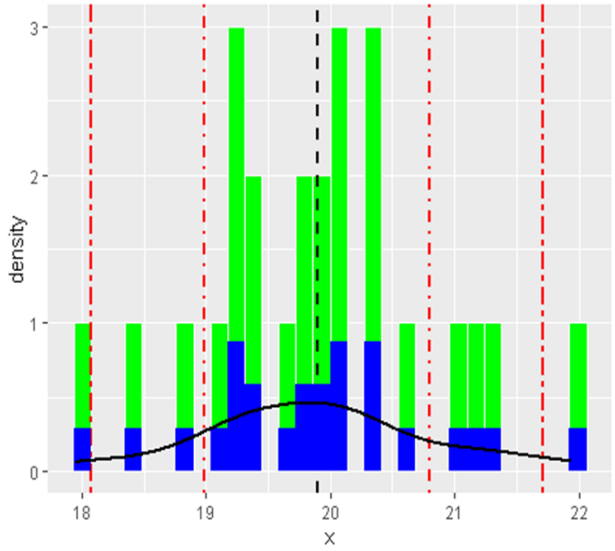


**Figure 8.** $x$ Variable: Density Function and Histogram

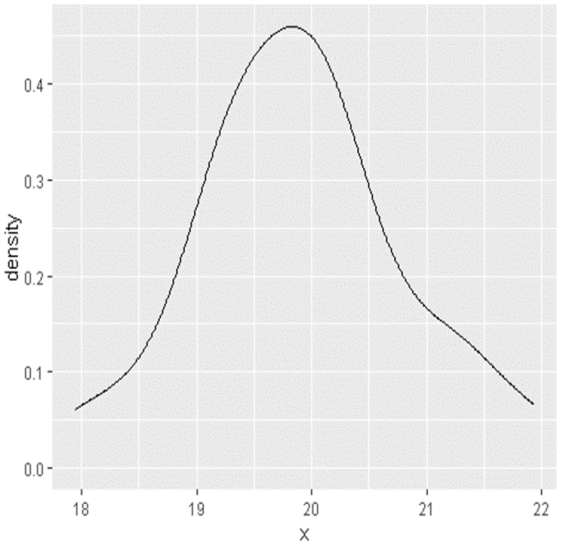


**Figure 9.** $x$ Variable: Density

The black vertical dotted line represents the mean value of 19.89, while the red dotted lines to the right are the 68% and 95% confidence interval (CI) values added to the mean, respectively. The red vertical values to the left show 68% and the 95% removed from the mean. These are one and two standard deviations away from the mean. Figure 9 shows the Density function plot of the x variable; this represents the spread about the mean. This is the skewness of the distribution.

### 3.3.1 Shapiro-Wilk normality Test

Since the Shapiro $p - value$ is greater than 0.05, that is, $W = 0.9871, p - value = 0.982$, the data is normally distributed. Similarly, $y_{it}$ has a mean of 79.81 and a standard deviation of 3.0376., this shows that 68% of the values generated were 79.81 plus or minus 3.04 (1 standard deviation away from the mean), 95% of these values were 79.81 plus or minus 6.08 (2 standard deviation away from the mean), and 99% of the values were 79.81 plus or minus 9.12 (3 standard deviation away from the mean). Therefore, data with a wider spread about the mean might not exhibit the same pattern as concluded in this study.

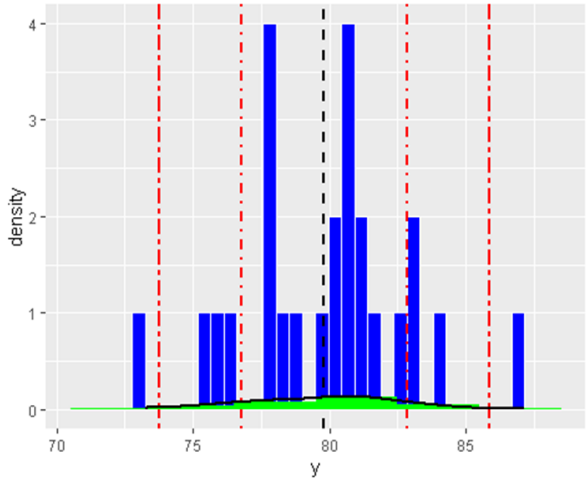


**Figure 10.** $y$-Variable: Density Function and Histogram

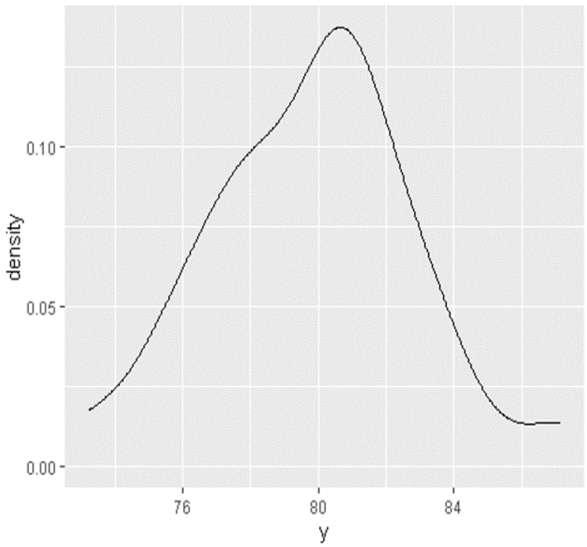


**Figure 11.** $y$-Variable: Density Function

As shown in Figure 10, the black vertical dotted line represents the mean value of 79.8, while the red dotted lines to the right are the 68% and 95% confidence interval (CI) values added to the mean, respectively. The red vertical values to the left show 68% and the 95% removed from the mean. These are one and two standard deviations away from the mean.

### 3.3.2 Shapiro-Wilk normality Test

The Shapiro value W is 0.9823 and the $p-value$ is 0.9266. Since the Shapiro $p-value$ is greater than 0.05, it implies that the data is normally distributed. However, the Between estimator might not be the best estimator to be considered if the data is not normally distributed.
Figure 11 shows the Density function plot for the Y variable; this represents the skewness of the distribution.

### 3.3.3 Interpretation of Result

Panel data refers to data sets consisting of multiple (repeated) observations on each sampling unit. To create a panel data set, we collect observations on the same set of entities over several periods of time. We could generate this by pooling time series observations across a variety of cross-sectional units, such as countries, states, regions, firms, or randomly sampled individuals or households. Panel data covers a much larger sample and is representative of all demographic groups.

Epidemiological data where there is evidence of epidemic, pandemic, or endemic cases. Cases where data values are close, and there is a dispersion between the response and the explanatory variables. Additionally, we can apply it to econometric data, where the data values exhibit minimal or no changes over time for various variants, and the observation follows a normal distribution. It can also be used in econometric data, where variant values generally have little change over time. Observing a country's standard of living using different metrics or indexes over a period of time and comparing it to other countries in a region might not be seen as time series or cross-sectional data, but as longitudinal data.

Furthermore, environmental degradation, which measures the number of trees degraded over a period, occurs after panel observations and data collection. However, the data follows a deterministic pattern, with the explanatory variable uniquely determining the response variable. The simulated data

exhibits relatively close values for both the response and explanatory variables. Specifically, the response variable falls between 73.23 and 87.16, while the explanatory variable falls between 17.95 and 20.35. We observe that the response variable is approximately four times larger than the explanatory variable. The chosen mean and standard deviation could explain the dispersion in the data, which renders discovering such data rare.

## 4. Conclusion

The estimators employed for a 5% increment, that is, 5%, 10%,15% and 20% of missingness attributed to unbalanced data and, the outcomes follow the same trend as when it was for the balanced panel data. Based on the results obtained from a Monte-Carlo simulation, it could be concluded that among the five estimators examined, the between estimator consistently outperforms its counterparts in managing unbalanced panel data under varying degrees of missingness. The findings offer valuable guidance for the application of these estimators in empirical research, reinforcing the significance of model selection and its impact on the validity of conclusions drawn from panel data analyses. Future research should continue to explore the robustness of these findings across different contexts and datasets to further establish best practices for panel data analysis.

## 5. Recommendations

For further investigation, two aspects are recommended to be considered, including:

- Monte Carlo study with different sample sizes to estimate mean values and standard deviations, as well as possible data transformation, is recommended for further studies.
- As a signal in the study, it is recommended that the Between estimator should be adopted for fitting the panel data models when evidence of missingness at different values of sample sizes (from small, moderate to large sizes) are observable in the panel data.

## 6. References

Allison, P. D. (2002). *Missing data*. Thousand Oaks: Sage Publications, Inc.

Amemiya, T. 1971: The estimation of the variances in a variance-components model, *International Economic Review*. 12: 1-13.

Arslanoglu, N. (2016). Empirical modeling of solar radiation exergy for Turkey Faculty of Engineering, Mechanical Engineering Department, Uludag University, Gorukle Campus, TR-16059 Bursa, Turkey

Arellano, M. (2003). Panel Data Econometrics.

Balogun, O. P., Yahya, W. B., & Umar-Mann A. (2022). Performance Evaluation of Some Estimators under Unbalanced Panel Data Models. *Professional Statisticians Society of Nigeria, Edited Proceedings of 6th International Conference. 6(1).*

Baltagi, B. H. (2005). *Econometric Analysis of Panel Data*, John Wiley and Sons, England.

Enders C. K. (2003). Using the Expectation Maximization Algorithm to Estimate Coefficient Alpha for Scales with Item-Level Missing Data. *Psychol Meth,* 8(3): 322–337. doi: 10.1037/1082-989X.8.3.322

Garba M. K., Oyejola, B.A. and Yahya, W. B. 2013: Investigations of Certain Estimators for Modeling Panel Data Under Violations of Some Basic Assumptions. 3(10).

Greene, W. H. (2008). Econometric Analysis.

Little, T. D., Jorgensen, T.J., Lang, K. M., Whitney, E., & Moore, G. (2013). On the Joys of Missing Data. *Journal of Pediatric Psychology* 39(2), 151–162. doi:10.1093/jpepsy/jst048.

Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data*. Vol. 793. John Wiley & Sons, Harvard.

Little, R. J. (2021) Missing Data Assumptions. The Annual Review of Statistics and Its Application is online at statistics.annualreviews.org. https://doi.org/10.1146/annurev-statistics-040720

Matyas, L. and Sevestre, P. (1992). *The Econometrics of Panel Data*, 46-71. Kluwer Academic Publish

Marsh, H. W. (1998). Pairwise deletion for missing data in structural equation models: Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes. *Structural Equation Modeling: A Multidisciplinary Journal*, 5: 22-36.

Bates, M. D., Papke, L. E., & Wooldridge, J. M. (2024). Nonlinear correlated random effects models with endogeneity and unbalanced panels. *Econometric Reviews*, 43(9): 713–732. https://doi.org/10.1080/07474938.2024.2357431

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74: 525-556.

Popovich, D. (2024). How To Treat Missing Data in Survey Research. *Journal of Marketing Theory and Practice*, 1–17. https://doi.org/10.1080/10696679.2024.2376052

Ren, L., Wang, T., Seklouli-Sekhari, A., Zhang,H., Bouras, A. . (2023) A review on missing values for main challenges and methods. *Information Systems*, 119, pp.102268. ff10.1016/j.is.2023.102268ff. ffhal-04426492f

Robeson. S. M., Willmott, C. J. (2023). Decomposition of the mean absolute error (MAE) into systematic and unsystematic components. *PLoS ONE* 18(2): e0279774. https://doi.org/10.1371/journal.pone.0279774.

Salgado, C.M., Azevedo, C., Proença, H., & Vieira, S.M. (2016). *Missing Data. In: Secondary Analysis of Electronic Health Records*. Springer, Cham. https://doi.org/10.1007/978-3-319-43742-2_13

Stewart, K. (2024). M*ean Squared Error*. *Encyclopedia Britannica*. https://www.britannica.com/science/mean-squared-error

Swarmy. P. A. V. B. & Arora. S. S. (1972). The Exact Finite Sample Properties of the Estimators of Coefficients in the Error Components Regression Models. *Econometrica*, 40(2): pp. 261-275

Veroniki, A. & Salanti. G. (2013). Methods to estimate the heterogeneity variance, its uncertainty and to draw inference on the meta-analysis summary effect. 2013. url: https://methods.cochrane.org/statistics/sites/methods.cochrane.org.statistics/files/public/uploads/VeronikiSalantiHeterogeneitySMGMeetingQuebec2013.pdf.

Wooldridge. J. M. (2012). *Introductory Econometrics*. A Modern Approach *Fifth Edition*