Mathematics

AIMS Mathematics, 9(12): 33551–33571.
DOI: 10.3934/math.20241601
Received: 06 August 2024
Revised: 21 October 2024
Accepted: 25 October 2024
Published: 26 November 2024

https://www.aimspress.com/journal/Math

Research article

Solving the incomplete data problem in Greco-Latin square experimental design by exact-scheme analysis of variance without data imputation

Kittiwat Sirikasemsuk[1,*], Sirilak Wongsriya[1] and Kanogkan Leerojanaprapa[2]

[1] Department of Industrial Engineering, School of Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand
[2] Statistics Department, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand

* **Correspondence:** Email: kittiwat.sirikasemsuk@gmail.com; Tel: +6623298339225.

Abstract: This study introduced a novel exact-scheme analysis of variance to tackle the challenge of incomplete data within the Greco-Latin square experimental design (GLSED), specifically for scenarios with a single missing observation across any treatment and block level, thus eliminating the need for conventional data imputation methods. This approach innovatively addresses and mitigates the bias in the treatment sum of squares, a significant drawback of traditional missing plot techniques, by providing a precise, exact-scheme-based formula for calculating the treatment sum of squares in fixed-effect GLSED contexts with unrecorded values. Moreover, it offers a method for correcting biased treatment sum of squares values, presenting an adjustment mechanism for instances where the least squares method was previously employed to estimate missing values. This comprehensive strategy not only enhances the methodological accuracy and integrity of GLSED studies but also contributes significantly to the field by offering a solution to navigate the complexities of incomplete datasets without resorting to data imputation, thus improving the rigor and validity of experimental designs in the face of missing data challenges.

Keywords: design of experiment; ANOVA; linear algebra equations; Greco-Latin square design; unrecorded observation; missing value
Mathematics Subject Classification: 15A06, 15A09, 62J10, 62K05, 62K10

1. Introduction

In business, the design of experiments (DOE) emerges as a focal scientific methodology that significantly enhances profitability through strategic manipulation of potential factors by management. This technique is instrumental in refining and optimizing both the design and the advancement of manufacturing processes. It achieves this by minimizing variation in the response variable, thereby facilitating the development of products and processes that are both reliable and efficient. Central to Fisher's approach to DOE is the employment of an analysis of variance (ANOVA), which relies on the F-test to assess the statistical significance among group means. The foundations of Fisher's DOE are bolstered by three critical principles: blocking, randomization, and replication, each serving to ensure the integrity and validity of experimental results [1]. In the context of DOE, experiments that allow for the investigation of a single potential factor under conditions where every observation is equally likely to participate in the trial are categorized as completely randomized designs. In contrast, designs that do not fully randomize the assignment of observations but still focus on a single potential factor—such as the randomized complete block design (RCBD), Latin square experimental design (LSED), and Greco-Latin square experimental design (GLSED)—are recognized for their structured approach to managing variability within experimental settings.

The Greco-Latin square experimental design (GLSED) is a structured methodology that assigns treatments within a two-dimensional matrix, with the stipulation that each treatment is represented exactly once in every row and column, thereby facilitating the control of three nuisance factors efficiently within the confines of limited resources and time constraints [2,3]. This design framework is especially advantageous in the fields of agriculture, medicine, and industry, where it is imperative to systematically test multiple variables to elevate the precision and reliability of research outcomes [4]. An exemplary case further elucidates the efficacy of GLSED, demonstrating its capacity to address complex experimental demands by ensuring comprehensive coverage and balanced representation of treatments across the experimental matrix, thus exemplifying its critical role in optimizing experimental strategies across diverse scientific disciplines.

According to Diawara et al. [5], the employment of a Greco-Latin square significantly streamlines the experimental process by reducing both the labor and costs associated with conducting experimental runs, thereby enabling a thorough and systematic investigation of all relevant factors. An illustrative analysis utilizing a square of order 3 accentuated the profound impact of variables such as flow rate, insulin type, pump type, and vibration on the quantity of insulin delivered, a finding of paramount importance for achieving precise dosing and averting long-term complications in diabetes patients. Further, Mahamud and Gomes [6] investigated the enzymatic saccharification of sugar cane bagasse using the Greco-Latin square design. They found optimal conditions at 2.5% substrates, 5.5 ml enzyme, pH 4.5, and 45°C, differing from previous conditions of 2.0% substrates, 5 ml enzyme, pH 5, and 50°C. This enhanced saccharification yields more fermentable sugar, improving the competitiveness of fuel ethanol production. In another application, Woodside and Pearce [7] examined the efficacy of shotblasting—a novel furnace tube cleaning technology—analyzing factors such as price, cleaning time, energy efficiency, and tube damage. Using a Greco-Latin square, the study adeptly condensed the experimental design from an unwieldy 81 scenarios in a full factorial setup to a more manageable nine-product design, revealing that all factors, except for cleaning time, exert a significant influence on market share. This insight offers strategic value to marketers seeking to enhance the industrial market penetration of shotblasting technology. Meanwhile, Tovar-Aguilar et al. [8] applied the Greco-

Latin square design to evaluate the effectiveness of safety eyewear among citrus harvest workers in Florida, addressing concerns over lens fogging and dew. The study distinguished one type of safety glasses as exhibiting superior fog resistance, demonstrating the design's utility in minimizing biases and enhancing the reliability of results across various research contexts.

In the domain of design of experiments (DOE), both completely and incompletely randomized frameworks can encounter unrecorded observations, leading to an incomplete-data design. This phenomenon may arise from two primary causes: (1) the intentional exclusion of observation data at the outset, necessitated by the scarcity of experimental units [9], or (2) the emergence of unexpected circumstances or the presence of outliers. In the former case, the arrangement of observation data often adheres to a balanced DOE approach, benefiting from established formulas for computing treatment and error sums of squares, with exemplars of such designs including the Youden square design [10], the balanced incomplete block design (BIBD) [11], and the balanced incomplete Latin square design (BILSD) [12]. Conversely, in scenarios marked by the latter cause, outliers are pinpointed and excised, resulting in an unbalanced DOE. Distinguished from its balanced counterpart by the absence of pre-existing ANOVA formulas to directly address the unbalanced nature of the data, unbalanced DOEs employ the missing plot technique to estimate the unrecorded observations. Subsequently, the treatment and error sums of squares are computed utilizing tailor-made ANOVA formulas specifically devised for these more complex and irregular datasets [13,14], reflecting the comprehensive and adaptive methodologies required to maintain the integrity and validity of experimental research under such conditions.

Table 1 presents a systematic compilation of scholarly literature focused on the estimation of missing observations within the context of unbalanced designs, utilizing the traditional missing plot technique. This methodological approach, crucial for addressing data gaps in experimental designs, undergoes a thorough examination across different instances of unbalanced designs of experiments (DOEs) in the comprehensive reviews detailed in references [15] and [16]. These sources collectively offer an in-depth exploration of the conventional missing plot technique, shedding light on its application, efficacy, and limitations within the framework of unbalanced DOEs, by this means contributing to the broader academic discourse on experimental design and data estimation strategies.

Table 1. Literature on the estimation of missing observations (unbalanced DOE) using the conventional missing plot technique.

Design strategy	Reference
RCBD	[17,18]
LSED	[17,18]
Greco-Latin square experimental design (GLSED)	[19]
BIBD	[20,21]
Youden design	[10]
Split plot design	[13]

Moreover, the analysis of covariance (ANCOVA) [22] has been employed as a statistical technique for estimating missing observations within unbalanced designs, with notable contributions from scholars such as Coons [23], Cochran [24], and Wilkinson [25]. This approach enhances the precision of data analysis by incorporating covariates that may influence the dependent variable, thus providing a more nuanced understanding of experimental results. In a specific advancement within this domain, Ogbonnaya and Uzochukwu [26] developed specialized formulas aimed at estimating missing observations in scenarios involving one-factor ANCOVA with a single covariate. Furthermore, in the context of the Greco-Latin square experimental design (GLSED), where the challenge of a single missing data value arises, Kupolusi and Ojo [19] introduced a formula that utilizes minimizing the error sum of squares to estimate this missing value. Subsequently, the analysis of variance is conducted as though the dataset were complete, maintaining the integrity of the experimental design and allowing for the comprehensive interpretation of results. This methodological innovation signifies a critical step forward in addressing the complexities associated with missing data in experimental research, ensuring robust and reliable analysis despite the challenges posed by unbalanced designs.

The reliance on the conventional missing plot technique and ANCOVA for the estimation of missing observations, as opposed to utilizing actual experimental data, introduces a notable bias in the treatment sum of squares, as highlighted in studies [13,14]. This inherent bias necessitates a thorough estimation process to accurately identify and subsequently subtract it from the treatment sum of squares, which are initially derived from pre-existing, ready-made formulas. This corrective measure is crucial for ensuring the integrity and accuracy of statistical analyses within experimental research, addressing the challenges posed by incomplete datasets and preserving the validity of conclusions drawn from such studies.

The exact-scheme analysis of variance emerges as a formidable resolution for handling designs plagued by incomplete data, obviating the necessity for the estimation of missing values, as substantiated by the literature [1,27,28]. Termed alternatively as the model comparison-based exact scheme, this methodology guarantees robustness in statistical analysis that traditional approaches might not offer. Specifically, Montgomery [1] delineates a precise methodology designed to tackle the issue of a single missing value within the domains of randomized complete block design (RCBD) and Latin square experimental design (LSED) through the application of exact-scheme analysis of variance. This approach strategically circumvented the conventional reliance on data imputation, thus presenting a more direct and potentially unbiased avenue for analyzing experimental data under conditions of incompleteness. This innovation not only enhances the accuracy of statistical analyses but also aligns with the broader goal of maintaining methodological integrity in the face of missing experimental data.

Sirikasemsuk et al. [29] expanded Montgomery's [1] research to analyze RCBD with missing data, where the missing values could occur in any cell of the observation table. They developed mathematical formulas for fitted parameters and the regression sum of squares using the exact-scheme analysis of variance. Similarly, Sirikasemsuk and Leerojanaprapa [30] addressed incomplete 4×4 Latin square designs with various patterns of two missing observations, simplifying parameter estimation and ANOVA calculations. Furthermore, Sirikasemsuk [31] and Sirikasemsuk and Leerojanaprapa [32] have shed light on the unbiased calculation of the treatment sum of squares within the context of LSED when confronted with a single missing value, utilizing the aforementioned method.

To illustrate these concepts, let us revisit a real example of the Greco-Latin square experimental design (GLSED). Subramani [33] referenced a 5×5 GLSED with incomplete data from Montgomery [34] to study the effects of raw material batches, acid concentrations, standing times, and catalyst concentrations on the yield of a chemical process. In this study, Subramani [33] estimated several missing values using non-iterative least squares estimation. However, such approaches can lead to biased analyses of the sum of squares and mean squares, which may result in incorrect conclusions about the factors affecting the response variable. This bias has been demonstrated in several studies, including those by Rangaswamy [14], Little and Rubin [16], and Ott and Longnecker [35]. Recent research by AlAita et al. [36] further emphasizes the significance of missing data as a critical challenge in analyzing Greco-Latin square designs, highlighting the risks of bias and reduced reliability in conclusions.

Given these challenges, an improved method is essential to eliminate bias and enhance the reliability of experimental results in the presence of missing data. Their exploratory work in gap analysis accentuates a noticeable void in the existing literature, specifically the absence of definitive evidence concerning the unbiased treatment sum of squares in GLSED. This identified lacuna not only delineates the scope but also establishes the central focus of the current research, indicating a targeted investigation aimed at addressing this gap and contributing novel insights into the methodology for ensuring unbiased statistical analysis in GLSED settings.

Nevertheless, in the conventional framework of the exact scheme, the estimation of fitted parameters—comprising the overall mean, treatment, and block effects—alongside the calculation of the regression sums of squares (RSS) for both full and reduced effect models, emerges as a notably time-intensive process. Addressing this efficiency concern, the present research introduces a novel, exact scheme-based formula for the instantaneous computation of RSS within the full model of a $P \times P$-dimension fixed-effect GLSED with a single missing observation. This innovation circumvents the traditional necessity for estimating fitted parameters, where P represents the levels of treatment and blocks. Table 2 provides a visualization of the GLSED table, underscoring the practical application of this formula. This advancement not only streamlines the analytical process but also enhances the methodological rigor by enabling more rapid derivation of results without compromising on the accuracy and integrity of the statistical analysis.

Table 2. Example of a *4 × 4* Greco-Latin square experimental design.

Colum Row		ψ_1	ψ_2	ψ_3	ψ_4	$y_{i\cdots}$
		1	*2*	*3*	*4*	
θ_1	*1*	$A \propto= y_{1111}$	$B\beta = y_{1222}$	$C\gamma = y_{1333}$	$D\delta = y_{1444}$	$\boldsymbol{y_{1\cdots}}$
θ_2	*2*	$B\delta = y_{2241}$	$A\gamma = y_{2132}$	$D\beta = y_{2423}$	$C \propto = y_{2314}$	$\boldsymbol{y_{2\cdots}}$
θ_3	*3*	$C\beta = y_{3321}$	$D \propto = y_{3412}$	$A\delta = y_{3143}$	$B\gamma = y_{3234}$	$\boldsymbol{y_{3\cdots}}$
θ_4	*4*	$D\gamma = y_{4421}$	$C\delta = y_{4342}$	$B \propto = y_{4213}$	$A\beta = y_{4124}$	$\boldsymbol{y_{4\cdots}}$
$y_{\cdots l}$		$\boldsymbol{y_{\cdots 1}}$	$\boldsymbol{y_{\cdots 2}}$	$\boldsymbol{y_{\cdots 3}}$	$\boldsymbol{y_{\cdots 4}}$	$\boldsymbol{y_{\cdots\cdot}}$

The structure of this research paper is organized as follows: Section 1 introduces the foundational concepts and the overarching aim of the study. Section 2 delineates the model comparison-based exact scheme, offering a detailed examination of its theoretical underpinnings. In Section 3, the focus shifts to linear algebra equations and the methodologies employed for estimating the fitted parameters, crucial for the analytical framework of the study. Section 4 explores the intricacies of the regression sum of squares, a key component in the evaluation of model efficacy. Section 5 proposes an innovative exact scheme-based instant formula for calculating the Latin letter treatment sum of squares, a significant contribution to the field of experimental design. Finally, Section 6 presents concluding remarks, encapsulating the findings and implications of the research. This logical progression through the sections ensures a coherent and comprehensive exploration of the subject matter, facilitating a deeper understanding of the proposed methodologies and their potential impact on experimental design analysis.

The notations and definitions used in this research are as follows:

y_{ijkl} the observation of row i, Latin letter treatment j, Greek letter *k,* and column *l*;

P the levels or size of the Greco-Latin square experimental design;
i row index ($i = 1, 2, 3, \ldots, P$);
j Latin letter treatment index ($j = 1, 2, 3, \ldots, P$);
k Greek letter index ($k = 1, 2, 3, \ldots, P$);
l column index ($l = 1, 2, 3, \ldots, P$);
μ the overall mean;
θ_i the row effect of level i;
τ_j the Latin letter treatment effect of level j;
ω_k the Greek letter effect of level k;
ψ_l the column effect of level l;
ε_{ijkl} the statistical error due to other sources of variability;
$y_{....}$ the grand total;
$y_{i...}$ the i^{th} row total;
$y_{.j..}$ the j^{th} Latin letter treatment total;
$y_{..k.}$ the k^{th} Greek letter total;
$y_{...l}$ the l^{th} column total;
r the row index of the missing observation;
n the Latin letter treatment index of the missing observation;
m the Greek letter index of the missing observation;
c the column index of the missing observation;
SS_R the row sum of squares;
SS_{Tr} the Latin letter treatment sum of squares;
SS_G the Greek letter sum of squares;
SS_C the column sum of squares;
SS_T the total sum of squares;
SS_E the error sum of squares;
$R(\mu, \theta, \tau, \omega, \psi)$ the overall regression sum of squares (ORSS) for the full model;
$R(\mu, \tau, \omega, \psi)$ the regression sum of squares for the reduced model, omitting the row effect;
$R(\mu, \theta, \omega, \psi)$ the regression sum of squares for the reduced model, omitting the Latin letter treatment effect;
$R(\mu, \theta, \tau, \psi)$ the regression sum of squares for the reduced model, omitting the Greek letter effect;
$R(\mu, \theta, \tau, \omega)$ the regression sum of squares for the reduced model, omitting the column effect.

2. The model comparison-based exact scheme

The Latin letter treatment sum of squares (SS_{Tr}) for the complete-data GLSED can be calculated by

$$SS_{Tr} = \frac{\sum_{j=1}^{P} y_{.j..}^2}{P} - \frac{y_{....}^2}{P^2}. \tag{2.1}$$

However, Eq (2.1) is not applicable to the incomplete-data (unbalanced) design. Therefore, the model comparison-based exact scheme is used to solve the incomplete-data design. SS_{Tr} of the exact scheme is expressed as

$$SS_{Tr} = R(\mu, \theta, \tau, \omega, \psi) - R(\mu, \theta, \omega, \psi). \tag{2.2}$$

It is noted that the Latin letter treatment sum of squares (SS_{Tr}) is calculated as the difference between the full-model regression sum of squares and the reduced-model regression sum of squares, omitting the treatment effect [1,9,14]. Likewise, the row, column, and Greek letter sums of squares are calculated as the discrepancies between the regression sums of squares of the full and reduced models, excluding the corresponding effects. Hence, the sums of squares for the GLSED are determined as

$$SS_R = R(\mu,\theta,\tau,\omega,\psi) - R(\mu,\tau,\omega,\psi), \tag{2.3}$$

$$SS_G = R(\mu,\theta,\tau,\omega,\psi) - R(\mu,\theta,\tau,\psi), \tag{2.4}$$

$$SS_C = R(\mu,\theta,\tau,\omega,\psi) - R(\mu,\theta,\tau,\omega). \tag{2.5}$$

3. Linear algebra equations and the fitted parameters

The GLSED full-model linear equation for y_{ijkl} is

$$y_{ijkl} = \mu + \theta_i + \tau_j + \omega_k + \psi_l + \varepsilon_{ijkl} \begin{cases} i = 1,2,\dots,P \\ j = 1,2,\dots,P \\ k = 1,2,\dots,P \\ l = 1,2,\dots,P \end{cases}. \tag{3.1}$$

In the fixed-effect GLSED, representing constraints, the sums of the fitted parameters can be expressed as

$$\textstyle\sum_{All\ i}^{P} \hat{\theta}_i = 0, \tag{3.2}$$

$$\textstyle\sum_{All\ j}^{P} \hat{\tau}_j = 0, \tag{3.3}$$

$$\textstyle\sum_{All\ k}^{P} \hat{\omega}_k = 0, \tag{3.4}$$

and

$$\textstyle\sum_{All\ l}^{P} \hat{\psi}_l = 0. \tag{3.5}$$

The regression sum of squares (RSS) for the full model of y_{ijkl} is expressed as [1],

$$R(\mu,\theta,\tau,\omega,\psi) = \hat{\mu}y_{....} + \textstyle\sum_{i=1}^{P} \hat{\theta}_i y_{i...} + \sum_{j=1}^{P} \hat{\tau}_j y_{.j..} + \sum_{k=1}^{P} \hat{\omega}_k y_{..k.} + \sum_{l=1}^{P} \hat{\psi}_l y_{...l}. \tag{3.6}$$

In estimating the parametric values of the full model ($\hat{\mu}$, $\hat{\theta}_i$, $\hat{\tau}_j$, $\hat{\omega}_k$, $\hat{\psi}_l$), a series of linear algebra equations for the $P \times P$ GLSED with one missing observation can be written as

$$\mu: \quad \begin{aligned} &(P^2-1)\hat{\mu} + P\textstyle\sum_{i=1,i\neq r}^{P} \hat{\theta}_i + (P-1)\hat{\theta}_r + P\sum_{j=1,j\neq n}^{P} \hat{\tau}_j + (P-1)\hat{\tau}_n \\ &+P\textstyle\sum_{k=1,k\neq m}^{P} \hat{\omega}_k + (P-1)\hat{\omega}_m + P\sum_{l=1,l\neq c}^{P} \hat{\psi}_l + (P-1)\hat{\psi}_c = y_{....}, \end{aligned} \tag{3.7}$$

$$\theta_r: \quad (P-1)\hat{\mu} + (P-1)\hat{\theta}_r + \textstyle\sum_{j=1,j\neq n}^{P} \hat{\tau}_j + \sum_{k=1,k\neq m}^{P} \hat{\omega}_k + \sum_{l=1,l\neq c}^{P} \hat{\psi}_l = y_{r...}, \tag{3.8}$$

$$\tau_n: \quad (P-1)\hat{\mu} + \textstyle\sum_{i=1,i\neq r}^{P} \hat{\theta}_i + (P-1)\hat{\tau}_n + \sum_{k=1,k\neq m}^{P} \hat{\omega}_k + \sum_{l=1,l\neq c}^{P} \hat{\psi}_l = y_{.n..}, \tag{3.9}$$

$$\omega_m: \quad (P-1)\hat{\mu} + \textstyle\sum_{i=1,i\neq r}^{P} \hat{\theta}_i + \sum_{j=1,j\neq n}^{P} \hat{\tau}_j + (P-1)\hat{\omega}_m + \sum_{l=1,l\neq c}^{P} \hat{\psi}_l = y_{..m.}, \tag{3.10}$$

$$\psi_c: \qquad (P-1)\hat{\mu} + \textstyle\sum_{i=1,i\neq r}^{P} \hat{\theta}_i + \sum_{j=1,j\neq n}^{P} \hat{\tau}_j + \sum_{k=1,k\neq m}^{P} \hat{\omega}_k + (P-1)\hat{\psi}_c = y_{\cdot\cdot\cdot c}, \tag{3.11}$$

$$\theta_i: \qquad P\hat{\mu} + P\hat{\theta}_i + \textstyle\sum_{j=1}^{P} \hat{\tau}_j + \sum_{k=1}^{P} \hat{\omega}_k + \sum_{l=1}^{P} \hat{\psi}_c = y_{i\cdot\cdot\cdot} \ when\ i \neq r, \tag{3.12}$$

$$\tau_j: \qquad P\hat{\mu} + \textstyle\sum_{i=1}^{P} \hat{\theta}_i + P\hat{\tau}_j + \sum_{k=1}^{P} \hat{\omega}_k + \sum_{l=1}^{P} \hat{\psi}_c = y_{\cdot j\cdot\cdot} \ when\ j \neq n, \tag{3.13}$$

$$\omega_k: \qquad P\hat{\mu} + \textstyle\sum_{i=1}^{P} \hat{\theta}_i + \sum_{j=1}^{P} \hat{\tau}_j + P\hat{\omega}_k + \sum_{l=1}^{P} \hat{\psi}_c = y_{\cdot\cdot k\cdot} \ when\ k \neq m, \tag{3.14}$$

$$\psi_l: \qquad P\hat{\mu} + \textstyle\sum_{i=1}^{P} \hat{\theta}_i + \sum_{j=1}^{P} \hat{\tau}_j + \sum_{k=1}^{P} \hat{\omega}_k + P\hat{\psi}_l = y_{\cdot\cdot\cdot l} \ when\ \ l \neq c. \tag{3.15}$$

Given Eqs (3.2)–(3.5) and the matrix form of Eqs (3.7)–(3.11), the fitted parameters for the P × P GLSED with one missing observation can be determined by

$$\hat{\mu} = \frac{(P-4)y_{\cdot\cdot\cdot\cdot} + y_{r\cdot\cdot\cdot} + y_{\cdot n\cdot\cdot} + y_{\cdot\cdot m\cdot} + y_{\cdot\cdot\cdot c}}{(P-3)(P-1)P}, \tag{3.16}$$

$$\hat{\theta}_r = \frac{y_{r\cdot\cdot\cdot} - y_{\cdot\cdot\cdot\cdot}}{P} + (P-1)\hat{\mu}, \tag{3.17}$$

$$\hat{\tau}_n = \frac{y_{\cdot n\cdot\cdot} - y_{\cdot\cdot\cdot\cdot}}{P} + (P-1)\hat{\mu}, \tag{3.18}$$

$$\hat{\omega}_m = \frac{y_{\cdot\cdot m\cdot} - y_{\cdot\cdot\cdot\cdot}}{P} + (P-1)\hat{\mu}, \tag{3.19}$$

$$\hat{\psi}_c = \frac{y_{\cdot\cdot\cdot c} - y_{\cdot\cdot\cdot\cdot}}{P} + (P-1)\hat{\mu}. \tag{3.20}$$

Likewise, the fitted parameters for $i \neq r,\ j \neq n,\ k \neq m,$ and $l \neq c$ can be rewritten as

$$\hat{\theta}_i = \frac{y_{i\cdot\cdot\cdot}}{P} - \hat{\mu}, \qquad for\ \ i \neq r, \tag{3.21}$$

$$\hat{\tau}_j = \frac{y_{\cdot j\cdot\cdot}}{P} - \hat{\mu}, \qquad for\ \ j \neq n, \tag{3.22}$$

$$\hat{\omega}_k = \frac{y_{\cdot\cdot k\cdot}}{P} - \hat{\mu}, \qquad for\ \ k \neq m, \tag{3.23}$$

and

$$\hat{\psi}_l = \frac{y_{\cdot\cdot\cdot l}}{P} - \hat{\mu}, \qquad for\ \ l \neq c. \tag{3.24}$$

4. The regression sum of squares

The full-model and reduced-model regression sums of squares (RSS) can be determined following Propositions 4.1 and 4.2, respectively.

Proposition 4.1. Given the P × P GLSED with one missing observation, the full-model RSS of y_{ijkl} can be calculated by

$$R(\mu, \theta, \tau, \omega, \psi) = \frac{\left(\sum_{All\ i} y_{i\cdot\cdot\cdot}^2 + \sum_{All\ j} y_{\cdot j\cdot\cdot}^2 + \sum_{All\ k} y_{\cdot\cdot k\cdot}^2 + \sum_{All\ l} y_{\cdot\cdot\cdot l}^2\right)}{P} + \frac{(1-P)\left(3y_{\cdot\cdot\cdot\cdot}^2 - y_{SUM}^2\right) + (y_{SUM} - 3y_{\cdot\cdot\cdot\cdot})^2}{P(P-3)(P-1)}, \tag{4.1}$$

when $y_{SUM} = y_{r\cdots} + y_{\cdot n\cdot\cdot} + y_{\cdot\cdot m\cdot} + y_{\cdots c}$.

Proof. Substituting Eqs (3.17)–(3.24) in Eq (3.6), we have

$$R(\mu,\theta,\tau,\omega,\psi) = \hat{\mu}y_{\cdots\cdot} + \frac{\left(\sum_{i=1,i\neq r}^{P} y_{i\cdots}^{2} + \sum_{j=1,j\neq n}^{P} y_{\cdot j\cdot\cdot}^{2} + \sum_{k=1,k\neq m}^{P} y_{\cdot\cdot k\cdot}^{2} + \sum_{l=1,l\neq c}^{P} y_{\cdots l}^{2}\right)}{P}$$

$$-\begin{pmatrix} \hat{\mu}\sum_{i=1,i\neq r}^{P} y_{i\cdots} + \hat{\mu}\sum_{j=1,j\neq n}^{P} y_{\cdot j\cdot\cdot} + \\ \hat{\mu}\sum_{k=1,k\neq m}^{P} y_{\cdot\cdot k\cdot} + \hat{\mu}\sum_{l=1,l\neq c}^{P} y_{\cdots l} \end{pmatrix} + \hat{\mu}(P-1)(y_{r\cdots} + y_{\cdot n\cdot\cdot} + y_{\cdot\cdot m\cdot} + y_{\cdots c})$$

$$+ \frac{(y_{r\cdots}^{2} + y_{\cdot n\cdot\cdot}^{2} + y_{\cdot\cdot m\cdot}^{2} + y_{\cdots c}^{2})}{P} - \frac{y_{\cdots\cdot}}{P}(y_{r\cdots} + y_{\cdot n\cdot\cdot} + y_{\cdot\cdot m\cdot} + y_{\cdots c}).$$

Modifying all the summation terms to account for all indexes (*i, j, k, l*),

$$R(\mu,\theta,\tau,\omega,\psi) = \frac{\left(\sum_{i=1}^{P} y_{i\cdots}^{2} + \sum_{j=1}^{P} y_{\cdot j\cdot\cdot}^{2} + \sum_{k=1}^{P} y_{\cdot\cdot k\cdot}^{2} + \sum_{l=1}^{P} y_{\cdots l}^{2}\right)}{P} + P\hat{\mu}(y_{r\cdots} + y_{\cdot n\cdot\cdot} + y_{\cdot\cdot m\cdot} + y_{\cdots c})$$

$$-\frac{y_{\cdots\cdot}}{P}(y_{r\cdots} + y_{\cdot n\cdot\cdot} + y_{\cdot\cdot m\cdot} + y_{\cdots c}) - 3\hat{\mu}y_{\cdots\cdot},$$

where $\sum_{i=1}^{P} y_{i\cdots} = \sum_{j=1}^{P} y_{\cdot j\cdot\cdot} = \sum_{k=1}^{P} y_{\cdot\cdot k\cdot} = \sum_{l=1}^{P} y_{\cdots l} = y_{\cdots\cdot}$.

For the sake of convenience,

$$y_{SUM} = y_{r\cdots} + y_{\cdot n\cdot\cdot} + y_{\cdot\cdot m\cdot} + y_{\cdots c}.$$

The full-model RSS of y_{ijkl} can be rewritten as

$$R(\mu,\theta,\tau,\omega,\psi) = \frac{\left(\sum_{i=1}^{P} y_{i\cdots}^{2} + \sum_{j=1}^{P} y_{\cdot j\cdot\cdot}^{2} + \sum_{k=1}^{P} y_{\cdot\cdot k\cdot}^{2} + \sum_{l=1}^{P} y_{\cdots l}^{2}\right)}{P} + P\hat{\mu}y_{SUM} - \frac{y_{\cdots\cdot}}{P}y_{SUM} - 3\hat{\mu}y_{\cdots\cdot}.$$

Substituting $\hat{\mu}$ (Eq (3.16)) into the above equation and rearranging, the full-model RSS can be derived as Eq (4.1).

This completes the proof.

Proposition 4.2. In the presence of a single missing observation in the *P* × *P* GLSED, the reduced-model regression sums of squares can be determined by

$$R(\mu,\tau,\omega,\psi) = \frac{\sum_{j=1}^{P} y_{\cdot j\cdot\cdot}^{2} + \sum_{k=1}^{P} y_{\cdot\cdot k\cdot}^{2} + \sum_{l=1}^{P} y_{\cdots l}^{2}}{P} + \frac{(1-P)(2y_{\cdots\cdot}^{2} - y_{SUM\_R}^{2}) + (y_{SUM\_R} - 2y_{\cdots\cdot})^{2}}{(P-2)(P-1)P}, \tag{4.2}$$

$$R(\mu,\theta,\omega,\psi) = \frac{\sum_{i=1}^{P} y_{i\cdots}^{2} + \sum_{k=1}^{P} y_{\cdot\cdot k\cdot}^{2} + \sum_{l=1}^{P} y_{\cdots l}^{2}}{P} + \frac{(1-P)(2y_{\cdots\cdot}^{2} - y_{SUM\_N}^{2}) + (y_{SUM\_N} - 2y_{\cdots\cdot})^{2}}{(P-2)(P-1)P}, \tag{4.3}$$

$$R(\mu,\theta,\tau,\psi) = \frac{\sum_{i=1}^{P} y_{i\cdots}^{2} + \sum_{j=1}^{P} y_{\cdot j\cdot\cdot}^{2} + \sum_{l=1}^{P} y_{\cdots l}^{2}}{P} + \frac{(1-P)(2y_{\cdots\cdot}^{2} - y_{SUM\_M}^{2}) + (y_{SUM\_M} - 2y_{\cdots\cdot})^{2}}{(P-2)(P-1)P}, \tag{4.4}$$

$$R(\mu,\theta,\tau,\omega) = \frac{\sum_{i=1}^{P} y_{i\cdots}^{2} + \sum_{j=1}^{P} y_{\cdot j\cdot\cdot}^{2} + \sum_{k=1}^{P} y_{\cdot\cdot k\cdot}^{2}}{P} + \frac{(1-P)(2y_{\cdots\cdot}^{2} - y_{SUM\_C}^{2}) + (y_{SUM\_C} - 2y_{\cdots\cdot})^{2}}{(P-2)(P-1)P}, \tag{4.5}$$

when $y_{SUM\_R} = y_{\cdot n\cdot\cdot} + y_{\cdot\cdot m\cdot} + y_{\cdots c}$, $y_{SUM\_N} = y_{r\cdots} + y_{\cdot\cdot m\cdot} + y_{\cdots c}$, $y_{SUM\_M} = y_{r\cdots} + y_{\cdot n\cdot\cdot} + y_{\cdots c}$, and $y_{SUM\_C} = y_{r\cdots} + y_{\cdot n\cdot\cdot} + y_{\cdot\cdot m\cdot}$.

Proof. The determination of the reduced-model regression sums of squares can be conducted in an identical fashion to the case of $R(\mu,\theta,\tau,\omega,\psi)$ as in Proposition 4.1. For example, let us consider a case of $R(\mu,\theta,\omega,\psi)$ (Eq (2.2)) omitting the Latin letter treatment effect, where the reduced-model linear equation for $y_{i(j)kl}$ is

$$y_{i(j)kl} = \mu^{NT} + \theta_i^{NT} + \omega_k^{NT} + \psi_l^{NT} + \varepsilon_{ijkl} \begin{cases} i = 1,2,\dots,P \\ j = 1,2,\dots,P \\ k = 1,2,\dots,P \\ l = 1,2,\dots,P \end{cases}.$$

It is presumed that Latin letter treatment effects (τ_j) are not accounted for across all values of j. Consequently, the estimated variables μ, θ_i, ω_k, and ψ_l in Eq (3.1) will be replaced with $\hat{\mu}^{NT}$, $\hat{\theta}_i^{NT}$, $\hat{\omega}_k^{NT}$, and $\hat{\psi}_l^{NT}$, respectively, as opposed to $\hat{\mu}$, $\hat{\theta}_i$, $\hat{\omega}_k$, and $\hat{\psi}_l$. The expression for $R(\mu,\theta,\omega,\psi)$ can be formulated as follows:

$$R(\mu,\theta,\omega,\psi) = \hat{\mu}^{NT} y_{....} + \sum_{i=1}^{P} \hat{\theta}_i^{NT} y_{i...} + \sum_{k=1}^{P} \hat{\omega}_k^{NT} y_{..k.} + \sum_{l=1}^{P} \hat{\psi}_l^{NT} y_{...l}.$$

The subsequent estimated model parameters, namely $\hat{\mu}^{NT}$, $\hat{\theta}_i^{NT}$, $\hat{\omega}_k^{NT}$, and $\hat{\psi}_l^{NT}$, were obtained through a sequence of linear algebraic equations, as outlined below:

$$\hat{\mu}^{NT} = \frac{(P-3)y_{....} + y_{r...} + y_{..m.} + y_{...c}}{(P-2)(P-1)P},$$

$$\hat{\theta}_r^{NT} = \frac{y_{r...} - y_{....}}{P} + (P-1)\hat{\mu}^{NT},$$

$$\hat{\omega}_m^{NT} = \frac{y_{..m.} - y_{....}}{P} + (P-1)\hat{\mu}^{NT},$$

$$\hat{\psi}_c^{NT} = \frac{y_{...c} - y_{....}}{P} + (P-1)\hat{\mu}^{NT},$$

$$\hat{\theta}_i^{NT} = \frac{y_{i...}}{P} - \hat{\mu}^{NT} \quad for \quad i \neq r,$$

$$\hat{\omega}_k^{NT} = \frac{y_{..k.}}{P} - \hat{\mu}^{NT} \quad for \quad k \neq m,$$

$$\hat{\psi}_l^{NT} = \frac{y_{...l}}{P} - \hat{\mu}^{NT} \quad for \quad l \neq c.$$

Replacing these values in the expression for $R(\mu,\theta,\omega,\psi)$ and after some algebraic simplification, we have Eq (4.3). This completes the proof.

5. The sum of squares and examples

Proposition 5.1. In the $P \times P$ GLSED with one missing experimental data, the unbiased Latin letter treatment sum of squares can be determined as

$$SS_{Tr} = \frac{\sum_{j=1}^{P} y_{.j..}^2}{P} + \frac{1}{(P-1)}\left[\frac{(y_{SUM} - y_{....})^2}{(P-3)} - \frac{\left(y_{SUM\_N} - y_{....}\right)^2}{(P-2)} + \frac{y_{....}(2y_{.n..} - y_{....})}{P}\right]. \tag{5.1}$$

Additionally, the determination of unbiased sums of squares for the rows, Greek letters, and columns can be established as follows:

$$SS_R = \frac{\sum_{i=1}^{P} y_{i...}^2}{P} + \frac{1}{(P-1)}\left[\frac{(y_{SUM} - y_{....})^2}{(P-3)} - \frac{\left(y_{SUM\_R} - y_{....}\right)^2}{(P-2)} + \frac{y_{....}(2y_{r...} - y_{....})}{P}\right], \tag{5.2}$$

$$SS_G = \frac{\sum_{k=1}^{P} y_{\cdot\cdot k\cdot}^2}{P} + \frac{1}{(P-1)}\left[\frac{(y_{SUM}-y_{\cdot\cdot\cdot\cdot})^2}{(P-3)} - \frac{\left(y_{SUM\_M}-y_{\cdot\cdot\cdot\cdot}\right)^2}{(P-2)} + \frac{y_{\cdot\cdot\cdot\cdot}(2y_{\cdot\cdot m\cdot}-y_{\cdot\cdot\cdot\cdot})}{P}\right], \tag{5.3}$$

$$SS_C = \frac{\sum_{l=1}^{P} y_{\cdot\cdot\cdot l}^2}{P} + \frac{1}{(P-1)}\left[\frac{(y_{SUM}-y_{\cdot\cdot\cdot\cdot})^2}{(P-3)} - \frac{\left(y_{SUM\_C}-y_{\cdot\cdot\cdot\cdot}\right)^2}{(P-2)} + \frac{y_{\cdot\cdot\cdot\cdot}(2y_{\cdot\cdot\cdot c}-y_{\cdot\cdot\cdot\cdot})}{P}\right]. \tag{5.4}$$

Proof. Equation (4.1) in Proposition 4.1 can be rearranged as:

$$\begin{aligned} R(\mu,\theta,\tau,\omega,\psi) &= \frac{\left(\sum_{All\,i}^{P} y_{i\cdot\cdot\cdot}^2 + \sum_{All\,j}^{P} y_{\cdot j\cdot\cdot}^2 + \sum_{All\,k}^{P} y_{\cdot\cdot k\cdot}^2 + \sum_{All\,l}^{P} y_{\cdot\cdot\cdot l}^2\right)}{P} \\ &\quad + \frac{\left(3y_{\cdot\cdot\cdot\cdot}^2 - y_{SUM}^2 - 3Py_{\cdot\cdot\cdot\cdot}^2 + Py_{SUM}^2\right) + \left(y_{SUM}^2 - 6y_{SUM}y_{\cdot\cdot\cdot\cdot} + 9y_{\cdot\cdot\cdot\cdot}^2\right)}{P(P-3)(P-1)} \\ &= \frac{\left(\sum_{All\,i}^{P} y_{i\cdot\cdot\cdot}^2 + \sum_{All\,j}^{P} y_{\cdot j\cdot\cdot}^2 + \sum_{All\,k}^{P} y_{\cdot\cdot k\cdot}^2 + \sum_{All\,l}^{P} y_{\cdot\cdot\cdot l}^2\right)}{P} + \frac{\left(Py_{SUM}^2 - 2Py_{SUM}y_{\cdot\cdot\cdot\cdot} + Py_{\cdot\cdot\cdot\cdot}^2\right)}{P(P-3)(P-1)} \\ &\quad + \frac{\left(2Py_{SUM}y_{\cdot\cdot\cdot\cdot} - 4Py_{\cdot\cdot\cdot\cdot}^2 - 6y_{SUM}y_{\cdot\cdot\cdot\cdot} + 12y_{\cdot\cdot\cdot\cdot}^2\right)}{P(P-3)(P-1)}, \end{aligned}$$

$$R(\mu,\theta,\tau,\omega,\psi) = \frac{\left(\sum_{All\,i}^{P} y_{i\cdot\cdot\cdot}^2 + \sum_{All\,j}^{P} y_{\cdot j\cdot\cdot}^2 + \sum_{All\,k}^{P} y_{\cdot\cdot k\cdot}^2 + \sum_{All\,l}^{P} y_{\cdot\cdot\cdot l}^2\right)}{P} + \left[\frac{(y_{SUM}-y_{\cdot\cdot\cdot\cdot})^2}{(P-3)(P-1)}\right] + \left[\frac{\left(2Py_{SUM}y_{\cdot\cdot\cdot\cdot} - 4y_{\cdot\cdot\cdot\cdot}^2\right)}{P(P-1)}\right]. \tag{5.5}$$

Equation (4.3) in Proposition 4.2 can be rearranged as:

$$\begin{aligned} R(\mu,\theta,\omega,\psi) &= \frac{\left(\sum_{All\,i}^{P} y_{i\cdot\cdot\cdot}^2 + \sum_{All\,k}^{P} y_{\cdot\cdot k\cdot}^2 + \sum_{All\,l}^{P} y_{\cdot\cdot\cdot l}^2\right)}{P} + \frac{\left(2y_{\cdot\cdot\cdot\cdot}^2 - y_{SUM\_N}^2 - 2Py_{\cdot\cdot\cdot\cdot}^2 + Py_{SUM\_N}^2\right) + \left(y_{SUM\_N}^2 - 4y_{SUM}y_{\cdot\cdot\cdot\cdot} + 4y_{\cdot\cdot\cdot\cdot}^2\right)}{P(P-2)(P-1)} \\ &= \frac{\left(\sum_{All\,i}^{P} y_{i\cdot\cdot\cdot}^2 + \sum_{All\,k}^{P} y_{\cdot\cdot k\cdot}^2 + \sum_{All\,l}^{P} y_{\cdot\cdot\cdot l}^2\right)}{P} + \frac{\left(Py_{SUM_N}^2 - 2Py_{SUM_N}y_{\cdot\cdot\cdot\cdot} + Py_{\cdot\cdot\cdot\cdot}^2\right)}{P(P-2)(P-1)} \\ &\quad + \frac{\left(2Py_{SUM\_N}y_{\cdot\cdot\cdot\cdot} - 3Py_{\cdot\cdot\cdot\cdot}^2 - 4y_{SUM\_N}y_{\cdot\cdot\cdot\cdot} + 6y_{\cdot\cdot\cdot\cdot}^2\right)}{P(P-2)(P-1)}, \end{aligned}$$

$$R(\mu,\theta,\omega,\psi) = \frac{\left(\sum_{All\,i}^{P} y_{i\cdot\cdot\cdot}^2 + \sum_{All\,k}^{P} y_{\cdot\cdot k\cdot}^2 + \sum_{All\,l}^{P} y_{\cdot\cdot\cdot l}^2\right)}{P} + \frac{\left(y_{SUM\_N}-y_{\cdot\cdot\cdot\cdot}\right)^2}{(P-2)(P-1)} + \frac{\left(2Py_{SUM\_N}y_{\cdot\cdot\cdot\cdot} - 3y_{\cdot\cdot\cdot\cdot}^2\right)}{P(P-1)}. \tag{5.6}$$

According to Eq (2.2), the Latin letter treatment sum of squares in Eq (5.1) can be deduced by subtracting Eq (5.5) from Eq (5.6). The calculations for the row, Greek letter, and column sums of squares are similarly performed for the Latin letter treatment sum of squares above. This completes the proof.

The technique employed for estimating missing data through the least squares method constitutes one of various approaches to facilitate the derivation of the analysis of variance based on the original formula. Before proceeding with additional result analysis, it is essential to rectify the bias through subtraction, as detailed in Proposition 5.2.

Proposition 5.2. In the $P \times P$ GLSED with a singular missing data point, the formula for bias adjustment concerning the sum of squares for the Latin letter treatment, applied subsequent to the estimation of missing data through the least squares method, is established as follows:

$$Bias = \frac{(y_{\cdot\cdot\cdot\cdot} - y_{SUM} - (p-3)y_{\cdot n\cdot\cdot})^2}{(P-3)^2(P-2)(P-1)}. \tag{5.7}$$

Proof. We have

$$Bias = SS_{Tr(bias)} - SS_{Tr(unbias)\_exact}, \tag{5.8}$$

where $SS_{Tr(unbias)\_exact}$ represents the treatment sum of squares without bias, calculated from Eq (5.1) in Proposition 5.1 and $SS_{Tr(bias)}$ denotes the treatment sum of squares with bias after estimating missing data via the least squares method, as outlined by Kupolusi and Ojo [19] in the following Eq (5.9) for the missing value (Z) estimate:

$$Z = \frac{Py_{SUM} - 3y_{....}}{(P-3)(P-1)}. \tag{5.9}$$

Hence, Eq (2.1) for the complete data can be rewritten as

$$SS_{Tr} = \frac{\sum_{j=1, j\neq n}^{P} y_{\cdot j\cdot\cdot}^{2} + (y_{\cdot n\cdot\cdot} + Z)^2}{P} - \frac{(y_{....} + Z)^2}{P^2}, \tag{5.10}$$

where the symbol $y_{....}$ represents the grand total, exclusive of the value represented by Z.

It is noted that the inaugural expression in Eq (5.1) is subtracted from the initial term in Eq (5.10), resulting in the value of $\frac{2Zy_{\cdot n\cdot\cdot} + Z^2}{P}$.

Substituting (5.1), (5.9), and (5.10) into (5.8), we obtain the following:

$$Bias = \frac{2y_{\cdot n\cdot\cdot}\left(\frac{Py_{SUM}-3y_{....}}{(P-3)(P-1)}\right)}{P} + \frac{\left(\frac{Py_{SUM}-3y_{....}}{(P-3)(P-1)}\right)^2}{P} - \frac{\left(y_{....}+\frac{Py_{SUM}-3y_{....}}{(P-3)(P-1)}\right)^2}{P^2}$$
$$- \frac{(y_{SUM}-y_{....})^2}{(P-3)(P-1)} + \frac{\left(y_{SUM_N}-y_{....}\right)^2}{(P-2)(P-1)} - \frac{y_{....}(2y_{\cdot n\cdot\cdot}-y_{....})}{(P-1)P}.$$

The initial three terms may be reformulated as follows:

$$\frac{2(P-3)(py_{SUM}-3y_{....})y_{\cdot n\cdot\cdot} + py_{SUM}^2 - 2py_{SUM}y_{....} - \left((p-7)p+9\right)y_{....}^2}{(P-3)^2(P-1)P}$$
$$= \frac{2(P-3)y_{\cdot n\cdot\cdot}y_{SUM} + y_{SUM}^2 - 2y_{SUM}y_{....} - (p-7)y_{....}^2}{(P-3)^2(P-1)} + \frac{-6(P-3)y_{\cdot n\cdot\cdot}y_{....} - 9y_{....}^2}{(P-3)^2(P-1)P}$$
$$= (P-2) * \left[\frac{2(P-3)y_{\cdot n\cdot\cdot}y_{SUM} + y_{SUM}^2 - 2y_{SUM}y_{....} - (p-7)y_{....}^2}{(P-3)^2(P-2)(P-1)} + \frac{-6(P-3)y_{\cdot n\cdot\cdot}y_{....} - 9y_{....}^2}{(P-3)^2(P-2)(P-1)P}\right].$$

The concluding trio of terms can be rephrased as follows:

$$\frac{(P-3)Py_{\cdot n\cdot\cdot}^2 - 2(P-3)(Py_{SUM}-2y_{....})y_{\cdot n\cdot\cdot} + P^2y_{....}^2 + P(-y_{SUM}^2 + 2y_{SUM}y_{....} - 6y_{....}^2) + 6y_{....}^2}{(P-3)(P-2)(P-1)P}$$
$$= \frac{(P-3)y_{\cdot n\cdot\cdot}^2 - 2(P-3)y_{\cdot n\cdot\cdot}y_{SUM} + Py_{....}^2 + (-y_{SUM}^2 + 2y_{SUM}y_{....} - 6y_{....}^2)}{(P-3)(P-2)(P-1)} + \frac{4(P-3)y_{\cdot n\cdot\cdot}y_{....} + 6y_{....}^2}{(P-3)(P-2)(P-1)P}$$
$$= (P-3) * \left[\frac{(P-3)y_{\cdot n\cdot\cdot}^2 - 2(P-3)y_{\cdot n\cdot\cdot}y_{SUM} + Py_{....}^2 + (-y_{SUM}^2 + 2y_{SUM}y_{....} - 6y_{....}^2)}{(P-3)^2(P-2)(P-1)} + \frac{4(P-3)y_{\cdot n\cdot\cdot}y_{....} + 6y_{....}^2}{(P-3)^2(P-2)(P-1)P}\right].$$

Finally, we have

$$Bias = \frac{P^2y_{\cdot n\cdot\cdot}^2 - 6Py_{\cdot n\cdot\cdot}^2 + 9y_{\cdot n\cdot\cdot}^2 + 2Py_{\cdot n\cdot\cdot}y_{SUM} - 4Py_{\cdot n\cdot\cdot}y_{....} - 6y_{\cdot n\cdot\cdot}y_{SUM} + 12y_{\cdot n\cdot\cdot}y_{....} + y_{SUM}^2 - 2y_{SUM}y_{....} + 4y_{....}^2}{(P-3)^2(P-2)(P-1)}$$
$$+ \frac{-3y_{....}^2 + 2(p-3)y_{\cdot n\cdot\cdot}y_{....}}{(P-3)^2(P-2)(P-1)}$$
$$= \frac{(y_{....}^2 - 2y_{SUM}y_{....} + y_{SUM}^2) - 2(p-3)y_{\cdot n\cdot\cdot}y_{....} + 2(p-3)y_{\cdot n\cdot\cdot}y_{SUM} + (P-3)^2y_{\cdot n\cdot\cdot}^2}{(P-3)^2(P-2)(P-1)}$$

$$=\frac{(y_{....}-y_{SUM})^2-2(y_{....}-y_{SUM})((p-3)y_{\cdot n\cdot\cdot})+((p-3)y_{\cdot n\cdot\cdot})^2}{(P-3)^2(P-2)(P-1)},$$

$$Bias=\frac{(y_{....}-y_{SUM}-(p-3)y_{\cdot n\cdot\cdot})^2}{(P-3)^2(P-2)(P-1)}.$$

This completes the proof.

The following content presents a comparison of two methods for solving the incomplete data problem: Exact-scheme analysis of variance without data imputation and estimating missing data through the least squares method. This comparison is based on the three case study examples.

Case Study 1: A *4 × 4* GLSED adapted from Subramani [33] examined the influence of four different television assembly methods on assembly time, as shown in Table 3. In this experiment, assembly methods were represented by Latin letter treatments, workstations by Greek letters, assembly order by rows, and workers by columns.

Case Study 2: A *5 × 5* GLSED adapted from Montgomery [34] examined the influence of five different time intervals on the yield of various chemical processes, as shown in Table 4. In this study, the time intervals were represented as Latin letter treatments, catalyst concentrations as Greek letters, raw materials as rows, and acid concentrations as columns.

Case Study 3: A *7 × 7* GLSED adapted from Hinkelmann and Kempthorne [37] examined the influence of seven different levels of lysine percentages in the diet on milk production in cows, as shown in Table 5. In this study, the percentage of lysine in the diet was represented by Latin letter treatments, the percentage of protein in the diet by Greek letters, cows by rows, and duration by columns.

Table 3. Data for the Greco-Latin square design in Case Study 1.

Column Row		ψ_1	ψ_2	ψ_3	ψ_4
		1	2	3	4
θ_1	1	$C\beta$ =11	$B\gamma$ =10	$D\delta$ =14	$A \propto$=8
θ_2	2	$B \propto$=8	$C\delta$ =12	$A\gamma$ =10	$D\beta$ =12
θ_3	3	$A\delta$ =9	$D \propto$=11	$B\beta$ =missing value	$C\gamma$ =15
θ_4	4	$D\gamma$ =9	$A\beta$ =8	$C \propto$= 18	$B\delta$ =6

Table 4. Data for the Greco-Latin square design in Case Study 2.

Column Row		ψ_1	ψ_2	ψ_3	ψ_4	ψ_5
		1	2	3	4	5
θ_1	1	$A \propto$=missing value	$B\beta$ =16	$C\gamma$ =19	$D\delta$ =16	$E\varepsilon$ =13
θ_2	2	$B\gamma$ =18	$C\delta$ =21	$D\varepsilon$ =18	$E \propto$=11	$A\beta$ =21
θ_3	3	$C\varepsilon$ =20	$D \propto$=12	$E\beta$ =16	$A\gamma$ =25	$B\delta$ =13
θ_4	4	$D\beta$ =15	$E\gamma$ =15	$A\delta$ =22	$B\varepsilon$ =14	$C \propto$=17
θ_5	5	$E\delta$ =10	$A\varepsilon$ =24	$B \propto$=17	$C\beta$ =17	$D\gamma$ =14

Table 5. Data for the Greco-Latin square design in Case Study 3.

Column		ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_7
Row		1	2	3	4	5	6	7
θ_1	1	$A \propto$=304	$B\varepsilon$ =436	$C\beta$ =350	$D\phi$ =504	$E\chi$ =417	$F\gamma$ =519	$G\delta$ =432
θ_2	2	$B\beta$ =381	$C\phi$ =505	$D\chi$ =425	$E\gamma$ =564	$F\delta$ =494	$G \propto$=350	$A\varepsilon$ =413
θ_3	3	$C\chi$ =432	$D\gamma$ =566	$E\delta$ =479	$F \propto$=357	$G\varepsilon$ =461	$A\beta$ =340	$B\phi$ =502
θ_4	4	$D\delta$ =442	$E \propto$=372	$F\varepsilon$ =536	$G\beta$ =366	$A\phi$ =495	$B\chi$ =425	$C\gamma$ =507
θ_5	5	$E\varepsilon$ =496	$F\beta$ =449	$G\phi$ =493	$A\chi$ =345	$B\gamma$ =509	$C\delta$ =481	$D \propto$=380
θ_6	6	$F\phi$ =534	$G\chi$ =421	$A\gamma$ =452	$B\delta$ =427	$C \propto$=346	$D\varepsilon$ =478	$E\beta$ =397
θ_7	7	$G\gamma$ =543	$A\delta$ =386	$B \propto$=435	$C\varepsilon$ =485	$D\beta$ =406	$E\phi$ =554	$F\chi$ =missing value

Using Eq (5.9), the least squares method for estimating missing data yields values of 15, 21, and 474.38 for Case Studies 1, 2, and 3, respectively. The comparison of the Latin letter treatment sum of squares between the exact method and the least squares method is shown in Table 6.

Table 6. Comparison of Latin letter treatment sum of squares for exact and least squares methods.

Case Study	Exact-scheme analysis	Estimating missing data via least squares method	Bias
1	$SS_{Tr(unbias)\_exact}$ = 59.333	$SS_{Tr(bias)}$ = 63.50	4.167
2	$SS_{Tr(unbias)\_exact}$ = 217.467	$SS_{Tr(bias)}$ = 282.80	65.333
3	$SS_{Tr(unbias)\_exact}$ = 32,704	$SS_{Tr(bias)}$ = 34,620	1916

Table 6 shows that the Latin letter treatment sum of squares calculated using the least squares method is biased. However, this bias can be corrected using a bias adjustment formula (Eq (5.7)), or the exact method can be used for an unbiased estimate.

From the three examples, the analysis of variance results can be demonstrated without bias, where the sums of squares are calculated from Proposition 5.1, as shown in Tables 7–9.

Table 7. Unbiased analysis of variance for Case Study 1.

Source of variation	Sum of squares	Degrees of freedom	Mean square	F_0
Latin letter treatment	59.333	3	19.778	2.55
Greek letter	2.833	3	0.944	
Rows	6.500	3	2.167	
Columns	30.833	3	10.278	
Error	15.500	2	7.750	
Total	136.933	14		

Table 8. Unbiased analysis of variance for Case Study 2.

Source of variation	Sum of squares	Degrees of freedom	Mean square	F_0
Latin letter treatment	217.467	4	54.367	9.81
Greek letter	17.917	4	4.479	
Rows	6.000	4	1.500	
Columns	22.317	4	5.579	
Error	38.800	7	5.543	
Total	355.333	23		

Table 9. Unbiased analysis of variance for Case Study 3.

Source of variation	Sum of squares	Degrees of freedom	Mean square	F_0
Latin letter treatment	32,704	6	5,450.7	9.28
Greek letter	155,215	6	25,869.1	
Rows	7,412	6	1,235.4	
Columns	1,270	6	211.7	
Error	13,515	23	587.6	
Total	213,217	47		

The exact method provides an unbiased estimate of both the Latin letter treatment sum of squares and the error sum of squares (Montgomery [1]). This method employs the general regression significance test, which is equivalent to the exact-scheme analysis of variance without data imputation. A more comprehensive validation of this approach was demonstrated through the following simulation studies.

Missing values within plausible ranges, based on each case study, are assumed and generated to estimate the sums of squares for Latin letter treatment and error as shown in Figures 1 and 2 for Case Study 1, Figures 3 and 4 for Case Study 2, and Figures 5 and 6 for Case Study 3. In Case Study 1 (Figure 1), the estimated missing value of 16.98, determined using simulation studies, results in the sum of squares for Latin letter treatment aligning with the exact-scheme method. However, the sum of squares for error is biased and does not match the exact-scheme results. For Figure 2, an estimated value of 15 (or derived using the least squares method) yields a sum of squares for error consistent with the exact-scheme analysis. Nonetheless, the sum of squares for Latin letter treatment remains biased and does not align with the exact-scheme results.

In Case Studies 2 (Figures 3 and 4) and 3 (Figures 5 and 6), similar patterns are observed. There is a high probability that replacing a single missing data introduces bias either in the Latin letter treatment sum of squares or in the error sum of squares.

Therefore, it remains challenging for various missing data estimation methods to consistently produce unbiased sums of squares for both Latin letter treatment and error that correspond to those calculated by the exact-scheme analysis of variance.

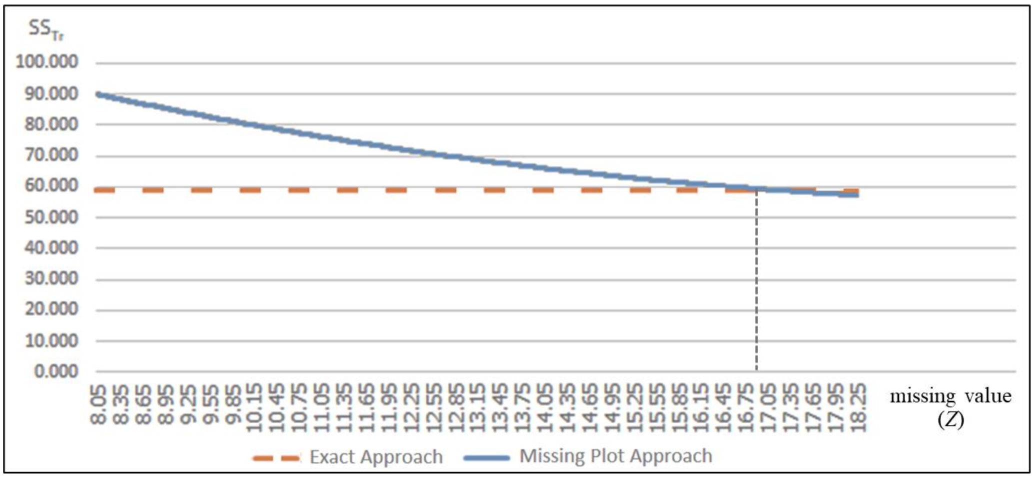

Figure 1. Trend of treatment sum of squares when simulating missing data for Case Study 1.

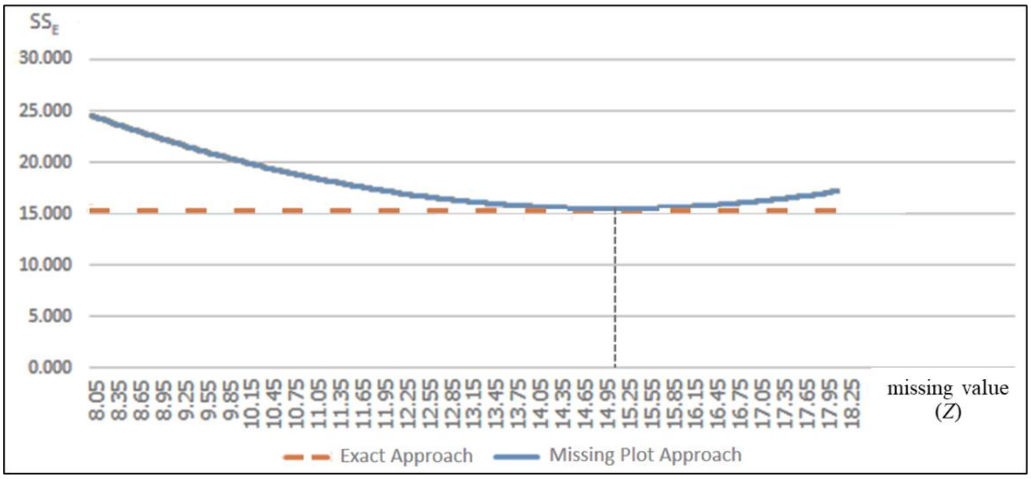

Figure 2. Trend of error sum of squares when simulating missing data for Case Study 1.

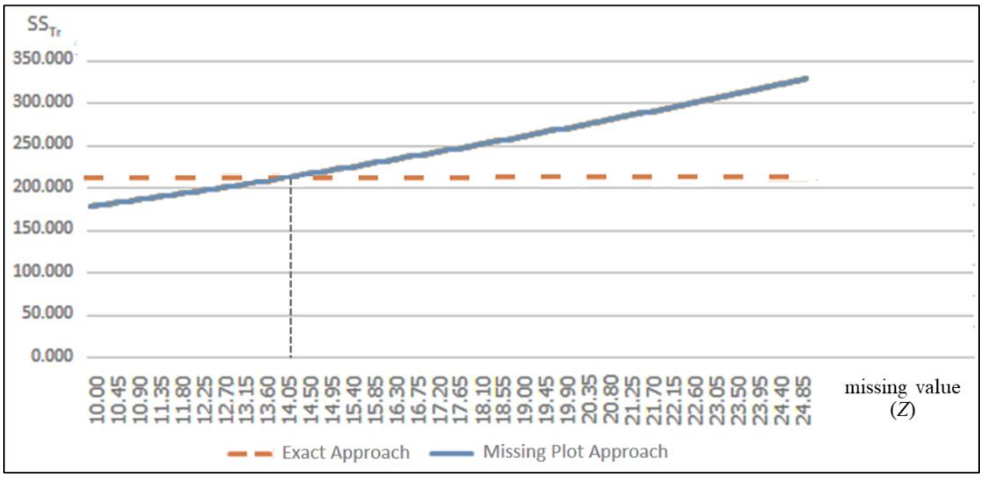

Figure 3. Trend of treatment sum of squares when simulating missing data for Case Study 2.

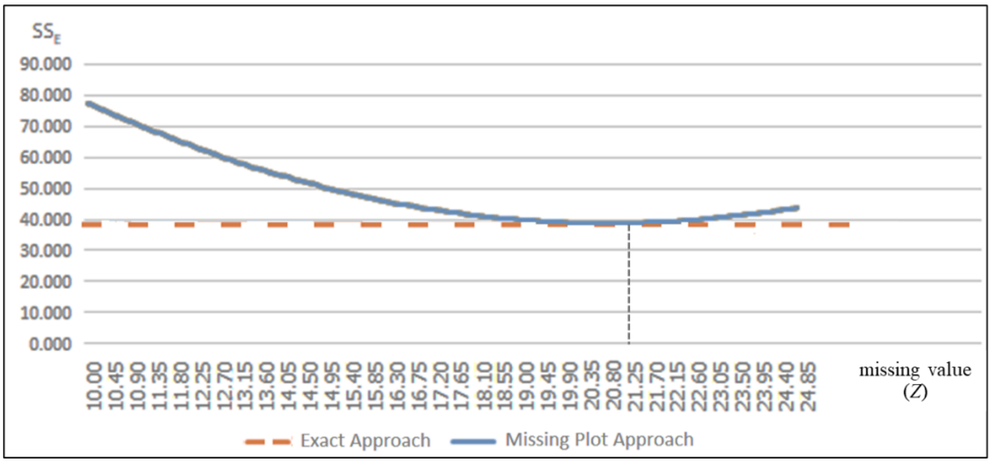

Figure 4. Trend of error sum of squares when simulating missing data for Case Study 2.

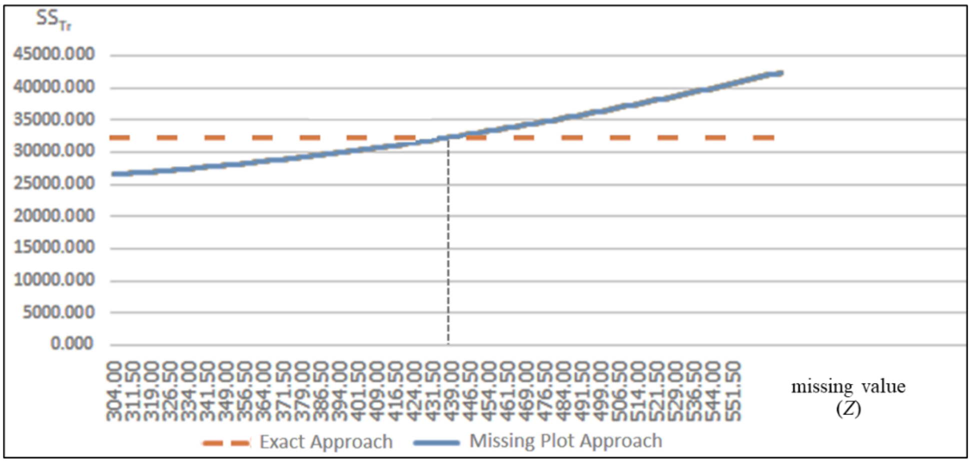

Figure 5. Trend of treatment sum of squares when simulating missing data for Case Study 3.

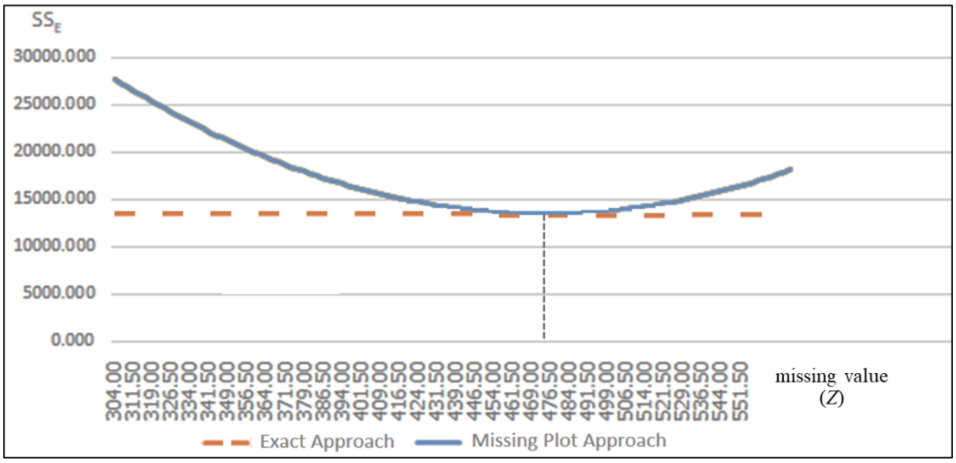

Figure 6. Trend of error sum of squares when simulating missing data for Case Study 3.

Handling multiple missing values requires developing linear algebra equations for fitted parameters in both reduced and full models, as well as recalculating regression sums of squares. Each scenario presents unique equations based on the number and pattern of missing data. We recommend this as an area for future research, as expanding the exact-scheme methodology to address multiple missing data points in GLSED necessitates comprehensive study.

6. Conclusions

In conclusion, the Greco-Latin square experimental design (GLSED) showcases substantial benefits across a variety of fields, including agriculture, medicine, and industry, affirming its effectiveness in conducting systematic tests on multiple variables within the confines of limited resources and time. Within the framework of Fisher's design of experiments (DOE), the prevalent use of the missing plot technique for estimating missing observations is acknowledged, although it is noted to introduce a positive bias in the treatment sum of squares. Addressing this limitation, the present study introduces a novel model comparison-based exact scheme specifically designed for GLSED configurations encountering a single missing observation, a methodology that is applicable regardless of the levels of treatment and blocks. This advancement not only enhances the accuracy of experimental analyses but also broadens the utility of GLSED by mitigating the inherent bias associated with traditional estimation techniques, thereby contributing a significant methodological innovation to the field of experimental design.

The introduction of the proposed exact scheme represents a significant advancement in the field of experimental design, specifically addressing the limitations inherent in GLSED with a single missing observation. This innovative approach effectively obviates the conventional necessity for estimating missing observations, thus significantly reducing the potential for bias in the treatment sum of squares. A critical gap in existing methodologies— the absence of a readily accessible exact-scheme formula tailored to GLSED scenarios with an unrecorded value—is bridged by this study through the development of an instant formula. This formula, grounded in the exact scheme, is adeptly designed for calculating the sums of squares in $P \times P$ fixed-effect GLSED configurations without necessitating the estimation of fitted parameters. Additionally, this research enriches the methodological arsenal available to researchers by providing formulas for both full-model and reduced-model regression sums of squares, facilitating the comprehensive development of sums of squares and ANOVA tables tailored to GLSED contexts with missing data. This methodological innovation not only enhances the precision of statistical analyses but also broadens the scope of experimental designs that can be accurately analyzed under conditions of incomplete data.

This paper articulates and implements a formulated bias adjustment, specifically targeting the sum of squares, after the estimation of missing data via the least squares method. Through a thorough mathematical exposition, it lays out a systematic methodology for rectifying the bias induced by missing data within the ambit of the GLSED model. It emphasizes the critical necessity for researchers and practitioners within the domains of design of experiments, analysis of variance, and experimental design to be fully aware of the limitations inherent in prevailing methods. Moreover, it highlights the unique advantages offered by the model comparison-based exact scheme. This approach not only addresses the bias in a rigorous and structured manner but also signifies a paradigm shift in how missing data are treated, moving toward a more accurate and reliable analysis that fundamentally enhances the integrity of experimental outcomes.

In random-effects models, the expected values of model parameters are zero; however, the sums of squares for random-effects models can be calculated in the same way as for fixed-effects models, yielding the same final formula as shown in Proposition 5.1. Thus, the proposed method in this study is applicable not only to random-effects models but also to mixed-effects models, as the sums of squares remain valid under these conditions. Furthermore, the Greco-Latin square experimental design (GLSED) does not account for interactions between blocks or between blocks and treatment factors, which means that the mean square formulas for random-effects models are identical to those for fixed-effects models (Freund et al. [38]). This simplifies the analysis, allowing the methodology to be applied consistently in both models without requiring further modifications.

While encountering more than one missing value is possible in well-designed experiments, it is less frequent than single missing values, as researchers typically aim to complete their experiments successfully for accurate analysis. This study focuses on the case of a single missing value due to the complexity of deriving unbiased sums of squares in the exact-scheme analysis. Expanding this methodology to accommodate multiple missing data points is a promising direction for future research, requiring the development of mathematical frameworks to handle various missing data patterns. Previous work by Sirikasemsuk and Leerojanaprapa [29], which provided formulas for unbiased treatment sums of squares in Latin square designs with two missing values, lays a strong foundation for these extensions.

Author contributions

Kittiwat Sirikasemsuk: Conceptualization, methodology, formal analysis, writing – original draft, supervision, writing – review & editing; Sirilak Wongsriya: Formal analysis, investigation, methodology, provided the example of comparison; Kanogkan Leerojanaprapa: Conceptualization, methodology, data curation, validation, writing – review & editing, provided the example of comparison. All authors have read and approved the final version of the manuscript for publication.

Acknowledgments

The authors sincerely thank Ms. Sirilak Wongsriya for her significant contributions to this research. Her 2018 master's thesis, "A Comparison of Treatment Sum of Squares for Greco-Latin Square Design with One Missing Observation," completed at King Mongkut's Institute of Technology Ladkrabang, Thailand, provided partial data and valuable insights that enhanced our study. Dr. Kittiwat Sirikasemsuk, her advisor, also offered critical guidance in developing her research. Ms. Wongsriya's dedication and analysis greatly enriched this work.

Conflict of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. D. C. Montgomery, *Design and analysis of experiments,* 10th Eds., John Wiley & Sons, 2019.
2. R. E. Kirk, *Experimental design: Procedures for the behavioral sciences,* 4th Eds., SAGE Publications Inc., 2013. https://doi.org/10.4135/9781483384733

3. R. A. Johnson, G. K. Bhattacharyya, *Statistics: Principles and methods*, 7th Eds., New Jersey: John Wiley & Sons, 2014.
4. G. Canavos, J. Koutrouvelis, *Introduction to the design & analysis of experiments*, 1st Ed., Pearson, 2008.
5. N. Diawara, A. Demuren, E. Gyuricsko, Impairment of continuous insulin delivery therapy and analysis from graeco-latin square design model, *J. Biosci. Med.*, **4** (2016), 40–51. https://doi.org/10.4236/jbm.2016.48006
6. M. R. Mahamud, D. J. Gomes, Enzymatic saccharification of sugar cane bagasse by the crude enzyme from indigenous fungi, *J. Sci. Res.*, **4** (2012), 227. https://doi.org/10.3329/jsr.v4i1.7745
7. A. G. Woodside, W. G. Pearce, Testing market segment acceptance of new designs of industrial services, *J. Prod. Innovat. Manag.*, **6** (1989), 185–201. https://doi.org/10.1111/1540-5885.630185
8. J. A. Tovar-Aguilar, P. F. Monaghan, C. A. Bryant, A. Esposito, M. Wade, O. Ruíz-Barzola, et al., Improving eye safety in citrus harvest crews through the acceptance of personal protective equipment, community-based participatory research, social marketing, and community health workers, *J. Agromedicine*, **19** (2014), 107–116. https://doi.org/10.1080/1059924x.2014.884397
9. R. Mead, S. G. Gilmour, A. Mead, *Statistical principles for the design of experiments: Applications to real experiments*, Cambridge University Press, 2012. https://doi.org/10.1017/CBO9781139020879
10. W. J. Youden, Use of incomplete block replications in estimating tobacco-mosaic virus, *Contrib. Boyce Thomps.*, **9** (1937), 41–48.
11. F. Yates, Incomplete randomized blocks, *Ann. Eugen.*, **7** (1936), 121–140. https://doi.org/10.1111/j.1469-1809.1936.tb02134.x
12. M. Ai, K. Li, S. Liu, D. K. J. Lin, Balanced incomplete Latin square designs, *J. Statist. Plann. Inference*, **143** (2013), 1575–1582. https://doi.org/10.1016/j.jspi.2013.05.001
13. R. L. Anderson, Missing-plot techniques, *Biometrics Bull.*, **2** (1946), 41–47. https://doi.org/10.2307/3001999
14. R. Rangaswamy, *A textbook of agricultural statistics*, 2nd Eds., New Age International, 2010.
15. K. Sirikasemsuk, A review on incomplete Latin square design of any order, *AIP Conf. Proc.*, **1775** (2016), 030022. https://doi.org/10.1063/1.4965142
16. R. J. A. Little, D. B. Rubin, *Statistical analysis with missing data*, 3rd Eds., John Wiley & Sons, 2019.
17. F. E. Allan, J. Wishart, A method of estimating the yield of a missing plot in field experimental work, *J. Agri. Sci.*, **20** (1930), 399–406. https://doi.org/10.1017/S0021859600006912
18. F. Yates, The analysis of replicated experiments when the field results are incomplete, *Emprie J. Exp. Agri.*, **1** (1933), 129–142.
19. J. A. Kupolusi, O. O. Ojo, One missing observation in graeco Latin square design: An approximate analysis of variance, *Amer. Based Res. J.*, **10** (2021), 1–8.
20. E. A. Cornish, The estimation of missing values in incomplete randomized block experiments, *Ann. Eugen.*, **10** (1940), 112–118. https://doi.org/10.1111/j.1469-1809.1940.tb02240.x
21. H. R. Baird, C. Y. Kramer, Analysis of variance of a balanced incomplete block design with missing observations, *J. Roy. Statist. Soc. Ser. C*, **9** (1960), 189–198. https://doi.org/10.2307/2985719
22. M. S. Bartlett, Some examples of statistical methods of research in agriculture and applied biology, *J. R. Stat. Soc.*, **4** (1937), 137–183. https://doi.org/10.2307/2983644

23. I. Coons, The analysis of covariance as a missing plot technique, *Biometrics*, **13** (1957), 387–405. https://doi.org/10.2307/2527922
24. W. G. Cochran, Analysis of covariance: Its nature and uses, *Biometrics*, **13** (1957), 261–281. https://doi.org/10.2307/2527916
25. G. N. Wilkinson, Estimation of missing values for the analysis of incomplete data, *Biometrics*, **14** (1958), 257–286. https://doi.org/10.2307/2527789
26. C. E. Ogbonnaya, E. C. Uzochukwu, Estimation of missing data in analysis of covariance: A least-squares approach, *Commun. Stat. Theory Methods*, **45** (2016), 1902–1909. https://doi.org/10.1080/03610926.2013.868000
27. M. H. Kutner, C. J. Nachtsheim, J. Neter, W. Li, *Applied linear statistical models*, 5th Eds., New York: McGraw-Hill Irwin, 2005.
28. G. P. Quinn, M. J. Keough, *Experimental design and data analysis for biologists*, 1st Ed., Cambridge University Press, 2002. https://doi.org/10.1017/CBO9780511806384
29. K. Sirikasemsuk, K. Leerojanaprapa, S. Sirikasemsuk, Regression sum of squares of randomized complete block design with one unrecorded observation, *AIP Conf. Proc.*, **2016** (2018), 020136. https://doi.org/10.1063/1.5055538
30. K. Sirikasemsuk, K. Leerojanaprapa, Analysis of two-missing-observation *4 × 4* Latin squares using the exact approach, In: *Recent advances in information and communication technology 2017*, Cham: Springer, **566** (2018), 69–81. https://doi.org/10.1007/978-3-319-60663-7_7
31. K. Sirikasemsuk, One missing value problem in Latin square design of any order: Regression sum of squares, In: *2016 Joint 8th international conference on soft computing and intelligent systems (SCIS) and 17th international symposium on advanced intelligent systems (ISIS)*, Japan: IEEE, 2016, 142–147. https://doi.org/10.1109/SCIS-ISIS.2016.0041
32. K. Sirikasemsuk, K. Leerojanaprapa, One missing value problem in Latin square design of any order: Exact analysis of variance, *Cogent Eng.*, **4** (2017), 1411222. https://doi.org/10.1080/23311916.2017.1411222
33. J. Subramani, Non-iterative least squares estimation of missing values in graeco-Latin square designs, *Biometrical J.*, **33** (1991), 763–769. https://doi.org/10.1002/bimj.4710330619
34. D. C. Montgomery, *Design and analysis of experiments*, John Wiley & Sons, 1984.
35. R. Ott, M. Longnecker, *An introduction to statistical methods and data analysis*, 7th Eds., Cengage Learning, 2021.
36. A. AlAita, M. Aslam, K. Al Sultan, M. Saleem, Analysis of graeco-latin square designs in the presence of uncertain data, *J. Big Data*, **11** (2024), 109. https://doi.org/10.1186/s40537-024-00970-1
37. K. Hinkelmann, O. Kempthorne, *Design and analysis of experiments: Introduction to experimental design*, John Wiley & Sons, 2007.
38. R. J. Freund, W. J. Wilson, D. L. Mohr, *Statistical methods, student solutions manual (e-only)*, Academic Press, 2010. Available from: http://www.sars-expertcom.gov.hk/english/reports/reports.html