Menemui Matematik
(Discovering Mathematics)

journal homepage: https://persama.org.my/dismath/home

Robust Modifications of Variance Homogeneity Tests in Two Factor Factorial Design: Applications in Response Surface Methodology (RSM)

Liao Penghui[1], Nazihah Mohamed Ali[2*] and Xue Haoying[2]

[1,2,3]*Department of Mathematics and Statistics, Faculty of Science, University Putra Malaysia, 43400 UPM Serdang, Selangor,*

[1]gs66360@student.upm.edu.my, [2]nazihanma@upm.edu.my, [3]gs66601@student.upm.edu.my
*Corresponding author

Received: 25 October 2025
Accepted: 10 December 2025

ABSTRACT

This study comprehensively evaluates the robustness of the Levene and O'Brien tests for homogeneity of variances, along with their robust modifications under different conditions of non-normal data distributions, heterogenous variance ratios, and varied experimental designs. The primary objective of this study is to provide clear guidance on selecting the most robust test when classical assumptions are violated, with a key contribution being the evaluation of modern robust modification against establish standards. The investigated tests include the Levene test modified with the modified one-step M-estimator and O'Brien's procedures modified with Winsorized variance and with Winsorized mean and variance. Extensive simulation studies demonstrates that while the original tests perform well under homogeneity, their robustness deteriorates with increasing variance heterogeneity or data skewness. For instance, in three group scenarios with a variance ratio of 1:1:4, meaning the population variance of the third group is four times larger than the that of the two groups the Levene test modified with modified one-set M-estimator demonstrates superior Type I error compared to the original Levene test. Furthermore, in more extreme four group scenario with the variance ratio of 1:1:1:9, where one group's variance is nine time larger than the others, the O'Brien procedure modified with Winsorized mean and variance emerges as the most robust method. The practical application of these findings is illustrated through an analysis of the real-world dataset, concretely highlighting the relative strengths and limitations of each method.

Keywords: Two-factor factorial design, MOM, MOM-H, Winsorized, Robustness.

INTRODUCTION

Testing for the equality of variances (homoscedasticity), is a critical concern in a statistical modelling, as it is a fundamental assumption underlying the validity of standard ANOVA and regression. Violations of this assumption, termed variance heterogeneity can severely inflate Type I error rates and undermine the reliability of statistical conclusions (Md Yusof et al., 2012a). When this assumption is violated, researchers are faced with a choice of traditional remedies. These include using nonparametric tests which often require large samples for adequate power (Wilcox & Keselman, 2003b) or variance stabilizing transformation (Djalic, I., & Terzić, S. 2021). However, these solutions often represent a compromise, sacrificing information, or interpretability,

particularly in the small sample scenarios common in practical research. This driven the development of robust tests design to reliably assess homoscedasticity itself.

Among test for homoscedasticity, Levene's test (1960) remains a cornerstone due to its robustness to non-normality. However, its well documented sensitivity to extreme skewness and kurtosis in small samples has spurred the development of robust alternatives like Brown-Forsythe test (Kulaksiz & Noyan, 2019; Öztuna, 2022). Similarly, O'Brien's (1984) procedure offers a powerful rank-based approach for multivariate settings but can exhibit inflated Type I error when its covariance assumptions are violated (Tilley & Woolson, 2005).

To directly address the issue outliers, which worsen the problems above, robust estimators have been developed. Technique like Winsorization preserve data integrity but introduce subjectivity in threshold selection (Mulry et al., 2016; Karakulak & Ergül, 2021)). In contrast, the modified one-step M-estimator (MOM) offers more adaptive trimming to improve Type I error control (Wilcox, 2012). Further extensions, such as the MOM-H statistic integrate M-estimators with robust tests but can suffer from suboptimal power and inconsistent error rates under skewed distributions (Othman et al., 2004; Alenazi, 2023).

Critically, a significant gap persists in the extension and integration of these robust method to complex experimental designs. While the aforementioned techniques are developed and evaluated primarily in single factor context, modern research often relies on multi-factorial designs for example two-way ANOVA (Bansal & Goyal, 2024). In these designs, interaction effects and unbalanced data compound the challenge of variance heterogeneity, and traditional robust method are not equipped to handle them effectively. Recent simulation studies continue to confirm the sensitivity of common tests to variance heterogeneity in complex design, underscoring the need for more robust solution (Garcia et al., 2022). There is pressing need for unified statistical approaches that can provide robust inference in multifactorial framework without sacrificing statistical power.

This study directly addressed this need by proposing and evaluating novel modifications of Levene and O'Briens tests, integrating them with advanced robust estimators by including Winsorized variance, MOM and MOM-H within permutation-based inference framework. The key advantage of our proposed methods is their ability to simultaneously handle the dual challenges of non-normality and variance heterogeneity in two factor factorial design with small samples. By leveraging permutation tests, we avoid restrictive parametric assumptions, ensuring better control of Type I error rates, an approach whose utility for robust variance testing is gaining renewed attention (Zygmont & Smith, 2014). Furthermore, our integration of adaptive estimators like MOM aims to maintain high statistical power where traditional methods fail. Through comprehensive simulation and real-data analysis, we demonstrate that our proposed framework offers a more reliable and powerful solution for assessing homoscedasticity in the complex, real-world faced by today's researchers.

METHODOLOGY

This study evaluates the robustness of test for homogeneity of variances within the framework of two-factor factorial design with fixed effects. This section details the statistical model, the test under investigation, and the analytical approach. Specifically, we evaluate and compare a range of test, including the classical Levene test, a MOM-based Levene test utilizing MOM-H statistic, the classical O'Brien test and modified O'Brien's test based on Winsorized mean and variance.

Two-Factor Factorial Design with Fixed Effect Model

The two-factor factorial design with fixed effect model is a cornerstone of experimental design, allowing for the investigation of the individual (main) and joint (interaction) effects of two categorical factors. The model is formally defined as:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \quad (1)$$

where α_i is the effect of factor A, β_j is the effect of factor B, $(\alpha\beta)_{ij}$ is the effect of factor A and factor B. The test statistics uses F test, which is the ratio of two variances. If the test statistics greater than critical value, the factor or interaction is significant.

Test for Homogeneity of Variance

This study evaluates the performance of several tests for accessing homoscedasticity in the two-factor factorial model specified by Equation (1). The formal hypotheses for these tests are:

H_0: All variances are equal ($\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$)
H_1: Not all group variances are equal

where k represent the number of different groups. Robust modification is required due to the sensitivity of traditional tests to non-normality and variance heterogeneity in two-factor design.

Modified Levene Test with Modified One-Step M-Estimator (MOM)

Levene's test assesses homoscedasticity by performing ANOVA on the absolute deviations of observations from their group means. The test statistics is:

$$W = \frac{N-k}{k-1} \times \frac{\sum_{i=1}^{k} N_i (Z_{i\cdot} - Z_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - Z_{i\cdot})^2} \quad (2)$$

where
k = number of different groups
N = number of total samples in all groups
N_i = number of samples in i^{th} group
$Z_{ij} = \left| Y_{ij} - \bar{Y}_{i\cdot} \right|$, $\bar{Y}_{i\cdot}$ is the mean of i^{th} group
$Z_{i\cdot} = \frac{1}{N_i} \sum_{j=1}^{N_i} Z_{ij}$.
$Z_{\cdot\cdot} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{N_i} Z_{ij}$.

$Z_{i\cdot}$ are the mean of Z_{ij} for i^{th} group and $Z_{\cdot\cdot}$ mean for the overall. The test statistic W is approximately distributed as F distribution with $(k-1,\ n-k)$ degree of freedom.

The Levene test enhances robustness against outliers by replacing the group mean, $\bar{Y}_{i.}$ in the deviation calculation with Modified one-step M-estimator (MOM), $\hat{\theta}_j$. The test statistic is then computed as in Equation (3).

$$W_{MOM} = \frac{N-k}{k-1} \times \frac{\sum_{i=1}^{k} N_i (Z_{i.} - Z_{..})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (|Y_{ij} - \hat{\theta}_j| - Z_{i.})^2} \quad (3)$$

where

$\hat{\theta}_j = \frac{1.28 MAD_{nj}(i_2 - i_1) + \sum_{i=i_1+1}^{n_j - i_2} Y_{(i)j}}{n_j - i_1 - i_2}$

$MAD_{nj} = \frac{MAD_j}{0.6745}$

$MAD_j = \text{median}\left|y_{ij} - \hat{M}_j\right|$

$\hat{M}_j =$ median of group j

$i_1 =$ the number of observations Y_{ij} such that $(Y_{ij} - \hat{M}_j) < -2.24(MAD_{nj})$

$i_2 =$ the number of observations Y_{ij} such that $(Y_{ij} - \hat{M}_j) > 2.24(MAD_{nj})$.

This study employs the H-statistic, introduced by Schrader and Hettmansperger (1980), a test statistic designed for use with any measure of central tendency. The corresponding MOM-H test is robust against outliers and non-normal distribution, leading to more reliable parameter estimation. Its H-statistics, is defined as:

$$H = \frac{1}{N} \sum_{j=1}^{j} n_j (\hat{\theta}_j - \hat{\theta}_.)^2 \quad (4)$$

where $N = \sum_{j=1}^{J} n_j$, total sample size and $\hat{\theta}_. = \frac{1}{J} \sum_{j=1}^{j} \hat{\theta}_j$, the average of $\hat{\theta}_j$, the average of the group MOM estimators.

Modified O'Brien Test with Winsorized Mean and Winsorized Variance

This study utilizes the classical O'Brien test, a global statistical hypothesis testing method designed to detect the overall differential effect between two groups or more groups across a set of correlated endpoints. The transformation parameter ω was specified as 0.05. The resulting transformed values, r_{ijk} are generated by using the equation:

$$r_{ijk} = \frac{(n_{ij} - 1.5) n_{ij} (y_{ijk} - \bar{y}_{ij})^2 - 0.5 s_i^2 (n_{ij} - 1)}{(n_{ij} - 1)(n_{ij} - 2)} \quad (5)$$

where $\bar{y}_{ij} = \frac{\sum y_{ijk}}{n_{ij}}$ and $s_i^2 = \frac{\sum (y_{ijk} - \bar{y}_{ij})^2}{(n_{ij} - 1)}$ are the mean and variance for i^{th} group, respectively.

The standard O'Brien test can be made more robust to outliers by incorporating Winsorized estimators. Using this approach, the original sample mean and variance are replaced by their Winsorized counterparts. The Winsorized mean is a robust measure of central tendency that mitigates the influence of extreme values by replacing the tails of the distribution. The Winsorized method can reduce the weight of outliers in the tail, which replaces the minimum and maximum observation with values close to the centre with a percentage$(\alpha\ \%)$. In this study we use $\alpha = 0.025$. Specifically, the smallest k observations are replaced by the value of the $(k+1)^{\text{th}}$ smallest

observation, and the largest k observations are replaced by the value of the $(n-k)^{\text{th}}$ largest observation as given in Equation (6) and (7). The mean is then calculated on this modified dataset. It is formally defined as:

$$\bar{y}_w(\alpha) = \frac{(k+1)y_{k+1}+y_{k+2}+\cdots+y_{n-k-1}+(k+1)y_{n-k}}{n} \quad (6)$$

where α is Winsorization ratio and $k = \alpha n$ represent the number of observations replaced at each end of the ordered sample.

To maintain consistency in the test statistic, the sample variance must also be calculated from the same Winsorized dataset. The Winsorized variance provides a more stable estimate of scale in the presence of outliers and is computed as:

$$s_w{}^2(\alpha) = \frac{\begin{matrix}(k+1)[y_{k+1}-(\bar{y}_w(\alpha))]^2+[y_{k+2}-(\bar{y}_w(\alpha))]^2+\cdots \\ +[y_{n-k-1}-(\bar{y}_w(\alpha))]^2+(k+1)[y_{n-k}-(\bar{y}_w(\alpha))]^2\end{matrix}}{n-1} \quad (7)$$

In robust O'Brien test, the transformed data values, r_{ijk} are calculated using Equation (5), but with the original mean, $\bar{y}_{ij}$ and variance, $s_i{}^2$ substituted by their robust equivalents, the Winsorized mean, $\bar{y}_w(\alpha)$ and Winsorized variance, $s_w{}^2(\alpha)$ respectively.

Response Surface Methodology (RSM)

Response Surface Methodology (RSM) is a collection of statistical and mathematical technique used for developing, improving and optimizing processes. Its core objective is to model and analyze the relationship between the independent variables (factors) and a response variable (output) of interest (Myers, Montgomery, & Anderson-Cook 2016). Since the true functional relationship between the factors and the response is typically complex and unknown, RSM employs a sequence of lower-order polynomial models to approximate the response surface within a specific region of interest. This empirical model-building approach is a fundamental principle of Response Surface Methodology (Box & Wilson, 1951).

The initial model used, especially when the data is collected from factorial design, is a first-order model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_q x_{iq} + \varepsilon_i \ \ (i = 1,2,\ldots,N) \quad (8)$$

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i \, x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i<j}^{k} \beta_{ij} x_i x_j + \varepsilon \quad (9)$$

These models are fitted using least squares, and the significance of terms is assessed to understand the factor effects and optimize the response.

SIMULATION STUDY

Definition of Type I Error

A type I error is the incorrect rejection of a true null hypothesis or called as a "false positive". In context of this study, it occurs when a test of homogeneity of variance incorrectly concludes that the group are significantly different when, in fact they are equal. The probability of committing

Type I error is denoted by α, which is the significance level of the test (typically set at 0.05). A test is considered robust if its empirical Type I error rate remains close to this nominal α level under various conditions of non-normality and sample size.

Simulation design

This simulation study evaluated the robustness of six tests for homoscedasticity by comparing their Type I error rates. The tests included the classical Levene test, classical O'Brien test, along with their respective modified tests in equation (3), (4), (6) and (7).

A two-factor factorial design was employed, with two specific layouts, that is a 3×3 design (Factor A: 3 levels, Factor B: 3 levels, resulting in 9 cells) and a 4×4 design (16 cells). The total sample size was fixed at $N = 36$ for 3×3 design and $N = 64$ for 4×4 design, ensuring a constant replication of $n = 4$ observation per cell in both configurations.

Two experimental factors were manipulated to access test performance under adverse conditions:

1. Variance Heterogeneity: The degree of variance across groups.
2. Distribution Shape: The form of the data distribution, including non-normal shapes.

The specific conditions of these factors are detailed in Table 1 and 2.

Table 1: Factors and Levels of Variance Conditions

	Levels of Design		**Description**
	3×3	4×4	
Ratio of Variance	(1:1:1)	(1:1:1:1)	Equal variances: Establish a baseline for Type I error under the null hypothesis.
	(1:1:4)	(1:1:1:4)	Moderate Heterogeneity: Test performance under a common level of violation
	(1:1:9)	(1:1:1:9)	Extreme Heterogeneity: Stresses the test to evaluate their breaking point and robustness limit.

Table 2: Distribution Used in the Simulation Study

Types of distributions	Normal distribution	Normal distribution as a baseline for optimal performance
	Chi-square distribution	Represent moderate right-skewness and kurtosis
	Exponential distribution	Represent a highly skewed and kurtotic distribution, challenging the test's robustness to severe non-normality.

Test robustness was assessed based on the Type I error rate near the nominal α level of 0.05. Following Bradley's (1978) liberal criterion, a test was considered robust if its empirical Type I error rate fell within the interval [0.025, 0.075].

SIMULATION RESULT

From Table 3, we can see the traditional Levene's test, its Type I error rates under equal variances (1:1:1) are 0.051 at Level 3 and 0.057 at Level 4, both falling within Bradley's robust range (approximately from 0.025 to 0.075). This indicates that Levene's test performs acceptably when variances are equal. However, as the variance ratio increases (e.g., 1:1:4 and especially 1:1:9), its Type I error rates drop significantly. For instance, under a variance ratio of 1:1:9, the error rates fall to 0.009 at Level 3 and 0.003 at Level 4, which are far below the robust range. This sharp decline suggests that Levene's test loses robustness under conditions of high variance heterogeneity, potentially leading to reduced statistical power.

Second, for the modified Levene-MOM test, its performance under equal variances is characterized by Type I error rates of 0.026 at Level 3 and 0.011 at Level 4. While the result for Level 3 is close to the lower bound of the robust range, the result for Level 4 is clearly below it, indicating overly conservative behavior in such scenarios. More critically, as the variance ratio increases (e.g., 1:1:9), Type I error rates rise sharply to 0.26 at Level 3 and 0.345 at Level 4, far exceeding the robust range. This pattern demonstrates that Levene-MOM also loses stability when variances become highly unequal. Notably, the Levene-MOM-H variant consistently produces a Type I error rate of zero across all tested conditions, suggesting that this approach fails to function effectively in practice.

Third, the O'Brien test performs well under equal variances (1:1:1), with Type I error rates for Factor A, Factor B, and their interaction (AB) ranging from approximately 0.026 to 0.031 across Levels 3 and 4 with all within Bradley's robust range. However, as variance heterogeneity intensifies (e.g., a ratio of 1:1:9), its robustness deteriorates significantly. For example, at Level 3 under this extreme condition, the Type I error rate for Factor A surges to 0.6482 which is well above acceptable levels, indicating substantial anticonservatism.

For the Winsorized O'Brien test, its performance under equal variances is similarly satisfactory; for instance, at Level 3 with a variance ratio of 1:1:1, the Type I error rate for Factor A is approximately 0.069. However, as variance heterogeneity increases (e.g., a ratio of 1:1:9), its Type I error rate rises dramatically to as high as 0.891 for Factor A at Level 3 that is far exceeding robust thresholds and reflecting compromised stability under extreme inequality conditions. Nonetheless, the Winsorized correction method demonstrates stronger adaptability in some extreme scenarios. For example, in a four-group design with a variance ratio of 1:1:1:9 at Level 4, the Type I error rate for AB is observed to be 0.053, closer to the nominal significance level of 0.05 than that of the standard O'Brien test under similar conditions.

Table 3: Simulation Results

	Variance Ratio	Levene	Levene-MOM	O'Brien			O'Brien Winsorized Variance			O'Brien Winsorized Mean and Variance		
				A	B	AB	A	B	AB	A	B	AB
Normal Level=3	1:1:1	**0.051**	**0.026**	**0.026**	**0.027**	**0.026**	**0.069**	**0.07**	**0.067**	0.014	0.013	0.014
	1:1:4	0.023	0.109	0.322	**0.034**	**0.046**	0.571	**0.068**	0.09	0.249	0.021	**0.032**
	1:1:9	0.01	0.259	0.648	**0.04**	**0.056**	0.891	**0.066**	0.099	0.536	**0.029**	**0.046**
Level=4	1:1:1:1	**0.057**	0.011	**0.03**	**0.032**	**0.026**	0.093	0.097	0.114	0.012	0.014	0.01
	1:1:1:4	0.019	0.109	0.564	**0.039**	**0.054**	0.093	0.097	0.114	0.446	0.019	**0.034**
	1:1:1:9	0.003	0.346	0.912	**0.044**	**0.073**	0.983	**0.08**	0.157	0.82	**0.027**	**0.053**

Chi-Squ Level=3	1:1:1	**0.026**	0.017	**0.0583**	**0.059**	**0.066**	0.136	0.138	0.153	0.023	0.024	**0.029**
	1:1:4	0.015	**0.066**	0.263	**0.055**	**0.065**	0.474	0.11	0.132	0.154	**0.027**	**0.036**
	1:1:9	0.008	0.199	0.559	**0.05**	**0.066**	0.799	0.086	0.117	0.404	**0.03**	**0.043**
Level=4	1:1:1:1	**0.067**	**0.026**	**0.074**	**0.074**	0.088	0.207	0.19	0.387	0.017	0.015	0.019
	1:1:1:4	**0.027**	0.103	0.398	**0.067**	0.086	0.622	0.158	0.278	0.186	0.021	**0.027**
	1:1:1:9	0.007	0.324	0.796	**0.053**	0.08	0.928	0.11	0.202	0.567	**0.027**	**0.044**
Exp Level=3	1:1:1	**0.034**	**0.026**	**0.052**	**0.053**	**0.058**	0.122	0.118	0.138	0.024	**0.026**	**0.027**
	1:1:4	0.01	**0.068**	0.205	**0.056**	**0.069**	0.457	0.118	0.140	0.107	**0.032**	**0.039**
	1:1:9	0.005	0.1	0.253	**0.056**	**0.071**	0.554	0.117	0.142	0.141	**0.032**	**0.041**
Level=4	1:1:1:1	**0.06**	**0.051**	**0.056**	**0.059**	**0.068**	0.169	0.168	0.291	0.017	0.017	0.021
	1:1:1:4	0.015	0.091	0.231	**0.059**	0.081	0.545	0.164	0.283	0.089	0.021	**0.032**
	1:1:1:9	0.008	0.116	0.273	**0.06**	0.082	0.627	0.162	0.284	0.110	0.021	**0.032**

Real data

The data utilized in this analysis are drawn from a 2009 study conducted by Espinós, Fernández-Abascal, and Ovejero, which investigated differences in nonverbal sensitivity, by using MiniPONS test across psychological health and age group. This dataset is ideal for validating our robust homoscedasticity tests because it represents a multifactorial design with potential for variance heterogeneity between the defined groups. The study employed a two-factor design with 3 level Factor A (Psychological Group) and 3 level of Factor B (Age Group) as shown in Table 4.

Table 4: Level of Factors in Real Data

Factor A	Bipolar Disorder (BD) Unipolar Depression (UD) Control Group
Factor B	Age of 20 - 39 Age of 40 - 55 Age of 56 – 70

We first assessed the MiniPONS score for suitability for parametric analysis. The descriptive statistics and assumption test are summarized below:

Table 5: Descriptive of Real Data

Real data	Skewness	Kurtosis	Shapiro-Wilk test	Durbin-Watson test
Value	0.6251	3.68	0.1597	0.1337

Real data shows the skewness of 0.6251 and kurtosis of 3.68 values indicate a unimodal distribution that is slightly right-skewed and more peaked (leptokurtic) than normal distribution. The non-significant Shapiro Wilk test statistic (p-value = 0.1597) suggest that the deviation from normality is not statistically severe. However, the leptokurtic nature suggests a higher propensity of outlier, which aligns with the goal of evaluating robust statistical methods. The Durbin Watson statistic value near 0.13 is low and would typically indicate positive autocorrelation; however, this test is not appropriate for data structured by factorial groups rather than sequential time. Therefore, the independence assumption is better assessed by the study's design, which utilized independent participant groups.

Table 6 shows the result of homogeneity of variances of tests. The presence of extreme values or outliers has a notable impact, the O'Brien test (including its correction method) tends to maintain its nominal Type I error rate, as evidenced by *p*-values generally remaining above 0.05. This indicates that the test is robust under such adverse conditions. In contrast, the Levene test and its correction are more sensitive to these extreme observations, often producing significant *p*-values (e.g., 0.0045 for Levene, 0.0022 for Levene-MOM) that may falsely indicate variance differences.

This assessment pattern suggests that the O'Brien test and its corrections offer a more reliable assessment of homogeneity of variance when outliers are present, aligning with previous simulation studies that have highlighted its robustness in the face of non-normality and skewed data distributions.

Table 6: *p*-value of Homogeneity Test

Test	*p*-value		
Levene	0.0045*		
Levene-MOM	0.0777		
Levene-MOM-H	0.0022*		
	A	B	AB
O'Brien	0.0674	0.4709	0.3705
O'Brien Winsorized Variance	0.0188*	0.3140	0.1923
O'Brien Winsorized Mean and Variance	0.0852	0.5046	0.4621

Table 7 shows the ANOVA and response surfaced method (RSM) results. Based on the ANOVA results, the effects of factor A and the interaction terms are statistically significant; therefore, the analysis was extended by incorporating RSM.

However, the validity of RSM, which is based on ordinary least square (OLS) regression. This relies on several key assumption, including homoscedasticity (constant variance of errors). A violation of this assumption, can severely undermine the analysis by producing biased an inefficient estimate of the model coefficients, affecting their precision. Furthermore, compromising the reliability of significant tests (*p*-values) for model terms, potentially leading to incorrect conclusions about which factors are important. Hence, resulting in misleading model predictions and unreliable optimizing process, as the error variance is not stable across the design space. This justifies the need for our robust approach that integrates a check of this critical assumption within the modelling process.

A first-order RSM model was fitted to the data, resulting the following equation:

$$Y = 34.806 + 4.625X_A - 1.5X_B.$$

The model shows that factor A is the main driving force, as changes in its level have a considerable positive impact on the system response, while factor B only makes a small contribution when considered alone. However, recognizing that factor B may still affect the system through interaction with factor A, and the R-square of the first-order model is 0.3871, which is a poor explanation of the model, a second-order RSM model is constructed to more fully capture the potential curvature and interaction effects. The model is represented as:

$$Y = 57.19 - 9.3X_A - 11.96X_B - 1.96X_A^2 + 1.08X_B^2 + 3.1X_AX_B.$$

Table 7: -value of ANOVA and RSM

Test	p value				
	A	B	AB	A^2	B^2
ANOVA	0.0004**	0.2008	0.004*		
Modified ANOVA	0.418	0.880	0.349		
RSM-First Order	0.00013***	0.1684			
RSM-Second Order	0.21	0.11	0.53	0.017*	0.26

The second-order analysis confirms that factor A is still the main influencing factor on the response, and its influence shows obvious nonlinearity. Although the independent contribution of factor B is small, its interaction with factor A cannot be ignored and may lead to a local optimum in the response surface. The second-order model indicate that the optimal system performance was achieve by maintaining factor A at an intermediate level (approximately A2) and factor B at a low level (approximately B1). This result offers a definitive for performance optimization under the experimental conditions.

CONCLUSION

This study systematically evaluates the performance of the robustness of several homogeneity of variance under various condition prevalent in real world research. The result provides critical insight into their statistical efficiency and practical utility for applied researchers.

Levene-MOM-H fails to provide meaningful result across most scenarios, rendering it unsuitable for practical application. Our results indicate that while Levene's test is robust under homoscedastic condition across all distribution, its Type I error rate inflates substantially with increasing variance ratios and distributional skewness. The Levene-MOM test provided improved error control when variances were equal but exhibit a strong tendency towards conversation under heteroscedasticity. Conversely, the Levene-MOM-H test proved ineffective, performing poorly across the majority of simulated conditions.

While the original O'Brien test was effective under the ideal assumption of equal variances, its performance deteriorated markedly with increasing variance heterogeneity and skewness. Replacing the variance estimator with Winsorized version (O'Brien Winsorized Variance) substantially improved robustness for moderately unequal variances. The most robust variant, which Winsorized both the mean and variance (O'Brien Winsorized Mean and Variance) successfully handled extreme condition but was less powerful than its counterpart in several instances. An analysis of real-world MiniPONS data provides critical empirical support for our simulation study. We confirm that the MOM correction is a n effective solution for variance heterogeneity, outperforming the ineffective Levene-MOM-H test. Notably, the O'Brien test and its corrections show higher robustness in handling real-world complexities, particularly when outliers or extreme values are present.

Through RSM analysis, Factor A (mental health diagnosis) is identified as the primary driver of system response, with its level changes having significant nonlinear effects. Factor B (age group) contributes less independently but interacts meaningfully with Factor A to create localized optimization effects. The second-order RSM model identifies optimal conditions for system performance when Factor A is at an intermediate level (A2) and Factor B is at a lower level (B1).

REFERENCES

Alenazi, L. H. (2025). Artificial Intelligence in Nursing Education: A Cross-sectional UTAUT Analysis Study. Journal of Clinical & Diagnostic Research, **19(1)**.

Gastwirth, J. L., Gel, Y. R., & Miao, W. (2009). The impact of Levene's test of equality of variances on statistical theory and practice. Statistical Science, **24(3)**: 343-360.

Glass, G. V. (1966). Testing homogeneity of variances. American Educational Research Journal, 3(3), 187-190.

Goyal, V., Asati, A. K., & Arora, A. (2023). An experimental and modelling study for a novel bank-type earth air heat exchanger for the summer season using full factorial design. *Journal of Thermal Science and Engineering Applications*, **15(2)**: 021006.

Hernández-García, Y., Melgar-Lalanne, G., Téllez-Medina, D. I., Ruiz-May, E., Salgado-Cruz, M. D. L. P., Andrade-Velásquez, A., & Santiago Gomez, M. P. (2022). Scavenging peptides, antioxidant activity, and hypoglycemic activity of a germinated amaranth (Amaranthus hypochondriacus L.) beverage fermented by Lactiplantibacillus plantarum. *Journal of food biochemistry*, **46(7)**: e14139.

Huang, P., Woolson, R. F., & O'Brien, P. C. (2008). A rank-based sample size method for multiple outcomes in clinical trials. *Statistics in medicine*, **27(16)**: 3084-3104.

Karakulak, F. S., Uzer, U., Kabasakal, H., & Namoğlu, İ. B. (2024). Confirmation of the presence of Helicolenus dactylopterus (Delaroche, 1809), in the Sea of Marmara with morphometrical and bioecological notes. *Ege Journal of Fisheries & Aquatic Sciences (EgeJFAS)/Su Ürünleri Dergisi*, **41(4)**.

Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. N. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. Psychological methods, **13(2)**: 110.

Kumar, P. (2020). Response Surface Methodology -A Review. International Journal for Scientific Research & Development.

Luh, W. M., & Guo, J. H. (2004). Improved robust test statistic based on trimmed means and Hall's transformation for two-way ANOVA models under non-normality. Journal of Applied Statistics, **31(6)**: 623-643.

Martinoz, C.F., Haziza, D., & Beaumont, J. (2015). A method of determining the Winsorization threshold, with an application to domain.

Melik, H. N., Ahad, N. A., & Yahaya, S. S. S. (2018). Modified One-Step M-Estimator with Robust Scale Estimator for Multivariate Data. Journal of Engineering and Applied Sciences, **13(24)**: 10396-10400.

Wilcox, R. R., & Keselman, H. J. (2003). Repeated measures one-way ANOVA based on a modified one-step M-estimator. British Journal of Mathematical and Statistical Psychology, **56(1)**: 15-25.

Wooldridge, J. M. (2001). Applications of generalized method of moments estimation. Journal of Economic perspectives, **15(4)**: 87-100

Xao, O. G., Yahaya, S. S. S., Abdullah, S., & Yusof, Z. M. (2014, July). H-statistic with Winsorized modified one-step M-estimator for two independent groups design. In AIP Conference Proceedings (Vol. 1605, No. 1, pp. 928-931). American Institute of Physics.

Yi, Z., Chen, Y. H., Yin, Y., Cheng, K., Wang, Y., Nguyen, D., ... & Kim, E. (2022). Brief research report: A comparison of robust tests for homogeneity of variance in factorial ANOVA. The Journal of Experimental Education, **90(2)**: 505-520.

Yusof, Z. M., Othman, A. R., & Yahaya, S. S. (2010). Comparison of type I error rates between T1 and Ft statistics for unequal population variance using variable trimming. Malaysian Journal of Mathematical Sciences.

Zhang, L., Liu, J., Shen, X., Li, S., Li, W., & Xiao, X. (2023). Response Surfaces Method and Artificial Intelligence Approaches for Modelling the Effects of Environmental Factors on Chlorophyll a in Isochrysis galbana. *Microorganisms*, **11(8)**: 1875.

Zygmont, C., & Smith, M. R. (2014). Robust factor analysis in the presence of normality violations, missing data, and outliers: Empirical questions and possible solutions. *The Quantitative Methods for Psychology*, **10(1)**: 40-55.