Statistica Sinica Preprint No: SS-2025-0061	
Title	Minimum Aberration Fractional Factorial Designs under Baseline Parameterization
Manuscript ID	SS-2025-0061
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0061
Complete List of Authors	Xinxin Xia, Fasheng Sun and Chunyan Wang
Corresponding Authors	Chunyan Wang
E-mails	chunyanwang@ruc.edu.cn

MINIMUM ABERRATION FRACTIONAL FACTORIAL DESIGNS UNDER BASELINE PARAMETERIZATION

Xinxin Xia[1], Fasheng Sun[1] and Chunyan Wang[2]

[1]*Northeast Normal University and* [2]*Renmin University of China*

Abstract: Fractional factorial designs under the baseline parameterization have received significant attention, with two-level designs being the most popular due to their simplicity. However, extending them to s-level designs for $s \geq 3$ introduces additional challenges. This paper explores the general theory of s-level baseline designs for any $s \geq 3$. Under the baseline parameterization, we demonstrate that orthogonal arrays maintain D_s- and G-optimality across all designs, while also achieving A_s-optimality among balanced designs. We also establish the connection between the wordlength pattern in orthogonal parameterization and the K-value sequence of the designs under the baseline parameterization. Finally, we propose a general method for minimum aberration baseline designs.

Key words and phrases: Baseline parameterization, Fractional factorial design, Minimum aberration, Orthogonal array.

Corresponding author: Chunyan Wang, Center for Applied Statistics and School of Statistics, Renmin University of China, Beijing 100872, China. E-mail: chunyanwang@ruc.edu.cn.

1. Introduction

Fractional factorial designs are widely recognized as one of the most effective tools for screening experiments. Traditionally, most research has focused on these designs under the orthogonal parameterization (OP), where factorial effects are defined through a set of orthogonal contrasts. However, baseline parameterization (BP) has recently attracted growing attention, especially in contexts where a clear null state or baseline level is naturally associated with each factor. The BP defines factorial effects with reference to intrinsic baseline levels of the factors, which can arise quite naturally in many applications (Mukerjee and Tang, 2012). For instance, in a toxicological study with binary factors representing the presence or absence of specific toxins, the absence of a toxin naturally serves as the baseline level for each factor.

Factorial designs under BP have been extensively studied in the context of cDNA microarray experiments by Yang and Speed (2002), Glonek and Solomon (2004), and Banerjee and Mukerjee (2008), all of whom focused on two-level full factorial designs. Mukerjee and Tang (2012) extended this work by investigating optimal two-level fractional factorial designs using the minimum aberration (MA) criterion under BP. The construction of MA baseline designs was further developed by Li, Miller and Tang (2014), Miller and Tang (2016), Mukerjee and Tang (2016), and Chen, Sun and

Tang (2021). More recently, Sun and Tang (2022) established a linear relationship between OP and BP, demonstrating its utility for design construction under BP with respect to estimability, optimality, and robustness. Chen and Tang (2023) proposed MA factorial designs under mixed parameterization, including both OP and BP, for experiments in which some factors have baseline levels while others do not. It is worth noting that all the above studies were limited to two-level designs due to their simple structure. However, real-world applications such as cDNA experiments (Banerjee and Mukerjee, 2008) and agricultural research often necessitate designs with three or more levels. For instance, in genetic studies analyzing cell lines over time, a temporal factor may involve three distinct levels (e.g., measurements at three developmental stages). Similarly, agricultural experiments investigating fertilizer efficacy typically require a three-level design to compare outcomes across no fertilizer, chemical fertilizer, and organic fertilizer treatments. These examples underscore the critical need for baseline designs with s-level factors, where $s \geq 3$. Nevertheless, extending traditional two-level methodologies to multi-level baseline designs introduces significant theoretical and computational complexities. Yan and Zhao (2024) first introduced the MA criterion under BP (BP-MA) for s-level designs with $s \geq 3$ and employed a complete search algorithm based on the BP-MA criterion to identify optimal designs under BP. However, this approach involves a substantial computational burden and gives limited consideration to the structural properties of the optimal designs.

This article advances the construction of optimal s-level baseline designs by studying the theoretical properties of the BP-MA criterion for any $s \geq 3$. First, we examine the relationship between OP and BP. Building on this, we establish that under the main-effect model with BP, orthogonal arrays maintain their status as D_s- and G-optimality within all designs. Further, we also demonstrate that orthogonal arrays are A_s- optimal among all balanced designs. Additionally, we explore the general theoretical properties of MA baseline designs within the BP framework. Our findings uncover connections between the wordlength pattern under OP and the K-value sequence under BP. Building on these theoretical insights, a method is proposed for constructing s-level MA baseline designs for any $s \geq 3$. Examples are given throughout to illustrate the results.

The remainder of this paper is organized as follows. Section 2 introduces some notation and definitions. Section 3 discusses the optimality and robustness of the orthogonal array under BP. Section 4 examines the properties of the BP-MA criterion. Section 5 studies baseline designs derived from regular designs, and proposes a construction method for s-level

BP-MA designs for any $s \geq 3$. Section 6 concludes the paper and offers a discussion. All proofs and some approximate BP-MA designs are provided in the Supplementary Material.

2. Notation and Preliminaries

Let Z_s denote a Galois field of order s, where s is a prime number or a prime power. Consider an s^n factorial that includes n factors $F_1, \ldots, F_n$, with levels taken in Z_s, where 0 is the baseline level. Let $\tau_{(i_1 \ldots i_n)}$ and $\theta_{i_1 \ldots i_n}$ be the treatment effect and factorial effect for treatment combination $i_1 \ldots i_n$, respectively. The baseline mean is defined as $\theta_{0 \ldots 0} = \tau_{(0 \ldots 0)}$, representing the response when all factors are at their baseline level. The main effect of F_j is represented by the $s-1$ parameters

$$\theta_{0 \ldots i_j \ldots 0} = \tau_{(0 \ldots i_j \ldots 0)} - \tau_{(0 \ldots 0)}, \quad i_j = 1, \ldots, s-1,$$

which quantify the effect when the jth factor is at level i_j while all other factors are held at the baseline level. The two-factor interaction of F_j and F_k is represented by the $(s-1)^2$ parameters

$$\theta_{0 \ldots 0 i_j 0 \ldots 0 i_k 0 \ldots 0} = \tau_{(0 \ldots 0 i_j 0 \ldots 0 i_k 0 \ldots 0)} - \tau_{(0 \ldots 0 i_j 0 \ldots 0)} - \tau_{(0 \ldots 0 i_k 0 \ldots 0)} + \tau_{(0 \ldots 0)},$$

which measures the two-factor interaction effect when the jth and kth factors are set at levels i_j and i_k, respectively, with all remaining factors fixed at the baseline level. Here $1 \leq j < k \leq n$ and $i_j, i_k \in \{1, \ldots, s-1\}$. Then, for $1 \leq j < k \leq n$ and $i_j, i_k \in \{1, \ldots, s-1\}$, we obtain

$$\tau_{(0\ldots i_j \ldots 0)} = \theta_{0\ldots i_j \ldots 0} + \theta_{0\ldots 0},$$

$$\tau_{(0\ldots 0 i_j 0 \ldots 0 i_k 0 \ldots 0)} = \theta_{0\ldots 0 i_j 0 \ldots 0 i_k 0 \ldots 0} + \theta_{0\ldots 0 i_j 0 \ldots 0} + \theta_{0\ldots 0 i_k 0 \ldots 0} + \theta_{0\ldots 0}.$$

Thus, for any treatment combination $i_1 \ldots i_n \neq 0 \ldots 0$, we have

$$\tau_{(i_1 \ldots i_n)} = \theta_{0\ldots 0} + \sum_{b=1}^{n} \sum_{h_1, \ldots, h_b \in \psi_b} \left(\prod_{w=1}^{b} j_{h_w} \right) \theta_{\prod_{w=1}^{b} g_{h_w}}.$$

Here, for $b = 1, \ldots, n$, ψ_b is the set of b-tuples $h_1, \ldots, h_b$ with $1 \leq h_1 < \cdots < h_b \leq n$. For $l = 1, \ldots, n$, if $i_l = 0$, then $j_l = 0$, otherwise $j_l = 1$. g_l consists of n elements, of which the lth element is the level of factor F_l, and the rest are 0. The product $\prod_{\omega=1}^{b} g_{h_w}$ is defined as $g_{h_1} + \cdots + g_{h_b}$. For example, consider the treatment combination 120 in a three-level baseline design with three factors, that is, $n = s = 3$, $i_1 = 1$, $i_2 = 2$, and $i_3 = 0$. Then we have $j_1 = 1$, $j_2 = 1$, $j_3 = 0$, $g_1 = 100$, $g_2 = 020$, $g_3 = 000$, and thus

$$\begin{aligned}
\tau_{(120)} &= \theta_{000} + \sum_{b=1}^{3} \sum_{h_1, \ldots, h_b \in \psi_b} \left(\prod_{w=1}^{b} j_{h_w} \right) \theta_{\prod_{w=1}^{b} g_{h_w}} \\
&= \theta_{000} + j_1 \theta_{g_1} + j_2 \theta_{g_2} + j_3 \theta_{g_3} \\
&\quad + j_1 j_2 \theta_{g_1 g_2} + j_1 j_3 \theta_{g_1 g_3} + j_2 j_3 \theta_{g_2 g_3} + j_1 j_2 j_3 \theta_{g_1 g_2 g_3} \\
&= \theta_{000} + \theta_{100} + \theta_{020} + \theta_{120}.
\end{aligned}$$

According to the effect hierarchy principle, the primary focus is on the main effects. Suppose all interactions can be ignored. Then $\theta_{g_1 \dots g_b} = 0$ for $b \geq 2$, and thus

$$\tau_{(i_1 \dots i_n)} = \theta_{0 \dots 0} + j_1 \theta_{g_1} + \cdots + j_n \theta_{g_n}.$$

For an $N \times n$ design Z, the main-effect model is

$$Y = W\theta + \epsilon = \mathbf{1}_N \theta_{0 \dots 0} + Z_1 \theta_1 + \epsilon. \tag{2.1}$$

Here Y is the observation vector. $W = (\mathbf{1}_N, Z_1)$, where $\mathbf{1}_N$ is an $N \times 1$ vector with all elements equal to one. Z_1 is an $N \times (s-1)n$ model matrix corresponding to all main effects. Specifically, each column of Z corresponds to the $s-1$ columns Z_1, and their relation is linked by a mapping shown in Table 1. $\theta = (\theta_{0 \dots 0}, \theta_1^T)^T$, where $\theta_1 = (\theta_{10 \dots 0}, \dots, \theta_{00 \dots s-1})^T$ is an

Table 1: The relation between entry of Z and that of Z_1.

Z		Z_1			
0	$\rightarrow$	0	0	$\cdots$	0
1	$\rightarrow$	1	0	$\cdots$	0
2	$\rightarrow$	0	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$s-1$	$\rightarrow$	0	0	$\cdots$	1

$(s-1)n \times 1$ vector consisting of all main effects. $\epsilon = (\epsilon_1, \ldots, \epsilon_N)^T$ is the vector of random errors that are uncorrelated and have a constant variance σ^2. Under model (2.1), the least square estimate of θ is $\hat{\theta} = (W^T W)^{-1} W^T Y$. The variance–covariance matrix of $\hat{\theta}$ is $\sigma^2 (W^T W)^{-1}$. For screening experiments, as the main interest lies in the estimation of the main effects rather than the intercept term, we consider $\mathrm{var}(\hat{\theta}_1) = \sigma^2 (W^T W)^{-1}_{(-1,-1)}$, where $A^{-1}_{(-1,-1)}$ is obtained from A^{-1} by deleting the first row and first column. To minimize $\mathrm{var}(\hat{\theta}_1)$, we aim to find a baseline design that minimizes $(W^T W)^{-1}_{(-1,-1)}$. There are various considerations for minimizing $(W^T W)^{-1}_{(-1,-1)}$, the most common of which is to minimize the determinant or the trace of $(W^T W)^{-1}_{(-1,-1)}$.

It is well known that in some practical applications, interactions cannot be completely ignored. Then, the true model is

$$Y = W\theta + Z_2\theta_2 + \cdots + Z_n\theta_n + \epsilon, \tag{2.2}$$

where Z_j is the model matrix associated with all j-factor interactions effects, and θ_j is the corresponding vector of unknown parameters for $j = 2, \ldots, n$. Then, the expected value of $\hat{\theta}$ under model (2.2) is

$$E(\hat{\theta}) = \theta + (W^T W)^{-1} W^T Z_2 \theta_2 + \cdots + (W^T W)^{-1} W^T Z_n \theta_n.$$

It is clear that $(W^T W)^{-1} W^T Z_j \theta_j$ represents the contribution to the bias in

$\hat{\theta}$ due to the j-factor interactions. Yan and Zhao (2024) proposed the MA criterion under BP by introducing

$$K_b = \|(W^TW)^{-1}_{-1}W^TZ_b\|_F^2, \tag{2.3}$$

where $b = 2, \ldots, n$ and $\|A\|_F^2$ represents the squared Frobenius norm of matrix A, which is computed as the sum of the squares of all its entries. Here, $(W^TW)^{-1}_{-1}$ is derived from $(W^TW)^{-1}$ by removing the first row. K_b measures the bias of the estimate of main effects due to all the b-factor interactions. The MA criterion under BP is given below.

Definition 1. Given two s-level designs d_1 and d_2, let r be the smallest integer such that $K_r(d_1) \neq K_r(d_2)$, then d_1 is said to have less aberration than d_2 if $K_r(d_1) < K_r(d_2)$ under BP. Furthermore, d_1 is called a MA design under BP if no other design has less aberration than d_1.

3. Optimal Baseline Designs for Main-Effect Model

To better explore the optimality of baseline designs, we first establish the relationship between OP and BP.

3.1 The relationship between OP and BP

Let $\tau = (\tau_{(0\ldots0)}, \ldots, \tau_{(s-1\ldots s-1)})^T$ and $\widetilde{\theta} = (\theta_{0\ldots0}, \ldots, \theta_{s-1\ldots s-1})^T$, where the elements of both vectors are arranged in Yates order. Then under BP, we have

$$\tau = (B_s \otimes \cdots \otimes B_s)\widetilde{\theta},$$

where B_s is repeated n times, and $B_s = \big(\mathbf{1}_s, (\mathbf{0}_{s-1}, I_{s-1})^T\big)$, with I_{s-1} being the identity matrix of order $s-1$. Similarly, let $\widetilde{\beta} = (\beta_{0\ldots0}, \ldots, \beta_{s-1\ldots s-1})^T$. Then under OP, we have

$$\tau = (P_s \otimes \cdots \otimes P_s)\widetilde{\beta}, \tag{3.1}$$

where P_s is an $s \times s$ column-orthogonal matrix ($P_s^T P_s = sI_s$) with the first column being $\mathbf{1}_s$, and P_s is repeated n times in (3.1). Thus, we obtain

$$\widetilde{\theta} = (B_s^{-1} \otimes \cdots \otimes B_s^{-1})\tau = (B_s^{-1}P_s \otimes \cdots \otimes B_s^{-1}P_s)\widetilde{\beta},$$

$$\widetilde{\beta} = (P_s^{-1} \otimes \cdots \otimes P_s^{-1})\tau = (P_s^{-1}B_s \otimes \cdots \otimes P_s^{-1}B_s)\widetilde{\theta}.$$

Lemma 1. *If only the main effects are active under OP, then only the main effects are active under BP and vice versa.*

In the following, Example 1 provides an illustration of Lemma 1.

Example 1. Let $n = 2$, $s = 3$, so $\widetilde{\theta} = (\theta_{00}, \theta_{01}, \theta_{02}, \theta_{10}, \theta_{11}, \theta_{12}, \theta_{20}, \theta_{21}, \theta_{22})$, $\widetilde{\beta} = (\beta_{00}, \beta_{01}, \beta_{02}, \beta_{10}, \beta_{11}, \beta_{12}, \beta_{20}, \beta_{21}, \beta_{22})$, and

$$B_3 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \quad P_3 = \begin{pmatrix} 1 & -\frac{\sqrt{6}}{2} & \frac{\sqrt{2}}{2} \\ 1 & 0 & -\sqrt{2} \\ 1 & \frac{\sqrt{6}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}.$$

According to $\widetilde{\theta} = (B_3^{-1}P_3 \otimes B_3^{-1}P_3)\widetilde{\beta}$ and $\widetilde{\beta} = (P_3^{-1}B_3 \otimes P_3^{-1}B_3)\widetilde{\theta}$, we obtain that

$$\theta_{11} = \frac{3}{2}(\beta_{11} - \sqrt{3}\beta_{12}) + \frac{3}{2}(3\beta_{22} - \sqrt{3}\beta_{21}), \quad \theta_{12} = 3\beta_{11} - 3\sqrt{3}\beta_{21};$$
$$\theta_{21} = 3\beta_{11} - 3\sqrt{3}\beta_{12}, \quad \theta_{22} = 6\beta_{11}, \quad \beta_{11} = \frac{1}{6}\theta_{22}, \quad \beta_{12} = \frac{\sqrt{3}}{18}(\theta_{22} - 2\theta_{21});$$
$$\beta_{21} = \frac{\sqrt{3}}{18}(\theta_{22} - 2\theta_{12}) \quad \beta_{22} = \frac{1}{9}(2\theta_{11} - \theta_{12}) + \frac{1}{18}(\theta_{22} - 2\theta_{21}).$$

Example 1 clearly demonstrates that each effect in $\widetilde{\theta}$ can be linearly represented by the effects in $\widetilde{\beta}$, and vice versa. Additionally, the two-factor interactions under BP are determined solely by the two-factor interactions under OP. Consequently, if all two-factor interaction effects are inactive under OP, the corresponding effects under BP will also be inactive.

According to Lemma 1, if the model under OP is the main-effect model

$$Y = X\beta + \epsilon = \mathbf{1}_N\beta_0 + X_1\beta_1 + \epsilon, \tag{3.2}$$

where $X = (\mathbf{1}_N, X_1)$, β_1 is the vector of main effects, and X_1 is the matrix of contrast coefficients for β_1. Then the model under BP is also the main-effect model.

3.2 Optimality and robust prediction of orthogonal arrays

Under model (3.2), a design is D_s-optimal if it minimizes the determinant of $(X^TX)^{-1}_{(-1,-1)}$, or A_s-optimal if it minimizes $\mathrm{tr}\{(X^TX)^{-1}_{(-1,-1)}\}$. Similarly, the D_s- and A_s-optimal criteria under model (2.1) can be obtained by replacing X with W. Moreover, a design achieves G-optimality if it minimizes the maximum prediction variance over the entire experimental region. Cheng (1980) showed that under OP, the orthogonal array of strength 2 is universally optimal under the main-effect model. An $N \times n$ array with entries from Z_s is said to be an orthogonal array of strength t, if every $N \times t$ subarray contains each t-tuple on Z_s with the same frequency (Hedayat, Sloane and Stufken, 1999). Without loss of generality, the strength of the orthogonal arrays mentioned below is assumed to be greater than 1. The following theorem fundamentally extends this optimality hierarchy from OP to BP.

Theorem 1. *Under model* (2.1), *orthogonal arrays are D_s- and G-optimal among all designs. Furthermore, they achieve A_s-optimality among all balanced designs.*

Theorem 1 demonstrates that orthogonal arrays, when employed as baseline designs, simultaneously minimize the determinant of the variance-covariance matrix for main effects estimation, minimize the maximum prediction variance across the experimental domain, and achieve minimal total variance for main effects estimation among all balanced designs. This unified optimality guarantees both efficient parameter estimation and robust predictive performance. Therefore, we exclusively focus on orthogonal arrays for subsequent sections.

4. Theory of Minimum Aberration Criterion

Let $\alpha(g_{k_1}\dots g_{k_b})$ denote the frequency of $11\dots1$ as a row in the $N\times b$ subarray given by the $\{(s-1)(k_i-1)+g_{k_i}[k_i], i=1,\dots,b\}$th columns of Z_1, where $g_i[i]$ denotes the ith element of g_i, and k_i denotes the position of the ith non-zero digit in $g_{k_1}\dots g_{k_b}$, counting from left to right. For example, when $g_{k_1}g_{k_2}=1020$, we have $k_1=1$, $k_2=3$. Further, let $\phi(g_{k_1}\dots g_{k_b})$ be an $(s-1)n\times 1$ vector with the jth element being $\alpha(\langle jg_{k_1}\dots g_{k_b}\rangle)$, where $\alpha(\langle jg_{k_1}\dots g_{k_b}\rangle)$ denotes the frequency of $11\dots1$ as a row in the $\big\{j,(s-1)(k_i-1)+g_{k_i}[k_i], i=1,\dots,b\big\}$th columns of Z_1. For example, let $s=3$ and let Z be the OA$(9,4,3,2)$ in the left part of Table 2, where the four factors are denoted by A,B,C, and D, respectively. Then, the model matrix corresponding to main effects, Z_1, is a 9×8 matrix, as shown in the right part of Table 2. Let A_1C_2 be an active effect with factorial effect θ_{1020}, where $g_{k_1} = g_1 = 1000$ and $g_{k_2} = g_3 = 0020$. Then $\alpha(g_{k_1}g_{k_2}) = \alpha(1020) = 1$ denotes the frequency of 11 as a row in columns 1 and 6 of Z_1. $\phi(g_{k_1}g_{k_2}) = (1,0,1,0,0,1,0,0)^T$, where the jth element, $\alpha(\langle j1020\rangle)$, is the frequency of 111 or 11 as a row in the $(j,1,6)$th columns of Z_1 for $j = 1,\ldots,8$. For example, the third element $\alpha(\langle 31020\rangle) = 1$ is the frequency of 111 as a row in the $(3,1,6)$th columns of Z_1.

Table 2: Baseline design Z and its first-order model matrix Z_1.

Z				Z_1							
A	B	C	D	A_1	A_2	B_1	B_2	C_1	C_2	D_1	D_2
0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	2	0	0	1	0	1	0	0	1
0	2	2	1	0	0	0	1	0	1	1	0
1	0	1	1	1	0	0	0	1	0	1	0
1	1	2	0	1	0	1	0	0	1	0	0
1	2	0	2	1	0	0	1	0	0	0	1
2	0	2	2	0	1	0	0	0	1	0	1
2	1	0	1	0	1	1	0	0	0	1	0
2	2	1	0	0	1	0	1	1	0	0	0

When an orthogonal array is used to create a baseline design, an additional derivation of K_b in (2.3) is provided below.

$$K_b = \sum_{g_{k_1}\cdots g_{k_b}\in\Phi_b} \xi(g_{k_1}\ldots g_{k_b})^T\xi(g_{k_1}\ldots g_{k_b}), \tag{4.1}$$

where $\Phi_b = \{\varphi_b \mid \varphi_b \text{ is a } 1\times n \text{ vector with } b \text{ non-zero entries from } Z_s\}$, $1 \le k_1,\ldots,k_b \le n$, and $\xi(g_{k_1}\ldots g_{k_b}) = s/N\{A_c\phi(g_{k_1}\ldots g_{k_b}) - \alpha(g_{k_1}\ldots g_{k_b})\mathbf{1}_{(s-1)n}\}$.

A_c is a block-diagonal matrix of order $(s-1)n$ with the diagonal block $H = I_{s-1} + J_{s-1}$. Here, I_{s-1} is an identity matrix of order $s-1$ and J_{s-1} is an all-one matrix of order $s-1$. The smaller the value of K_b, the smaller the bias of b-factor interactions in estimating the main effects is.

We now derive a new expression for K_b in (4.1), which is given in Lemma 2 and plays a key role in the subsequent theoretical results.

Lemma 2. *For* $b = 2, \ldots, n$,

$$K_b = \frac{s^2}{N^2} \sum_{\varphi_b \in \Phi_b} \left(bT_1^{\varphi_b} + T_2^{\varphi_b}\right),$$

where

$$T_1^{\varphi_b} = \|\beta(g_{k_1} \ldots g_{k_b})^T H - \alpha(g_{k_1} \ldots g_{k_b}) \mathbf{1}_{s-1}^T\|_F^2, \text{ and}$$

$$T_2^{\varphi_b} = \sum_{j \in V} \|\beta(jg_{k_1} \ldots g_{k_b})^T H - \alpha(g_{k_1} \ldots g_{k_b}) \mathbf{1}_{s-1}^T\|_F^2.$$

$\beta(g_{k_1} \ldots g_{k_b})$ *is an* $(s-1) \times 1$ *vector, with one element being* $\alpha(g_{k_1} \ldots g_{k_b})$, *and the remaining elements being 0.* $V = \{1, \ldots, n\} \backslash \{k_1, \ldots, k_b\}$. *For* $l = 1, \ldots, s-1$, $j_l = (s-1)(j-1)+l$, *and* $\beta(jg_{k_1} \ldots g_{k_b}) = \big(\alpha(\langle j_1 g_{k_1} \ldots g_{k_b}\rangle), \ldots, \alpha(\langle j_{s-1} g_{k_1} \ldots g_{k_b}\rangle)\big)^T$.

Lemma 2 provides a new expression for K_b. Using Lemma 2, we derive several significant theoretical results.

Theorem 1 demonstrates that the orthogonal array minimizes the variance of the main effects estimates from various aspects. Next, we examine the bias of these estimates.

Theorem 2. *If an $N \times n$ s-level orthogonal array of strength $t \geq 2$ is used to create a baseline design, then the $(K_2, \ldots, K_t)$ sequence satisfies:*

(1). *for* $2 \leq v \leq t-1$, $K_v = \frac{v(s-1)^v}{s^{2v-2}} \binom{n}{v}$,

(2). $K_t = \frac{t(s-1)^t}{s^{2t-2}} \binom{n}{t} + s^2 J_t$, *where* $J_t = \frac{1}{N^2} \sum_{\varphi_t \in \Phi_t} T_2^{\varphi_t}$.

Remark 1. Theorem 2 generalizes Theorem 1 in Miller and Tang (2016), which corresponds to the special case of $s = 2$, and thus allows for the study of general baseline designs.

According to Theorem 2, all s-level orthogonal arrays of strength t have identical $K_2, \ldots, K_{t-1}$. Moreover, orthogonal arrays can be classified into regular and nonregular designs, among which regular designs have specific algebraic structures and are the most widely used. Next, we further consider K_t and K_{t+1} based on regular designs.

5. Construction of MA Baseline Designs

5.1 Baseline designs from regular designs

An OA$(s^{n-p}, m, s, 2)$ with levels from GF(s) is said to be regular and denoted as an s^{n-p} design, if its runs are the solution to the system of equations

$A^T x = 0$, where $A = (a_1, \ldots, a_p)$, and $a_1, \ldots, a_p$ are linearly independent n-dimensional column vectors. Let $R(A)$ be the p-dimensional space generated by $a_1, \ldots, a_p$, and call it the defining contrast subgroup. The aliasing can be captured by the wordlength pattern $(A_3, A_4, \ldots)$, where A_j represents the number of words of length j in $R(A)$. The resolution of design D is the smallest integer j such that $A_j(D) > 0$ (Cheng, 2014). Note that for a regular design, if its resolution is $t+1$, then its strength is t.

As discussed, s^{n-p} designs of resolution $t+1$ have identical K_v values for $2 \leq v \leq t-1$. The following theorem shows that for these designs, K_t (equivalent to J_t) can be minimized by minimizing A_{t+1}.

Theorem 3. *If an s-level regular design of resolution $t+1$ is used to create a baseline design, we have*

$$J_t = \frac{t+1}{s^{2t}} \left\{ (s-1)\gamma_{s1}^t + \gamma_{s2}^t \right\} A_{t+1},$$

where $s \geq 3$ is a prime power, $\gamma_{s1}^t = (-1)^t(s-1)/s + (s-1)^t/s$, and $\gamma_{s2}^t = (s-1)^t - \gamma_{s1}^t$.

Next, let us see an example for illustration.

Example 2. Consider a 5^{4-2} design of resolution 3 with factors A, B, C, and D, where the defining contrast subgroup is

$$I = ABC^4 = AB^3D^4 = ACD^3 = BC^3D^2 = AB^2C^2D^2 = AB^4C^3D.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0061

Here $t = 2$, $A_3 = 4$, $\gamma_{51}^2 = 4$, $\gamma_{52}^2 = 12$, thus $J_2 = \frac{12}{625}(4 \times 4 + 12) = 0.5376$.

Now, we further consider K_{t+1} for s^{n-p} designs of resolution $t + 1$. For any word $F_1^{a_1} F_2^{a_2} \ldots F_k^{a_k}$ of length k, define its degenerate word as $F_1 F_2 \ldots F_k$, where $a_i \in GF(s)$ for $i = 1, \ldots, k$. For example, the degenerate words of words $AB^2C^2D^2$ and AB^4C^3D are both $ABCD$. Further let A_{t+2}^* denote the number of degenerate words of length $t + 2$, and A_{t+2}^1 denote the number of degenerate words of length $t + 2$ that share $t + 1$ common factors with some degenerate word of length $t + 1$. To provide a clearer understanding of these symbols, consider the following example.

Example 3. Consider two 5^{5-2} designs, denoted as Design I and Design II, both incorporating factors A, B, C, D, and E. For these designs, the words of lengths 3 and 4 are given as follows.

$$\text{Design I} : I = ABD^4 = AB^2CE^4 = BCDE^4 = AC^4D^3E.$$

$$\text{Design II} : I = ABD^4 = AB^2E^4 = AD^3E = BDE^4 = AB^4D^2E^2 = AB^3DE^3.$$

For both Designs I and II, we have $t = 2$. Let A_4^* be the number of degenerate words of length 4, and A_4^1 be the number of degenerate words of length 4 that share 3 common factors with some degenerate word of length 3. Design I has three distinct length-4 words: $AB^2CE^4, BCDE^4$ and AC^4D^3E. These three words correspond to three distinct degenerate words: $ABCE$, $BCDE$ and $ACDE$, i.e., $A_4^* = 3$. Moreover, there is no degenerate word of length 4 that shares 3 common factors with any degenerate word of length 3, i.e., $A_4^1 = 0$. Design II has two distinct length-4 words: $AB^4D^2E^2$ and AB^3DE^3. These two words correspond to the same degenerate word $ABDE$, where $ABDE$ shares 3 common factors with some degenerate words of length 3. Thus, we have $A_4^* = A_4^1 = 1$.

Theorem 4. *If an s^{n-p} design of resolution $t+1$ is used to create a baseline design, we obtain*

$$K_{t+1} = \frac{1}{s^{2t}}\{C_1(t+1)A_{t+1} + C_2(t+2)A_{t+2}^1 + C_3(t+2)A_{t+2}^2 + C_4\}, \tag{5.1}$$

where $C_1 = s^2\gamma_{s2}^t + \{(s-1)\gamma_{s2}^t + (s-1)^2\gamma_{s1}^t\}(n-t-1) - (s-1)^{t+1}$, $C_2 = \{s^2-(s-1)(t+1)\}\gamma_{s2}^t + \{s^2(s-2)-(s-1)^2(t+1)\}\gamma_{s1}^t$, $C_3 = \gamma_{s2}^{t+1} + (s-1)\gamma_{s1}^{t+1}$, $C_4 = \frac{(s-1)^{t+1}(t+1)n!}{(t+1)!(n-t-1)!}$, *and* $A_{t+2}^2 = A_{t+2}^* - A_{t+2}^1 > 0$.

Remark 2. When comparing regular designs of resolution $t+1$ with identical K_t values, it becomes necessary to further compare their K_{t+1} values. In this regard, Theorem 4 provides valuable guidance. Furthermore, Theorem 4 extends Theorem 2 of Miller and Tang (2016) to accommodate designs with $s > 2$. This extension broadens its applicability to a wider range of factorial experiments.

Specifically, when $t = 2$, we have $\gamma_{s1}^2 = s-1$, $\gamma_{s2}^2 = (s-1)(s-2)$,

$\gamma_{s1}^3 = (s-1)(s-2)$, and $\gamma_{s2}^3 = (s-1)^3 - (s-1)(s-2)$. Then we obtain the following result.

Corollary 1. *When $t = 2$, we have $A_4^2 = A_4 - (s-3)A_4^1$, and*

$$K_3 = \frac{1}{s^4}\big(3C_1A_3 + 4C_3A_4 + 4C_5A_4^1 + C_4\big),$$

where $C_1 = (s-1)^2(2s-3)n + s^2(s-1)(s-2) - (7s-10)(s-1)^2$, $C_3 = (s-1)^3 + (s-1)(s-2)^2$, $C_5 = C_2 - (s-3)C_3 = (2s-6)(s-1)^2$, and $C_4 = n(n-1)(n-2)(s-1)^3/2$.

From Theorem 3 and Corollary 1, we obtain that the sequential minimization of (K_2, K_3) is equivalent to the sequential minimization of $(A_3, C_3A_4 + C_5A_4^1)$. Let us consider the following example.

Example 4. Based on enumeration, there are nine combinatorially non-isomorphic classes of 4^{7-4} designs. One representative design is selected from each class and labeled as Designs I through IX. These nine 4^{7-4} designs involve factors A, B, C, D, E, F, and G, with their defining words and corresponding index values summarized in Table 3.

By examining Table 3, we observe that when A_3 and A_4 are sequentially minimized, K_2 and K_3 are also sequentially minimized. Moreover, when A_3 remains the same, K_2 does not change.

Table 3: The defining words and associated index values for nine designs.

Design	D	E	F	G	A_3	A_4	A_4^*	A_4^1	A_4^2	K_2	K_3
I	AB	AC	AB^2C^2	AB^3C^3	3	23	23	0	23	32.06	33.84
II	AB	AC	BC^2	AB^2C^2	4	19	19	0	19	34.88	34.32
III	AB	AC	BC	ABC^2	5	15	15	0	15	37.69	34.80
IV	AB	AB^2	AB^3C	AB^2C^2	5	19	19	1	18	37.69	37.52
V	AB	AB^2	AC	BC^2	6	15	15	1	14	40.50	38.00
VI	AB	AC	BC	ABC	7	7	7	0	7	43.31	35.77
VII	AB	AB^2	AC	BC	7	11	11	1	10	43.31	38.48
VIII	AB	AB^2	AC	AC^2	8	11	11	2	9	46.13	41.68
IX	AB	AB^2	AB^3	AC	11	11	11	5	6	54.56	51.28

Remark 3. Extensive simulations reveal that for designs with the same value of A_3, minimizing A_4^1 also leads to the minimization of A_4. This finding suggests that when searching for the MA design under BP, one can start with the MA design under OP.

Given the frequent occurrence for $s = 3, 4$, and 5 in practical applications, we derive specific expressions for K_{t+1} with $t = 2$ for these representative cases. The detailed formulas are provided below.

Corollary 2. (1). *For a 3^{n-p} design of resolution 3, we have*

$$K_3 = \frac{1}{81} \left\{ (36n - 78)A_3 + 40A_4 + 4n(n-1)(n-2) \right\}.$$

(2). *For a 4^{n-p} design of resolution 3, we have*

$$K_3 = \frac{1}{256}\left\{(135n-198)A_3 + 156A_4 + 72A_4^1 + \frac{27}{2}n(n-1)(n-2)\right\}.$$

(3). *For a 5^{n-p} design of resolution 3, we have*

$$K_3 = \frac{1}{625}\left\{(336n-300)A_3 + 400A_4 + 256A_4^1 + 32n(n-1)(n-2)\right\},$$

where A_4^1 is the number of degenerate words of length 4 that share 3 common factors with some degenerate word of length 3.

We next illustrate Corollary 2 with two examples.

Example 5. Consider a 3^{5-2} design of resolution 3 with factors A, B, C, D, and E, where the defining contrast subgroup is

$$I = ABD^2 = CDE^2 = ABCE^2 = ABC^2DE.$$

Here, $n = 5$, $A_3 = 2$, $A_4 = 1$, thus $K_2 = 12.89$ and $K_3 = 5.98$.

Example 6. 5^{5-2} designs have four combinatorially non-isomorphic classes based on enumeration. Select one design from each of these four classes and denote the resulting designs as I, II, III, and IV, respectively. The five factors are denoted as A, B, C, D, and E. Table 4 presents the definition relationships for these four designs. The words of length 3 and 4 in each design can be obtained, as shown below.

$$\text{Design I}: I = ABCD^4 = AB^2C^3E^4 = BC^2DE^4 = AC^4D^3E = AB^3DE^3.$$

$$\text{Design II}: I = ABD^4 = AB^2CE^4 = BCDE^4 = AC^4D^3E.$$

$$\text{Design III}: I = ABD^4 = ACE^4 = BC^4D^4E.$$

$$\text{Design IV}: I = ABD^4 = AB^2E^4 = AD^3E = BDE^4 = AB^4D^2E^2 = AB^3DE^3.$$

As shown in Table 4, when A_3 and A_4 are minimized sequentially, K_2 and K_3 are also minimized sequentially.

Table 4: Some associated index values for designs in Example 6.

Design	D	E	A_3	A_4	A_4^*	A_4^1	A_4^2	J_2	K_2	K_3
I	ABC	AB^2C^3	0	5	5	0	5	0	12.80	6.27
II	AB	AB^2C	1	3	3	0	3	0.13	16.16	7.20
III	AB	AC	2	1	1	0	1	0.27	19.52	8.13
IV	AB	AB^2	4	2	1	1	0	0.54	26.24	13.59

5.2 A general construction method for MA baseline designs

Building on the definition of baseline isomorphism (Yan and Zhao, 2024) and the non-exchangeability of the baseline level with other levels under BP, it is sufficient to consider s distinct level permutations for an s-level design. In this paper, the following s level permutations are considered.

$$\begin{aligned} &\{0,1,2,\ldots,s-1\} \rightarrow \{0,1,2,\ldots,s-1\},\\ &\{0,1,2,\ldots,s-1\} \rightarrow \{1,0,2,\ldots,s-1\},\\ &\qquad\vdots\\ &\{0,1,2,\ldots,s-1\} \rightarrow \{1,2,\ldots,0,s-1\},\\ &\{0,1,2,\ldots,s-1\} \rightarrow \{1,2,\ldots,s-1,0\}. \end{aligned} \tag{5.2}$$

For example, when $s=3$, we only need to consider the following three level permutations.

$$\{0,1,2\} \rightarrow \{0,1,2\},\ \{0,1,2\} \rightarrow \{1,0,2\},\ \{0,1,2\} \rightarrow \{2,1,0\}.$$

Therefore, based on the theoretical properties of the MA criterion under BP, we propose a method of s-level MA designs under BP based on MA designs under OP and level permutations in (5.2), as given in Algorithm 1.

Algorithm 1

Step 1 Given N and n, we list all s-level regular MA designs under OP with resolution $t+1$, suppose there are p such arrays.

Step 2 For each of the p MA designs in Step 1, obtain s^n designs by applying level permutations in (5.2) on one or more columns. For each of the resulting $s^n p$ designs, calculate the K_r for $r=t,\ldots,n-1$ and then find the minimum aberration designs.

When $s = 5$, Table 5 demonstrates the K_2 and K_3 values for the 125-run designs with n factors generated by the regular MA design (RMA), as well as the K_2 and K_3 values derived from applying level permutations to the regular MA design (PMA). As shown in Table 5, by applying level permutations to regular MA designs under OP, baseline designs with smaller K_2 and K_3 values can be obtained. Note that these designs may no longer be regular designs.

Table 5: The comparison of K_2 and K_3-values of RMA and PMA.

n	3	4	5	6	7	8	9	10
(RMA)K_2	3.84	7.68	12.80	19.20	33.60	49.28	69.60	124.80
(PMA)K_2	3.84	7.68	12.80	19.20	32.88	47.84	67.08	117.60
(RMA)K_3	0.31	1.87	6.27	15.74	34.60	64.49	110.07	201.24
(PMA)K_3	0.31	1.87	6.27	15.74	34.61	64.43	109.63	193.39

Remark 4. Theorems 3 and 4 provide a theoretical basis for constructing baseline designs from MA regular designs, where the detailed method is given in Algorithm 1. In fact, the level permutation can be extended into nonregular designs to identify approximate BP-MA designs. Please refer to the supplementary materials for details. It is shown that the proposed method performs well for nonregular cases as well.

Remark 5. In Step 2 of Algorithm 1, evaluating the K_r for $r = t, \ldots, n-1$

across all $s^n p$ candidate designs becomes computationally intractable for large $s/n/p$. To address this combinatorial explosion, we adopt a randomized subsampling strategy. Specifically, we randomly select a subset of designs from the $s^n p$ candidate designs. For the selected designs, we compute K_r for $r = t, \ldots, n-1$ to identify approximately minimum aberration designs. Further details can be found in the supplementary material.

6. Conclusion and Discussion

Fractional factorial designs under BP have garnered significant attention. However, the theory of s-level fractional factorial designs with $s \geq 3$ under BP has not been thoroughly explored. This paper investigates the general theory of s-level baseline designs for any $s \geq 3$. We first establish the relationship between OP and BP. Subsequently, we show that orthogonal arrays retain their D_s- and G-optimality among all designs. Moreover, we demonstrate that orthogonal arrays achieve A_s-optimality within the class of balanced designs. Furthermore, we explore the theory of the BP-MA criterion and its connection to the MA criterion under OP. Finally, we propose a method for constructing BP-MA designs based on these theoretical properties. Additionally, we provide theoretical support for the algorithm proposed by Yan and Zhao (2024) and supplement it with designs for cases where $s = 5$ and the number of runs is 25, 50, 75, 100, and 125 in the supplementary material.

This paper may inspire further exploration in related research areas. Theorem 1 demonstrates that orthogonal arrays are D_s-, G-, and A_s-optimal among all balanced designs. This suggests an interesting and meaningful direction for future research — exploring whether orthogonal arrays are also ϕ-optimal among all balanced designs under any concave and signed permutation invariant criterion $\phi(\cdot)$, as proposed by Peng, Mukerjee and Lin (2019). In addition to this theoretical inquiry, another promising direction lies in extending the construction of s-level baseline designs beyond balanced settings. For instance, the compromise design constructed by Karunanayaka and Tang (2017) and Li, Liu and Tang (2022) in the case of two levels. Building on this, we aim to construct compromise designs for $s \geq 3$ levels in the future. To further broaden the applicability of baseline designs, it is also essential to explore their development in more complex experimental frameworks, such as block experiments, which warrant further investigation.

Supplementary Material

Supplementary material presents the proofs of theoretical results and lists 5-level approximate BP-MA designs and their K_2 and K_3 values for runs of 25, 50, 75, 100, and 125.

Acknowledgments

The authors thank Editor Yi-Hau Chen, an associate editor, and two referees for their valuable comments and suggestions. This work was supported by the National Natural Science Foundation of China (Grant Nos. 12301323 and 12371259), the Fundamental Research Funds for the Central Universities, China (2412023YQ003), and MOE Project of Key Research Institute of Humanities and Social Sciences (22JJD110001). Chunyan Wang and Fasheng Sun are the corresponding authors.

References

Banerjee, T. and Mukerjee, R. (2008). Optimal factorial designs for cDNA microarray experiments. *Ann. Appl. Stat.* **2**, 366–385.

Chen, A., Sun, C. Y. and Tang, B. (2021). Selecting baseline designs using a minimum aberration criterion when some two-factor interactions are important. *Stat. Theory and Relat. Fields* **5**, 95–101.

Chen, G. Z. and Tang, B. (2023). Minimum aberration factorial designs under a mixed parametrization. *Statist. Sinica* doi: 10.5705/ss.202023.0165

Cheng, C. S. (1980). Orthogonal arrays with variable numbers of symbols. *Ann. Statist.* **8**, 447–453.

Cheng, C. S. (2014). *Theory of Factorial Design: Single- and Multi-Stratum Experiments*. Boca Raton: CRC Press.

Glonek, G. F. V. and Solomon, P. J. (2004). Factorial and time course designs for cDNA microarray experiments. *Biostatistics* **5**, 89–111.

Hedayat, A. S., Sloane, N. J. and Stufken, J. (1999). *Orthogonal Arrays: Theory and Applications*. Springer Series in Statistics. New York: Springer-Verlag.

Karunanayaka, R. C. and Tang, B. (2017). Compromise designs under baseline parameterization. *J. Statist. Plann. Inference.* **190**, 32–38.

Li, P., Miller, A. and Tang, B. (2014). Algorithmic search for baseline minimum aberration designs. *J. Statist. Plann. Inference.* **149**, 172–182.

Li, W. L., Liu, M. Q. and Tang, B. (2022). A systematic construction of compromise designs under baseline parameterization. *J. Statist. Plann. Inference.* **219**, 33–42.

Miller, A. and Tang, B. (2016). Using regular fractions of two-level designs to find baseline designs. *Statist. Sinica* **26**, 745–759.

Mukerjee, R. and Tang, B. (2012). Optimal fractions of two-level factorials under a baseline parameterization. *Biometrika* **99**, 71–84.

Mukerjee, R. and Tang, B. (2016). Optimal two-level regular designs under baseline parameterization via cosets and minimum moment aberration. *Statist. Sinica* **26**, 1001–1019.

Peng, J. Y., Mukerjee, R. and Lin, D. K. (2019). Design of order-of-addition experiments. *Biometrika* **106**, 683–694.

Sun, C. Y. and Tang, B. (2022). Relationship between orthogonal and baseline parameterizations and its applications to design constructions. *Statist. Sinica* **32**, 239-250.

Yang, Y. H. and Speed, T. (2002). Design issues for cDNA microarray experiments. *Nature Genet.* **3**, 579–588.

Yan, Z. H. and Zhao, S. L. (2024). Optimal s-level fractional factorial designs under baseline parameterization. *J. Statist. Plann. Inference.* **236**, 106242.

Xinxin Xia

KLAS and School of Mathematics and Statistics, Northeast Normal University, Changchun, Jilin 130024, China

E-mail: xiaxx818@nenu.edu.cn

Fasheng Sun

KLAS and School of Mathematics and Statistics, Northeast Normal University, Changchun, Jilin 130024, China

E-mail: sunfs359@nenu.edu.cn

Chunyan Wang

Center for Applied Statistics and School of Statistics, Renmin University of China, Beijing 100872, China

E-mail: chunyanwang@ruc.edu.cn