A Sliced Design Approach for Conducting Online Experiments with Four Platforms, with Application to an Industry Email Campaign

Soheil SADEGHI, Tzu-Hsiang HUNG, Peter CHIEN*, and Neeraj ARORA

Abstract

Multivariate testing is a popular method to improve the effectiveness of digital marketing in industry. Online campaigns are often conducted across multiple platforms, such as desktops, tablets, smart phones, and smart watches. We propose minimum sliced aberration designs to accommodate online experiments with four platforms. This approach provides important insights into how different sets of design factors work differently across the four platforms, which can be used by industry for optimizing many forms of digital marketing. The effectiveness of the proposed approach is illustrated by an industrial email campaign with four platforms.

KEYWORDS AND PHRASES: Multivariate testing, Factorial designs, Sliced designs, Email campaign.

1. INTRODUCTION

Digital marketing through emails, social media, webinars, podcasts and other forms is commonly used in all industries. For example, email marketing is a powerful marketing tool used by many Business-to-Business and Business-to-Consumer companies and about 87% percent of marketers use it to disseminate their content. Email marketing has a high return on investment (ROI) ($42 for every $1 investment, on average); about 4 billion individuals send about 300 billion emails each day and these figures are expected to grow. About 80% of small and medium sized business rely on emails for customer acquisition and retention and a wide spectrum of industries including software and technology, hospitality, entertainment, retail, and consumer goods depend on emails as the primary form of promotion [7].

In light of widespread use of emails, it is not surprising that experiments are often used to improve email effectiveness. Common factors tested in email experiments include plain text vs. HTML, image A vs. B, the location of an image (e.g. right vs. left aligned), template design C vs. D, day of the week, time of the day, personalization (e.g. first name vs. no name), image call to action (CTA) vs. text CTA, familiar tone vs. professional tone, long vs. short emails, etc.

Statistical design (A/B or multivariate testing) plays a key role in email marketing. While A/B tests are effective in assessing the performance of one factor at a time, multivariate tests are far more powerful because they can be used to determine the optimal combination of several factors at the same time. It is also possible to assess interaction effects of email factors in a multivariate experiment. Online testing is a popular method to improve the layout of digital products such as a website and an app. It is usually conducted for the

purpose of increasing the engagement and conversion metrics, e.g., page visits, click-through rate, and purchase. In its general form, online testing includes multiple attributes of a digital product and the effects of these attributes are studied on a response variable simultaneously. Factorial designs are increasingly used to perform online testing; for example, see [3]. As a unique challenge in digital spaces, online testing is conducted across multiple platforms including desktops, tablets, smart phones, and smart watches. A customer can interact with an application on one of these platforms, and a different set of attribute combinations may optimize his/her engagement metric for each platform. For example, although the presence of multiple images may work the best for an application on a tablet, a series of links might be the best for the same application on a smart-watch.

Recent research in marketing points to the fact that potential buyers follow different paths to purchase [4] that may involve different devices. For example, a person may initiate a purchase process on their smart-phone at home, continue to evaluate alternatives at their desktop computer during lunch, and may purchase a product on their laptop or tablet at home. This multi-device path to purchase requires that marketers ensure that their website is optimized for the user experience from a variety of device types (smart-phone, tablet, laptop, and desktop). Display advertising [2] and retargeted display advertising [5] copy that potential buyers see, should be optimized for the different device types. Such optimization is not limited to different device types alone. Variations also occur because of four browser types (Chrome, Internet Explorer, Firefox, and Safari). Moreover, marketers may want to optimize their display advertising campaigns across the four social media outlets that include Facebook, Instagram, Twitter, and LinkedIn.

*Corresponding author.

[8] introduced a sliced version of the minimum aberration criterion to accommodate online experiments with two platforms. This article extends this method to construct sliced factorial designs for online experiments with four platforms using the method of replacement from [1] and [11]. The proposed designs are applied to an industrial email campaign by a network company. The goal of the campaign is to identify which attributes of the campaign are the most effective to impact the measured outcome (e.g. open rate). The email design team of the company identified six binary design factors for the multivariate test for four platforms: Android, iOS, Windows, and macOS.

The remainder of the article is organized as follows. Section 2 introduces the email campaign problem faced by the company. Section 3 provides a design solution to this campaign using sliced factorial designs for the four platforms and generalizes the method to any number of design factors. Section 4 gives results of the application of this design in the email campaign. Section 5 concludes with discussions.

2. MOTIVATING EXAMPLE

The network company launched an email blast to identify which among the six attributes are the most effective to impact the measured outcome. The email design team in the company sent an email to its customers with brief information on the market research report. To maintain the confidentiality of the company, we have masked parts of the email that may reveal the company name. There are six binary design factors of the multivariate test for four platforms P_1, P_2, P_3, and P_4. Platform P_1 refers to Android, P_2 refers to iOS, P_3 refers to Windows, and P_4 refers to macOS. The slice factor **S** is defined as a four-level factor where the jth level of **S** represents P_j. The six binary design factors are thumbnail, subject line, asset type, header image, preview text, and content display. If a full factorial design was needed, we would have to create 2^6 versions for each of the four platforms. Blocking is a common method to form blocks of homogeneous units in a factorial design. While this method works well for agriculture and engineering applications where treatment-blocking interaction is negligible [10], it is ill-suited for online experiments with multiple platforms [8]. If one uses the slice factor **S** as a block factor to construct a blocked factorial design d with blocks $d_1, \ldots, d_4$, —for example, a 32-run 2^{6-1} fractional factorial design with generators $6 = 12345$ and $B_1 = 134$ and $B_2 = 234$ —then S would be aliased with the higher-order interaction effects of the design factors. This assumes that the slice factor **S** has a negligible interaction with the design factors. This assumption contradicts the primary goal of how the six design factors effect may interact with the four platforms. Practical constraints such as budgets of extensive programming limit the number of versions. The company we work with can only afford up to eight versions for each of the four platforms and is interested in modeling the interaction between the design factors and the four platforms in addition to the factorial effects of the design factors. None of the aforementioned designs fit the requirement. We decided to use a 2^{6+2-3} minimum sliced aberration design constructed in Section 3 for the email campaign.

Using the design to be generated in Section 3, we created 2^3 versions to perform the multivariate testing. We use the 2^{6+2-3} minimum sliced aberration design for four platforms. Table 1 lists six binary design factors identified for this study. These factors are 1: thumbnail, 2: subject line, 3: asset type, 4: header image, 5: preview text, and 6: content display. For each factor, we label the two levels as $+$ and $-$.

Each platform has eight versions to perform this multivariate testing. The eight versions form a 2^{6-3} fractional factorial design. The first version of our design is the version with all six design factors at $-$ levels, presented in Table 3. Version two has factors 1, 4, and 5 that are at $+$ levels and the remaining three factors are at $-$ levels. Similarly, version three has factors 2, 4, and 6 at $+$ levels although the other three factors are at $-$ levels. Table 2 lists the description of the eight versions. Tables 3 and 4 include the email of all eight versions used in the campaign.

3. SLICED FACTORIAL DESIGNS WITH FOUR PLATFORMS

We discuss how we constructed the sliced design in Section 2 for the email campaign. Using the same notation as [8], we cast our email campaign as a multi-platform experiment with four platforms: Android, iOS, Windows, and macOS. For readers who are unfamiliar with design of experiments, please refer to Appendix A and [10].

3.1 Four-Platform Experiment: Android, iOS, Windows, and macOS

Consider the four-platform experiment involving Android, iOS, Windows, and macOS discussed above. Denote the six two-level design factors by $1, \ldots, 6$ on the four platforms denoted by $P_1, \ldots, P_4$. The complete design d of the experiment consists of four sub designs, $d_1, \ldots, d_4$, with d_j associated with P_j. To quantify the difference among the platforms, let **S** denote a categorical factor, called the slice factor, with four levels. The jth level of **S** is associated with P_j.

We consider the following properties from [8] to guide the construction of our design:

Property 1. For $j = 1, \ldots, 4$, the sub design d_j should achieve desirable estimation capacity for the design factors on platform P_j.

Property 2. Combined together, the complete design d should achieve desirable estimation capacity for the slice

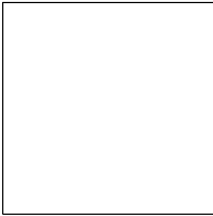Table 1. Six binary design factors for an industrial email blast.

Factor	+	−
1 Thumbnail	Yes	No

Factor	+	−
2 Subject Line	Direct	Indirect
	Juniper Is a Leader...	Take me to your Leader
3 Asset Type	With	Without
	Report:	
4 Header Image	Including	No

Factor	+	−
5 Preview Text	No	Including
6 Content Display	Paragraph in the body	Bullet Points

Thanks to continually working with our customers in applying automation to increase efficiency to their data centers,   have just announced this year's   for   . And the best part? They've named Juniper Networks as a Leader, for the third consecutive year.

Read the full report to see exactly what   has to say about our hardware and software and how we believe Juniper is using AI to improve data center networking efficiency.

has just announced this year's   for   . And the best part? They've named Juniper Networks a Leader – which we believe we couldn't have done without our amazing customers.

Juniper's solutions can improve data center and cloud networking with:

- Innovative hardware and software automation capabilities
- AI-driven data center networking efficiencies
- Far reaching Juniper solutions that stretch from management to integration

Table 2. Description of eight versions of the email study.

Version	Attribute					
	Thumbnail	Subject Line	Asset Type	Header Image	Preview Text	Content Display
1	No	Indirect	Without	No	Including	Bullet Points
2	With	Direct	With	No	Including	Bullet Points
3	With	Indirect	Without	Including	No	Bullet Points
4	No	Direct	Without	Including	Including	Paragraph
5	No	Indirect	With	No	No	Paragraph
6	With	Direct	Without	No	No	Paragraph
7	With	Indirect	With	Including	Including	Paragraph
8	No	Direct	With	Including	No	Bullet Points

factor **S** and the two-way interactions between **S** and the design factors.

As a result of Property 1, each sub design d_j estimates the effects of design factors on platform P_j, and according to effect hierarchy [10, p. 168], the focus of estimation is on the lower-order effects – main effects and two-way interactions. Property 2 suggests that the complete design d focuses on the estimation of the slice factor **S** and its two-way interactions with the design factors. This requires a different ordering of effects than the effect hierarchy for the complete design d in which **S** is more likely to be important than the main effects of the design factors, and two-way interaction effects of **S** with the design factors are more likely to be important than the two-way interaction effects of the design factors. [8] proposed the *sliced effect hierarchy* for the complete design d in order to accommodate Property 2. To formally define this ordering of effects for the design d in our experiment, let E_I be the set of all effects that exclude the slice factor **S** and E_S be the set of all effects that include the slice factor **S**. [8] defined the sliced effect hierarchy as follows:

Table 3. Versions one, two, three, and four of the email study.

Version one:	Version two:
Take me to your Leader	Report: Juniper is a Leader...
Version three:	Version four:
Take me to your Leader	Juniper Is a Leader...

Sliced Effect Hierarchy.

- For E_I or E_S, the lower-order effects are more likely to be important than the higher-order effects.
- For E_I or E_S, effects of the same order are equally likely to be important.
- Any effect in the set E_S is likely to be more important than an effect in E_I that is of the same order.
- Any effect in the set E_S is likely to be less important than an effect in E_I that is of a lower order.

In this experiment, the slice factor differs from the design factors in two ways. First, our four-platform experiment aims to detect what level of the design factors should be chosen for each platform and is not trying to select between platforms. Second, according to the sliced effect hierarchy, the importance of the effects related to the slice factor is higher than the importance of the same-order effects of the design factors. A design of the experiment should distinguish between the slice factor effects and the effects of the design factors.

We wanted to use the *sliced factorial designs* in [8] for our experiment. In a sliced factorial design, each sub design d_j follows the effect hierarchy and the complete design d follows the sliced effect hierarchy. Unfortunately, [8] only constructed such designs for *two* platforms. Since our problem consists of four platforms, we cannot use that method directly. Below we discuss a solution by extending the method in [8] to accommodate our four platforms: Android, iOS, Windows, and macOS.

Our solution generates a design d with 2^{6+2-p} runs for

Table 4. Versions five, six, seven, and eight of the email study.

Version five:	Version six:
Report: Take me to your Leader	Report: Juniper is a Leader...

Version seven:	Version eight:
Report: Take me to your Leader	Report: Juniper is a Leader...

our experiment, which is a $(\frac{1}{2})^p$ fraction of a 2^{6+2} factorial design. First, we describe the construction of a full factorial design d for the experiment. Consider a saturated 2^{N-1} design with $N = 2^8$ runs. We represent the $N-1$ columns of this design by eight independent columns denoted by $1, \ldots, 8$, and their interactions of order two to eight, $12, 13, \ldots, 12 \cdots 8$ [11]. Any three columns of the form (a, b, ab), where ab is the interaction column between columns a and b, can be used to represent the levels of the slice factor $\mathbf{S}$ without affecting orthogonality [1]. This replacement can be done according to the rule in Table 5.

Next, we discuss details of our design d with 2^{6+2-p} runs. Consider a full factorial design with 2^{6+2-p} runs, with the 4-level column represented by $\mathbf{S} = (s_1, s_2, s_3)$ with $s_3 = s_1 s_2$, and the 2-level columns represented by $1, \ldots, 6 - p$. The remaining p columns, $6 - p + 1, \ldots, 6$, can be generated as

Table 5. Rule for replacing any three columns of the form (a, b, ab) by the 4-level column S.

a	b	ab		4-level column S
0	0	0		0
0	1	1	$\longrightarrow$	1
1	0	1		2
1	1	0		3

interactions of the first $6 - p + 2$ columns. How to pick these p columns determines the generators and the defining relation of the design d. For a two-platform experiment, [8] defined the sliced wordlength pattern to accommodate the aliasing relation of the slice factor $\mathbf{S}$. For a four-platform experiment, this definition does not work as the slice factor $\mathbf{S}$ has three

aliasing relations of s_1, s_2, and s_3, respectively. The aliasing relation of s_j is obtained by multiplying the defining relation of d by s_j. Therefore, a word W in the defining relation of d appears in the three aliasing relations for the slice factor **S** as $s_1 W$, $s_2 W$, and $s_3 W$. We extend [8]'s definition of the sliced wordlength pattern to cover the minimum length of $s_1 W$, $s_2 W$, and $s_3 W$. This extension minimizes the number of the shortest length of a sliced wordlength pattern. Defining the sliced wordlength pattern over the minimum length of $s_1 W$, $s_2 W$, and $s_3 W$ ensures that the minimum sliced aberration protects against the worst-case scenario.

We use [11]'s definition of wordlength pattern for designs with two-level and four-level factors to define the sliced wordlength pattern. The design d with 2^{6+2-p} runs has two types of words in its defining relation. The first, called type 0, involves only the design factors $1, \ldots, 6$, and the second, called type 1, involves one of the s_j's and some of the design factors $1, \ldots, 6$. Because of the relation $s_1 s_2 s_3 = I$, any two s_j's appearing in a word can be replaced by the third s_j. Therefore, these two types cover all the possible combinations. Following [11], the vector

$$W(d) = ([A_{i0}(d), A_{i1}(d)])_{i \geq 3}, \tag{3.1}$$

is the wordlength pattern of d in which $A_{i0}(d)$ and $A_{i1}(d)$ are the numbers of type 0 and type 1 words of length i in the defining relation of d, respectively. The term $[A_{20}(d), A_{21}(d)]$ is not considered in (3.1) because any design d with a positive value of $[A_{20}(d), A_{21}(d)]$ is not useful as two of its main effects are aliased. We define the sliced wordlength pattern of a design d for a four-platform experiment as follows: For a design d with the wordlength pattern $W(d) = ([A_{i0}(d), A_{i1}(d)])_{i \geq 3}$ for the four-platform experiment under consideration, we define the *sliced wordlength pattern* to be the vector $SW(d) = ([SA_{i0}(d), SA_{i1}(d)])_{i \geq 2}$, where

- $SA_{i0}(d) = A_{(i+1)1}(d)$ for $i \geq 2$;
- $SA_{i1}(d) = A_{(i-1)0}(d)$ for $i \geq 4$;
- $SA_{i1}(d) = 0$ for $i = 2, 3$.

A type 0 word W in the defining relation of d appears as a type 1 word in the aliasing relations of the sliced factor S. It is counted as a type 1 word in the sliced wordlength pattern, resulting in $SA_{i1}(d) = A_{(i-1)0}(d)$. A type 1 word W in the defining relation of d appears as a type 1 word in the aliasing relations of two s_j's and as a type 0 word in the aliasing relation of the third s_j. It is counted as a type 0 word in the sliced wordlength pattern with $SA_{i0}(d) = A_{(i+1)1}(d)$ because the sliced wordlength pattern is defined over the minimum length of a word in the three aliasing relations.

The sliced resolution of d is defined to be the smallest i for which at least one of $SA_{i0}(d)$ and $SA_{i1}(d)$ is positive. Additional discrimination among designs with the same sliced resolution is covered by the following minimum sliced aberration. The two types of words of the design d are not treated the same. According to the sliced effect hierarchy, a type 1

word in the aliasing relations of the slice factor **S** is more serious because it involves one s_j. This is consistent with [11]'s result that ranks a type 0 word in the defining relation of d more important than a type 1 because a type 0 word in the defining relation appears as a type 1 word in the aliasing relations of the slice factor **S**. Therefore, it is more important to require a smaller $SA_{i1}(d)$ than a smaller $SA_{i0}(d)$ for the same i. We define minimum sliced aberration designs for a four-platform experiment as follows:

Definition 1 (Minimum Sliced Aberration Designs). Suppose that, for our experiment, two designs $d^{(1)}$ and $d^{(2)}$ with 2^{6+2-p} runs are to be compared. Let r be the smallest integer such that $[SA_{r0}(d^{(1)}), SA_{r1}(d^{(1)})] \neq [SA_{r0}(d^{(2)}), SA_{r1}(d^{(2)})]$. If $SA_{r1}(d^{(1)}) < SA_{r1}(d^{(2)})$, or $SA_{r1}(d^{(1)}) = SA_{r1}(d^{(2)})$ but $SA_{r0}(d^{(1)}) < SA_{r0}(d^{(2)})$, then $d^{(1)}$ is said to have less sliced aberration than $d^{(2)}$. If there is no design with less sliced aberration than $d^{(1)}$, then $d^{(1)}$ is called a minimum sliced aberration design.

For our experiment with six design factors, let s_1, s_2, 1, 2, and 3 be the five independent columns of the 32-run 2^5 design. Consider two designs:

$$d^{(1)} : S, 1, 2, 3, 12, 13, 23$$
$$d^{(2)} : S, 1, 2, 3, 13s_2, 23s_2, 123s_1$$

where $\mathbf{S} = (s_1, s_2, s_3 = s_1 s_2)$ and the last three columns represent the last three design factors. For example, $4 = 12$, $5 = 13$ and $6 = 23$ in $d^{(1)}$ and $4 = 13s_2$, $5 = 23s_2$ and $6 = 123s_1$ in $d^{(2)}$. Therefore, the defining relations of $d^{(1)}$ and $d^{(2)}$ are:

$$d^{(1)} : I = 124 = 135 = 236 = 2345 = 1346 = 1256 = 456$$
$$d^{(2)} : I = 134s_2 = 235s_2 = 1236s_1 = 1245 = 246s_3 = 156s_3$$
$$= 3456s_1.$$

The defining relation of $d^{(1)}$ has seven words of type 0: four of length three and three of length four. The wordlength pattern of $d^{(1)}$ is $W(d^{(1)}) = ([4, 0]_3, [3, 0]_4)$. Multiplying the defining relation of $d^{(1)}$ by s_j's provides the following three aliasing relations of the slice factor **S**:

$$s_1 = 124s_1 = 135s_1 = 236s_1 = 2345s_1 = 1346s_1 = 1256s_1 = 456s_1$$
$$s_2 = 124s_2 = 135s_2 = 236s_2 = 2345s_2 = 1346s_2 = 1256s_2 = 456s_2$$
$$s_3 = 124s_3 = 135s_3 = 236s_3 = 2345s_3 = 1346s_3 = 1256s_3 = 456s_3.$$

Each type 0 word in the defining relation of $d^{(1)}$ appears as type 1 word in all three aliasing relations of s_j's. The sliced wordlength pattern of $d^{(1)}$ is $SW(d^{(1)}) = ([0, 0]_2, [0, 0]_3, [0, 4]_4, [0, 3]_5)$. The defining relation of $d^{(2)}$ has one word of type 0 of length four and six words of type 1: four of length four and two of length five. The wordlength pattern of $d^{(2)}$ is $W(d^{(2)}) = ([0, 0]_3, [1, 4]_4, [0, 2]_5)$. Multiplying the defining

Table 6. Minimum sliced aberration design for six design factors with 32 runs, $\mathbf{S} = (s_1, s_2, s_1 s_2)$.

Design d	$SW(d)_{i \geq 4}$
$S, 1, 2, 3, 4 = 12, 5 = 13, 6 = 23$	$([0,4]_4, [0,3]_5)$

relation of $d^{(2)}$ by s_j's provides the following three aliasing relations of the slice factor $\mathbf{S}$:

$$s_1 = 134s_3 = 235s_3 = \underline{1236} = 1245s_1 = 246s_2 = 156s_2 = \underline{3456}$$
$$s_2 = \underline{134} = \underline{235} = 1236s_3 = 1245s_2 = 246s_1 = 156s_1 = 3456s_3$$
$$s_3 = 134s_1 = 235s_1 = 1236s_2 = 1245s_3 = \underline{246} = \underline{156} = 3456s_2.$$

The type 1 word $134s_2$ in the defining relation of $d^{(2)}$ appears as type 0 word of length three in the aliasing relation of s_2 because $s_2 s_2 = I$ and as type 1 word of length four in the aliasing relations of s_1 and s_3 because $s_1 s_2 = s_3$ and $s_3 s_2 = s_1$. It has length three of type 0 in the sliced wordlength pattern. Similar explanations can be used for the other six words in the defining relation of $d^{(2)}$. The sliced wordlength pattern of $d^{(2)}$ is $SW(d^{(2)}) = ([0,0]_2, [4,0]_3, [2,0]_4, [0,1]_5)$. Between the two designs $d^{(1)}$ and $d^{(2)}$, 3 is the smallest integer such that $[0,0]_3(d^{(1)}) \neq [4,0]_3(d^{(2)})$. The design $d^{(1)}$ has less sliced aberration than $d^{(2)}$ because $SA_{31}(d^{(1)}) = SA_{31}(d^{(2)}) = 0$ and $SA_{30}(d^{(1)}) = 0 < 4 = SA_{30}(d^{(2)})$. We will show later that $d^{(1)}$ is a minimum sliced aberration design with six design factors and 32 runs. Here $d^{(2)}$ is a minimum aberration design with 32 runs from [11], which is inferior to a minimum sliced aberration design for a four-platform experiment.

Equipped with a suitable design criterion for our experiment, we are now ready to construct the corresponding minimum sliced aberration designs given in Section 2. Theorem 1 below guides the construction of the minimum sliced aberration designs using readily available minimum aberration designs of fewer numbers of factors.

Theorem 1. *A minimum sliced aberration design as defined above corresponds to a defining relation in which all words are type* 0.

As a result of Theorem 1, constructing a minimum sliced aberration design entails a search among possible designs for which all the words are type 0 in the defining relation. Therefore, minimizing the number of the shortest length in the sliced wordlength pattern of d with 2^{6+2-p} runs is equivalent to minimizing the number of the shortest length in the wordlength pattern of a 2^{6-p} fractional design consisting of design factors only. We use Theorem 1 to generate the minimum sliced aberration design given in Table 6.

The minimum sliced aberration designs in Theorem 1 have a cross array structure similar to product parameter design [10].

3.2 Generalization to a General Number of Factors

The aforementioned theoretical results and the construction method work for the general k number of factors by changing six to k. Following the general case of Theorem 1, constructing a sliced minimum aberration design entails search among possible designs for which all the words are type 0 in the defining relation. Therefore, minimizing the number of the shortest length in the sliced wordlength pattern of d with 2^{k+2-p} runs is equivalent to minimizing the number of the shortest length in the wordlength pattern of a 2^{k-p} fractional design consisting of design factors only. For a four-platform experiment, we use Theorem 1 to provide sliced minimum aberration designs with 16, 32, and 64 runs in Tables 7-9, respectively.

Table 7. Sliced minimum aberration designs with 16 runs, $\mathbf{S} = (s_1, s_2, s_1 s_2)$.

k	Design d	$SW(d)_{i \geq 4}$
3	$S, 1, 2, 12$	$([0,1]_4)$

4. RESULTS

In this section, we discuss the results of the application of our design to the email campaign under consideration.

4.1 Summary of the Design

Using the design criterion in Section 3, we create 2^3 versions to perform the multivariate testing. Each platform has eight versions and we generate them based on the criterion in Section 3 to perform this multivariate testing. This design is a 2^{6+2-3} minimum sliced aberration design for four platforms. By Theorem 1, the sub design for each platform is a 2^{6-3} minimum aberration design with the three generators 124, 135, and 236 [10, p.252]. The sliced word length pattern is $([0,4]_4, [0,3]_5)$. More details of this design were discussed in Section 2.

4.2 Data Display and Summary

The response variable in the study is the email open rate. As the data are aggregated across users exposed to each version, how the response variable varies within a version is unknown to us. We use the Lenth's method [6] to identify significant factors, which is specifically designed for testing effects in experiments for which variance estimates are not available.

Table 10 includes some descriptive statistics of the study. The total number of recipients is 139033, which are divided into roughly equal eight sets receiving the eight versions of the email. Table 11 is a two-way table providing the number of opened emails in each combination of operating systems and email versions.

Table 8. Sliced minimum aberration designs with 32 runs, $\mathbf{S} = (s_1, s_2, s_1 s_2)$.

k	Design d	$SW(d)_{i \geq 4}$
4	S, 1, 2, 3, 123	$([0,0]_4, [0,1]_5)$
5	S, 1, 2, 3, 12, 13	$([0,2]_4, [0,1]_5)$
6	S, 1, 2, 3, 12, 13, 23	$([0,4]_4, [0,3]_5)$
7	S, 1, 2, 3, 12, 13, 23, 123	$([0,7]_4, [0,7]_5, [0,0]_6, [0,0]_7, [0,1]_8)$

Table 9. Sliced minimum aberration designs with 64 runs, $\mathbf{S} = (s_1, s_2, s_1 s_2)$.

k	Design d	$SW(d)_{i \geq 4}$
5	S, 1, 2, 3, 4, 1234	$([0,0]_4, [0,0]_5, [0,1]_6)$
6	S, 1, 2, 3, 4, 123, 124	$([0,0]_4, [0,3]_5)$
7	S, 1, 2, 3, 4, 123, 124, 134	$([0,0]_4, [0,7]_5)$
8	S, 1, 2, 3, 4, 123, 124, 134, 234	$([0,0]_4, [0,14]_5, [0,0]_6, [0,0]_7, [0,0]_8, [0,1]_9)$
9	S, 1, 2, 3, 4, 123, 124, 134, 234, 1234	$([0,4]_4, [0,14]_5, [0,8]_6, [0,0]_7, [0,4]_8, [0,1]_9)$
10	S, 1, 2, 3, 4, 123, 124, 134, 234, 1234, 34	$([0,8]_4, [0,18]_5, [0,16]_6, [0,8]_7, [0,8]_8, [0,5]_9)$
11	S, 1, 2, 3, 4, 123, 124, 134, 234, 1234, 34, 24	$([0,12]_4, [0,26]_5, [0,28]_6, [0,24]_7, [0,20]_8, [0,13]_9, [0,4]_{10})$
12	S, 1, 2, 3, 4, 123, 124, 134, 234, 1234, 34, 24, 14	$([0,16]_4, [0,39]_5, [0,48]_6, [0,48]_7, [0,48]_8, [0,39]_9, [0,16]_{10},$ $[0,0]_{11}, [0,0]_{12}, [0,1]_{13})$
13	S, 1, 2, 3, 4, 123, 124, 134, 234, 1234, 34, 24, 14, 23	$([0,22]_4, [0,55]_5, [0,72]_6, [0,96]_7, [0,116]_8, [0,87]_9, [0,40]_{10},$ $[0,16]_{11}, [0,6]_{12}, [0,1]_{13})$
14	S, 1, 2, 3, 4, 123, 124, 134, 234, 1234, 34, 24, 14, 23, 13	$([0,28]_4, [0,77]_5, [0,112]_6, [0,168]_7, [0,232]_8, [0,203]_9, [0,112]_{10},$ $[0,56]_{11}, [0,28]_{12}, [0,7]_{13})$
15	S, 1, 2, 3, 4, 123, 124, 134, 234, 1234, 34, 24, 14, 23, 13, 12	$([0,35]_4, [0,105]_5, [0,168]_6, [0,280]_7, [0,435]_8, [0,435]_9, [0,280]_{10},$ $[0,168]_{11}, [0,105]_{12}, [0,35]_{13}, [0,0]_{14}, [0,0]_{15}, [0,1]_{16})$

Table 10. Recipients and opened emails in each version.

	V1	V2	V3	V4	V5	V6	V7	V8
Recipients	17295	16920	17452	17306	17362	17558	17582	17558
Opened	1458	1436	1446	1178	1337	1195	1336	1234

Table 11. Two-way frequency tables: operating system vs. version.

	V1	V2	V3	V4	V5	V6	V7	V8
Android	100	61	100	61	75	58	87	55
iOS	243	236	242	172	210	180	204	226
Windows	804	803	766	636	704	637	741	662
macOS	128	141	116	130	139	135	118	100

4.3 Identification of Platform-Specific Significant Effects

Since there are eight versions for each operating system (platform), seven effects of design factors can be estimated per platform. Table 12 includes the aliased effects within each platform. For convenience, we label each set of aliased effects.

In the sliced factorial design framework, slices are used for analyzing data of the four operating systems together. Tables 13 to 16 include the effects of design factors that are estimated using the design in Table 2 on each platform. The Lenth's method is used to test the significance of effects and to report the *p*-values. The same method is done for each operating system to estimate the effect of design factors

within the platform.

Comparing Tables 13-16 indicates that effect $\mathbf{B}$ is significant on P_1 (Android), two effects $\mathbf{B}$ and $\mathbf{F}$ are significant on P_2 (iOS), and effect $\mathbf{D}$ is significant on P_4 (macOS) although no effect is significant on P_3 (Windows). Table 12 reveals that effect $\mathbf{B}$ is the sum of the following aliased effects $\mathbf{2}$, $\mathbf{14}$, $\mathbf{36}$, $\mathbf{345}$, $\mathbf{156}$, $\mathbf{2456}$, $\mathbf{1235}$, $\mathbf{12346}$. As the slices follow the effect hierarchy principle, $\mathbf{B}$ can be viewed to represent effect $\mathbf{2}$ by assuming that all higher-order aliased effects are negligible. The main takeaway for the Android system from Table 13 is that using the direct subject line will likely decrease the open rate and other factors are not expected to decrease or increase the metric. Similar arguments can be made for the other three platforms P_2, P_3, and P_4. For P_2

Table 12. Aliased effects.

Labels	Aliased effects
A	$\mathbf{1 = 24 = 35 = 346 = 256 = 1236 = 1456 = 12345}$
B	$\mathbf{2 = 14 = 36 = 345 = 156 = 2456 = 1235 = 12346}$
C	$\mathbf{3 = 15 = 26 = 245 = 146 = 1234 = 3456 = 12356}$
D	$\mathbf{4 = 12 = 56 = 235 = 136 = 1345 = 2346 = 12456}$
E	$\mathbf{5 = 13 = 46 = 126 = 234 = 1245 = 2356 = 13456}$
F	$\mathbf{6 = 23 = 45 = 134 = 125 = 1246 = 1356 = 23456}$
G	$\mathbf{16 = 34 = 25 = 145 = 246 = 356 = 123 = 123456}$

Table 13. Results for Android.

Effect	Estimate	P-value	
A	2.07e-4	> 0.2	
B	-1.80e-3	0.015	$*$
C	-5.84e-4	0.158	
D	8.13e-5	> 0.2	
E	-3.44e-4	> 0.2	
F	-5.38e-4	0.18	
G	-3.42e-6	> 0.2	

Note: Any *p*-value less than 0.1 is considered statistically significant which is indicated by $*$.

Table 14. Results for iOS.

Effect	Estimate	P-value	
A	1.78e-4	> 0.2	
B	-1.15e-3	0.074	$*$
C	6.03e-4	> 0.2	
D	-5.16e-4	> 0.2	
E	-1.14e-4	> 0.2	
F	-2.71e-3	0.014	$*$
G	-2.68e-4	> 0.2	

Note: Any *p*-value less than 0.1 is considered statistically significant which is indicated by $*$.

Table 15. Results for Windows.

Effect	Estimate	P-value
A	2.07e-3	> 0.2
B	-3.72e-3	> 0.2
C	1.11e-3	> 0.2
D	-2.57e-3	> 0.2
E	-3.60e-3	> 0.2
F	-4.95e-3	0.183
G	-1.51e-3	> 0.2

Note: Any *p*-value less than 0.1 is considered statistically significant which is indicated by $*$.

Table 16. Results for macOS.

Effect	Estimate	P-value	
A	7.76e-5	> 0.2	
B	2.30e-4	> 0.2	
C	-1.17e-5	> 0.2	
D	-1.10e-3	0.061	$*$
E	-3.66e-4	> 0.2	
F	3.46e-4	> 0.2	
G	-6.36e-4	0.195	

Note: Any *p*-value less than 0.1 is considered statistically significant which is indicated by $*$.

from Table 14, using the direct subject line and displaying content in paragraph is expected to decrease the open rate. The remaining factors will not likely impact the metric in any way. The main takeaway for the Windows system from Table 15 is that no factor will likely affect the metric. The main takeaway for the macOS system from Table 16 is that including header image is expected to decrease the open rate and other factors are not expected to have any effect on the response. In summary, a comparison of Tables 13 to 16 indicates that different sets of design factors work differently for these multiple operating systems.

4.4 Calculation of the Factorial Effects for the Multiple Operating Systems

From Table 11, the open rate for Windows is extremely larger than that of other platforms. As no factor will affect the open rate for the Windows system, the operating system might impact the metric. It is important to figure out whether different operating systems have significant interaction with the platform-specific significant effects. To compare the results of the four platforms, the complete design d is used to estimate the slice factor and its interaction with the platform-specific significant effects. The slice factor $\mathbf{S}$ is represented by $\mathbf{S} = (s_1, s_2, s_3)$ with $s_3 = s_1 s_2$. Table 17 describes the relation between P_i's and (s_1, s_2, s_3).

We use Lenth's method to test the significance of the effects. Because the slice factor $\mathbf{S}$ is four-level, the effect $\mathbf{S}$ contains three effects s_1, s_2 and s_3. Take the interaction

Table 17. Relation between P_j and (s_1, s_2, s_3).

s_1	s_2	$s_3 = s_1 s_2$	Platform
−	−	+	P_1
−	+	−	P_2
+	−	−	P_3
+	+	+	P_4

between **S** and **2** for example. It includes three effects $\mathbf{2}s_1$, $\mathbf{2}s_2$, and $\mathbf{2}s_3$. Table 18 includes the 12 effects for **S** and the interactions between the slice factor **S** and three platform-specific significant effects **2**, **4**, and **6**. We now focus on the first three effects s_1, s_2, and s_3 in Table 18. The magnitude of the effect of Android is 2.41e-2, the counterpart of the effect of iOS is 0.79e-2, the counterpart of Windows is 5.01e-2, and the counterpart of macOS is 1.81e-2. The magnitude of the effect of Windows is about two times larger than the effect of Android, six times larger than the effect of iOS, and three times larger than the effect of macOS, which explains why the open rate for Windows is extremely larger than that of other platforms. Further, the magnitude of the effects s_1, s_2, and s_3 are around ten to hundred times larger than the effects of design factors. This finding is consistent with the sliced effect hierarchy principle. The other effects except for s_1, s_2 and s_3 in Table 18 uncover the way the effects **2**, **4**, and **6** interact with the slice factor **S**, respectively, meaning how these effects differentially affect the open rate from P_1 to P_4. Only $\mathbf{6}s_3$ is significant, implying that the differential effect of **6** on the open rate from P_2 and P_3 to P_1 and P_4 is significant. As a result, one version that adopts displaying content in paragraph is expected to decrease the open rate in P_2 and P_3. However, the same version will likely increase the metric in P_1 and P_4.

Table 18. Slice factor behavior.

Effect	Estimate	P-value	
s_1	1.60e-2	< 0.001	∗
s_2	−1.30e-2	< 0.001	∗
s_3	−2.11e-2	< 0.001	∗
$\mathbf{2}s_1$	−1.34e-4	> 0.2	
$\mathbf{2}s_2$	1.15e-3	0.193	
$\mathbf{2}s_3$	8.24e-4	> 0.2	
$\mathbf{4}s_1$	−8.09e-4	> 0.2	
$\mathbf{4}s_2$	2.18e-4	> 0.2	
$\mathbf{4}s_3$	5.17e-4	> 0.2	
$\mathbf{6}s_1$	−3.39e-4	> 0.2	
$\mathbf{6}s_2$	7.82e-4	> 0.2	
$\mathbf{6}s_3$	1.87e-3	0.046	∗

Note: Any *p*-value less than 0.1 is considered statistically significant which is indicated by ∗.

In conclusion, the application of the sliced design in Section 3 to the email campaign shows that different sets of design factors would increase the open rate for each of the four operating systems. From Tables 13 to 16, each design factor significant on the corresponding platform has a negative effect, which means factors **2**, **4**, and **6** should be at the − level. See Table 1 for the information of the design factors. Because we do not know in advance what platform a particular user will use to open the email, it is desirable to choose a version that has factors **2**, **4**, and **6** at the − level. This hypothesis should be tested by looking at the interactions between **S** and the three design factors **2**, **4**, and **6** from the complete design *d*. Table 18 indicates the effect of $\mathbf{6}s_3$ is significant, implying that **6** at the − level is expected to increase the open rate in P_2 and P_3 but would decrease the metric in P_1 and P_4. In order to test this hypothesis, we fit a regression model using the open rate as the response and s_1, s_2, s_3, **2**, **4**, **6**, and $\mathbf{6}s_3$ as the covariates. The average open rate is estimated by

$$\begin{aligned} \text{Average Open Rate} = \\ 0.0163 + 0.0080s_1 - 0.0065s_2 - 0.0105s_3 \\ - 0.0008\mathbf{B} - 0.0005\mathbf{D} - 0.0010\mathbf{F} + 0.0009\mathbf{F}s_3. \end{aligned} \qquad (4.1)$$

We use (4.1) to compare the average open rate between the two versions: design factors **2**, **4**, **6** are at − level and design factors **2**, **4** are at − level but **6** is at + level for each of the four platforms. The result is given in Table 19.

Table 19. Comparison between two versions: design factors **2**, **4**, **6** at − level vs. design factors **2**, **4** at − level but **6** at + level for each of the four platforms.

	Version A $(\mathbf{B}, \mathbf{D}, \mathbf{F}) = (-, -, -)$	Version B $(\mathbf{B}, \mathbf{D}, \mathbf{F}) = (-, -, +)$
P_1	0.00566	0.00556
P_2	0.01556	0.01173
P_3	0.04464	0.04081
P_4	0.00867	0.00858

Table 19 indicates that the average open rate of version A is larger than that of version B for each of the four operating systems. To conclude, we recommend the following changes for the network company to increase the open rate: (i) use the indirect subject line, (ii) drop the header image, and (iii) display content in bullet points. These statistics-guided recommendations can help the network company optimize email campaigns and increase the ROI of its marketing efforts.

5. DISCUSSION AND CONCLUSIONS

Email marketing is a big business (US spend of 2.84 billion dollars in 2020). The average email open rate is 21% and welcome emails have a much higher open rate of about 82%. Many customers seek emails (e.g. for coupons and sales) and targeted and personalized emails are known to have higher open rates (50%) and more effective. [9] find that adding the

name of the message recipient to the email's subject line increased the probability of the recipient opening from 9% to 11% and an increase in sales leads from 0.39% to 0.51%.

We successfully applied a sliced design solution to an industry email campaign with four platforms. This application revealed interesting insights into how different sets of design factors work differently for the four operating systems. We identified the best version for the four platforms. Our statistics-guided recommendations can help the network company and the industry, in general, optimize email campaigns and increase the ROI of marketing efforts.

There are many possible directions for future explorations. It will be of interest to apply the proposed method for marketing campaigns with multiple popular social networking services: Facebook, YouTube, WhatsApp, and Instagram. Another possibility is to apply the method to general multivariate testing problems in industry, such as web design, parameter tuning of deep learning and test and learn programs in insurance and finance.

APPENDIX A. REVIEW OF FACTORIAL DESIGNS AT TWO LEVELS

We provide a brief introduction of two-level factorial designs for readers who are unfamiliar with the topic. The material is adopted by [10] and [8]. *A full factorial design* requires all 2^k level combinations of the k factors. For large k, a fraction of a full factorial design, called a *fractional factorial design* denoted by 2^{k-p}, is often used. In general, 2^{k-p} denotes a $(\frac{1}{2})^p$ fraction of a 2^k factorial design. The optimal fraction can be selected according to resolution and aberration based criteria [10].

To construct a 2^{k-p} fractional factorial design, first let **1**, ..., **k − p** denote the $k − p$ independent columns that generate the 2^{k-p} factorial design. The remaining p columns, **k − p + 1**, ..., **k** can be generated as interactions of the first $k−p$ columns. Choice of these p columns determines the generators and the defining relation of the design. The defining relation of the design consists of the identity element **I** plus the group formed by the p generators (2^{p-1} words in the group). For a 2^{k-p} design, let A_i be the number of words of length i in its defining relation. The *wordlength pattern* of the design is

$$W = (A_3, \ldots, A_k). \tag{A.1}$$

The resolution of a 2^{k-p} design is defined to be the smallest r such that $A_r \geq 1$ which is the length of the shortest word in the defining relation. In general, a design of resolution R is one in which no p-factor effect is aliased with any other effect containing less than $R − p$ factors.

The maximum resolution design is the 2^{k-p} design with the highest resolution. However, resolution is not always enough to select the best design. Consider two 2^{7-2} designs $d_1 : \mathbf{I} = \mathbf{4567} = \mathbf{12346} = \mathbf{12357}$ and $d_2 : \mathbf{I} = \mathbf{1236} = \mathbf{1457} = \mathbf{234567}$. The word **12357** is simply obtained by multiplying the two generators **4567** and **12346**

of d_1. The defining relation of d_2 is obtained by a similar mechanism. The wordlength pattern $W(d_1) = (0, 1, 2, 0, 0)$ is different from $W(d_2) = (0, 2, 0, 1, 0)$ although they both have resolution IV. Since d_1 has one word of length 4, it has three aliased pairs of two-factor interactions (**45 = 67**, **46 = 57**, **47 = 56**). In contrast, d_2 has six aliased pairs of two-factor interactions as it has two words of length 4 (**12 = 36**, **13 = 26**, **16 = 23**, **14 = 57**, **15 = 47**, **17 = 45**). Suppose two 2^{k-p} designs d_g and d_h are to be compared. Let r be the smallest integer such that $A_r(d_g) \neq A_r(d_h)$. Design d_g is said to have less aberration if $A_r(d_g) < A_r(d_h)$. If there is no design with less aberration than d_g, then d_g is called the minimum aberration design [10]. For a given pair of k and p, a minimum aberration design always exists. The minimum aberration criterion can be used to rank any two designs.

APPENDIX B. PROOF OF THEOREM 1

We restate the theorem below.

Theorem. *A minimum sliced aberration design as defined above corresponds to a defining relation in which all words are type* 0.

Proof. It suffices to prove that any defining relation with at least a type 1 word is inferior to a defining relation in which all words are type 0. Any design with a type 1 word has at least one s_j involved in the generators of the design. We prove for the case where one generator uses one s_j. The proof can be easily generalized to the case with more than one generator using s_j's. Consider a design d with 2^{k+2-p} runs that has $p−1$ generators not involving s_j's and one generator g involving s_j. It suffices to prove that a design with all generators excluding s_j's is better according to the minimum sliced aberration criterion. Form a new design d_{new} by removing s_j from g. Call the new generator g_{new}. As s_j only appears in g, the product of g_{new} with other generators will result in type 0 words in the defining relation of d_{new}. Therefore, the length of all type 0 words formed by g_{new} in d_{new} has decreased by one compared with the length of all type 1 words formed by g in d. As a result, all these words formed by g_{new} of d_{new} appear as type 1 words in defining relations of s_j's and are recorded with higher length in the sliced wordlength pattern compared to the ones in d. The lengths of all other words of d_{new} not formed by g_{new} remain the same as the ones in d not formed by g in their sliced wordlength patterns. Therefore, d_{new} is better according to the minimum sliced aberration design. □

ACKNOWLEDGEMENTS

REFERENCES

[1] ADDELMAN, S. Techniques for constructing fractional replicate plans. *Journal of the American Statistical Association* **58**(301) 45–71 (1963). MR0150901

[2] GOLDFARB, A. and TUCKER, C. Online display advertising: Targeting and obtrusiveness. *Marketing Science* **30**(3) 389–404 (2011).

[3] HAIZLER, T. and STEINBERG, D. M. Factorial designs for online experiments. *Technometrics* **63**(1) 1–12 (2021). https://doi.org/10.1080/00401706.2019.1701556. MR4205687

[4] KANNAN, P., REINARTZ, W. and VERHOEF, P. C. (2016). The path to purchase and attribution modeling: Introduction to special section.

[5] LAMBRECHT, A. and TUCKER, C. When does retargeting work? information specificity in online advertising. *Journal of Marketing research* **50**(5) 561–576 (2013).

[6] LENTH, R. V. Quick and easy analysis of unreplicated factorials. *Technometrics* **31**(4) 469–473 (1989). https://doi.org/10.2307/1269997. MR1041567

[7] RYAN, D. *Understanding digital marketing: marketing strategies for engaging the digital generation*. Kogan Page Publishers (2016).

[8] SADEGHI, S., CHIEN, P. and ARORA, N. Sliced designs for multi-platform online experiments. *Technometrics* **62**(3) 387–402 (2020). https://doi.org/10.1080/00401706.2019.1647288. MR4125504

[9] SAHNI, N. S., WHEELER, S. C. and CHINTAGUNTA, P. Personalization in email marketing: The role of noninformative advertising content. *Marketing Science* **37**(2) 236–258 (2018).

[10] WU, C. J. and HAMADA, M. S. *Experiments: Planning, Analysis, and Optimization*, Volume **552**. John Wiley & Sons (2011). MR2583259

[11] WU, C. J. and ZHANG, R. Minimum aberration designs with two-level and four-level factors. *Biometrika* **80**. 203–209 (1993). https://doi.org/10.1093/biomet/80.1.203. MR1225225

Soheil Sadeghi. University of Wisconsin-Madison, Madison, WI, USA. E-mail address: soheil.sadeghy@gmail.com

Tzu-Hsiang Hung. University of Wisconsin-Madison, Madison, WI, USA. E-mail address: thung6@wisc.edu

Peter Chien. University of Wisconsin-Madison, Madison, WI, USA. E-mail address: peter.chien@wisc.edu

Neeraj Arora. University of Wisconsin-Madison, Madison, WI, USA. E-mail address: neeraj.arora@wisc.edu