Statistical Methods in the Search for a Dominant Cause of Variation

by

Mahsa Panahi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2023

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Marcus Perry Professor, Culverhouse College of Business, University of Alabama
Supervisors:	Stefan H. Steiner Professor, Dept. of Statistics and Actuarial Science, University of Waterloo
	Jeroen de Mast Professor, Amsterdam Business School, University of Amsterdam
Internal Members:	Ryan Browne Associate Professor, Dept. of Statistics and Actuarial Science, University of Waterloo
	Nathaniel T. Stevens Assistant Professor, Dept. of Statistics and Actuarial Science, University of Waterloo
Internal-External Member:	Fatma Gzara Professor, Dept. of Management Science and Engineering, University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The chapters of this thesis describe and contain Mahsa Panahi's work, under the supervision of Professors Stefan H. Steiner and Jeroen de Mast, that has been published or submitted in the following peer-reviewed journals.

Chapter 2: Panahi, M., De Mast, J., and Steiner, S. H. (2021). Identifying dominant causes using leveraged study designs. *Quality Engineering*, 33(4):581–593.

Chapter 3: Panahi, M., Steiner, S. H., and De Mast, J. (2023). Identifying the major cause of variation using component swapping. Submitted to *IISE Transactions*.

Chapter 4: Panahi, M., Steiner, S. H., and De Mast, J. (2023). Verifying a dominant cause of output variation. *Quality Engineering*. In press. doi: 10.1080/08982112.2023.2253303.

Abstract

Excessive variation in critical-to-quality characteristics, referred to as process outputs in this thesis, is a common issue in manufacturing industries. Most variation reduction frameworks initially investigate the process to identify the cause(s) of output variation and then seek a solution to eliminate the effect of the identified cause(s). However, among all causes, usually, only a few have a large contribution to the overall variability. The literature refers to them as the dominant cause(s).

Identifying the dominant cause(s) is an effective and recommended initial step in reducing output variation; however, it is often not straightforward. An effective strategy for this step is to employ the method of elimination, i.e., starting with a large number of suspect causes and progressively, after each investigation, eliminating groups of suspects, thereby homing in on the identity of the actual dominant cause(s). Once the dominant cause(s) is identified, verifying it before proceeding with corrective actions is crucial.

Although identifying and verifying a dominant cause is the recommended first step in variation reduction projects, we believe some of the employed statistical tools for these purposes are not the most efficient and lack a thorough scientific analysis. This thesis aims to bridge this gap by proposing study designs and analysis methods that retain valuable ideas from the existing literature but are better suited for the goal of identifying or verifying the dominant causes(s). Our objective is to contribute to the enrichment of the field of statistics in problem-solving and variation reduction, which needs further development.

This thesis is structured in an integrated format, comprising five chapters: the introduction, three papers, and the conclusion.

Chapter 1 is devoted to a literature review and provides some background on some important variation reduction approaches such as the Taguchi method, Six Sigma, the Shainin System™, and the Statistical Engineering algorithm. The focus and main interest of this thesis are on the Statistical Engineering algorithm and the Shainin System™.

Chapter 2 is devoted to a critical examination of *group comparison*, an investigation type often used in the method of elimination to help identify the dominant cause(s). With group comparison, we select two groups of six or more parts, one group consisting of parts with large output characteristic values and the other group consisting of parts with low output characteristic values. For these selected parts, we measure as many input characteristics as possible that are still suspect dominant causes (and possible to determine after observing the output). If an input is a dominant cause, its values must differ substantially between the two groups. The existing analysis procedures frame the group comparison investigation as a hypothesis test, which we demonstrate is unreliable and inefficient. Instead, we frame

the question as an estimation problem based on maximum likelihood. A critical evaluation reveals that our proposed method is superior.

Chapter 3 is devoted to a critical assessment of *component swapping*, another investigation type often used in the method of elimination. The component swapping investigation is applicable when assembled products can be disassembled and reassembled without significant damage. Component swapping consists of a series of studies to determine whether the dominant cause(s) acts within the assembly process or within one or more of the components. It selects two products, one with a high and one with a low output value, and then conducts a two-phased investigation to identify the home of the dominant cause. Although the investigation plan is valuable, we demonstrate that the existing analysis procedures are unreliable. This chapter explores an improved plan and analysis procedure. We proposed a reliable alternative analysis procedure based on maximum likelihood. It also addresses a critical gap in the existing literature by effectively alerting users to possible important interactions either between assembly and components or among individual components.

Chapter 4 investigates how to *verify* a dominant cause effectively. All existing analysis procedures only use a randomized controlled experiment for the verification study. However, we demonstrate that experimental studies alone cannot provide all the required information, and we also require observational studies. This chapter lists some viable composite study designs, assesses their relative merits, and recommends proper sample sizes. It also investigates how to systematically conduct a verification study in the era of smart manufacturing.

In Chapter 5 we conclude and present some directions for future research.

Acknowledgements

I am deeply grateful to my respected supervisors, Prof. Stefan H. Steiner and Prof. Jeroen de Mast, for their constant support and invaluable guidance, without which this academic journey would not have been possible. I would also like to extend my heartfelt appreciation to Prof. R. Jock MacKay, whose generous assistance and insights have consistently enriched my work. Lastly, I must acknowledge the immeasurable support, enthusiasm, and encouragement provided by my husband, Hamed, and my parents, Rasoul and Maryam, throughout this journey.

Dedication

To the love of my life, Hamed Khorrami.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In today's intensely competitive world, companies across the globe are trying to ensure the quality of their services or manufactured products. Although no universal definition exists for quality, the American National Standards Institute and the American Society for Quality Control (1978) define it as "the totality of features and characteristics of a product or service that bears on its ability to satisfy given needs".

According to Juran (1989), ensuring quality in processes involves the execution of three fundamental activities: quality planning, quality control, and quality improvement. Quality planning encompasses activities such as establishing quality targets, developing processes, identifying customers, and determining their needs. Quality control involves implementing operational strategies to maintain quality standards by evaluating actual performance. Meanwhile, quality improvement entails making continuous and systematic beneficial changes to enhance process quality. This thesis primarily focuses on quality improvement, explicitly emphasizing the application of statistical methods for investigating the causes of quality deficiencies.

An important class of quality improvement efforts is reducing the process output variation. This addresses a common challenge industrial engineers face, particularly in contemporary manufacturing, where extremely tight tolerances are imposed on critical-to-quality characteristics such as dimensions. The rationale behind these strict standards is that excessive variation in these quality characteristics can produce defective items (necessitating scrap and rework) and low-quality products, ultimately leading to low overall performances and customer dissatisfaction.

In the context of variation reduction, the established literature recommends following Juran's diagnostic and remedial journeys (Juran and Gryna (1980); Smith (1988); Wagner (1993); Ho and Sculli (1997); MacDuffie (1997)). In other words, we first investigate the process to identify the cause(s) of variation (the diagnosis), and then, we seek a solution to eliminate or mitigate the effect of the identified cause(s) (the remedy). This is a recommended approach since, with knowledge of the important cause(s) of variation, it is usually easier/cheaper to reduce the process output variation in a sustainable way.

Process output variation usually arises from a large number of sources (i.e., process inputs), such as variability in materials, manufacturing conditions, operators, etc. However, among all these causes of variation, typically only a few greatly impact the output variation, while many have only a marginal contribution (De Mast et al. (2019)). Gryna and Juran (1988) called this principle after Pareto, and they referred to them as the "vital few" causes, which are few in number but account for almost all of the overall variation, and the "trivial many" causes, which are large in number, but even their combined contribution is often negligible.

Steiner and MacKay (2005) refer to the one or a few sources of variation with the most substantial impact on the overall variability as the *dominant cause(s)*. Note that we are not primarily interested in identifying statistically significant causes, but rather in identifying the dominant causes. The reason is that with large sample sizes, causes with only small contributions to the output variation may also be statistically significant. These minor causes are typically a mere distraction to practitioners as addressing them is not a very effective way to reduce the output variation.

Identifying the dominant cause(s) of process output variation is often challenging, requiring a systematic approach. The challenge arises from the typically complex and extensive search space, consisting of numerous potential causes, some of which may be inadequately defined or unidentified properties of the process. Traditional strategies, such as brainstorming about suspect dominant causes followed by experiments to establish their effects, can be overwhelming or easily lead to an incorrect search space (Mooren et al. (2012); De Mast et al. (2019)). This thesis focuses on identifying the dominant causes of excessive process output variations (i.e., the diagnostic journey) and assumes that these dominant causes exist.

Diagnosing and remedying variation problems in manufacturing processes is easier when we are systematic. In the following, we provide a concise introduction to some process improvement strategies in the existing literature.

1.2 Background and Literature Review

Four of the most important statistical process improvement strategies in the industrial statistics literature are the Taguchi method (Taguchi (1986); Nair (1992)), Six Sigma's DMAIC method (Hahn et al. (1999, 2000); De Mast and Lokkerbol (2012)), the Shainin System™ (Shainin (1993b); Bhote and Bhote (2000); Steiner et al. (2008a)), and the Statistical Engineering algorithm (Steiner and MacKay (2005)). These strategies have different tools, techniques, and terminologies; however, they have many similarities.

While Statistical Process Monitoring and Control (SPC) charts can be considered a variation reduction tool, this thesis does not regard it as a practical variation reduction strategy. The reason is that SPC is very passive, i.e., we must wait until a control chart signals. Moreover, it does not provide much guidance on corrective actions once a control chart signals.

In the following, we briefly introduce and compare the four strategies. For a comprehensive comparison, see Ledolter and Swersey (1997), Vining and Meyers (1990), De Mast et al. (2000), and De Mast (2004).

1.2.1 The Taguchi Method

Genichi Taguchi, a Japanese quality engineer, introduced a novel perspective, arguing that when a product's quality characteristic deviates from its intended target, even if it falls within the specification limits, there is a substantial loss to society. This loss encompasses various dimensions, including financial costs associated with redesign and rework, delay in project timelines, material wastage, and harm to the company's brand reputation. It also considers the impact on customers, who may experience dissatisfaction due to receiving suboptimal quality products.

Taguchi proposed using an experimental design aimed at optimizing product or process parameters to minimize the loss that a customer is likely to experience and enhance overall quality. Therefore, rather than a quality characteristic itself, he focused on a loss function of the characteristic. Although there are different loss functions in the literature, a standard function for a product with two-sided specification limits is the quadratic loss function

$$L(y) = k(y - t)^2,$$

where $L(y)$ denotes the loss associated with the observed quality characteristic y, the constant k is the quality loss coefficient, and t is the target value. This loss function

illustrates that the minimum loss occurs when the observed quality characteristics precisely matches the target value, and deviation from the target, in either direction, increases the loss, even if y falls within the specification limits.

The experiment proposed by Taguchi distinguishes between control and noise variables. Setting control variables (i.e., fixed inputs) to a desired value is possible and feasible. In contrast, noise variables (i.e., varying inputs) are not controlled, and their values change haphazardly. Taguchi's idea revolves around selecting the settings for control variables (based on the experiment results) such that the process becomes robust against the effect of variation in the noise variables. After that, the method targets the process mean by manipulating control variables affecting the process mean (and not the variance). For further variation reduction, practitioners can use tolerance design. To do so, they narrow the tolerance limit of the important noise variables. This leads to a variation reduction of the noise variables, and as a result, the variation of the quality characteristic decreases.

The Taguchi method exploits the advantages of experimental investigation and recognizes the importance of understanding and managing interaction effects to improve the process. However, instead of understanding the system, Taguchi focuses on finding optimal settings for the process. This is a consequence of studying the loss function instead of defining the problem in terms of a directly measurable characteristic. Moreover, this method picks the optimal variables' combination based on only a *one-shot* experiment (De Mast (2004)). However, design of experiment experts usually recommend starting with a small experiment, and after that, conducting a sequence of experiments based on the findings of the earlier ones (Nair (1992)).

Moreover, Taguchi provides limited guidance for how to *identify* the potential causes. This limited guidance is based primarily on basic tools such as brainstorming, fishbone diagrams, and flowcharting (Ross (1988)). These basic tools list all potential dominant causes based on the engineers' knowledge. As a consequence, the actual dominant cause may be overlooked (Mooren et al. (2012)).

1.2.2 Six Sigma

Six Sigma is a company-wide philosophy aimed at improving quality. Linderman et al. (2003) define it as "an organized and systematic method for strategic process improvement and new product and service development that relies on statistical methods and the scientific method to make dramatic reductions in customer-defined defect rates."

Six Sigma is a comprehensive program that enhances organizational structure and has two primary aspects. The first aspect provides structures and metrics for strategic co-

ordination, which is more relevant to quality management than statistical improvement. The second aspect focuses on significantly improving a defined process, often referred to as the "Breakthrough Cookbook" or the "DMAIC-loop" method (Harry (1997)). In this discussion, we specifically focus on the quality improvement aspect of Six Sigma.

Six Sigma encompasses several variants (e.g., see Harry (1997), Breyfogle (1999), and Pyzdek (2001)), each outlining different descriptions of steps and recommended tools. However, most variants adhere to the DMAIC phase structure, which represents a stepwise strategy involving five phases: Define, Measure, Analyze, Improve, and Control. DMAIC incorporates a comprehensive set of industrial statistics techniques and tools, including design of experiments, control charts, tolerance design, and robust design (De Mast et al. (2000)). It is also possible to apply the Six Sigma improvement strategies to a new process using Six Sigma's DMADV, which stands for Define, Measure, Analysis, Design, and Verify (Pyzdek (2001)).

Six Sigma's DMAIC is one of the most complete statistical improvement strategies. This program strongly emphasizes aligning quality characteristics with customer demands related to quality, delivery, or cost. However, similar to the Taguchi method, it focuses on modelling the effect of candidate causes once identified. It does not offer an efficient and systematic method for *identifying* candidate causes in the first place. Six Sigma employs techniques and tools such as brainstorming, flowcharting, fishbone diagrams, statistically designed experiments, and multivari charts. As a consequence, Six Sigma (similar to the Taguchi method) tends to create too many candidate causes or fail to identify important causes as candidates (Mooren et al. (2012); De Mast and Lokkerbol (2012)). Furthermore, the center of Six Sigma is on experimentation; however, to improve existing processes, we can generate valuable clues about the dominant causes using observational studies.

1.2.3 The Shainin System™

The Shainin System™, introduced by Dorian Shainin, is a coherent stepwise variation reduction strategy with several problem-solving techniques developed for manufacturing environments (Shainin (1993b)). This system revolves around a set of easy-to-apply and easy-to-understand tools without using advanced techniques. It is "[...] developed for and is best suited to problem-solving on operating, medium to high volume processes where data are cheaply available" (Steiner et al. (2008a)). The primary implementation of the Shainin System™ is in parts and assembly operations.

Shainin (1993a) believed that subjective problem-solving strategies such as fishbone diagrams and brainstorming have no place in discovering the potential causes of serious

problems. Instead, he proposes inductive reasoning based on observed process data and their patterns. This approach proves effective, especially when no prior knowledge of potential causes exists. He proposed a sequential approach called *the method of elimination* to search for the dominant cause. This method begins with a large pool of suspect dominant causes, and after each investigation, eliminates groups of suspects, thereby homing in on the identity of the actual dominant cause(s).

The method of elimination consists of a series of often observational tools. Steiner et al. (2008a) provide a critical review of some notable elimination tools within the Shainin System^{TM}, such as Group Comparison, Component Search^{TM}, Variable Search^{TM}, and the Multivari chart. See Bhote and Bhote (2000) for a more comprehensive list. In subsequent chapters of this thesis, we will delve deeper into the first two tools mentioned above. However, in what follows, we discuss some of the limitations of the Shainin System^{TM}.

Shainin (1993b) employs statistically simple tools that usually require relatively small sample sizes. While straightforward and integrative tools are highly recommended, the analysis associated with most of these tools relies on graphical approaches or nonparametric tests, which can be nonintuitive and weak in practical applications (Steiner et al. (2008a)). With modern statistical software readily available, calculation complexity is less of a concern, easily allowing for the seamless integration of standard and more straightforward analyses alongside graphical tools. Furthermore, the Shainin System^{TM} solely concentrates on the diagnosis journey, assuming that the solution is obvious once the dominant cause is identified. We believe this assumption is unrealistic. In addition, in some circumstances, it may not be feasible or cost-effective to identify the dominant cause(s) of output variation, a scenario not addressed by the Shainin System^{TM}.

1.2.4 The Statistical Engineering Algorithm

Shainin (1993b) suggests a valuable system designed to address a variety of variation reduction problems. He deserves credit for adeptly combining known statistical methods into sequential strategies in an understandable and coherent way. However, some of the statistical tools associated with the Shainin System^{TM} are not necessarily the most efficient options (see Steiner et al. (2008a)). In order to use the best elements of the Shainin System^{TM}, Steiner and MacKay (2005) propose an enhanced alternative approach called the Statistical Engineering algorithm. This algorithm, similar to the Shainin System^{TM}, advocates the use of simple-to-understand and easy-to-apply variation reduction techniques. Nevertheless, it enhances certain tools employed by the Shainin System^{TM} by introducing more standard alternative analysis methods. Moreover, unlike the Shainin System^{TM}, the

Statistical Engineering algorithm avoids employing strange hyperbole with trademarks or service marks.

As previously mentioned, the Shainin System™ solely focuses on the diagnostic journey. In contrast, the Statistical Engineering algorithm takes a more comprehensive approach, encompassing both remedial and diagnostic journeys. It offers seven variation reduction approaches that allow improvement: fixing the obvious, desensitization, feedforward control, feedback control, making a process robust, 100% inspection, and moving the process center. The first three approaches require an identified dominant cause, and the latter four are designed to work in cases where identifying a dominant cause is not possible or cost-effective. These different potential remedies represent a unique and advantageous feature of the algorithm, enhancing its efficiency.

1.3 Goal and Outline

In manufacturing industries, many processes suffer from excessive process output variation. While Section 1.2 briefly discussed some widely-used approaches (namely, the Taguchi method, Six Sigma, the Shainin System™, and the Statistical Engineering algorithm), they mainly emerged in practice and in the field by practitioners. Consequently, some of their suggested tools lack a comprehensive scientific analysis.

This thesis aims to bridge this gap by proposing enhanced methods with superior efficiency and reliability. Our objective is to contribute to the enrichment of the field of statistics in problem-solving and variation reduction, which needs further development.

As mentioned, following Juran's diagnostic and remedial journeys is a recommended variation reduction approach. For the diagnostic journey, practitioners typically begin by preparing a list of potential causes, and subsequently, they employ statistically designed experiments to pinpoint the dominant cause(s) of variation among these candidates. Similar to many other statistical techniques, the Taguchi method and Six Sigma offer minimal guidance for the identification of candidate causes. Their search for potential dominant causes is limited to subjective techniques such as brainstorming. Nevertheless, brainstorming is inefficient since it can easily lead to being stuck in the wrong search space or overlooking the actual dominant cause.

This thesis centers on the diagnostic journey (i.e., discovering the potential causes of variation), which has ample room for improvement and requires more attention. Our focus and interest are mainly on the Statistical Engineering algorithm and the Shainin System™. Instead of relying on brainstorming, these strategies use inductive reasoning

from observational data and their patterns to discover the potential causes of output variation. However, some of their proposed statistical tools do not currently represent the most efficient options. Therefore, this thesis aims to identify valuable process improvement ideas currently proposed in conjunction with poorly chosen statistical analysis and integrate and reconcile them with traditional statistical methods.

In Chapter 2, we assess whether the group comparison procedure used by the Statistical Engineering algorithm and the Shainin System™ is reliable. Group comparison is a way to identify the dominant cause(s) of process output variation after using the method of elimination as much as possible so that the list of potential dominant causes is relatively short. Group comparison is an investigation plan that exploits the idea of leveraging, that is, selecting parts with low and high output values. Its requirement is being able to measure the input values later than the output values. Our main finding is that Shainin's group comparison procedure is unreliable and inefficient in identifying the dominant cause. Therefore, we propose a new efficient and reliable analysis procedure based on maximum likelihood. A critical evaluation reveals that our proposed method is superior. We also provide a tangible example and compare the outcomes of both methods.

Chapter 3 is dedicated to a critical assessment of the Component Search™ procedure proposed by the Shainin System™ and the Statistical Engineering algorithm. Component Search™, also referred to as component swapping, is another method of elimination tool that effectively narrows down the search space for the dominant cause of process output variation in an assembly operation. This approach consists of a series of investigations aimed at determining whether the dominant cause operates within the assembly process or within one or more of the product's components. Component swapping is applicable to assembled products where parts can be disassembled and reassembled without any components or subassemblies facing significant change or damage. While the investigation plan based on leveraging is valuable, current analysis procedures are unreliable in identifying the dominant cause(s) due to the utilization of poorly chosen statistical tools. Chapter 3 introduces a superior and reliable alternative analytical approach based on maximum likelihood. Furthermore, it addresses a critical gap in the existing literature by effectively alerting users to interactions between assembly and components or among components.

Chapter 4 centers on the critical task of effectively verifying a dominant cause, a question that has not been properly answered in the existing literature. Both the Statistical Engineering algorithm and the Shainin System™ emphasize the importance of verifying that we have identified the true (dominant) cause(s) before moving on to the remedial journey. In other words, we should make sure that the suspect(s) is not only a *cause* of variation but also *dominant*. This is important because a misidentified (dominant) cause can obstruct the process improvement efforts. While it might initially seem that a ran-

domized controlled experiment is enough to verify a dominant cause, we demonstrate that experimental studies alone cannot provide all the required information. A carefully planned experiment can identify whether a suspect is a *cause* of variation; however, we also require observational studies to investigate whether it is also *dominant*. Chapter 4 proposes some viable composite study designs, assesses their relative merits, and recommends proper sample sizes. It also investigates how to systematically conduct a verification study in the era of smart manufacturing. Moreover, this chapter provides a tangible example of how our proposed procedure works.

Chapter 5 provides a conclusion and some potential future work.

Chapter 2

Identifying Dominant Causes using Group Comparison

Dorian Shainin's group comparison procedure is a way to identify the dominant cause(s) of variation in a process based on an investigation plan that exploits the idea of leveraging. In this chapter, we study whether Shainin's procedure is sound. Our main finding is that Shainin's procedure is unreliable in identifying the dominant cause, and in addition, the procedure is inefficient. We propose a new reliable and efficient analysis procedure based on the method of maximum likelihood estimation. A critical evaluation reveals our proposed method is superior. We also provide a tangible example and compare both methods' outcomes.

2.1 Introduction

One of the tools associated with the Shainin System™ is paired comparison, also called group comparison (Bhote and Bhote (2000)). The goal of group comparison is to help identify suspect dominant causes when we have exploited the method of elimination as much as possible and are conducting investigations focused on considering individual inputs. With group comparison, we select two groups of six or more (often eight) parts, one group (called the "best of the best" (BOB) in Shainin et al. (1997)) consisting of parts with large values for the quality characteristic Y, and the other group consisting of parts with low values for Y (called the "worst of the worst" (WOW) in Shainin et al. (1997)). For the parts in both groups, we measure as many input characteristics as possible that

are still suspect dominant causes. Note that in many variation reduction problems, the terminology of BOB and WOW is confusing since it may not be the case that higher and lower output values imply best or worse parts. For this reason, in what follows, we use the neutral terms "Upper" and "Lower" to refer to the two groups.

Group comparison, like many other of Shainin's tools, uses the idea of leveraging, that is comparing extremes (Bhote and Bhote (2000)) as defined by the "Upper" and "Lower" groups. Leveraging is an excellent way of gaining information in the search for a dominant cause with relatively small sample sizes. Leveraging works because if a candidate is a dominant cause, its value must substantially differ between the "Upper" and "Lower" groups. Note that it is somewhat misleading to claim that a group comparison investigation requires measurements from only a small sample of parts (typically 12 to 16) with extreme values. To select these extreme parts, we must measure Y for a large number of parts. We call this the baseline data.

The purpose of this chapter is to provide a critical assessment of group comparison as advocated by Shainin (1993b) and Bhote and Bhote (2000), and to propose an improved analysis procedure. We are interested to learn whether Shainin's group comparison method is reliable and effective, and whether it can be improved. Moreover, we believe that the study-design principle of leveraging may have broader applicability and could possibly be a valuable addition to the literature on the design of statistical studies (Browne et al. (2009b, 2010a)). Therefore, we are also interested to learn how well leveraging works in the group comparison procedure.

In Section 2.2, we provide an example to explain the group comparison procedure and provide a critical assessment. We demonstrate that the investigation plan is valuable, but the proposed analysis is less than ideal. To address these limitations, in the next two sections, we propose a new efficient and reliable analysis procedure based on maximum likelihood estimation, and we assess its merits. Note that our proposed procedure retains the valuable ideas of leveraging and the elimination strategy but alters the suggested analysis for the group comparison investigation. In Section 2.3, we consider the case of a continuous input X and use simulation studies to compare the new analysis approach with the previously promoted approach. Section 2.4 is similar to Section 2.3 but addresses binary inputs. In Section 2.5, we apply the new procedure to the example from Section 2.2 and compare the results with the ones from Bhote and Bhote (2000) and Shainin (1993b) procedure. We provide a summary and discussion in Section 2.6.

2.2 Shainin and Bhote's Group Comparison Procedure

In this section, we describe the group comparison procedure as presented by Shainin (1993b), and Bhote and Bhote (2000). We also provide an example to demonstrate how group comparison works. Note that while the example is based on a real case, we have changed many details to fit the expository purpose of this chapter better. Next, we formulate points of criticism of Shainin's group comparison procedure and identify where the procedure can be improved.

2.2.1 Example

Mooren et al. (2012) describe an example in a machining process for large casted metal parts, where the monitoring system frequently stops the process due to the imminent drill break signals. Once the line stops, suspect drill bits will be replaced. The frequent out-of-control signals lead to excessive line downtime and the replacement of drill bits far before their specified lifespan. The engineers found that the line stoppages were triggered by high peaks in the drill bits' torque. Since none of the analyzed drill bits showed any sign of wear-out, the large variation in torque was caused by another phenomenon. At first, the engineers tried to identify the cause(s) of significant torque variation using a brainstorming session, planning to use an experiment to identify the primary cause(s). However, since the engineers and operators alike did not have the faintest idea about the cause(s), the result of the brainstorming was a long list of hunches, most of which were too unspecific to test in an experiment. The engineers concluded that they needed a more systematic search strategy. To begin with, they defined a scale for rating the severity of the torque peaks, ranging from $Y = 1.0$ (no peaks in torque visible) to $Y = 5.0$ (high peaks in torque). Note that this scale is continuous.

A group comparison study brought a breakthrough. The engineers observed the process over four days, during which 52 drill bits were used as the baseline, and they followed the procedure described in Bhote and Bhote (2000). Based on this procedure, eight drill bits with the highest torque peaks, called the "Upper" group, and eight drill bits with the least visible torque peaks, called the "Lower" group, were selected. Then, the sixteen selected drill bits were divided into eight pairs, each consisting of one drill from the "Upper" and one drill from the "Lower" group. The most extreme drill bits with the highest and lowest quality characteristic values are denoted Pair 1, the drill bits with the second highest and second lowest quality characteristic values are Pair 2, and so on. The objective of a group

comparison is to identify the process inputs (X's) that could explain the difference in torque behavior in the "Upper" and "Lower" groups of drill bits. To identify such X's, the engineers checked all drill bits on several characteristics. Some of them were measured on a continuous scale (e.g., dimension), while others were visual characteristics judged on a two- or three-point scale. Table 2.1 shows how the ten candidate X's were measured or judged for both drill bits in each of the eight pairs.

To decide which of the candidate X's appear to explain the differences in torque behavior, Bhote and Bhote (2000) and Shainin (1993b) use the Tukey end-count procedure to compare X values in the "Lower" and "Upper" groups. The end-count procedure is a nonparametric test for comparing two samples introduced by Tukey (1959), which we explain in more detail in Section 2.2.2. The end-count values for each X are given at the bottom of Table 2.1.

The test finds that the *Cutting Edge Appearance* and *Sagging* values are significantly different for the "Upper" and "Lower" groups. Therefore, these candidates are identified as dominant causes of variation in torque behaviour. The end-count test finds no significant differences for the other eight candidate X's, thus ruling these out for further consideration. Note that pairing does not influence the calculated end-count. We discuss the rationale for pairing in Section 2.2.2.

2.2.2 Group Comparison Procedure

The example illustrates the goal and use of a group comparison investigation. Given that there is excessive variation in a quality characteristic Y, the purpose is to identify which (if any) out of a number of candidate causes is a dominant cause of variation. Group comparison starts from a baseline sample of n_b randomly selected parts, for which the quality measurements Y_i, $i = 1, \ldots, n_b$ are given. The obvious way to look for a dominant cause would be to measure all candidate causes for all n_b parts in the baseline sample. Group comparison, however, instead exploits the idea of *leveraging*, which means we measure the X's only for those parts having extreme output values, as these parts are likely to contain the most information about the dominant cause. The "Lower" and "Upper" groups of parts are defined as follows:

- Let the "Lower" group $\mathcal{L}$ be the n_l parts having the smallest output values, that is, $\mathcal{L} = \{(1), (2), \ldots, (n_l)\}$, where $(1), (2), \ldots$ represent the parts having the smallest output value, second smallest output value, and so on.

Quality Characteristics (Y)			Candidate (X's)									
			Top Angle	Side Angle	Sagging	Dimension A	Dimension B	Width	Diameter at Point P	Stains Near Top	Discoloration	Cutting Edge Appearance
Pair 1	Upper	5.0	125	78	0.11	7	34	8	467	Yes	Mild	Artifact
	Lower	1.0	130	80	0.12	7	30	9	360	No	No	OK
Pair 2	Upper	4.9	110	80	0.14	8	33	7	404	No	Yes	Artifact
	Lower	1.2	105	80	0.09	5	35	8	460	No	Yes	OK
Pair 3	Upper	4.8	110	80	0.12	7	33	9	432	No	No	Artifact
	Lower	1.2	115	75	0.09	8	32	8	564	Yes	No	OK
Pair 4	Upper	4.7	130	85	0.11	9	30	8	476	No	No	Artifact
	Lower	1.3	115	75	0.13	5	31	9	454	No	Mild	OK
Pair 5	Upper	4.7	125	80	0.12	6	30	8	576	No	No	Artifact
	Lower	1.4	120	75	0.09	7	31	8	489	No	Yes	OK
Pair 6	Upper	4.6	115	80	0.14	7	34	8	372	No	Mild	Artifact
	Lower	1.5	120	88	0.10	5	32	7	447	Yes	No	OK
Pair 7	Upper	4.5	110	78	0.12	6	33	9	471	No	Yes	Artifact
	Lower	1.5	135	80	0.10	8	33	8	503	Yes	Mild	OK
Pair 8	Upper	4.5	125	80	0.12	7	34	8	453	No	Mild	OK
	Lower	1.6	115	88	0.09	9	31	8	422	Yes	No	Artifact
Tukey end-count			2	5	8	4	2	3	2	3	3 [a]	14

Table 2.1: Results of group comparison of eight "Upper" and eight "Lower" drill bits.

[a] It is not clear how to extend the end-count method to an ordinal input with three levels. However, based on the observed results, it does not appear that *Discoloration* is an important input.

- Let the "Upper" group $\mathcal{U}$ be the n_u parts having the largest output values, that is, $\mathcal{U} = \{(n_b - n_u + 1), (n_b - n_u + 2), \ldots, (n_b)\}$.

With this plan, group comparison needs to measure the X's for only $n_l + n_u (< n_b)$ parts. Therefore, leveraging is beneficial when measuring suspect dominant causes is costly or time-consuming, as is usually the case. However, the necessary condition for this approach is the possibility of measuring the X values at a later stage than the Y values. This condition would not be met if, for example, a process temperature is a candidate X, unless the process temperatures were logged at the production time.

Bhote and Bhote (2000) and Shainin (1993b) describe the procedure as a *paired* comparison procedure. As illustrated in Table 2.1, parts are studied in pairs, each consisting of a part from the "Upper" and a part from the "Lower" group. Although the pairing is often arbitrary, we see the potential value of comparing paired upper and lower parts one against the other, as this may stimulate the identification of new candidate X's that would not have been considered before. As a matter of fact, in the example, it was such a direct visual comparison of bad drill bits against good drill bits that made the team aware of the artifacts in the *Cutting Edges Appearance*, as the differences stood out.

In the group comparison procedure, the end-count is calculated for each of the candidate X's separately, as presented in Table 2.1. To do so, parts should be ranked by their input values from smallest to largest, regardless of whether the part is from the "Upper" or "Lower" group. For example, the first candidate cause, *Top Angle*, gives 105 $(L; 1.2)$, 110 $(U; 4.5)$, 110 $(U; 4.8)$, 110 $(U; 4.9)$, 115 $(L; 1.2)$, 115 $(L; 1.3)$, 115 $(L; 1.6)$, 115 $(U; 4.6)$, 120 $(L; 1.4)$, 120 $(L; 1.5)$, 125 $(U; 4.5)$, 125 $(U; 4.7)$, 125 $(U; 5.0)$, 130 $(L; 1.0)$, 130 $(U; 4.7)$, 135 $(L; 1.5)$, with U and L indicating parts' membership in the "Upper" and "Lower" groups and the parts' output values given in parentheses. The end-count is the number of values in the one group exceeding all values in the other, plus the number of values in the other group falling below all those in the first group, where neither value can be zero (Tukey (1959)). For example, the end-count for *Top Angle* is $1 + 1 = 2$. Bhote and Bhote (2000) offer a table specifying critical values for the end-count needed to achieve certain levels of significance for testing whether X has an effect on Y. For $n_l = n_u = 8$, end-counts of 7, 10, and 13 are significant at the confidence levels of 95%, 99%, and 99.9%, respectively. Thus, an end-count of two is not significant, and we conclude that *Top Angle* is not a dominant cause.

On the other hand, *Cutting Edge Appearance* has a sorted sequence of OK $(L; 1.0)$, OK $(L; 1.2)$, OK $(L; 1.2)$, OK $(L; 1.3)$, OK $(L; 1.4)$, OK $(L; 1.5)$, OK $(L; 1.5)$, Artifact $(L; 1.6)$, OK $(U; 4.5)$, Artifact $(U; 4.5)$, Artifact $(U; 4.6)$, Artifact $(U; 4.7)$, Artifact $(U; 4.7)$, Artifact $(U; 4.8)$, Artifact $(U; 4.9)$, and Artifact $(U; 5.0)$, with an end-count of 14, which is significant

at the 99.9% level. Tukey (1959) does not explain how to deal with binary inputs, so we always sort the sequence using the output values that gives us the maximum end-count. From Table 2.1, the second highest end-count is for *Sagging*. Since the end-count of eight is significant at the 95% level, *Sagging* is also identified as a dominant cause. The end-count values for all the other characteristics are less than seven, meaning that they are not significant causes of variation. Therefore, the procedure identifies *Cutting Edge Appearance* and *Sagging* as dominant causes.

2.2.3 Critique of Group Comparison

The first criticism is that the implementation of the end-count method does not work well for discrete inputs, particularly when the quality characteristic Y is discrete as well. One issue is that there are many ties in the rank order when the input is binary. Moreover, it is unclear how to extend the end-count analysis to the case where the inputs are categorical variables with more than two values.

A more substantial criticism concerns the efficiency of the procedure. The analysis based on end-count values ignores all the observed output values in the baseline for parts not included in the "Upper" and "Lower" groups. As mentioned before, to select the extreme parts, we need to determine the Y values for a large number n_b of parts. Thus, it is wasteful to use the data for only the extreme parts. Moreover, Bhote and Bhote (2000) and Shainin (1993b) procedure discards observed output values for the extreme parts, and instead, only uses the "Upper" and "Lower" group membership information. It is inefficient not to use the continuous Y values and instead, only use the dichotomized "Upper" and "Lower" labels.

However, the most important point of criticism is that the described procedure frames the problem as *whether* a candidate X affects Y, whereas the question should be whether X is a *dominant* cause of variation in Y. In the following, we demonstrate that the Bhote and Bhote (2000) and Shainin (1993b) end-count-based analysis, which is supposed to identify *dominant* causes of variation, is likely to identify minor causes. This issue becomes more likely as the baseline sample size increases.

To quantify this claim and allow for a more precise discussion of these ideas, assume that the effect of a candidate cause X on the quality characteristic Y is approximately linear with intercept α and slope β, i.e.,

$$Y = \alpha + \beta X + \epsilon, \tag{2.1}$$

where the error, denoted ϵ, has a mean zero and includes noise (e.g., measurement variation) and the effects of other causes. Denoting $Var(X) = \sigma_X^2$, $Var(\epsilon) = \sigma_\epsilon^2$, and assuming X and ϵ are independent, we have $\sigma_Y^2 = Var(Y) = \beta^2\sigma_X^2 + \sigma_\epsilon^2$. Then, X is a large cause of variation if holding it fixed would substantially reduce the output variation σ_Y^2. A candidate cause X is strictly a dominant cause of variation only if $\beta^2\sigma_X^2 > \sigma_\epsilon^2$, as this implies that the variation transmitted to Y from X is larger than the variation due to noise and other causes. However, we are likely interested in identifying any cause with a large effect. In what follows, it is convenient to consider the squared correlation between X and Y given by

$$\rho^2 = \frac{\beta^2\sigma_X^2}{\beta^2\sigma_X^2 + \sigma_\epsilon^2}, \tag{2.2}$$

where $0 \leq \rho^2 \leq 1$ (the required calculations are available in Appendix A.1). Using this parameterization, X is strictly a dominant cause if $\rho^2 > 0.5$, since then $\beta^2\sigma_X^2 > \sigma_\epsilon^2$. Note that ρ^2 is the population-parameter version of the R^2 value (coefficient of determination) of a linear regression fit of Y on X.

In the following, for simplicity, we add a normality assumption for X and ϵ in Model 2.1, i.e., $X \sim N(\mu_X, \sigma_X^2)$ and $\epsilon \sim N(0, \sigma_\epsilon^2)$. To illustrate the issue, we calculate the probability of having end-count values equal to, or greater than, ten (the 99% significance level). For the simulation, we select $n_l = n_u = 8$, which is a common choice in the group comparison investigation (Shainin (1993b); Bhote and Bhote (2000)). Also, without loss of generality, we simulate data with $\alpha = 0$, $\sigma_\epsilon = 1$, $\mu_X = 0$, and $\sigma_X = 1$. We also set β so that $\rho^2 \in \{0.0, 0.1, \ldots, 0.5\}$. Figure 2.1 compares the simulation results for different baseline sample sizes ($n_b \in \{50, 100, 400, 1000, 5000\}$) using 5000 simulation runs.

Figure 2.1 reveals that the analysis based on the end-count test (with $n_l = n_u = 8$) depends strongly on the baseline sample size. For a given value of ρ^2, as n_b increases, the probability of observing end-counts higher than ten increases. This increase is especially clear when $0.1 \leq \rho^2 \leq 0.3$. However, note that X is actually not a dominant cause in this range since $\rho^2 < 0.5$. Therefore, Figure 2.1 nicely reveals the problem that even when X is not a dominant cause, with a large n_b, the end-count method often will mistakenly conclude that it is. This is undesirable for many reasons as it may lead us in many unproductive directions and waste resources in future studies. We ran the same simulation study for different n_l and n_u sizes. Increasing n_l and n_u, the relative positions of the lines for a fixed ρ^2 do not change, but the probability of observing a statistically significant result reaches one faster, and the same problem still exists. Note that for Figure 2.1, we could have used a different threshold for the end-count values, like 7 or 13 instead of 10, but the qualitative conclusions remain the same.

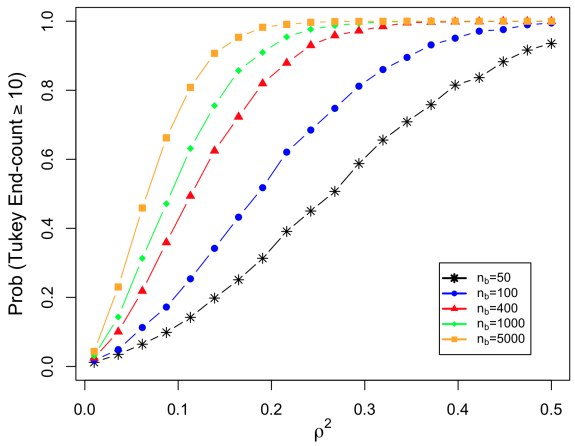

Figure 2.1: Probability of having Tukey end-count equal to or greater than ten vs. ρ^2 for different baseline sample sizes with $n_l = n_u = 8$.

In summary, as Figure 2.1 illustrates, the effect of misidentifying a non-dominant cause as the dominant cause is even more pronounced for larger baseline sample sizes. Although X's with $\rho^2 \leq 0.1$ do not have a big contribution to the output variation, if n_b is large, the analysis based on the end-count test might misidentify them as interesting, spurring subsequent effort.

Identifying the dominant cause(s) involves estimating the relative size of the effects of X's onto Y. Therefore, we propose that the problem should be framed as an estimation problem, which is more informative, rather than as a hypothesis-testing problem on ρ^2. The next section outlines our procedure for estimating ρ^2.

2.3 Proposed Group Comparison Analysis Procedure for Continuous Inputs

In this section, we first explain our proposed group comparison analysis procedure when X is continuous. Then, we compare the proposed approach with Shainin's procedure via simulation. Finally, we evaluate the performance of our procedure in comparison with two competing study plans. The case where inputs are binary is described in Section 2.4.

2.3.1 Statistical Modelling and Proposed Procedure

Here, the statistical model and assumptions are the same as in Section 2.2.3. For all n_b parts in the baseline, we observe the quality characteristic values $y_i, i = 1, 2, \ldots, n_b$. For the parts in the "Upper" and "Lower" groups, we have, in addition, the input values. Thus, our available data are:

- y_i for $i \in \{1, \ldots, n_b\} - \{\mathcal{U} \cup \mathcal{L}\}$
- (y_i, x_i) for $i \in \{\mathcal{U} \cup \mathcal{L}\}$

Corresponding to this data and the assumed model, the log-likelihood function is

$$\begin{aligned}
l &= \log\big(P(X_i = x_i \text{ for } i \in \{\mathcal{U} \cup \mathcal{L}\}; Y_1 = y_1, \ldots, Y_{n_b} = y_{n_b})\big) \\
&= \log\big(P(X_i = x_i | Y_i = y_i \text{ for } i \in \{\mathcal{U} \cup \mathcal{L}\}) \times P(Y_1 = y_1, \ldots, Y_{n_b} = y_{n_b})\big) \\
&= \log\Big(\prod_{i \in \{\mathcal{U} \cup \mathcal{L}\}} P(X_i = x_i | Y_i = y_i) \times \prod_{i=1}^{n_b} P(Y_i = y_i)\Big) \\
&= \log\Big(\prod_{i \in \{\mathcal{U} \cup \mathcal{L}\}} P(X_i = x_i | Y_i = y_i)\Big) + \log\Big(\prod_{i=1}^{n_b} P(Y_i = y_i)\Big) \\
&= \sum_{i \in \{\mathcal{U} \cup \mathcal{L}\}} \log\big(P(X_i = x_i | Y_i = y_i)\big) + \sum_{i=1}^{n_b} \log\big(P(Y_i = y_i)\big), \qquad (2.3)
\end{aligned}$$

with

$$Y_i \sim N(\alpha + \beta\,\mu_{_X}, \beta^2\sigma_X^2 + \sigma_\epsilon^2).$$

Based on our assumptions, (X_i, Y_i) have the following bivariate normal distribution

$$BVN\left(\begin{bmatrix} \mu_{_X} \\ \alpha + \beta\,\mu_{_X} \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \rho\,\sigma_{_X}\sqrt{\beta^2\sigma_X^2 + \sigma_\epsilon^2} \\ \rho\,\sigma_{_X}\sqrt{\beta^2\sigma_X^2 + \sigma_\epsilon^2} & \beta^2\sigma_X^2 + \sigma_\epsilon^2 \end{bmatrix}\right),$$

and $(X_i | Y_i = y_i)$ follows a normal distribution with $E\,(X_i | Y_i = y_i) = \mu_{_X} + \frac{\sigma_{_X}}{\sqrt{\beta^2\sigma_X^2 + \sigma_\epsilon^2}}\,\rho\,(y_i - \alpha - \beta\mu_{_X})$ and $Var(X_i | Y_i = y_i) = \sigma_X^2(1 - \rho^2)$ (Rao (1973)). We suggest using maximum likelihood to estimate the model parameters α, β, $\mu_{_X}$, σ_X^2, and σ_ϵ^2. Finally, an estimate for ρ^2 is obtained by plugging the estimated model parameters into Equation 2.2.

2.3.2 Comparison of the Proposed Procedure to Shainin's Procedure

Figure 2.1 demonstrates that Shainin's end-count-based procedure does not take into account the baseline sample size and, in some cases, has a large probability of misidentifying a cause as dominant even when it is not. Our proposed analysis procedure frames group comparison as an estimation problem rather than a hypothesis test. If desired, we could translate our estimates into decisions about whether we believe a given candidate cause is dominant or not in a variety of ways. One option is to simply compare our point estimate for ρ^2 to a given value, say 0.5 (or any other threshold). A more sophisticated alternative is to make a decision considering the confidence interval for our estimate. Note that we discuss confidence intervals for our estimates in Section 2.5. We leave it up to practitioners to interpret the procedure's estimates (and uncertainty bounds) in a way that makes sense in their context (this could depend on other factors such as cost).

To be able to directly compare our proposed procedure to Shainin's one, we have to add some decision rules to our procedure. Therefore, for illustration, we select a decision threshold of 0.26, i.e., if $\hat{\rho}^2 > 0.26$, we conclude the candidate cause is *dominant*. We select this low threshold intentionally to ensure that for the case of $n_l = n_u = 8$ and $n_b = 100$, the probability of identifying a candidate cause with an actual value of $\rho^2 = 0.5$ as the dominant cause equals 0.994 for both our proposed and the end-count procedures. This way, we can directly and fairly compare Shainin's procedure and our maximum likelihood estimation-based approach. Note that we would likely not recommend such a low threshold in applications. Figure 2.2 compares the results.

Figure 2.2 demonstrates the enhanced reliability of our procedure, as it is less likely to misidentify non-dominant causes as dominant. For example, when $\rho^2 = 0.25$, the probability of misidentification is 72% for Shainin's group comparison procedure and 58% for our proposed procedure. While not shown here, recall that our proposed procedure considers the baseline sample size in the analysis. So, we will not see arbitrarily large probabilities of concluding a non-dominant cause is dominant, as we see in Figure 2.1 for the end-count analysis, even if the baseline size is large.

2.3.3 Evaluation of the Proposed Procedure for Continuous Inputs

Here, we demonstrate the efficiency of leveraging in our estimation-based procedure through a comparison of two competing study plans. The main goal is to illustrate the

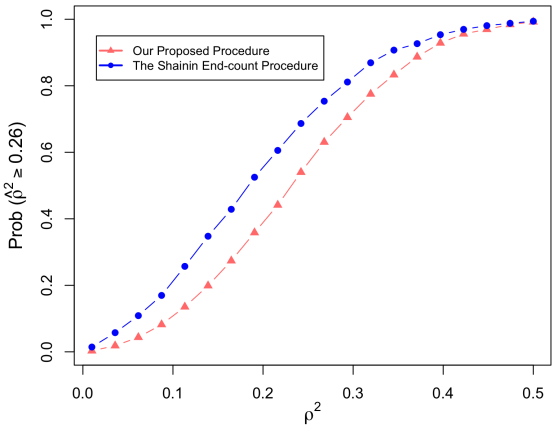

Figure 2.2: Probability of having $\hat{\rho}^2 \geq 0.26$ vs. ρ^2 for our proposed procedure and Shainin's procedure when $n_l = n_u = 8$ and $n_b = 100$.

value of selecting extreme parts (i.e., leveraging). For the proposed plan, we measure the X values only for the $n_l + n_u$ baseline parts with *extreme* output values, and we call this the "Leveraged" plan. Alternatively, we could measure the X values for all baseline parts. Then, the data are (y_i, x_i) for all n_b parts, and we call this the "Full" plan. A third approach is to measure the X values for $n_l + n_u$ *randomly* selected parts from the baseline, and we call this the "Random" plan.

Via simulation, we compare the three plans by calculating their standard deviation of $\hat{\rho}^2$, denoted by $SD(\hat{\rho}^2)$, for $\rho^2 \in \{0, 0.1, 0.2, \ldots, 0.9\}$. Note that we have five parameters in Model 2.1, namely α, β, σ_ϵ, μ_X, and σ_X, but ρ^2 depends only on β, σ_ϵ and σ_X. Without loss of generality, we simulate data with $\mu_X = 0$, $\sigma_X = 1$, and $\alpha = 0$. Also, we fix $\sigma_\epsilon = 1$ and determine the corresponding β so that $\rho^2 \in \{0, 0.1, 0.2, \ldots, 0.9\}$. In each simulation run, we estimate all model parameters (α, β, σ_ϵ, μ_X, and σ_X) by maximizing the likelihood, and we find the corresponding realization of $\hat{\rho}^2$. Then, via 1000 simulation runs, we estimate $SD(\hat{\rho}^2)$ for all nine possible combinations of $n_b \in \{100, 400, 1000\}$ and $n_l = n_u \in \{5, 8, 16\}$. Figure 2.3 presents the results.

As we would expect, Figure 2.3 demonstrates that the "Full" plan gives the lowest $SD(\hat{\rho}^2)$ for all examined combinations. It is also clear that the "Leveraged" plan always lies about halfway between the "Random" and "Full" plans. This reveals that leveraging is a valuable idea, as measuring the X values for the $n_l + n_u$ *extreme* parts yields a more precise estimate of ρ^2 than measuring the X values for $n_l + n_u$ *randomly* selected parts. Recall that while the "Leveraged" plan is not better than the "Full" plan, it requires much less effort since we only measure the X values for a small number of parts. Moreover,

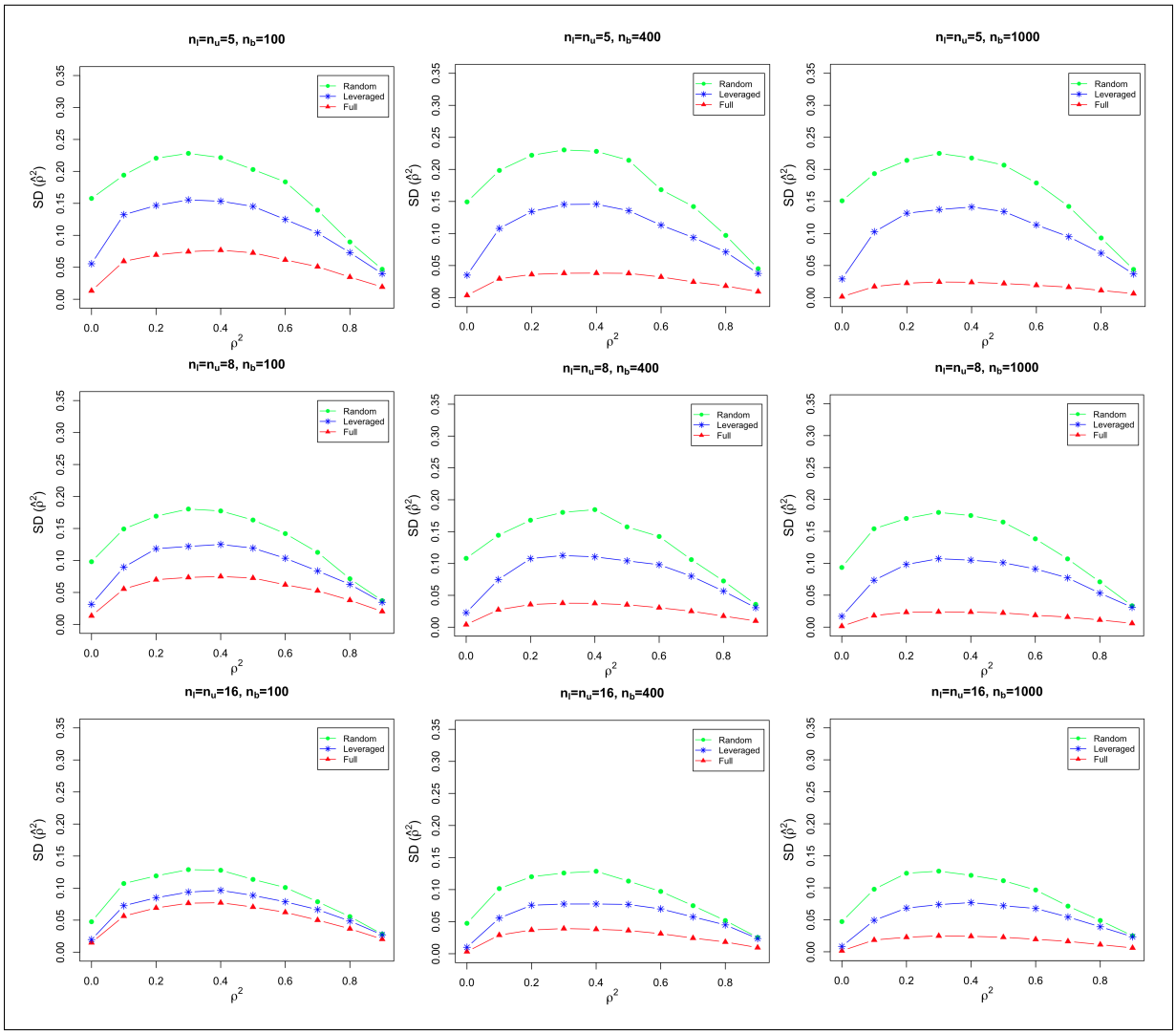

Figure 2.3: $SD(\hat{\rho}^2)$ vs. ρ^2 for $n_l = n_u \in \{5, 8, 16\}$ and $n_b \in \{100, 400, 1000\}$ for all "Full", "Leveraged", and "Random" plans when inputs are continuous.

$SD(\hat{\rho}^2)$ decreases as n_l, n_u, and n_b increase. However, n_l and n_u have a larger contribution in this decrease than n_b. Figure 2.3 also reveals that increasing n_b does not change $SD(\hat{\rho}^2)$ considerably. Thus, if it is an effort to increase the baseline sample, it is not worth it.

To give an idea about suitable sample sizes, we discuss the situation where $n_b = 400$. Our focus is on the situation that $\rho^2 \approx 0.5$ since we are mainly interested in estimating

the effects of X's with ρ^2 close to 0.5 or higher. Increasing n_l and n_u from five to eight reduces $SD(\hat{\rho}^2)$ from approximately 0.136 to 0.104, which can be beneficial in many cases. However, increasing the sample size further to $n_l = n_u = 16$, i.e., a further doubling of the amount of effort, yields only marginal improvement in $SD(\hat{\rho}^2)$.

In conclusion, unlike the end-count analysis, which is framed as a hypothesis test, our proposed likelihood-based analysis estimates the magnitude of the effect of X onto Y. Thus, to identify a dominant cause, our proposed estimation-based approach is preferred. Compared to the end-count analysis, this new approach is less likely to misidentify a non-dominant cause as the dominant cause.

2.4 Proposed Group Comparison Analysis Procedure for Binary Inputs

In practical applications, there may be a combination of continuous and binary candidate inputs. While a standard Tukey end-count test, as described by Bhote and Bhote (2000), is not applicable to discrete inputs, we can reverse the process by considering the actual values of the quality characteristic, as briefly explained for *Top Angle* in Section 2.2.2. Instead of the end-count procedure, this section proposes a maximum-likelihood method similar to Section 2.3.1 for binary inputs. To distinguish the type of inputs, we denote binary inputs with X^*. Note that there are also other possible extensions that one could consider, including the case where the inputs are ordinal.

2.4.1 Statistical Modelling and Proposed Procedure for Binary Inputs

Assume that the effect of the binary input X^* onto the continuous quality characteristic Y is approximately linear given by

$$Y = \alpha + \beta X^* + \epsilon,$$

where X^* and ϵ are independent, and the distribution of X^* is

$$X^* = \begin{cases} -1 & \text{with probability } q, \\ +1 & \text{with probability } 1-q, \end{cases}$$

where $0 < q < 1$. Note that in some applications, we may assume q is known. For example, suppose X^* represents two parallel processing streams. Then, due to equal volumes in the two streams, $q \approx 0.5$.

To make the conclusion more general, we continue with the assumption that q is unknown. Given that setup, $Var(X^*) = 4\,q\,(1-q)$ and $Var(Y) = \beta^2\,(4\,q)(1-q)+\sigma_\epsilon^2$. Similar to the continuous case, it is convenient to consider the squared correlation between Y and X^* given by

$$\rho^{*2} = \frac{\beta^2(4q)(1-q)}{\beta^2(4q)(1-q)+\sigma_\epsilon^2}, \tag{2.4}$$

where $0 \leq \rho^{2^*} \leq 1$ (the calculations are similar to those in Appendix A.1, but with slight modifications). Using this parameterization, X^* is strictly a dominant cause if $\rho^{*2} > 0.5$.

Based on the assumptions, the marginal distribution of Y is

$$P(Y_i = y_i) = q\;P(Y_i = y_i|X_i^* = -1) + (1-q)\;P(Y_i = y_i|X_i^* = +1), \tag{2.5}$$

with $(Y_i|X_i^* = \pm 1) \sim N(\alpha \pm \beta, \sigma_\epsilon^2)$.

To estimate the parameters (namely, α, β, σ_ϵ, and q), we select a subsample of size $n_l + n_u$ from the baseline and determine the corresponding values of x^*. Note that the conditional distribution of $(X^*|Y = y)$ only depends on the y value and not on how the part is selected.

Similar to the continuous case, the likelihood has two components, one comes from the baseline data, and the other comes from the selected parts with measured x^* values. Therefore, the log-likelihood is the same as in Equation 2.3, but with four parameters (α, β, σ_ϵ, and q) instead of five, and X^* instead of X. The conditional distribution of $(X^*|Y = y)$ and the required calculations to achieve the log-likelihood are available in Appendix A.2.

Given data, we suggest using maximum likelihood to derive the estimates $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}_\epsilon^2$, and $\hat{q}$. Finally, an estimate for ρ^{*2} is obtained by plugging the estimates into Equation 2.4.

2.4.2 Evaluation of the Performance of the Procedure for Binary Inputs

Similar to Section 2.3.3, we conduct a simulation study to estimate $SD(\hat{\rho^*}^2)$ for $\rho^{*2} \in \{0, 0.1, 0.2, \ldots, 0.9\}$ for the "Full", "Random", and "Leveraged" plans. Here, we have four parameters in the model (α, β, σ_ϵ, and q); however, ρ^{*2} depends only on β, σ_ϵ, and q. We

simulate data with $q = 0.5$ and $\alpha = 0$. We also set $\sigma_\epsilon = 1$, leading to $\beta = \sqrt{\frac{\rho^{*2}}{1-\rho^{*2}}}$. Similar to Section 2.3.3, we estimate $SD(\hat{\rho^*}^2)$ using 1000 simulation runs for all nine possible combinations of $n_b \in \{100, 400, 1000\}$ and $n_l = n_u \in \{5, 8, 16\}$, and in each simulation run, we estimate α, β, σ_ϵ, and q. Figure 2.4 presents the results.

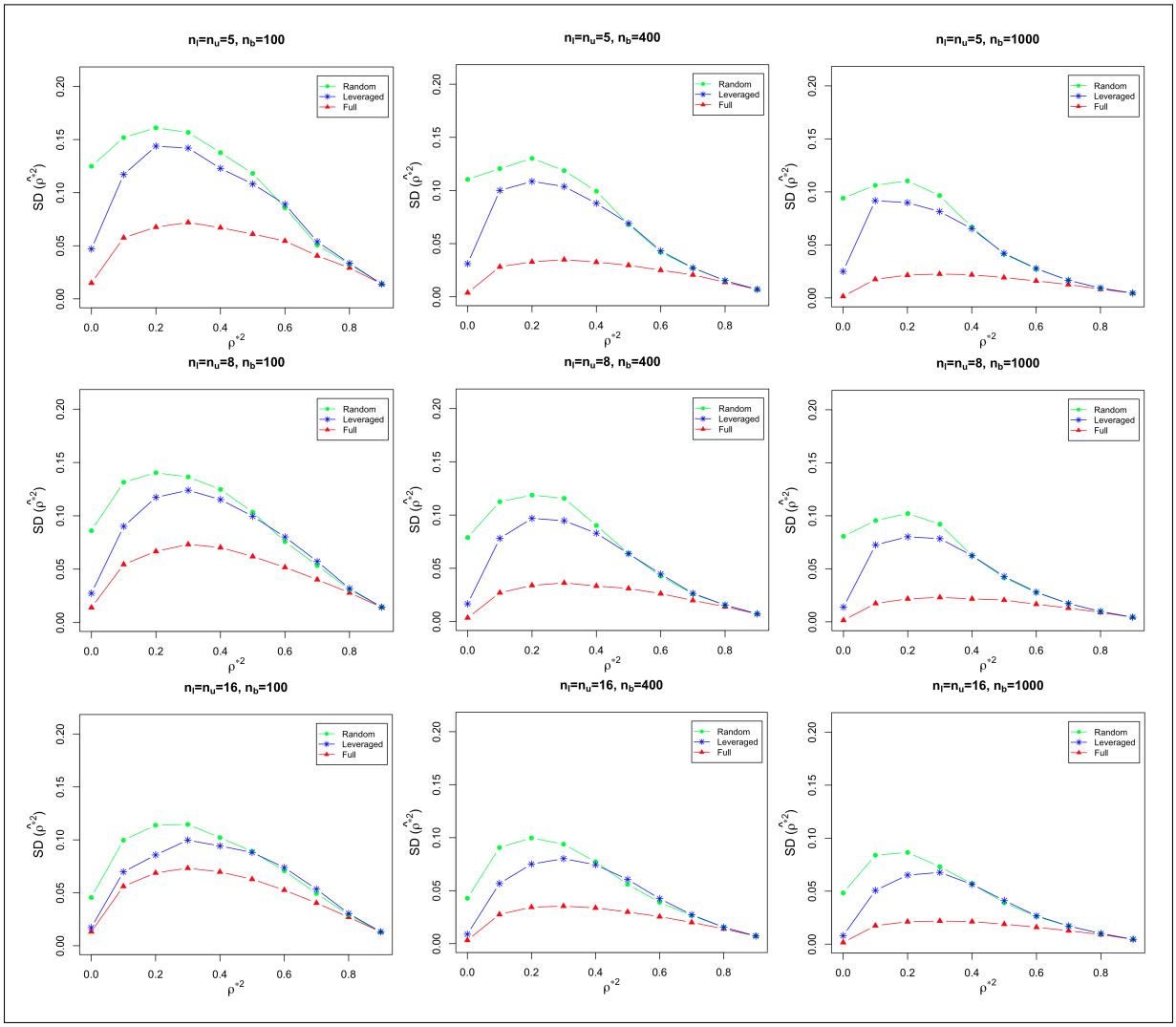

Figure 2.4: $SD(\hat{\rho^*}^2)$ vs. ρ^{*2} for $n_l = n_u \in \{5, 8, 16\}$ and $n_b \in \{100, 400, 1000\}$ for all "Full", "Leveraged", and "Random" plans when inputs are binary.

Figure 2.4, as we would expect, shows that the "Full" plan leads to the lowest $SD(\hat{\rho^*}^2)$ for all examined combinations. It is also clear that the "Leveraged" plan almost always lies between the "Random" and "Full" plans. However, the difference between the "Leveraged" and "Random" plans is very small when ρ^{*2} is large. The reason is that a large ρ^{*2} means X^* is clearly a dominant cause, and the distribution of Y on $X^* = -1$ and $X^* = 1$ is non-overlapping. Moreover, $SD(\hat{\rho^*}^2)$ decreases as n_b, n_u, and n_l increase. This is even more visible with larger values of n_l and n_u.

For the results shown in Figure 2.4, we estimate q as one of the model's parameters to calculate $SD(\hat{\rho^*}^2)$. It is noteworthy that the results are almost the same whether we assume q (say, $q = 0.5$) or estimate it.

2.5 Performance of the Proposed Procedure on the Example

Sections 2.2.1 and 2.2.2 demonstrate that the Bhote and Bhote (2000) and Shainin (1993b) procedure identifies *Sagging* and *Cutting Edge Appearance* as dominant causes, with end-count values of 8 and 14, respectively. In the example, the team looked into the *Sagging* issue, but this line of inquiry did not produce any interesting conclusions. The *Cutting Edge Appearance*, however, put the team on the right track. The cutting edges of the drill bits in the "Upper" group showed no irregularities, while the drill bits in the "Lower" group showed visible grinding artifacts. These imperfections seemed to improve the performance of the drill bits to the extent that their absence in the "Upper" group drill bits appeared to be responsible for the high peaks in torque, triggering line stoppages from the monitoring system. This could be understood as these artifacts could function as chip breakers, resulting in smaller chips that can be removed more easily by the cooling liquid. The team reasoned that the absence of such artifacts on the cutting edges would result in larger chips, which may create obstructions and cause peaks in torque. This conclusion suggested redesigning the drill bits to include similar artifacts that function as chip breakers. This remedy was implemented in the process, which solved the problem.

Using the likelihood-based proposed procedure, our conclusion is in line with what the team concludes. For *Sagging*, we have $\hat{\sigma}_Y = \sqrt{\hat{\beta}^2\hat{\sigma}_X^2 + \hat{\sigma}_\epsilon^2} = \sqrt{0.385 + 1.14} = 1.23$ (since the estimates of $\hat{\beta}$, $\hat{\sigma}_\epsilon$, and $\hat{\sigma}_X$ are 41.4, 1.07, and 0.015, respectively). This means that if our efforts would achieve the complete elimination of the variability in torque due to *Sagging*, then $\hat{\sigma}_Y = \sqrt{1.14} = 1.07$. So, the potential reduction in σ_Y is no more than 14%,

and *Sagging* with $\hat{\rho}^2 = 0.255$ is not a dominant cause. *Cutting Edge Appearance*, however, is identified as a dominant cause with $\hat{\rho}^2 = 0.662$. For the remaining eight variables, we find low values of $\hat{\rho}^2$, meaning that they are not dominant causes of variation. The estimated ρ^2 values and their 95% confidence intervals for all variables are reported in Table 2.2.

Candidate (X's)	Top Angle	Side Angle	Sagging	Dimension A	Dimension B	Width	Diameter at Point P	Stains Near Top	Discoloration	Cutting Edge Appearance
$\hat{\rho}^2$ (Equation 2.2 or 2.4)	0.001	0.001	0.255	0.018	0.034	0.001	0.001	0.076	0.003	0.662
95% C.I. on ρ^2	(0.000, 0.162)	(0.000, 0.156)	(0.062, 0.565)	(0.000, 0.204)	(0.000, 0.305)	(0.000, 0.166)	(0.000, 0.155)	(0.000, 0.615)	(0.000, 0.201)	(0.138, 0.848)
Tukey End-count	2	5	8	4	2	3	2	3	3	14
95% Bootstrap Interval for Tukey End-count	$(2, 7)$	$(2, 8)$	$(3, 16)$	$(2, 9)$	$(2, 10)$	$(2, 7)$	$(2, 8)$	-	-	-

Table 2.2: Results of Tukey end-count, $\hat{\rho}^2$, and their 95% intervals for the drill bits example.

Table 2.2 uses the fractional-random-weight bootstrap method to determine the confidence intervals (Xu et al. (2020)). Although some of the confidence intervals are somewhat wide, the precision of our procedure is dramatically better than the end-count procedure. To see the point, we determined 95% bootstrap intervals for end-count when inputs are continuous by using a standard parametric bootstrap. It is not clear how to implement that for the noncontinuous inputs; hence, there are three empty entries in the last row of Table 2.2. Note that for all continuous candidate causes, the critical value (at the 95% significance level) of seven is included within the 95% interval. This demonstrates that the Bhote and Bhote (2000) and Shainin (1993b) procedure, which discards all observed output values that are not in the "Upper" and "Lower" groups, discretizes the remaining ones, and uses nonparametric statistics, is quite inefficient. Note that the baseline sample size n_b is uncharacteristically small in the example, and the inefficiency will be even more extreme for more common baseline sizes.

2.6 Summary and Discussion

This chapter provides a critical assessment of the Bhote and Bhote (2000) and Shainin (1993b) group comparison investigation used as a part of the method of elimination to help identify the dominant cause(s) of the variation in a quality characteristic of interest.

The strategy of group comparison based on the idea of leveraging appears to be highly effective. It yields more precise estimates compared to a random sampling strategy, while significantly reducing the required effort compared to a full sampling plan.

The specific implementation of the end-count-based strategy proposed by Bhote and Bhote (2000) and Shainin (1993b), however, was found to be suboptimal. First, it frames the problem as a significance test (*whether* candidate inputs affect Y), whereas the goal of the procedure is to identify the *dominant* causes. This is much better captured by framing it as an estimation problem. As given, the procedure is likely to identify many minor causes unjustly as dominant causes.

Second, the analysis based on end-count values makes inefficient use of the data. Output values not in the "Upper" and "Lower" groups are discarded, and the remaining output values are reduced to a dichotomy ("Upper" or "Lower"). Our maximum-likelihood procedure makes full use of the available information.

Note that the idea of leveraging does not work well when the quality characteristic is binary, particularly when the input is binary as well. Even with a continuous quality characteristic and a binary input, there are many ties in the rank order for calculating the end-count values. However, we can extend the proposed end-count analysis and utilize the actual quality characteristic values for the selected parts instead of just labelling them as in the "Upper" and "Lower" groups. That way, we can rank the parts in terms of the quality characteristic and calculate the end-count. However, it is unclear how to extend the end-count analysis when the discrete input is ordinal or categorical with more than two levels.

Besides the evaluation and improvement of the group comparison procedure itself, we contribute a study of the idea of leveraging. This design principle is useful for studies that aim to estimate the correlation between input and quality characteristic variables. Leveraging may make the study more efficient if a large baseline sample is available whose Y values have been determined but whose input values are yet to be measured. The idea of leveraging is to select a subsample from the baseline sample consisting of parts having the largest and smallest output values. These parts give the most information about the desired correlation. Our study found that, in the group-comparison application, a

"Leveraged" sample is more efficient than a "Random" sample of parts from the baseline. On the other hand, "Leveraged" samples were found to give about half the estimation precision as a "Full" sample, but with only a small fraction of the input measurements. Note that we do not necessarily have to select the most extreme parts to have an effective leveraging. In some situations, it may be desirable not to select wildly outlying parts for fear that they may be due to a cause that rarely acts.

Later, in Chapter 3, we discussed Shainin's Components Search™ procedure, which also exploits the idea of leveraging (Bhote and Bhote (2000); Shainin and Shainin (1988); Steiner and MacKay (2005)). Moreover, Browne et al. (2009a,b, 2010a,b) applied the idea of leveraging in the context of Gauge R&R studies. The precision of measurement systems is often expressed in the Gauge R&R statistic γ, defined as the correlation of repeated measurements of the same part. Leveraging allows efficient estimation of a correlation coefficient. In this case, the baseline data consists of a large number of parts measured once. Then, only for the leveraged sample, which consists of parts with extremely large and small values in the baseline data, we collect the repeated measurements required to estimate γ.

As discussed, this study assumes that the effect of the input on the quality characteristic is approximately linear. If we are not sure about the linearity assumption, we recommend checking it by measuring a small number of input values for the parts with intermediate output values. Since our point is clue generation, linearity is often a reasonable initial assumption to keep the problem simple.

Moreover, we considered having only one dominant cause in Model 2.1. This has the advantage of being easy to understand and representative of how the idea works. Although in real-world problems, there might be multiple large causes, and they could even be correlated with other inputs or the dominant cause might be an interaction, we made this simplifying assumption to be able to focus on the main idea of the new procedure.

Chapter 3

Identifying Dominant Causes using Component Swapping

Component swapping is a valuable approach for identifying problems in assembled products. Shainin's component-swapping procedure is a widely utilized systematic strategy for identifying the dominant cause(s) of excessive output variation of an assembly process. Despite its practical popularity and intuitive appeal, the existing literature lacks a comprehensive investigation into its merits and drawbacks. Our critical evaluation reveals that although this procedure incorporates valuable ideas, it is unreliable in identifying the dominant cause(s) of variation due to the utilization of poorly chosen statistical tools. We introduce an alternative analytical approach based on more sophisticated new and existing statistical techniques. We substantiated the effectiveness and superiority of our proposed method through extensive simulation studies. Furthermore, our proposal addresses a gap present in the literature by effectively alerting users to important interactions between assembly and component, or among components.

3.1 Introduction

Component swapping is a well-known problem-solving strategy. It is an efficient strategy for identifying the cause(s) of problems with assembled products. A tangible illustration of component swapping is when we connect a laptop to a projector using a cable, and the image is not projected. To diagnose the problem, the initial step is to ascertain whether disconnecting and reconnecting the projector and laptop resolves the issue. If this action

is ineffective, the next phase is comparing the malfunctioning configuration (consisting of three components: laptop, projector, and cable) with a functional configuration (laptop, projector, and cable). The aim is to find the specific component(s) responsible for the malfunction. This may be achieved by swapping components between the two configurations until the malfunctioning system begins to function correctly while the previously functional system begins to malfunction. For instance, one could start by swapping the projectors between the two configurations. If the malfunctioning system now works and the functional system fails, the issue is in the projector, not in the cable or laptop. Then, one could apply component swapping recursively by swapping components between the good and defective projectors to identify the component(s) within the projector causing the problem. On the other hand, if swapping the projectors has no effect, the investigation would move on to the next component, such as the cable. We swap components between the two systems until we identify the component(s) that causes the functional system to malfunction while allowing the defective system to function.

Component swapping is an example of the general problem-solving strategy of hierarchical search or the method of elimination, which motivates why component swapping can be an efficient search strategy (Chittaro and Ranon (2004); De Mast (2011, 2013); Shainin (1993b); Steiner et al. (2008a,b)). Although most descriptions of the strategy are in loose and general terms, the strategy was formalized for a specific type of problem, which is identifying the dominant cause(s) of variation in assembled products. Excessive variation in critical-to-quality characteristics of assembled products is a common issue extensively studied in quality and industrial engineering. For this type of problem, a few concrete operationalizations of component swapping have been proposed, such as Shainin and Shainin (1988), Bhote and Bhote (2000), and Steiner and MacKay (2005). The literature in quality and industrial engineering has documented many applications of component-swapping studies in variation-reduction projects (e.g., Kiatcharoenpol et al. (2023); Panchal et al. (2020); Pietraszek et al. (2016); Ghurka and Pawar (2015); Mooren et al. (2012)). This chapter discusses component swapping in the context of trying to identify the dominant cause(s) of excessive variation.

Proponents of the component-swapping strategy claim that it is more efficient for identifying causes compared to alternative approaches, such as designed experiments. Although the literature has discussed the component-swapping strategy (e.g., Dasgupta et al. (2011); Amster and Tsui (1993); Logothetis (1990); Steiner et al. (2008a,b)), a systematic examination of its merits and drawbacks is still lacking. This chapter has four objectives:

(i) To evaluate known implementations of the component-swapping strategy, particularly those of Shainin and Shainin (1988) and Bhote and Bhote (2000);

(ii) To propose an alternative operationalization driven by sound statistical techniques and subsequently assess its efficiency;

(iii) To provide well-grounded recommendations for users, allowing them to apply the component-swapping strategy safely; and

(iv) To identify study-design principles that may hold broader interest.

The structure of this chapter is as follows. In Section 3.2, we present and critically assess the component-swapping procedure proposed by Shainin and Shainin (1988) and Bhote and Bhote (2000). Section 3.3 is devoted to explaining our proposed procedure and our recommendations. In Section 3.4, we evaluate our proposed procedure and demonstrate its superiority. We provide discussion and conclusions in Sections 3.5 and 3.6, respectively.

3.2 Shainin's Component-Swapping Procedure

To the best of our knowledge, among the various implementations of the general component-swapping strategy discussed in the literature, Shainin and Shainin (1988) is the only one that operationalizes the strategy into a concrete procedure, albeit for a specific type of problem. It is commonly used in quality engineering and covered in many variation-reduction and statistical-engineering courses. Moreover, they provide the clearest and most specific procedure, including decision boundaries with a statistical motivation. However, the paper is unclear about a stopping rule for the procedure once the dominant cause is identified. Therefore, we consider a slightly modified procedure incorporating the stopping criterion from Bhote and Bhote (2000). This latter reference has the procedure proposed by Shainin and Shainin (1988) as its origin. Thus, we refer to this slightly modified procedure as "Shainin's procedure". Appendix B.1 provides an overview of various alternative implementations in the peer-reviewed literature. In this section, we present and critically assess Shainin's component-swapping procedure.

3.2.1 Outline of the Procedure

The application context of the component-swapping procedure is to identify the dominant sources of variation in a critical-to-quality characteristic Y for assembled products. We assume the assembled product has at least two components or subassemblies that can be disassembled and reassembled without significantly changing or damaging the product.

As commonly recognized in the quality engineering literature, excessive variation in the critical process output, denoted Y, is typically caused by many sources. However, usually, only a few of them have a disproportionate impact on the overall variation, while many others contribute only marginally. Juran refers to the important causes as the "vital few", which are few in number but account for almost all of the overall variation, and the unimportant causes as the "trivial many", which are large in number but collectively contribute negligibly (Gryna and Juran (1988); Panahi et al. (2021, 2023)). In this chapter, we refer to the "vital few" causes of variation as the *dominant causes*.

The component-swapping procedure investigates whether the excessive variation in Y can be attributed to one or a few components or, alternatively, whether it is the process of assembling products that substantially contributes to the variation in Y. We define ρ_A^2 as the contribution of the assembly process to the overall variation in Y by

$$\rho_A^2 = \frac{\sigma_A^2}{\sigma_Y^2}, \tag{3.1}$$

where $\sigma_Y^2 = Var(Y)$, and $\sigma_A^2 = \sigma_Y^2 - Var(Y|A)$ is the variance in Y due to (causes in) the assembly process. Moreover, we define $\rho_{C_i}^2$ as the contribution of a component C_i to the overall variation by

$$\rho_{C_i}^2 = \frac{\sigma_{C_i}^2}{\sigma_Y^2}, \tag{3.2}$$

where $\sigma_{C_i}^2 = \sigma_Y^2 - Var(Y|C_i)$ is the variance in Y due to (causes in) C_i. Note that C_i could be a single or a group of components. The objective of the component-swapping procedure is to identify the component(s) with the largest contribution of ρ_C^2 in σ_Y^2, or alternatively, to identify whether the contribution of the assembly process ρ_A^2 is large enough to be considered a *dominant* cause.

Shainin's procedure starts by selecting *two* assembled products from a baseline of previously assembled products, whose output values have been measured. From this baseline, we select the two products with high and low y values, effectively representing the excessive output variation. Selecting extreme products, since they contain more information than average ones, is a commonly employed study-design principle in variation reduction known as *leveraging* (Panahi et al. (2021); Steiner et al. (2008a,b)).

The component-swapping procedures, then, consist of two phases. Phase I investigates whether the assembly process is the dominant cause of variation. This involves disassembling and reassembling the selected individual products multiple times and re-measuring the output value. By doing so, we can determine the extent of variation attributed to the assembly process. If Phase I suggests that assembly is not a dominant cause of output

variation, we proceed to Phase II, where we subsequently swap components between the two extreme products until we identify the component(s) that account for the majority of the output variation. Below, we present and assess Shainin's implementation of these two phases.

3.2.2 Phase I of Shainin's Procedure (Disassembling and Reassembling)

The procedure requires a baseline sample of n_b products for which the quality (output) measurements Y_i, $i = 1, \ldots, n_b$ are given. We assume that the baseline data are readily available from previous investigations or can be collected at a relatively low cost. Then, in Phase I, we select the two products with the lowest and highest y values from the baseline. We refer to them as the "Low" and "High" products and denote their original output values (as observed in the baseline) by y_0^L and y_0^H.

Subsequently, we disassemble and reassemble each of these two products r times (with the implementation of Shainin and Shainin (1988), $r = 2$) and remeasure the output values after each disassembly/reassembly. This yields the results given in Table 3.1.

	Low product	High product
y values from the baseline	y_0^L	y_0^H
y values after disassembling and reassembling r times	y_1^L $\vdots$ y_r^L	y_1^H $\vdots$ y_r^H

Table 3.1: Data collection and notation for Phase I of the component-swapping procedure with Shainin's choice of selecting two extreme products.

Shainin and Shainin (1988) recommend $r = 2$, and their procedure concludes that the assembly process is not the dominant cause of variation if both the following are satisfied:

- $Max(y_0^L, y_1^L, y_2^L) < Min(y_0^H, y_1^H, y_2^H)$ (this criterion is an application of Tukey's end-count test with complete separation; Tukey (1959));
- $D > 1.07\,\bar{R}$, where $D = Median(y_0^H, y_1^H, y_2^H) - Median(y_0^L, y_1^L, y_2^L)$, and $\bar{R} = [Range(y_0^H, y_1^H, y_2^H) + Range(y_0^L, y_1^L, y_2^L)]/2$.

The constant 1.07 tunes the decision threshold for D to a 95% confidence level of misidentifying assembly as a cause if in fact it has no effect. If *either* of the above two

criteria is not satisfied, we conclude assembly is the dominant cause, and the procedure stops. Otherwise, we proceed to Phase II.

3.2.3 Phase II of Shainin's Procedure (Swapping Components)

In Phase II, drawing from engineering knowledge, we assess the components of the product according to their plausibility of being the dominant cause of variation. Let the components' descending ranking be C_1, C_2, C_3, ..., and denote the components that cannot be swapped (e.g., part housing) or are no longer under suspicion by C_R. Starting with the top-ranked component C_1, we swap components between the low and high products one by one while remeasuring the output values after each swap. This yields the results presented in Table 3.2. For instance, $y^L_{C_1=H}$ represents the output value in the low product after swapping component C_1, that is replacing component C_1 in the low product with the one that was originally in the high product.

	Low product	High product
y values from the baseline	y^L_0	y^H_0
y values after swapping component C_1	$y^L_{C_1=H}$	$y^H_{C_1=L}$
y values after swapping component C_2	$y^L_{C_2=H}$	$y^H_{C_2=L}$
⋮	⋮	⋮
y values after swapping components C_1 and C_2	$y^L_{C_1,C_2=H}$	$y^H_{C_1,C_2=L}$
⋮	⋮	⋮

Table 3.2: Notation for Phase II of the component-swapping procedure.

To determine whether component (or a group of components) C_i is the dominant cause, after each swap we compare $y^L_{C_i=H}$ and $y^H_{C_i=L}$ with the decision intervals (calculated from the results of Phase I) given by $DI^L = Median(y^L_0, y^L_1, \ldots, y^L_r) \pm t_{(2r,0.95)}\bar{R}/d_2(r+1)$ and $DI^H = Median(y^H_0, y^H_1, \ldots, y^H_r) \pm t_{(2r,0.95)}\bar{R}/d_2(r+1)$, where $d_2(r+1)$ is a known constant from the control-charting literature (Ryan (1989)). In Shainin's implementation, $r = 2$, and we have $t_{(4,0.95)} = 2.776$ and $d_2(3) = 1.693$.

After each swap, Bhote and Bhote (2000) make a conclusion based on the following criteria:

- If $y^H_{C_i=L} \in DI^H$ and $y^L_{C_i=H} \in DI^L$, there are only *minor changes* in the output values after swapping C_i, i.e., C_i is an unimportant component. Thus, we eliminate C_i from further consideration and swap the next ranked component.

- If $y^{H}_{C_i=L} < Max(DI^{L})$ and $y^{L}_{C_i=H} > Min(DI^{H})$, there are *complete changes* in the output values after swapping C_i, i.e., C_i is an important component and explains the variation we have seen in the output. Thus, we identify C_i as the sole dominant cause of variation and stop the component-swapping process.
- Otherwise, there are *partial changes* in the output values after swapping C_i, i.e., C_i is important, but does not on its own explain the output variation. Thus, we retain C_i, swap the next ranked component, and continue exploring other components that also cause a partial change.

Figure 3.1 presents some fictitious examples and their corresponding conclusion.

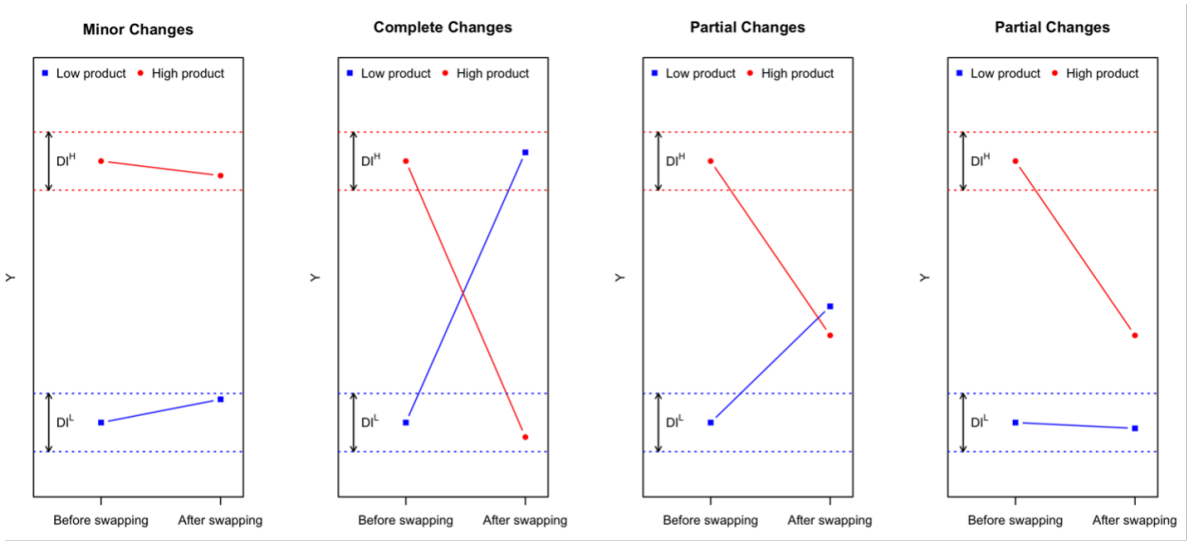

Figure 3.1: Some fictitious examples of different conclusions (following by Bhote and Bhote (2000)).

In situations where we identify more than one component with partial changes, it is necessary to conduct a capping run, where we simultaneously swap all the components that led to partial changes and investigate whether their combined effect is the dominant cause. The outcomes of these capping runs are also incorporated into Table 3.2. For instance, assuming both C_1 and C_2 individually result in partial swaps, $y^{L}_{C_1,C_2=H}$ represents the output value in the low product after simultaneously swapping both C_1 and C_2. Note that by concluding the combined effect as the dominant cause, the dominant cause could be the sum of the two main effects $C_1 + C_2$ or their interaction effect C_1C_2. However, we use the notation $C_1 + C_2$ here, because Shainin's procedure does not distinguish between the

two cases. We stop the procedure if this simultaneous swap results in complete changes. Otherwise, we continue swapping the next-ranked components until we either identify all the important causes or eliminate all the ranked components. If the latter occurs, the dominant cause lies among the remaining components C_R, or we have made an incorrect conclusion somewhere along the way. Figure 3.2 provides the flowchart of Phase II in Shainin's procedure.

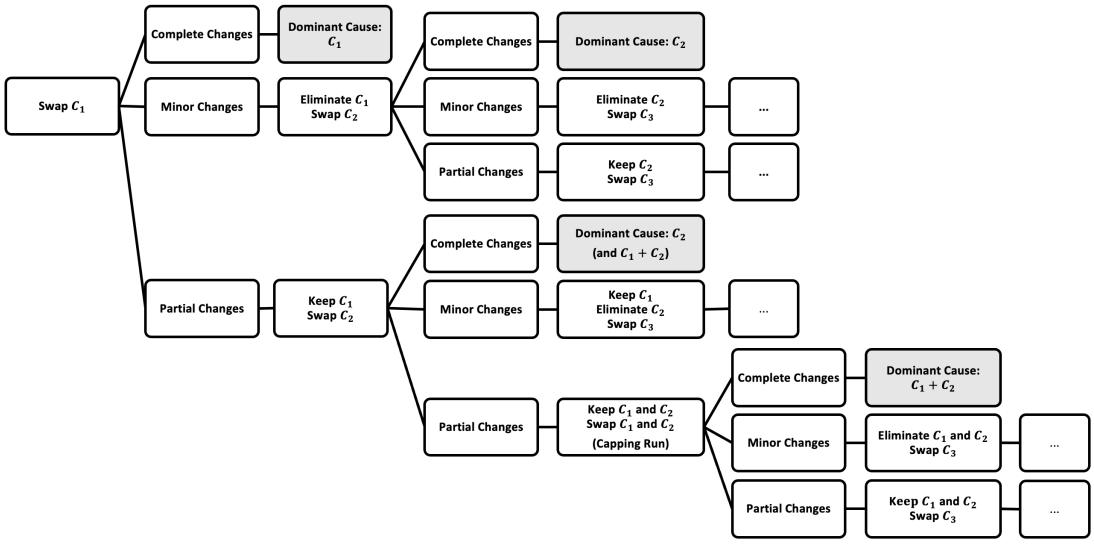

Figure 3.2: Flowchart of Shainin's procedure in Phase II, where the ranked components are C_1, C_2, C_3, etc.

3.2.4 Shainin's Procedure: Issues and Points for Improvement

We evaluate the performance of Shainin's procedure using extensive systematic simulation studies, with details presented in Section 3.4. These simulations reveal that Shainin's Phase I procedure is unreliable in identifying assembly as the dominant cause when it is indeed the dominant cause. For instance, with a baseline sample size $n_b = 1000$ and following their recommendation of selecting two extreme products and $r = 2$, Shainin's procedure identifies assembly as the dominant cause only 20% of the time, when its actual contribution to the overall variation in Y is 50% (i.e., $\rho_A^2 = 0.5$). Moreover, Shainin's Phase II procedure is unreliable in identifying the dominant cause among the components. It often identifies the combined effect of two or more components as the dominant cause, even if only a single component is the dominant cause. For instance, with their recommendation

of $r = 2$, assuming only two components, and setting $n_b = 1000$, Shainin's procedure misidentifies the combined effect of C_1 and C_2 as the dominant cause approximately 85% of the times, when C_1 individually is the dominant cause with 75% contribution to the overall variation in Y ($\rho^2_{C_1} = 0.75$, $\rho^2_{C_2} = 0.20$, and $\rho^2_A = 0.05$). Moreover, as mentioned before, after conducting a capping run, Shainin's procedure fails to differentiate whether the involved components *together* are the dominant cause, or if it is the *interaction* effect between them.

The surprisingly poor performance of Shainin's procedure appears to stem partly from the use of unsophisticated estimators, such as $\bar{R}$, to estimate variation. However, a more fundamental problem appears to arise from their use of decision criteria that are derived from statistical significance tests. The decision criteria in both Phase I (the Tukey end-count test and the decision threshold for D) and Phase II (DI^L and DI^H) identify assembly or components if they are *significant* causes of variation (with 95% confidence). This is obviously at odds with the authors' own claims that they intend to identify *dominant* causes (Shainin and Shainin (1988), Bhote and Bhote (2000)). This creates an issue because a statistically significant cause does not necessarily imply a dominant cause.

3.3 Proposed Procedure

Considering the abovementioned issues, Shainin's implementation is a problematic basis for evaluating the merits of component-swapping strategies. This section introduces an alternative to Shainin's analysis procedure. While Shainin's procedure is based on significance-testing decision intervals, our proposed analysis aims to *estimate* the proportion of variation based on more sophisticated estimators.

3.3.1 Proposed Phase I Analysis (Disassembling and Reassembling)

In this subsection, our focus is on Phase I of the component-swapping procedure. We first explore how best to estimate ρ^2_A, and then we provide our recommendations for investigating irregularities to further improve the conclusion of the procedure.

3.3.1.1 Estimation of ρ_A^2

We propose an analysis that aims to *estimate* the proportion of variation ρ_A^2 in Y due to assembly; however, note that we cannot separately estimate the output variation due to the assembly process and the variation due to measurement error. We can only estimate the proportion of variation in Y due to the combined effect of all *assembly- and measurement-related causes*. However, we assume that as a standard early step in many variation reduction projects, we have previously assessed the measurement system and have found that the measurement variability is small (Steiner and MacKay (2005)). For simplicity, we refer to the variation due to assembly- and measurement-related causes as the variation due to assembly, denoted by σ_A^2, and estimate its contribution ρ_A^2 using Equation 3.1.

If assembly accounts for the dominant proportion of the overall variation in Y (e.g., $\rho_A^2 > 0.5$), we conclude that assembly is the dominant cause that should be addressed, and we stop the procedure. To estimate σ_Y^2 and σ_A^2, we can use the baseline data $y_1, \ldots, y_{n_b}$ and the results of the Phase I investigation (i.e., disassembly/reassembly) where we obtain y_j^i, with $j = 0, 1, \ldots, r$ (see Table 3.2). In Shainin's procedure, where we select two extreme products, we have $i = L$ or H (the low and high products selected from the baseline). However, in what follows, we allow for the selection of more than two products and denote the number of products selected as k. For instance, as we shall suggest, one could select the low and high products, resulting in $k = 2$, or also include the product with the median y value, resulting in $k = 3$.

Our proposed estimation approach for ρ_A^2 is inspired by Browne et al. (2009b, 2010a), who explored the estimation of proportions of variances from similar study designs in the context of measurement-system assessment. Browne et al. (2009b, 2010a) proposed four approaches: a maximum likelihood estimator, a regression estimator, an ANOVA estimator, and a combined estimator. To select the most suitable estimator for our context, we conducted extensive simulation studies whose results are given in Appendix B.3.1. Based on the outcomes of these studies, we recommend the following combined estimator as it has a closed form, low bias, and low standard deviation compared to the alternatives:

$$\hat{\rho}_A^2 = 1 + \frac{q(1-\hat{\rho}^2_{A_{ANV}} - r^{-1}) - v_F(2-\hat{\rho}^2_{A_{Reg}})}{2\,(v_F - q)} + \frac{\sqrt{\Big(q(1-\hat{\rho}^2_{A_{ANV}} - r^{-1}) - v_F(2-\hat{\rho}^2_{A_{Reg}})\Big)^2 - 4(v_F - q)\Big(v_F(1-\hat{\rho}^2_{A_{Reg}}) + qr^{-1}(1-\hat{\rho}^2_{A_{ANV}})\Big)}}{2\,(v_F - q)},$$

where $\hat{\rho}^2_{A_{ANV}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{r}\left(y_j^i - \bar{y}_{PhI}^i\right)^2}{k(r-1)s_b^2}$, $\hat{\rho}^2_{A_{Reg}} = 1 - \frac{\sum_{i=1}^{k}(\bar{y}_{PhI}^i - \bar{y}_b)(y_0^i - \bar{y}_b)}{\sum_{i=1}^{k}\left(y_0^i - \bar{y}_b\right)^2}$, $v_F = \frac{2(n_b-1)^2(k(r-1)+n_b-3)}{k(r-1)(n_b-3)^2(n_b-5)}$, $q^{-1} = \sum_{i=1}^{k}\frac{\left(y_0^i - \bar{y}_b\right)^2}{s_b^2}$, and $\bar{y}_b$ and s_b^2 represent the sample mean and variance of the baseline data. Later, in Section 3.4, we demonstrate the good performance of this estimator.

3.3.1.2 Check for Irregularities

The analysis presented in Section 3.3.1.1 relies on certain assumptions, which may be compromised by variance heterogeneity or interaction effects (between components and assembly). We recommend conducting a variance equality test such as Bartlett's or Levene's test (Bartlett (1937), Levene (1960)) as part of the Phase I analysis. We test whether the variance of y_j^i for $j = 1, \ldots, r$ is equal across the products $i = 1, \ldots, k$.

If equality of variances is rejected, the heteroscedasticity may be due to an interaction effect between the assembly process and one or more components. Such an interaction implies that the results of disassembly and reassembly vary depending on the specific component within the product, making reassembly easier for some products but more challenging for others. Detecting evidence of such irregularities can be challenging, especially when the study involves only extreme products, as in Shainin's procedure with $i = L$ or $i = H$. To gain more insight into more common products, it is beneficial to also include a baseline product with a median y value (Prashar (2016), Cox (2011)). Appendix B.3.2 reveals that in some scenarios, selecting a median product considerably increases the probability of identifying interaction effects between the assembly process and components.

3.3.2 Proposed Phase II Analysis (Swapping Components)

In this subsection, we focus on Phase II of the component-swapping procedure. While we extended the notation and analysis of Phase I to allow the selection of $k \geq 2$, in Phase II, we only use the two extreme products, i.e., $i = L$ and $i = H$. We first explore how best to estimate ρ_C^2, and then we provide our recommendations for investigating irregularities to further improve the conclusion of the procedure.

3.3.2.1 Estimation of ρ_C^2

Phase II aims to estimate the proportion of variation in Y due to components. Therefore, after swapping component C_i, we estimate its contribution $\rho_{C_i}^2$ to the output variation as

defined in Equation 3.2. Then, if desired, we could translate our estimates into decisions about whether we believe C_i is a dominant cause by comparing $\hat{\rho}^2_{C_i}$ to a given threshold, say 0.5 (or any other threshold). Below, we introduce two viable estimators for $\rho^2_{C_i}$: the ANOVA and LVR estimators. With each estimator, we estimate $\rho^2_{C_i}$ after each swap while using all the Phase I and II results available at that time (except the Phase I results for the median product).

ANOVA Estimator

Table 3.3 presents the situation where we have swapped components C_1 and C_2, and we have also conducted a capping run where C_1 and C_2 are swapped simultaneously. We denoted the components from the low and high product as -1 and 1, respectively. The column labelled C_1C_2 represents the interaction between C_1 and C_2.

Phase	Results	C_1	C_2	C_1C_2	C_R
Baseline	y_0^L	-1	-1	1	-1
	y_0^H	1	1	1	1
Phase I	y_1^L	-1	-1	1	-1
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	y_r^L	-1	-1	1	-1
	y_1^H	1	1	1	1
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	y_r^H	1	1	1	1
Phase II (Swap C_1)	$y^L_{C_1=H}$	1	-1	-1	-1
	$y^H_{C_1=L}$	-1	1	-1	1
Phase II (Swap C_2)	$y^L_{C_2=H}$	-1	1	-1	-1
	$y^H_{C_2=L}$	1	-1	-1	1
Phase II (Swap C_1 and C_2)	$y^L_{C_1,C_2=H}$	1	1	1	-1
	$y^H_{C_1,C_2=L}$	-1	-1	1	1

Table 3.3: Component-swapping procedure presented as runs with factors and levels.

The ANOVA estimator uses the observed data in Phases I and II but ignores the original baseline values y_0^L and y_0^H. The first results in Phase II come from swapping the top-ranked component C_1. After obtaining these results, we fit the ANOVA model $Y \sim C_1 + C_R$. At this stage, attempting to include C_2 or C_1C_2 in the model would be futile, as these effects would be confounded with C_R. Our estimator of $\rho^2_{C_1}$ is based on the sums-of-squares in

the analysis of variance. Let SS_{C_1} be the sum-of-squares of the C_1 factor, and SS_T be the total sum-of-squares. Then, $\hat{\rho}^2_{C_1}$ is given by

$$\hat{\rho}^2_{C_1} = \frac{SS_{C_1}}{SS_T}.$$

Note that this value is commonly reported as R^2 in a standard analysis of variance. If the estimated contribution of C_1 is dominant (e.g., $\hat{\rho}^2_{C_1} \geq 0.5$), we stop the procedure. However, if C_1 is a minor cause, i.e., a substantial proportion of the output variation is still unaccounted for (e.g., $\hat{\rho}^2_{C_1} < 0.5$), we proceed to swap C_2 and fit the model $Y \sim C_1 + C_2 + C_R$ to all the now available data. Similarly, we derive $\hat{\rho}^2_{C_1}$ and $\hat{\rho}^2_{C_2}$ from the new sum-of-squares. If the estimated contribution of C_2 is dominant (e.g., $\hat{\rho}^2_{C_2} \geq 0.5$), we stop the procedure. If both estimates account for a reasonably large amount of the output variation but neither is individually dominant (e.g., $0.25 < \{\hat{\rho}^2_{C_1}, \hat{\rho}^2_{C_2}\} < 0.5$), we conduct a capping run (i.e., swap C_1 and C_2 together) and fit the model $Y \sim C_1 + C_2 + C_1C_2 + C_R$, enabling us to estimate the contribution $\hat{\rho}^2_{C_1C_2}$ of the interaction effect. If C_1C_2 is also identified as a minor cause, we continue with C_3 and further down the ranked list of components.

Leveraged Variance Ratio (LVR) Estimator

The ANOVA estimator uses the data from the $2r$ replications in Phase I and the results in Phase II (e.g., $y^L_{C_1=H}$ and $y^H_{C_1=L}$ when we swap C_1). However, it does not incorporate the observed output values in the baseline. The LVR estimator, on the other hand, also incorporates the baseline values for the selected extreme products, i.e., y^L_0 and y^H_0. This may be helpful since it defines the range of the n_b baseline data, conveying information about the variation in Y. The term *leveraged* in the name of this estimator refers to our use of the extreme values y^L_0 and y^H_0. Note that the LVR estimator is a novel approach not previously documented in the literature.

After swapping component C_i between the two extreme observations from the baseline, we introduce the following estimator

$$\hat{\rho}^2_{C_i} = Min\left(\frac{Var(\bar{y}^L_{PhI}, y^L_{C_i=H}) + Var(\bar{y}^H_{PhI}, y^H_{C_i=L})}{2Var(y^L_0, y^H_0)}, 1\right), \tag{3.3}$$

where $\bar{y}^L_{PhI} = (y^L_1 + \cdots + y^L_r)/r$ and $\bar{y}^H_{PhI} = (y^H_1 + \cdots + y^H_r)/r$.

After obtaining $\hat{\rho}^2_{C_i}$, we can determine whether C_i is the dominant cause by comparing it to a threshold, say 0.5, as before. If C_i is not the dominant cause, we proceed to

swap the next-ranked component. If none of the swapped components is identified as the dominant cause, we conclude that the dominant cause lies among C_R and assign the remaining proportion of variation to C_R.

In capping runs, multiple components are simultaneously swapped, and since we also have data from previous individual component swaps, Equation 3.3 changes. In this case, to estimate the proportion of variation due to a swapped component, we use the average of the variances when the setting of the swapped component changes while other components remain unchanged, and we pool it across different combinations of the levels of the other components. For instance, if we simultaneously swap C_i and C_j, we obtain the following estimates:

$$\hat{\rho}^2_{C_i} = Min\left(\frac{Var(\bar{y}^L_{PhI}, y^L_{C_i=H}) + Var(y^L_{C_j=H}, y^L_{C_i,C_j=H}) + Var(y^H_{C_i,C_j=L}, y^H_{C_j=L}) + Var(y^H_{C_i=L}, \bar{y}^H_{PhI})}{4\,Var(y^L_0, y^H_0)}, 1\right),$$

$$\hat{\rho}^2_{C_j} = Min\left(\frac{Var(\bar{y}^L_{PhI}, y^L_{C_j=H}) + Var(y^L_{C_i=H}, y^L_{C_i,C_j=H}) + Var(y^H_{C_i,C_j=L}, y^H_{C_i=L}) + Var(y^H_{C_j=L}, \bar{y}^H_{PhI})}{4\,Var(y^L_0, y^H_0)}, 1\right),$$

$$\hat{\rho}^2_{C_R} = Min\left(\frac{Var(\bar{y}^L_{PhI}, y^H_{C_i,C_j=L}) + Var(y^L_{C_j=H}, y^H_{C_i=L}) + Var(y^L_{C_i=H}, y^H_{C_j=L}) + Var(y^L_{C_i,C_j=H}, \bar{y}^H_{PhI})}{4\,Var(y^L_0, y^H_0)}, 1\right),$$

and $\hat{\rho}^2_{C_iC_j} = Max\left(1 - (\hat{\rho}^2_A + \hat{\rho}^2_{C_i} + \hat{\rho}^2_{C_j} + \hat{\rho}^2_{C_R}), 0\right)$.

To illustrate why these estimators make sense, consider the formula for the first equation, $\hat{\rho}^2_{C_i}$, and note that it is obtained by averaging the variances under the conditions listed in Table 3.4.

Variance term	C_i changes from	C_j remains unchanged at	C_R remains unchanged at
$Var(\bar{y}^L_{PhI}, y^L_{C_i=H})$	Low to High	Low	Low
$Var(y^L_{C_j=H}, y^L_{C_i,C_j=H})$	Low to High	High	Low
$Var(y^H_{C_i,C_j=L}, y^H_{C_j=L})$	High to Low	Low	High
$Var(y^H_{C_i=L}, \bar{y}^H_{PhI})$	High to Low	High	High

Table 3.4: Conditions under each variance composites of the $\hat{\rho}^2_{C_i}$ formula.

3.3.2.2 Check for Irregularities

Our simulations illustrate that both the ANOVA and LVR estimators become unreliable when there are strong interaction effects between components (see Appendix B.4.2 for a

detailed discussion). We propose the incorporation of the following two criteria into the procedure for alerting the user to such situations:

- "Partial criterion": $|Var(\bar{y}^H_{PhI}, y^H_{u=L}) - Var(\bar{y}^L_{PhI}, y^L_{u=H})|/Var(y^L_0, y^H_0)$ (all from Equation 3.3) exceeds a given threshold, say 0.2 for $u \in \{C_i, C_j\}$. This criterion is satisfied when swapping component u has a large effect on one product, but not the other. This is a classic sign of an important interaction. Note that the threshold of 0.2 is arbitrary; however, our simulation studies show its effectiveness in delivering favorable results.

- "Extreme criterion": Swapping one or more components yields output values that are more extreme than the most extreme values observed in baseline and Phase I, i.e., $y^H_{u=L} > Max(y^H_0, y^H_1, \ldots, y^H_r)$, or $y^L_{u=H} < Min(y^L_0, y^L_1, \ldots, y^L_r)$ for component or components u.

We recommend verifying the above criteria after each swap to check for evidence of interactions between components. If such interactions exist, these criteria will often detect evidence of them. See Appendix B.4.2 for a more detailed discussion of how well these criteria work.

3.3.3 Proposed Approach

Based on the outcomes of our simulation studies, we recommend the following study design:

1. Select $k = 3$ baseline products: one with the lowest, one with the highest, and one with a median output value.

2. In Phase I, disassemble and reassemble each selected product $r = 5$ times. The choice of $r = 5$ and $k = 3$ is based on Monte Carlo simulations. It provides a balanced trade-off between estimation precision, the ability to detect irregularities, and the amount of effort. Further motivation for this design choice is given in Appendix B.3.3.

3. Use the combined estimator from Section 3.3.1.1 to estimate ρ^2_A.

4. Perform a variance equality test such as Bartlett's and/or Levene's test to detect irregularities. Note that while Bartlett's test identifies irregularities more often (see Appendix B.3.2), it is more sensitive to the normality assumption (Levene (1960)). Further investigations are required if there is evidence of interaction between assembly and components.

5. If $\hat{\rho}_A^2$ is large (say greater than 0.5, and there are no irregularities), identify assembly as the dominant cause and stop the procedure; otherwise, proceed to Phase II.

6. In Phase II, rank components in descending order of plausibility based on engineering knowledge (or combine components into subassemblies if ranking is challenging or if there are many components). Select the two products with the most extreme y values from the $k = 3$ products used for Phase I, and label them the low and high products.

7. Swap the next top-ranked component C_i between the low and high products, and estimate $\rho_{C_i}^2$ using LVR.

8. Check for irregularities using the *Extreme* and *Partial* criteria. If there is evidence of interaction between two or more components, keep C_i as potentially being part of a dominant cause.

9. If $\hat{\rho}_{C_i}^2$ is large (say greater than 0.5, and there are no irregularities), identify C_i as the dominant cause and stop the procedure. Otherwise, if $\hat{\rho}_{C_i}^2$ is small (say less than 0.25), eliminate C_i from the suspect list, whereas if $\hat{\rho}_{C_i}^2$ is of a moderate size (say between 0.25 and 0.5), keep C_i as potentially being part of a dominant cause.

10. If we have so far identified two or more components as potentially being part of a dominant cause, swap them all simultaneously to see if their combined effect is a dominant cause. If this capping run identifies the dominant cause, stop the procedure; otherwise, keep the components involved in the capping run as potentially being part of a dominant cause.

11. Repeat steps 7 to 10 until we either identify the dominant cause or have no further swaps to try.

3.4 Evaluation of the Proposed Procedure

Both Shainin's procedure and our procedure proposed above were evaluated by means of comprehensive Monte Carlo simulation studies. This section summarizes our main findings. Detailed information regarding these studies can be found in Appendix B. Below, we outline the objectives of each study:

- Appendix B.2 evaluates Shainin's procedure using simulation studies. Appendices B.2.1 and B.2.2 are devoted to Phase I and Phase II, respectively. The evaluation

is based on the probability of identifying the actual dominant cause for different ρ_A^2 and ρ_C^2 values.

- Appendix B.3 outlines the proposed Phase I setup and analysis. In Appendix B.3.1, we introduce four viable estimators for ρ_A^2 and compare their corresponding bias and standard deviation, denoted as $Bias(\hat{\rho}_A^2)$ and $SD(\hat{\rho}_A^2)$, across different ρ_A^2 values. In Appendix B.3.2, we examine how well we can identify the interaction between assembly and component(s) with or without a median product. In Appendix B.3.3, we assess the choices of design parameters for Phase I, namely, r and k, based on $SD(\hat{\rho}_A^2)$.

- Appendix B.4 explores the proposed Phase II setup and analysis. In Appendix B.4.1, we compare the ANOVA and LVR estimators using bias, standard deviation, and mean square error of $\hat{\rho}_C^2$, denoted as $Bias(\hat{\rho}_C^2)$, $SD(\hat{\rho}_C^2)$, and $MSE(\hat{\rho}_C^2)$, across different ρ_C^2 values. In Appendix B.4.2, we examine how well we can identify the interaction between two or more components using the Partial and Extreme criteria for different magnitudes of interaction effects.

The main findings are presented below for each possible scenario.

3.4.1 Scenario: Assembly is the Dominant Cause

Our proposed procedure is more reliable than Shainin's Phase I procedure in identifying assembly as the dominant cause when it is indeed the dominant cause. Figure 3.3 compares the probability of identifying assembly as the dominant cause for different ρ_A^2 values using Shainin's procedure (with their recommendation of $k = r = 2$), our proposed estimator with Shainin's recommended sample size of $k = r = 2$, and our proposed estimator with our recommended sample size of $k = 3$ and $r = 5$. Moreover, we use a threshold of 0.5, that is, we identify assembly as the sole dominant cause if $\hat{\rho}_A^2 \geq 0.5$.

Figure 3.3 reveals what we noted earlier, namely, that Shainin's procedure has a small chance of identifying assembly as the dominant cause, even when it is indeed the dominant cause. For instance, when $\rho_A^2 = 0.6$, Shainin's procedure identifies assembly as the dominant cause only 30% of the time. However, our combined estimator with the same sample size as Shainin's one (i.e., $k = r = 2$) performs much better and identifies assembly as the dominant cause 68% of the time for $\rho_A^2 = 0.6$. With our recommended sample size ($k = 3$ and $r = 5$), the proposed estimator has a detection probability of 82% for $\rho_A^2 = 0.6$.

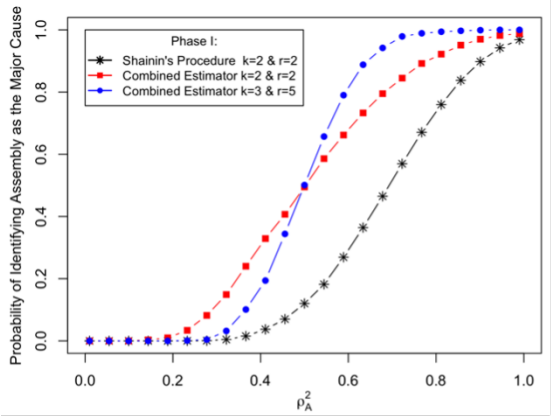

Figure 3.3: Probability of identifying assembly as the dominant cause vs. ρ_A^2 using Shainin's procedure, our proposed estimator with Shainin's recommended sample size, and our proposed procedure with $k = 3$ and $r = 5$.

3.4.2 Scenario: Interaction between Assembly and Components is the Dominant Cause

Shainin's procedure has no effective means of identifying an interaction effect between assembly and components. Our proposed procedure identifies such interactions using an equal variances test, recommended as part of the Phase I analysis. Especially when a median product is included in Phase I (as recommended), the proposed procedure reliably identifies assembly by component interactions. For instance, consider the scenario described in detail in Appendix B.3.2, where there was a large interaction between assembly and component(s). Then, by only selecting two extreme products, the probability of detecting the interaction is 12% and 8% using Bartlett's and Levene's tests, respectively. However, by also including the median product, we increase the probabilities to 90% and 20%, respectively. See Appendix B.3.2 for a detailed discussion.

3.4.3 Scenario: One Component is the Dominant Cause

The LVR and ANOVA estimators are more reliable than Shainin's Phase II procedure in identifying a single component as the dominant cause when it is indeed the sole dominant cause. Figure 3.4 compares the probability of identifying a single component C_1 as the dominant cause for different $\rho_{C_1}^2$ values using Shainin's procedure, ANOVA, and LVR when only two components exist, and $\rho_{C_2}^2 = 1 - \rho_{C_1}^2 - \rho_A^2$, and $\rho_A^2 = 0.05$. Note that in this

simulation, for Shainin's procedure, we follow their recommendation of $r = 2$, and for ANOVA and LVR estimators, we follow our recommendation of $r = 5$. Moreover, we use a threshold of 0.5, that is, we identify component C_1 as the sole dominant cause if $\hat{\rho}^2_{C_1} \geq 0.5$.

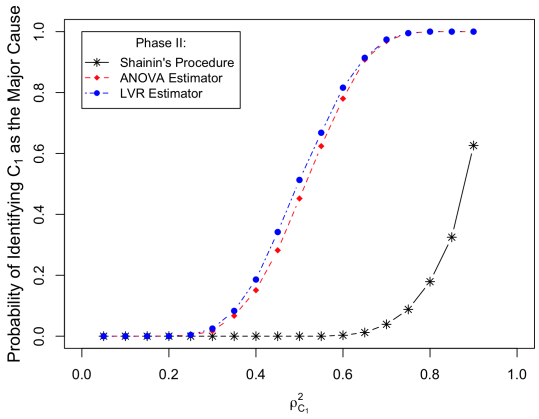

Figure 3.4: Probability of identifying C_1 as the dominant cause vs. $\rho^2_{C_1}$ using Shainin's procedure, ANOVA, and LVR estimators when there are only two components.

Figure 3.4 reveals that Shainin's procedure is unreliable when the dominant cause is a single component. For instance, when $\rho^2_{C_1} = 0.6$, the probability of identifying C_1 as a dominant cause is as low as 0.005% using Shainin's procedure. Therefore, as Appendix B.2.2 reveals, Shainin's procedure often identifies the combined effect of two or more components as the dominant cause, even when it is actually a single component.

Figure 3.4 also reveals that both LVR and ANOVA estimators are more reliable than Shainin's procedure in correctly identifying a single dominant cause. The performances of the LVR and ANOVA estimators are quite similar; however, Appendix B.4.1 reveals that the LVR estimator is more robust against large assembly variation. Therefore, we recommend using the LVR estimator for Phase II of component swapping.

3.4.4 Scenario: Interaction between two Components is the Dominant Cause

Shainin's procedure cannot distinguish between scenarios where the dominant cause is the additive effect of two components, i.e., $C_i + C_j$, and cases where the dominant cause is due to an interaction between these components, i.e., $C_i C_j$. Our proposed Phase II procedure addresses this issue by supplementing the LVR or ANOVA analysis with the Partial and

Extreme criteria. If either the Partial or Extreme criteria is satisfied, we conclude there is evidence of an interaction. This approach often detects evidence of such interactions, as elaborated in detail in Appendix B.4.2.

3.5 Discussion

The efficiency of component swapping is commonly attributed to its implementation of leveraging, ranking components, and hierarchical study design. Employing the principle of leveraging reduces the number of runs by focusing only on the most informative products, i.e., the ones with extreme output values. Moreover, the procedure swaps components in a user-defined order (based on engineering insights). If the initial ranking is reasonably successful, it substantially reduces the total number of required runs, particularly when the product has a large number of components (Dasgupta et al. (2011)). Furthermore, many systems can be decomposed into a *hierarchical structure* (Chittaro and Ranon (2004); De Mast (2011, 2013)), where products consist of subassemblies, and these subassemblies, in turn, consist of individual components (or smaller subassemblies). Following this principle, we enhance the efficiency and narrow the search space by first identifying the subassembly associated with the dominant cause, and then searching for the dominant cause only among the components of the identified subassembly.

At first glance, a readily apparent alternative to the component-swapping procedure for identifying the dominant cause(s) of variation is a factorial experiment with the components treated as factors. Despite the resemblance, there is a fundamental distinction between the presented component-swapping procedure and a factorial experiment. Namely, the -1 and $+1$ values in Table 3.3 signify whether components originate from products with extremely low and high output values in the baseline. These values do not correspond to the low and high settings that an experimenter selected for a factor (some unknown characteristic of a component) for the experiment. As a result, the standard design of experiments framework is not directly applicable within the context of component swapping. To illustrate the point, we note that factorial designs, lauded for being able to identify interactions in the normal context, cannot identify interactions when applied in the context of component swapping. For example, consider a scenario with only two components, C_1 and C_2, where a large interaction between them is the dominant cause, as depicted in Figure 3.5. This figure presents a classic interaction plot, showing output values for the four combinations of two possible values for C_1 and C_2, denoted as $-$ and $+$ to represent small and large values, respectively. In this scenario, combinations $(C_1, C_2) = (-, +)$ [denoted "*b*" in Figure 3.5] and $(C_1, C_2) = (+, -)$ [denoted "*c*" in Figure 3.5] both yield output values (Y) near the

bottom of the baseline distribution for Y. Similarly, $(C_1, C_2) = (-,-)$ [denoted "d" in Figure 3.5] and $(C_1, C_2) = (+,+)$ [denoted "a" in Figure 3.5] both yield output values (Y) near the top of the baseline distribution. In this scenario, when we select a pair of extreme products from the baseline for the component swap investigation, we are likely to select some combination of products a or d (for the high product) and b or c (for the low product). However, note that in a component swap investigation, the true relationship between C_1, C_2, and Y (as shown in Figure 3.5) is unknown to us. Our awareness is limited to identifying which components (C_1 and C_2) originate from the low and high products (and accordingly, we label them in our analysis). Suppose we happen to select products a and b in Figure 3.5 as the two extremes to use in the component swapping. For these two assemblies, C_2 is at virtually the same level, while C_1 is very different. Thus, swapping components C_1 will seem to explain all the variation in Y, and we will misidentify C_1 (rather than the C_1C_2 interaction) as the dominant cause. Alternatively, we could just as likely select another extreme pair, say products a and c in Figure 3.5. With this pair, on the other hand, C_2 seems to explain all the variation in Y, and we will misidentify C_2 as the dominant cause.

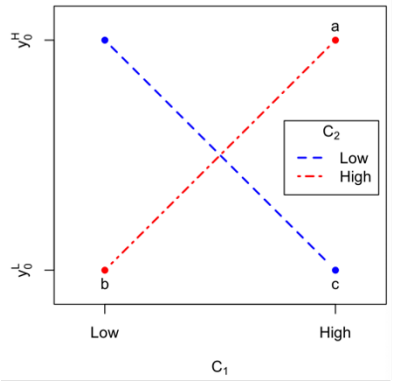

Figure 3.5: Interaction plot of the artificial example where the dominant cause is a large interaction between C_1 and C_2.

Figure 3.5 addresses the pure interaction case. However, we also fail to identify other types of interactions. This is not surprising due to the problem setup, that is how we select the extreme products for component swapping investigation and assign the levels for the components. To explain, consider the situation where we only swap component C_1. In this case, if we include the Phase I data, we have results from four different combinations of components C_1 and all the remaining components (denoted as C_2 here for simplicity). This appears to correspond to a 2^2 factorial experiment with some constraints. In particular, we have $y^L_{C_1=L} \approxeq y^L_0$, $y^H_{C_1=H} \approxeq y^H_0$, and $y^L_0 \leq y^L_{C_1=H}, y^H_{C_1=L} \leq y^H_0$. That is, we expect the

results from swapping C_1 to almost always lie between the results obtained from the two extreme products selected from the baseline. A consequence of this constraint is that the traditional estimate of the interaction for a 2^2 factorial experiment is largest when it equals the two main effects (see Appendix B.4.3). As a result, the interaction effect cannot be larger than 33% of the total. Therefore, as Appendix B.4.2.3 shows, without checking for irregularities (as recommended), neither the ANOVA nor LVR methods identify evidence of such large interaction.

3.6 Conclusions

Component swapping is a widely used strategy to identify the causes of problems in assembled products. Despite its practical popularity and intuitive appeal, there has been a noticeable absence of a comprehensive and systematic study of its merits and drawbacks. In this chapter, while keeping valuable ideas from the existing literature, we introduced our proposed procedure. We compared our proposal to one of the most well-known alternatives, namely Shainin's component-swapping procedure. Our evaluation demonstrated that Shainin's procedure suffers from poorly chosen statistical tools and analysis, and it frames the analysis as a significance test rather than an estimation problem. We demonstrated how unreliable these tools are, as they often result in incorrect conclusions. Moreover, whereas the literature offers no effective means of identifying interaction effects, our proposal signals to users that interactions may affect the results. Notably, we distinguish between interactions among assembly and components and interactions among two or more components.

Chapter 4

Verification of a Dominant Cause of Output Variation

Finding the dominant cause(s) of variation in process improvement projects is an important task. Before trying to reduce variation in the dominant cause or mitigate the effect of variation in the dominant cause to reduce output variation, it is strongly recommended that we verify we have identified the true (dominant) cause. This chapter is about how best to verify we have correctly identified a dominant cause, as the existing literature does not properly answer this question. Although it may seem that a randomized controlled experiment is sufficient for this purpose, we show that experimental studies alone cannot provide all the required information. A carefully planned experiment can identify whether a suspect is a cause of variation; however, we also require additional information (from observational studies) to determine whether it is dominant and not just significant. This chapter lists some viable composite study designs, assesses their relative merits, and recommends proper sample sizes. We also investigate how to systematically conduct a verification study in the era of smart manufacturing. Moreover, we provide a tangible example to illustrate our proposed procedure.

4.1 Introduction

The search for the dominant cause(s) is typically difficult due to the large number of possible important causes, some of which are poorly defined or unidentified properties of the process. Thus, traditional strategies such as brainstorming about the suspect dominant

causes followed by experiments to establish the effect of each of these causes can be overwhelming or lead to an incorrect search space (Mooren et al. (2012); De Mast et al. (2019)). To overcome this problem, Statistical Engineering and the Shainin System^{TM} propose a sequential approach called the method of elimination, which usually consists of a series of *observational* studies (e.g., multivari study (De Mast et al. (2001)), variation transmission (Steiner and MacKay (2005)), group comparison (Panahi et al. (2021))) or off-line experimental studies (e.g., component swap (Steiner and MacKay (2005))). The idea is to start with a large number of suspect dominant causes and to eliminate groups of suspects after each investigation, thereby homing in on the actual dominant cause(s).

After shortlisting the suspect dominant causes to only one or just a few process inputs (using the method of elimination or other approaches), Steiner and MacKay (2005) strongly recommend *verifying* that we have identified the true (dominant) cause of the output variation, i.e., making sure that the suspect(s) is both a *cause* of variation in the output and also *dominant* before we move to the remedial journey. We require a verification study because in the search for the dominant cause using the method of elimination, often many investigations are *observational*. Although these observational studies are appropriate for clue generation, they typically lack the rigor and systematic design that allows for strong causal inferences. If we misidentify the (dominant) cause, it will be challenging to successfully improve the process successfully, and we may waste considerable resources in a futile search for a way to improve the process. This chapter focuses on how best to verify the suspected dominant cause(s).

We believe the proper design of verification studies is substantially more challenging than statisticians and process engineers may believe, which motivates this chapter. One issue is that to verify a dominant cause, it is insufficient to merely estimate a variance component (Searle et al. (1992)). Instead, it is necessary to establish the causal mechanism that produces variation in the process output, and therefore, a verification study must allow causal inference (De Mast et al. (2023)). Second, the challenge in a verification study is not in the statistical models and their analysis (which can be straightforward), but instead in collecting a useful combination of experimental and observational data. Therefore, interest lies in the design of experiments, the design of observational studies, and a strong awareness of what can and cannot be estimated from each.

The structure of this chapter is as follows. In Section 4.2, we define the notation and the goal of the verification study. Section 4.3 discusses different types of experimental and observational study designs. Section 4.4 explains some viable combinations of study designs in the context of verifying a dominant cause. It also compares the relative merits of each design through simulation studies. Section 4.5 applies the proposed procedure to a realistic example. Section 4.6 discusses how to verify a dominant cause in the era of smart

manufacturing. We provide a conclusion and discussion in Section 4.7.

4.2 Motivation and Definitions

This section defines the goal of the verification study. However, we first present a motivating example to demonstrate the problem's complexity and importance.

Consider the production of agricultural produce, where a quality characteristic Y must be maintained within the tolerance limits. Excessive variation in Y is primarily attributed to the use of different types of fertilizers across various plots of land (mostly type A, but for some plots, types B or C). To investigate the situation, statisticians would typically collect data on the corresponding Y from plots of land treated with different fertilizer types (A, B, and C) and fit an ANOVA model such as

$$Y_{ij} = \mu_i + \epsilon_{ij},$$

where μ_i represents the fixed effect of the fertilizer type i for $i = 1, 2, 3$ (A, B, C). Additionally, ϵ_{ij} represents an i.i.d. $N(0, \sigma_\epsilon^2)$ error component that accounts for all other sources of variation. An analysis of variance allows us to determine whether the fertilizer type is a significant source of variation in Y, and a decomposition of sums of squares allows an assessment of whether this factor appears to be a *dominant* component of variation. However, even if in the collected data, fertilizer type accounts for a substantial part of the variation in Y, this does not demonstrate that the type of fertilizer *causes* this variation, and therefore, intervention in this factor (e.g., using only a single type of fertilizer) may not result in smaller variation in Y. The reason is that the observed fertilizer effect may be spurious due to a confounding factor. For instance, more than one farmer may work the land, each using a different type of fertilizer, but the true cause of variation in Y is due to other differences in the farmers' working methods unrelated to the type of fertilizer. Observational studies and variance decompositions are unsuited to distinguish spurious effects caused by confounders from true causal effects; therefore, in isolation, they are insufficient for the verification study.

The go-to approach for studying causal effects has long been randomized controlled experiments (including screening experiments), as they can demonstrate cause-and-effect relationships when designed and executed carefully (Holland (1986); Rubin (2005); Pearl (2009)). Suppose that, in addition to fertilizer type, crop density (i.e., number of plants per square meter) is a suspected cause of variation in Y. Then, for example, a 3^2 factorial experiment could be conducted with both *fertilizer* and *crop density* set at three levels

each and four replications (36 total runs). For causal inference purposes, plots of land should be *randomly* assigned to each run. From the experiment, a model such as

$$Y_{ij} = \mu_i + \beta\, X_j + \epsilon_{ij}$$

can be fitted, where crop density $X_j \sim N(\mu_X, \sigma_X^2)$, and β represents its effect on Y. Let p_1, p_2, and p_3 denote the probabilities that a customer gets a product treated with fertilizers A, B, or C. Then, the quality variation of a randomly selected product is $Var(Y) = \sum_{i=1}^{3} p_i(\mu_i - \sum p_j\mu_j)^2 + \beta^2\sigma_X^2 + \sigma_\epsilon^2$. The proportion of variation in Y attributed to crop density would then be $\beta^2\,\sigma_X^2/\mathrm{Var}(Y)$. From the experiment, the estimated β would represent the causal effect of X onto Y (the possible spurious effects created by confounders are likely minimized by randomization). However, the problem is that σ_X^2 (i.e., the variance of the crop density in the fields) cannot be estimated from the experiment. The reason is that the crop densities used in the 36 runs were not randomly drawn from $N(\mu_X, \sigma_X^2)$, but instead were set by the experimenter according to the chosen low, middle, and high levels of the 3^2 factorial design (the *controlled* part in a randomized controlled experiment). For the same reason, the experimental data also cannot be used to estimate $\mathrm{Var}(Y)$. Therefore, to estimate $\beta^2\,\sigma_X^2/\mathrm{Var}(Y)$ in the population, the experimenter must also consider previously collected observational data or collect additional observational data. This could be a sample of x values representative of $N(\mu_X, \sigma_X^2)$, but it could also be a sample of paired (x, y) observations. Therefore, a verification study requires a combination of data collected in a statistically designed experiment and additional observational data drawn from the data-generating distribution (further possibilities will be discussed in Section 4.6).

In what follows, we introduce our notation and definitions, as well as formalize the goal of the verification study. For simplicity, consider the situation where we have only one suspect dominant cause to verify. Note that since we assume that the method of elimination has been utilized to narrow down the possible suspect dominant causes (see Steiner and MacKay (2005) for further discussion), having only a small number of remaining suspects is typical in the context of a verification study. However, we also briefly discuss verification studies with multiple suspect dominant causes in Section 4.7.

To facilitate a more precise discussion of these concepts, assume that the effect of a continuous suspect dominant cause X on the output Y is approximately linear, with intercept α and slope β, i.e.

$$Y = \alpha + \beta\, X + \epsilon, \tag{4.1}$$

where ϵ is an i.i.d. error component including noise (e.g., measurement variation) and the effects of all causes other than X. Since X is the suspect dominant cause of variation in Y, it is a process input that naturally varies as the process operates under normal operating

procedures, including process controls [1]. Denoting $E(X) = \mu_X$, $Var(X) = \sigma_X^2$, $Var(\epsilon) = \sigma_\epsilon^2$, and assuming that X and ϵ are independent, we obtain $\sigma_Y^2 = Var(Y) = \beta^2\,\sigma_X^2 + \sigma_\epsilon^2$. Then, X is a large cause of variation if holding it fixed would substantially reduce the output variation σ_Y^2. A suspect cause X is strictly a *dominant cause* of variation only if $\beta^2\,\sigma_X^2 > \sigma_\epsilon^2$. In what follows, it is convenient to consider the squared correlation between X and Y, given by

$$\rho^2 = \frac{\beta^2\,\sigma_X^2}{\beta^2\,\sigma_X^2 + \sigma_\epsilon^2}, \tag{4.2}$$

where $0 \leq \rho^2 \leq 1$. Using this parameterization, X is strictly a dominant cause if $\rho^2 > 0.5$, since then $\beta^2\sigma_X^2 > \sigma_\epsilon^2$. Note that here we deliberately propose a simple yet reasonable model to keep the focus on the study design. Nevertheless, the conclusion drawn from this study can be readily extended to more complex models.

To verify X as a dominant cause in Model 4.1, the goal is to determine whether:

> The *causal* contribution of X to the output variation (i.e., $\beta^2\sigma_X^2$) is large compared to the variation due to noise and other causes (i.e., σ_ϵ^2).

Meeting this goal requires estimating three parameters: β, σ_X, and σ_ϵ, or two of these three parameters along with σ_Y, given that $\sigma_Y^2 = \beta^2\sigma_X^2 + \sigma_\epsilon^2$. Note that besides these parameters, we also need to estimate the nuisance parameters α and μ_X to fit Model 4.1.

Plugging in the estimated parameters into Equation 4.2, we obtain an estimate for ρ^2. Subsequently, we can compare $\hat{\rho}^2$ to a predetermined threshold, such as 0.5 (or any other threshold if we wish to verify large but not necessarily strictly dominant causes), to determine whether X is a dominant (important) cause. A more sophisticated alternative involves using the confidence interval for the estimate of ρ^2 to make a decision. We leave it up to the practitioner to interpret the procedure's estimates (and uncertainty bounds) in a way consistent with their context (this could depend on other factors such as cost).

As the motivating example illustrates, we need observational data to estimate σ_X, and an experiment alone does not provide sufficient information. Moreover, it is important that the estimated β represents the *causal* effect of X onto Y, and is not affected by spurious correlations. Therefore, two potential approaches to estimate β (and thus σ_ϵ) are randomized controlled experiments and causal inference from observational data. In the next section, we focus on study designs that combine randomized controlled experiments

[1]Note that we treat X as a random variable. Therefore, we implicitly assume that even if X is deliberately changed on occasion (e.g., due to process controllers), we look at the process over a long enough time to justify modelling X as a random variable.

and observational data. In Section 4.6, we discuss study designs that allow causal inference from observational data alone.

4.3 Study Designs and Models

This section outlines some potential observational and experimental study designs that could be utilized to achieve the verification study goal and identifies which parameters can be estimated from each design. We will use the notation $(x, y)_E$ to denote data consisting of tuples of paired x, y values from an experiment, whereas, for example, we use $(y)_O$ to denote data consisting of only output values from an observational investigation. Later in Section 4.4, we determine some viable data combinations to verify a dominant cause.

In the following, for simplicity, we add a normality assumption for X and ϵ so that $X \sim N(\mu_X, \sigma_X^2)$ and $\epsilon \sim N(0, \sigma_\epsilon^2)$ in Model 4.1. Then, $Y \sim N(\alpha + \beta\mu_X, \beta^2\sigma_X^2 + \sigma_\epsilon^2)$, and (X, Y) has a bivariate normal distribution with means and variances as given above and a correlation ρ, where ρ^2 is given by Equation 4.2. As mentioned earlier, we deliberately assume simple models to focus attention on the design questions. Section 4.7 discusses other model assumptions and multiple suspect dominant causes in the verification experiment. At the end of each of Subsections 4.3.1 through 4.3.4, we provide the corresponding log-likelihoods, which will be used in the simulation study presented in Section 4.4 to evaluate the relative merits of the viable study designs.

4.3.1 Experimental $(x, y)_E$ Data

An obvious way to check the causal link between X and Y and estimate β is to conduct an experiment, in which we deliberately manipulate the levels of X and apply experimental design principles such as randomization, replication, blocking, and balance. Note that since we use experimental designs, we assume it is possible to set X to any desired level. This may not be straightforward in this context since X is a process input that naturally varies as the process operates (and this is why the effect of varying X results in variation in a process output).

For the verification experiment, when X is continuous and the relationship between X and Y is linear, we recommend selecting only two levels at the extreme ends of the X values that we expect to encounter in the regular operation of the process. If X is continuous and the relationship between X and Y is nonlinear, we probably want to select three or more levels of X. While it is desirable to select extreme levels of X (to obtain more precise

estimates of β), it is important to avoid choosing extreme levels that exceed the typical range observed in the process. The reason is to prevent any process equipment potential damage/break or to violate the presumed linear/nonlinear relationship.

Assuming Model 4.1, we conduct a simple two-level experiment with the low and high settings of X, and $\frac{n_E}{2}$ replications at each level, where n_E is the total number of experimental runs. Also, we order the replicates randomly. Then, we have $(x_i, y_i)_E$ data where x_i is either the low or high setting for $i = 1, 2, \ldots, n_E$. With this experiment, we can estimate β and σ_ϵ in Model 4.1. Then, to establish a causal relationship, we want to see a small p-value for the significance test of $H_0 : \beta = 0$ vs. $H_A : \beta \neq 0$ from the experimental data. Recall that with the experimental data, since we deliberately set the levels of X, we cannot estimate σ_X, and thus, ρ^2. For a similar reason, the coefficient of determination R^2 calculated from the experimental data is not a reliable estimator for ρ^2. In summary, $(x, y)_E$ data are sufficient to estimate β and σ_ϵ, but not σ_X.

By considering Model 4.1 and the normality assumptions for ϵ, we have $Y|x \sim N(\alpha + \beta x, \sigma_\epsilon^2)$ for any fixed x. The log-likelihood when the data come from such an experiment depends only on α, β, and σ_ϵ, and is

$$l_E = -\frac{n_E}{2} \log(2\pi\sigma_\epsilon^2) + \sum_{i=1}^{n_E} -\frac{1}{2\sigma_\epsilon^2}\big(y_i - (\alpha + \beta x_i)\big)^2. \tag{4.3}$$

4.3.2 Observational Paired $(x, y)_O$ Data

Suppose we select a representative sample of paired X, Y values collected from the process operating under regular conditions. These data might already be available from previous method of elimination investigations. However, if such data are unavailable, we could collect them by measuring X, tracing parts through the process, and then measuring their corresponding Y values. Note that part tracing may be difficult and expensive in some processes. Using data from such a study and Model 4.1, we can estimate β, σ_X, and σ_ϵ. However, when a confounding variable, denoted C, exists, the estimated β and σ_ϵ from $(x, y)_O$ data may fail to accurately reflect the direct causal effect of X onto Y. To better illustrate this issue, suppose we observe a correlation between X and Y in an observational data set, while the true causal effects are as depicted by the arrows in the causal diagram in Figure 4.1.

In Figure 4.1, the correlation between X and Y in $(x, y)_O$ data aliases the direct effect of X onto Y with the spurious correlation induced by the confounding variable C. Consequently, reliable estimation of β from $(x, y)_O$ data is not possible unless we can either

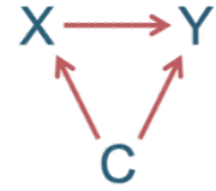

Figure 4.1: A possible causal diagram between X, Y, and C.

assume no C exists, or we also have data on the confounder C. Generally, the causal mechanisms involved are often complex, creating myriad confounding effects that make the estimation of β from $(x, y)_O$ data treacherous. One potential solution is combining the observational and experimental data, as we discussed in this and the following section. An alternative approach is to determine the *right* confounders as far as possible and collect $(x, y, c)_O$ data. By including the right confounders as explanatory variables in the analysis, spurious correlations can be eliminated. This approach is discussed in Section 4.6. In summary, if we assume no C exists, $(x, y)_O$ data are enough to estimate β, $\sigma_{_X}$, and σ_ϵ. However, if confounder issues are not clear and effectively addressed, we must treat $(x, y)_O$ data as $(x)_O\&(y)_O$ data for the verification study.

Since (X, Y) has a bivariate normal distribution as discussed in Section 4.3, the log-likelihood with n_O observations of $(x, y)_O$ data is

$$\begin{aligned} l_O = &- n_O \log\left(2\pi\sigma_{_X}\sqrt{(\beta^2\,\sigma_{_X}^2+\sigma_\epsilon^2)(1-\rho^2)}\right) \\ &+ \frac{1}{2(1-\rho^2)}\sum_{i=1}^{n_O}\left(\frac{(x_i-\mu_{_X})^2}{\sigma_{_X}^2}+\frac{(y_i-\alpha-\beta\mu_{_X})^2}{\beta^2\,\sigma_{_X}^2+\sigma_\epsilon^2}-\frac{2\,\rho\,(x_i-\mu_{_X})(y_i-\alpha-\beta\mu_{_X})}{\sigma_{_X}\sqrt{\beta^2\,\sigma_{_X}^2+\sigma_\epsilon^2}}\right). \end{aligned} \tag{4.4}$$

4.3.3 Observational $(x)_O$ Data

Another type of data is a representative sample of only X values. This data is usually less expensive to gather than $(x, y)_O$ data, and in some situations, $(x)_O$ data may already be available because it was collected for another purpose. From $(x)_O$ data, we can estimate $\sigma_{_X}$, but clearly not σ_ϵ and β. In summary, $(x)_O$ data alone are not enough to verify a dominant cause, but they may complement data from another study.

Since $X \sim N(\mu_{_X}, \sigma_{_X}^2)$, the corresponding log-likelihood with n_{Ox} observations of $(x)_O$

data is

$$l_{Ox} = -\frac{n_{Ox}}{2}\log(2\pi\sigma_X^2) + \sum_{i=1}^{n_{Ox}} -\frac{1}{2\sigma_X^2}(x_i - \mu_X)^2. \tag{4.5}$$

4.3.4 Observational $(y)_O$ Data

We could also select a representative sample of only Y values. This data is also usually less expensive to gather than $(x, y)_O$ data, and in fact, are often collected in the first step of variation reduction projects to establish the problem baseline. From $(y)_O$ data, we can estimate $\sigma_Y^2 = \beta^2\,\sigma_X^2 + \sigma_\epsilon^2$, but clearly not any of the three parameters individually. Similar to $(x)_O$ data, $(y)_O$ data alone are not enough to verify a dominant cause, but they may complement data from another study.

Since $Y \sim N(\alpha + \beta\mu_X, \beta^2\sigma_X^2 + \sigma_\epsilon^2)$, the corresponding log-likelihood with n_{Oy} observations of $(y)_O$ data is

$$l_{Oy} = -\frac{n_{Oy}}{2}\log\left(2\pi(\beta^2\sigma_X^2 + \sigma_\epsilon^2)\right) + \sum_{i=1}^{n_{Oy}} -\frac{1}{2\left(\beta^2\sigma_X^2 + \sigma_\epsilon^2\right)}\left(y_i - \alpha - \beta\mu_X\right)^2. \tag{4.6}$$

4.4 Some Viable Composite Study Designs

In practice, we may have different combinations of data from different study designs. Table 4.1 summarises the notation and what we can estimate from the four study designs discussed in Section 4.3.

Data	Sample size	What can we estimate?	Log-likelihood function
$(x, y)_E$	n_E	β, σ_ϵ	l_E
$(x, y)_O$	n_O	β, σ_X, σ_ϵ (assuming no C)	l_O
$(x)_O$	n_{Ox}	σ_X	l_{Ox}
$(y)_O$	n_{Oy}	$\beta^2\sigma_X^2 + \sigma_\epsilon^2$	l_{Oy}

Table 4.1: Different verification study designs and notation.

In this section, we present four viable composite study designs appropriate for verifying a dominant cause and briefly explain them. When combining the data from various sources, it is of course important to establish that all data we collected are in the same reference period.

(i) $(x, y)_O$ and $(x, y)_E$
In this case, we can estimate σ_X and $\beta^2\,\sigma_X^2 + \sigma_\epsilon^2$ using only $(x, y)_O$ data. To estimate β and σ_ϵ, we can use both $(x, y)_O$ and $(x, y)_E$ data, assuming no confounder exists. However, if a confounder exists, the estimated β and σ_ϵ from $(x, y)_O$ data may not accurately reflect the direct causal effect of X onto Y. Thus, with this combination of data sources, we suggest testing the hypothesis of whether β from $(x, y)_O$ data is equal to β from $(x, y)_E$ data (Cohen et al. (2003) pp. 46-67). If the hypothesis is not rejected, we suggest combining $(x, y)_O$ and $(x, y)_E$ data to estimate β and σ_ϵ. As we typically have more observational than experimental data, using the pooled estimate of σ_ϵ in Equation 4.2 leads to more precise estimate of ρ^2. However, if the hypothesis test is rejected, since the β and σ_ϵ estimated from $(x, y)_O$ data do not reliably reflect the causal effect, we must treat $(x, y)_O$ data as $(x)_O\&(y)_O$ data for the verification study. This scenario is discussed in situation (ii).

(ii) $(x)_O\&(y)_O$ and $(x, y)_E$
In this case, we can estimate σ_X from $(x)_O$ data and $\sigma_Y^2 = \beta^2\sigma_X^2 + \sigma_\epsilon^2$ from $(y)_O$ data. We can also estimate β and σ_ϵ from $(x, y)_E$ data. This enables us to estimate ρ^2.

(iii) $(x)_O$ and $(x, y)_E$
In this case, we can estimate σ_X from $(x)_O$ data as well as β and σ_ϵ from $(x, y)_E$ data. This enables us to estimate ρ^2.

(iv) $(y)_O$ and $(x, y)_E$
In this case, we can estimate $\sigma_Y^2 = \beta^2\,\sigma_X^2 + \sigma_\epsilon^2$ from $(y)_O$ data as well as β and σ_ϵ from $(x, y)_E$ data. In this case, we can solve for σ_X by plugging in the available estimates of β, σ_Y, and σ_ϵ. This enables us to estimate ρ^2.

Above, we identified four composite study designs suitable for verifying a dominant cause. In the following, we determine suitable sample sizes for each design, thereby establishing their relative required effort. To determine whether the relationship between X and Y is causal, we assess the effect of different experimental (n_E) based on the power of the hypothesis test $H_0 : \beta = 0$ vs. $H_A : \beta \neq 0$. Figure 4.2 represents the results of analytical power calculation (details are given in Appendix C.1) for $n_E \in \{6, 8, 10, \ldots, 32\}$ when $\rho^2 = 0.5$ and the two levels of X are selected as $\mu_{_X} \pm 2\sigma_X$. Figure 4.2 suggests that even with small sample sizes such as $n_E = 6$, we have strong power, and with $n_E = 8$, we achieve a power of almost one. In this context, having a high power is important because mistakenly eliminating X as the cause of variation in Y when it is the actual cause can waste considerable time and effort. Also, note that in this study, our focus is on $\rho^2 = 0.5$ as we are mainly interested in estimating the effects of strictly dominant inputs (i.e., $\rho^2 \geq 0.5$), and

$\rho^2 = 0.5$ has the lowest power among them. For other ρ^2 values, although the results are slightly different, the overall conclusions remain the same. Moreover, the power increases if we select more extreme levels of X for our experiment.

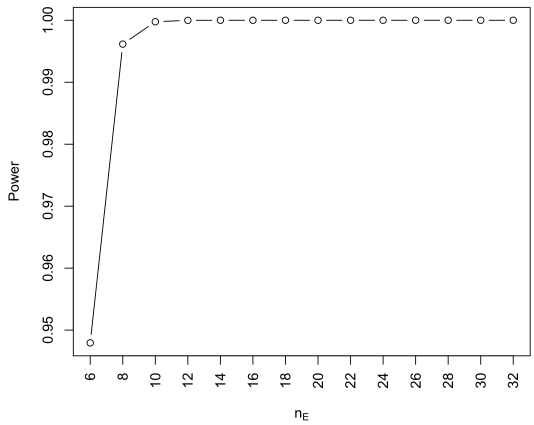

Figure 4.2: Power of the test $H_0 : \beta = 0$ vs. $H_A : \beta \neq 0$ for $n_E \in \{6, 8, 10, \ldots, 32\}$ when $\rho^2 = 0.5$ and the X levels are $\mu_X \pm 2\sigma_X$.

Next, through simulation, we investigate how well we can estimate ρ^2 for each viable composite study design. Since in these cases we always have $(x, y)_E$, we will refer to the different combinations by only the type of observational data they contain.

In the simulation, for each study design, we calculate the standard deviation of the estimated ρ^2 values for different sample sizes, assuming $\rho^2 = 0.5$. In Model 4.1, we have five parameters to estimate, namely α, β, σ_ϵ, μ_X, and σ_X, but ρ^2 depends only on β, σ_ϵ, and σ_X. In the simulation, without loss of generality, we generate data with $\alpha = 0$, $\mu_X = 0$, and $\sigma_X = 1$. Also, we fix $\beta = 1$ and determine the corresponding σ_ϵ so that $\rho^2 = 0.5$, i.e., we consider $\sigma_\epsilon = 1$. In each simulation run, we first estimate all the five parameters, and then we obtain $\hat{\rho}^2$ by plugging the estimates into Equation 4.2. Using 2000 simulation runs, we estimate the bias and standard deviation of $\hat{\rho}^2$, denoted by $Bias(\hat{\rho}^2)$ and $SD(\hat{\rho}^2)$, where $Bias(\hat{\rho}^2) = \hat{\rho}^2 - \rho^2$.

To estimate the parameters in Model 4.1 when we have $(x, y)_O$ or $(x)_O \& (y)_O$ data, we use maximum likelihood estimation. Assuming the independence of different parts used in each study, the overall log-likelihood can be written as the sum of the log-likelihoods with non-zero corresponding sample sizes provided by Equations 4.3 to 4.6.

We propose a similar but slightly different approach in cases with only $(x)_O$ or only $(y)_O$ data. Here, only $(x, y)_E$ data provide information about α, β, and σ_ϵ. In these cases,

if we use maximum likelihood to estimate these parameters since n_E is typically relatively small, the obtained $\hat{\sigma}_\epsilon$ will be biased. To correct the bias, we suggest instead to estimate σ_ϵ^2 with a $n_E - 2$ divisor, i.e., we multiply the estimate of σ_ϵ^2 from the maximum likelihood by $\frac{n_E}{n_E-2}$. If we only have $(x)_O$ data, μ_X and σ_X are estimated using the sample mean and sample variance of $(x)_O$ data. Similarly, if we only have $(y)_O$ data, μ_X and σ_X^2 are estimated using $\frac{(\bar{y}_O)-\hat{\alpha}}{\hat{\beta}}$ and $\text{Max}\big(\frac{s^2_{(y)_O}-\hat{\sigma}^2_\epsilon}{\hat{\beta}^2}, 0\big)$, respectively, where $(\bar{y})_O$ and $s^2_{(y)_O}$ are the sample mean and sample variance of $(y)_O$ data.

First, we investigate the relative merits of the viable composite study designs for $n_E = 8$ (recall that this experimental sample size gives very high power for the hypothesis test $H_0 : \beta = 0$ vs. $H_A : \beta \neq 0$ from $(x, y)_E$ data), where the observational sample sizes are $\{50, 100, 150, \ldots, 1000\}$. Figure 4.3 presents the results for $Bias(\hat{\rho}^2) = \hat{\rho}^2 - \rho^2$ and $SD(\hat{\rho}^2)$.

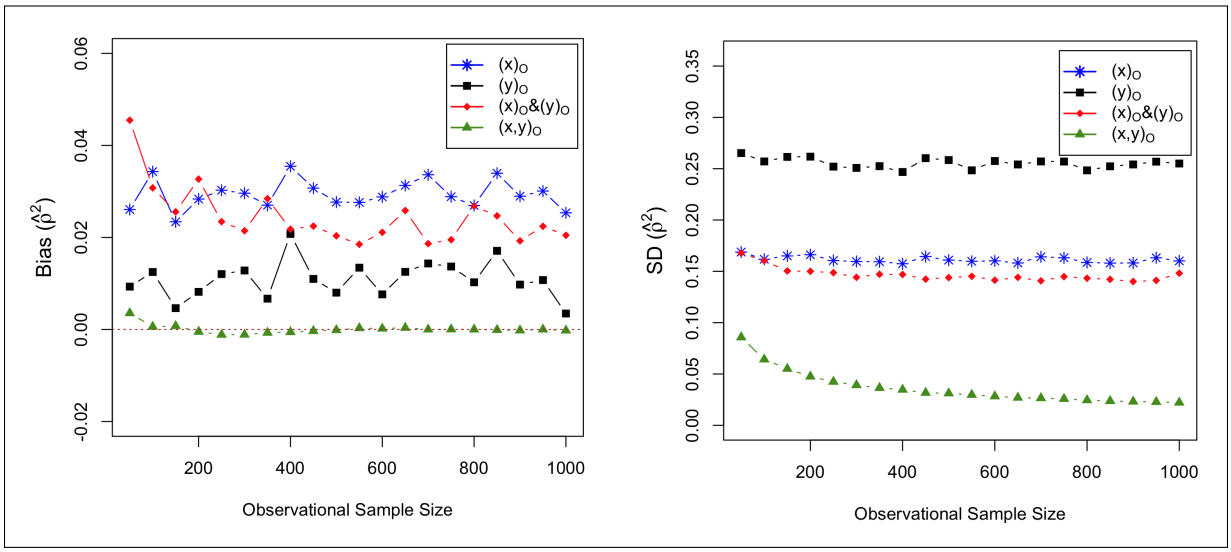

Figure 4.3: $Bias(\hat{\rho}^2)$ (left panel) and $SD(\hat{\rho}^2)$ (right panel) for different viable combinations of data when $\rho^2 = 0.5$ and $n_E = 8$.

The left panel of Figure 4.3 shows that the bias for all combinations is fairly small. As expected, the right panel reveals that the most precise estimates arise when we supplement the experiment with $(x, y)_O$ data. However, recall that to use such data, we must assume no confounder with a large influence exists. Figure 4.3 also reveals that only having $(y)_O$ data does not provide very valuable information in terms of $SD(\hat{\rho}^2)$, whereas $(x)_O$ data help considerably more. Having $(x)_O\&(y)_O$ data is only slightly better than $(x)_O$ data because $(y)_O$ data are not very informative, even though $(x)_O\&(y)_O$ data represent twice

as many observational data. Figure 4.3 also reveals that by increasing the observational sample sizes to more than roughly 200, $SD(\hat{\rho}^2)$ will not reduce much. We also investigated the mean squared error of $\hat{\rho}^2$ and the conclusions remain consistent.

Second, we investigate the relative merits of the viable composite study designs for n_O, n_{Ox}, or $n_{Oy} = 200$, when $n_E \in \{6, 8, 10, \ldots, 32\}$. Figure 4.4 presents the results for $Bias(\hat{\rho}^2) = \hat{\rho}^2 - \rho^2$ and $SD(\hat{\rho}^2)$.

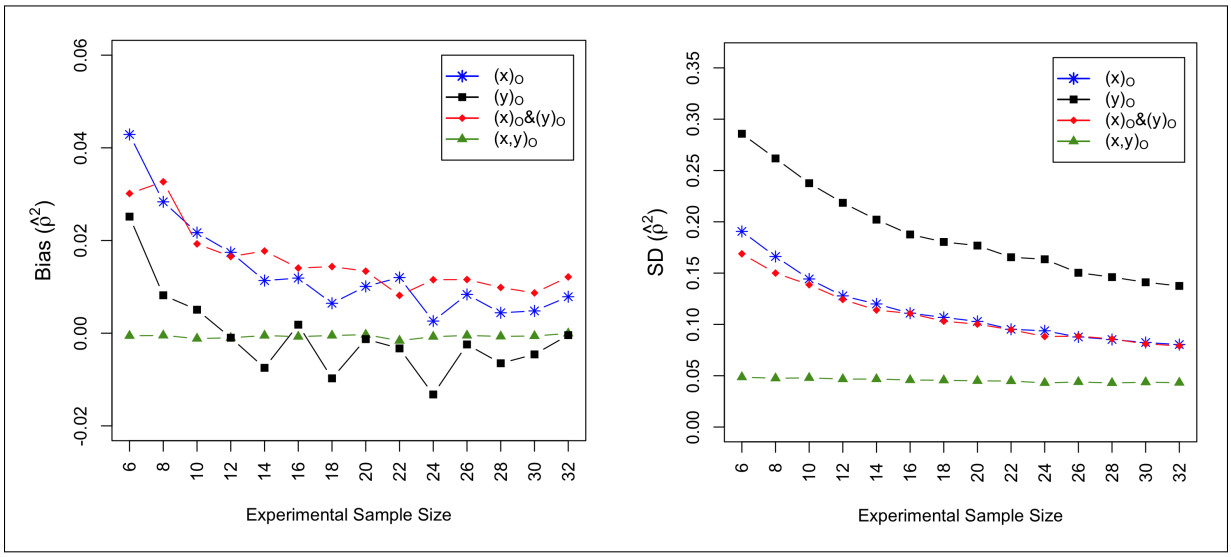

Figure 4.4: $Bias(\hat{\rho}^2)$ (left panel) and $SD(\hat{\rho}^2)$ (right panel) for different viable combinations of data when $\rho^2 = 0.5$ and there are 200 observational data.

The left panel of Figure 4.4 shows that the bias for all combinations is fairly small, where again $(x, y)_O$ data result in the least bias. The right panel reveals that when we have $(x)_O$, $(y)_O$, or $(x)_O\&(y)_O$ data, as n_E increases, $SD(\hat{\rho}^2)$ decreases. However, the results from $(x, y)_O$ data show that having more experimental data does not help in terms of $SD(\hat{\rho}^2)$. Therefore, in the absence of $(x, y)_O$ data, although the recommendation made by Shainin (1993b) to use $n_E = 6$ in a verification experiment results in high power for the hypothesis test, it is too small to precisely estimate ρ^2. In these cases, we leave it up to the practitioner to decide on the appropriate size of n_E in a way that the corresponding $SD(\hat{\rho}^2)$ makes sense in their context. We also investigated the mean squared error of $\hat{\rho}^2$ and the conclusions remain consistent.

4.5 Performance of the Proposed Procedure on an Example

Steiner and MacKay (2005, 2006) describe a project whose goal was to reduce the high *Crossbar Dimension* variation in a mold manufacturing context. If the crossbar was too long, components pressure fitted into the base tended to move or fall out. If the crossbar was too short, there was breakage during insertion.

To quantify the problem, they planned and executed a baseline investigation (with our notation, this corresponds to $(y)_O$ data) in which six consecutive parts were systematically sampled from the process each hour for five days. Note that the team did not use random selection because they were interested in how the process varied over time. The histogram of results (not shown here) demonstrates the full range of *Crossbar Dimension* seen in the baseline is -0.25 to 2.1, and the dominant cause acts hour-to-hour with some evidence of smaller day-to-day differences. From the baseline, the team estimated the overall standard deviation of the dimension (σ_Y) to be 0.45. The goal was to reduce it to less than 0.25.

The team tried to identify the cause of significant variation in *Crossbar Dimension* using a systematic search strategy. After ensuring that the measurement system was highly capable, they eliminated all process inputs that varied from one part to the next. They could identify only five inputs in the process that changed hour-to-hour. The team conducted an investigation in which 40 parts were selected systematically, four per hour over a two-day period. For each part, the team measured the *Crossbar Dimension* and recorded the five inputs: *Die Temperature*, *Nozzle Temperature*, *Barrel Temperature*, *Hydraulic Pressure*, and *Cavity Pressure*.

Using scatter plots, the team found a strong relationship between *Barrel Temperature* and *Crossbar Dimension* (shown in Figure 4.5), while there was little evidence of a relationship between *Crossbar Dimension* and the other inputs (not shown here). Since the range of dimension in the observational study matched that seen in the baseline, shown by the dashed lines in Figure 4.5, the team concluded that the dominant cause had acted. Hence, *Barrel Temperature* was a suspect dominant cause.

To verify the suspect dominant cause, the team conducted an experiment with the low level for *Barrel Temperature* as 75° and the high level as 80°. *Barrel Temperature* was difficult to hold fixed in normal production; however, it could be controlled for the experiment and changed in a few minutes. The team set the *Barrel Temperature*, made 25 parts to ensure a stabilized temperature, and measured the next ten parts. Then, they quickly changed the *Barrel Temperature* for the second run. There were two runs with

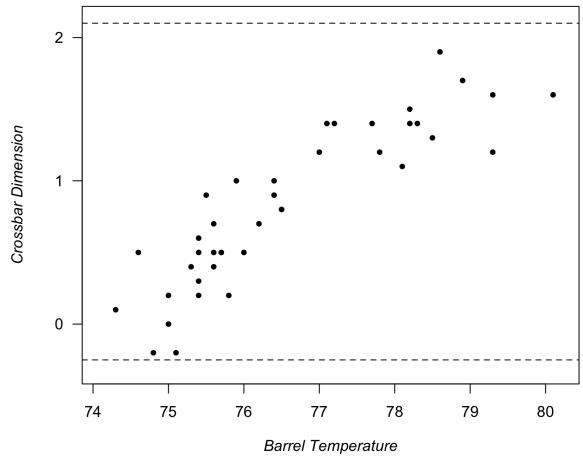

Figure 4.5: Observational data for *Crossbar Dimension* vs. *Barrel Temperature*.

ten repeats per run and no replication. Figure 4.6 presents the *Crossbar Dimension* values from the experiment.

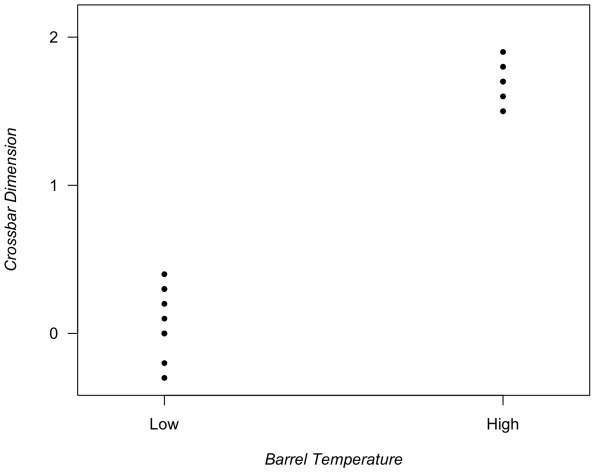

Figure 4.6: Experimental data for *Crossbar Dimension* vs. *Barrel Temperature*.

From Figure 4.6, the team informally verified that *Barrel Temperature* was the dominant cause of variation in *Crossbar Dimension* since the selected *Barrel Temperature* levels in the experiment were previously seen in the process and the *Crossbar Dimension* values matched the process baseline. However, we should be cautious about drawing a conclusion since we may worry about confounding between *Barrel Temperature* and some other input in the observational study. Also, in other similar situations, the results may not be as

clear. As a result, it would be helpful to have a more formal and systematic way to draw conclusions from a verification study.

To implement our proposed procedure to this example, we first perform a two-sample t-test on the experimental data. The p-value indicates a causal relationship between *Crossbar Dimension* and *Barrel Temperature*. Next, since we have both $(x, y)_O$ and $(x, y)_E$ data (situation (i) in Section 4.3), we need to investigate whether the estimated β from the observational data is close to the one from experimental data. In this example, they are 0.325 and 0.320, respectively. A formal hypothesis test on the two slopes fails to reject the hypothesis that the two β's are equal (Cohen et al. (2003), pp. 46-67). Thus, there is no evidence that *Barrel Temperature* correlates with some other input(s) (even ones not measured) that strongly affects *Crossbar Dimension*. Therefore, we can use either of the two β estimates, or better yet, use a pooled estimate. Using maximum likelihood estimation, we estimate the model parameters as $\hat{\mu}_X = 76.585$, $\hat{\sigma}_X^2 = 2.208$, $\hat{\alpha} = -22.906$, $\hat{\beta} = 0.309$, and $\hat{\sigma}_\epsilon^2 = 0.065$. Plugging the estimates into Equation 4.2, we have $\hat{\rho}^2 = 0.765$. Since $\hat{\rho}^2$ is very large, we have verified that *Barrel Temperature* is the dominant cause.

Note that in the previous sections, we discuss the situations where the suspect dominant cause follows a normal distribution. However, one can easily extend the idea to other distributions. For instance, Appendix C.2 discusses the model, results, and our recommendations for binary suspect dominant causes in detail.

4.6 Verification Experiments in the Era of Smart Manufacturing

The study designs discussed so far are a combination of observational and experimental data. Two recent developments open the way for alternative approaches, namely, the recent spur of new ideas and techniques in causal inference from observational data and the large-scale adoption of new digital technologies for collecting and storing process data. In this section, we briefly introduce these new data-collection systems, and we describe how they may offer an easier way to do a verification study, exploiting recent advances in causal modelling.

As mentioned before, to estimate σ_X and/or σ_Y, we require an observational study where the X and/or Y values are collected from the regular process. The traditional way of collecting such data is to conduct a separate study where we systematically observe/collect information regarding the selected X and/or Y values. In modern manufacturing, however,

we may have access to observational data collected automatically from the production process using inexpensive sensors. For example, currently, many manufacturing processes have one or some of the following systems: Enterprise Resource Planning (ERP), Manufacturing Execution Systems (MES), and Supervisory Control and Data Acquisition (SCADA). These business management systems help an organization collect, store, manage, and interpret data from many business activities. Since they operate in (or near) real-time and use a common database maintained by a database management system, they provide an integrated and continuously updated view of core business processes. In the following, we briefly explain each system.

As the name suggests, the ERP system helps plan resources in an organization. This high-level system provides information regarding the production schedule (e.g., product P is scheduled to be produced in the specific time T by machine M). Modern ERP systems may include material purchase and inventory management, production and operations planning, and logistics management. They can also include accounting, sales planning, and engineering tools.

MES translates the production schedule (from the ERP) to instructions for individual machines (e.g., we should produce A amount of product P, and to do so, we need V parts). This level-two system helps plan and execute process commands for the machines. It aims to maintain the proper quality of the products using maintenance of the inputs and quality control.

SCADA is a level-three control system that provides more detail than ERP and MES. SCADA aims to translate the instructions of MES to sensors. In other words, SCADA uses a network of computers, Programmable Logic Controllers (PLCs), sensors, and graphical user interfaces to create high-level supervisory management and control for operators controlling a large process plant or machinery. Note that ERP, MES, and SCADA systems work with relational databases, i.e., instead of having observational data in one table, there are lots of linked tables. To achieve the tidy data format, for data prepossessing, say we can run a query in SQL.

These systems could provide $(x, y)_O$ data with many X's simultaneously at virtually no added expense. In addition to making data collection easier and automatic, the availability of such large data sets on many variables in the production process may obviate the need for a randomized controlled experiment. Namely, in some circumstances, it may be possible to demonstrate that X and Y are *causally* related from the large amount of observational data readily collected by ERP, MES, or SCADA systems.

Making causal inferences from observational data has drawn a lot of attention in the recent academic literature (see e.g., Lederer et al. (2019) and Hernan et al. (2019)). The

structural causal modelling proposed by Pearl (2009) integrates and generalizes earlier approaches, such as path analysis, structural equations modelling, and the potential-outcome framework proposed by Rubin (2005). The danger is that an observed correlation between X and Y in the observational data is induced by confounding variables, without X and Y being directly causally related. The idea of causal inference from observational data is that sometimes, it is possible to deconfound X and Y by including such confounders C in the regression analysis. When done right, the inclusion of C as additional explanatory variables ensures that the estimated relation between X and Y only reflects the causal effect of X onto Y. The approach requires subtlety because including extra variables C in the analysis could also *create* a confounding problem. The naïve idea of simply including all possible variables in the analysis in the hope this will eliminate all confounding problems is therefore misguided. Pearl's *backdoor criterion* (Pearl (2009)) offers a simple graphical approach to determine a *suitable* set of variables C that deconfound X and Y, and assess the adequacy of controlling for a particular covariate set.

For a full treatise on causal inference from observational data, we refer the reader to Pearl and Mackenzie (2018), which is an excellent and easily read treatise on causal inference from observational data. Note that if we only have data on X and Y, there is no way to tell whether an observed correlation between X and Y is causal or spurious.

Provided that a *suitable* set C of deconfounders is included in the analysis (as demonstrated by the "backdoor criterion", for instance), then with $(x, y, c)_O$ data, we can fit the model

$$Y = \alpha + \beta X + \gamma\, C + \epsilon, \tag{4.7}$$

where C is independent of ϵ, but not necessarily from X. Note that we deliberately kept Model 4.7 simple to illustrate the idea; however, C and γ could be vectors. Using Model 4.7, we can demonstrate the causal relationship between X and Y by testing $H_0 : \beta = 0$ vs. $H_A : \beta \neq 0$. This data is also sufficient to estimate the size of the causal relationship since it also allows us to estimate the squared correlation between X and Y (in presence of confounders C), given by

$$\rho^2_C = \frac{\beta^2\sigma^2_X + \gamma^2\sigma^2_{XC}/\sigma^2_X + 2\beta\gamma\,\sigma_{XC}}{\beta^2\sigma^2_X + \gamma^2\sigma^2_C + 2\beta\gamma\,\sigma_{XC} + \sigma^2_\epsilon}, \tag{4.8}$$

where σ_C is the standard deviation of C and σ_{XC} is the covariance between X and C, and $0 \leq \rho^2_C \leq 1$ (the derivation of Equation 4.8 is available in Appendix C.3). Using this parameterization, X is strictly a dominant cause if $\rho^2_C > 0.5$. Moreover, the inclusion of C as an explanatory variable ensures that $\hat{\beta}$ reflects the *causal* effect of X onto Y, because the spurious part of the correlation is now accounted for in $\hat{\gamma}$.

In summary, causal inference based on observational data is a thriving field with substantial recent advances. The above may give the reader an idea of the new possibilities for verification studies that emerge in modern, data-rich manufacturing environments based on ERP, MES, SCADA, and similar systems.

4.7 Conclusion and Discussion

This chapter provides a systematic way to verify the dominant cause of the process output variation. At first glance, we may believe that a formal experiment can be used to verify a dominant cause; however, we show that an experiment alone cannot provide all the required information for the verification study. From experimental data, we can establish a causal effect between X and Y, but the coefficient of determination R^2 from experimental data is not a reliable estimator for the relative effect size of X on Y. Observational studies, on the other hand, are appropriate for clue generation and estimation of the effect size of X onto Y; however, they typically lack the rigor and systematic design that allows for strong causal inferences. As explained, we suggest two approaches for providing all the required information to verify a dominant cause. The first approach involves a combination of $(x, y)_E$ experimental with either $(x, y)_O$, $(x)_O$, or $(y)_O$ data. The second approach entails using $(x, y, c)_O$ observational data. The existing literature on verification studies has not previously considered any of these combinations.

This chapter in addition to listing some viable composite study designs for the verification study, compares their relative merits via simulation. The simulation results reveal that the most precise estimates for the effect size of X onto Y arise when we supplement the experiment with observational $(x, y)_O$ data. However, in this case, the estimated β from $(x, y)_O$ data may not reflect the direct strength of the causal effect of X onto Y due to possible confounding. To overcome this problem, we suggest first testing whether $\hat{\beta}$ from $(x, y)_O$ data is equal to $\hat{\beta}$ from $(x, y)_E$ data, and then follow our provided guideline.

Sections 4.3 and 4.4 investigate the case where the suspect dominant cause is continuous and follows a normal distribution. The supplementary material discusses binary suspect dominant causes. Although we only provide the model and results for these two common cases, one can easily extend the idea to other distributions for X. We recommend a robust study for future work. Before fitting any formal model, we recommend examining the data graphically and checking the model assumptions.

Moreover, we consider having only one suspect in Model 4.1. This has the advantage of being easy to understand and representative of how the idea works. However, in real-world problems, there might be more than one suspect dominant cause, and they could

even be correlated with other X's in this study, or the suspect dominant cause might be an interaction. However, since we use the method of elimination before moving to a verification experiment, the number of remaining suspect dominant causes should be relatively small. With two or more suspect dominant causes, we recommend conducting a full factorial verification experiment incorporating all the suspect dominant causes simultaneously. After that, for each suspect dominant cause that shows a significant causal link to the output, we supplement the analysis with some observational data (usually already collected) to investigate whether the causal link is dominant or not.

Chapter 5

Conclusion and Future Work

5.1 Summary and Conclusion

Excessive variation in critical-to-quality characteristics is an important challenge in manufacturing industries. The recommended approach to reduce this variation involves first identifying the causes of this variation and then seeking solutions to eliminate or mitigate the effect of the identified causes. However, our focus is only on identifying the dominant cause(s) rather than any causes. The reason is that minor causes are often a mere distraction for practitioners. This thesis focused on identifying and verifying the dominant cause(s) of process output variation.

Following the introduction and literature review, our exploration unfolds across three key chapters: two chapters contributed to a deeper understanding of the challenges and solutions associated with identifying dominant cause(s), and one chapter dedicated to the ones associated with verifying the identified dominant cause(s).

Chapter 2 provided a critical assessment of the group comparison investigation proposed by Bhote and Bhote (2000) and Shainin (1993b), an effective method of elimination tool for identifying the dominant causes of output variation. We first demonstrated that their strategy based on the idea of leveraging is highly effective, particularly when measuring input characteristics is costly or time-consuming. However, our critical assessment reveals that their analysis procedure is inefficient and, more importantly, unreliable in identifying the actual dominant causes. In other words, due to framing the analysis as a significance test rather than an estimation problem, the analysis often misidentifies minor causes as the dominant ones, leading to unproductive directions and wasting resources for

future studies. While retaining the valuable idea of leveraging from the existing literature, we propose an alternative likelihood-based analysis procedure. This novel approach substantially enhances the reliability and efficiency of identifying dominant causes. We also provided a tangible example and compared the outcomes of both methods.

Chapter 3 extended our exploration to the component swapping investigation, another effective method of elimination tool to identify dominant causes of variation. Similar to Chapter 2, we retain the valuable idea of leveraging from the existing literature while proposing our superior estimation-based procedure. We compared our proposal to one of the most well-known alternatives, Shainin's component-swapping procedure (Shainin (1993b)). Our evaluation demonstrated that Shainin's component swapping procedure suffers from poorly chosen statistical tools and analysis, as it again frames the analysis as a significance test rather than an estimation problem. We demonstrated the unreliability of these tools, as they often lead to incorrect identifications of dominant causes. Moreover, whereas the literature lacks effective means of identifying interaction effects, our proposal signals to users that important interactions may affect the results. Notably, we distinguished between interactions among assembly and components and interactions among two or more components.

We devoted Chapter 4 to introduce a systematic approach for verifying the dominant cause of the process output variation. Conducting a verification study (i.e., verifying that we have identified the true dominant *cause*) is strongly recommended before proceeding with the remedial journey. This is because, during the search for the dominant cause using the method of elimination, many investigations are *observational*. While these observational studies are appropriate for generating clues, they often lack the rigor and systematic design required for making strong causal inferences. We demonstrated that although it may initially seem that a formal experiment can be used to verify a dominant cause, an experiment alone cannot provide all the required information for the verification study. An experiment determines whether a suspect is a cause of variation, but additional information from observational studies is needed to determine whether it is dominant. This chapter listed some viable composite study designs and assessed their relative merits. We also delved into how to systematically conduct a verification study in the era of smart manufacturing. Moreover, we provide a tangible example to illustrate our proposed procedure.

5.2 Future Work

This thesis has uncovered various opportunities for future studies and advancements in the field of statistical process improvement and variation reduction. In the following, we

outline several promising avenues for further investigation.

In this thesis, we made certain simplifying assumptions for the sake of clarity and representation of how the proposed study designs operate. Nonetheless, real-world manufacturing processes often present greater complexities, and researchers can extend these concepts to more intricate scenarios. For instance, we primarily focused on having a single dominant cause, whereas in the real world, processes may have multiple large causes, some of which could be correlated with other inputs. However, given the nature of the method of elimination, the number of dominant causes should generally remain relatively small. Furthermore, the dominant cause might be an interaction between two or more inputs. While we addressed interaction identification in Chapter 3, there is potential to explore this challenge in other tools associated with the method of elimination. Additionally, we assumed an approximately linear relationship between input and output characteristics. While linearity is a reasonable initial assumption for clue generation, future studies could delve into nonlinear models.

This thesis limited the simulation studies to some specific distributions for input characteristics, namely, normal and binary distributions in Chapters 2 and 4, and normal distribution in Chapter 3. Although we do not limit our proposed study designs to a specific distribution, we added these distribution assumptions to be able to assess our maximum likelihood-based approaches. Future studies can explore other distributional assumptions and provide recommended sample sizes for different data types.

In Chapters 2 and 3, we used the idea of leveraging to identify the dominant causes and demonstrated its efficiency. In addition to group comparison and component swapping investigations, we believe that leveraging can be advantageous in the context of variation transmission studies. A variation transmission study is another effective tool associated with the method of elimination, aimed at determining whether variation observed in an intermediate operation is transmitted through to the downstream operation or if the downstream operations add substantial variation. To do so, Steiner and MacKay (2005) propose selecting a sample of products over a sufficiently extended timeframe that allows the dominant cause to act. Then, they measure the output characteristics of all samples after each process stage and analyze the resulting data to determine the stage where the dominant cause resides. However, we believe that leveraging is a more efficient and cost-effective approach to conducting variation transmission studies, where we select some extreme products and measure only their output characteristics in each process stage. Subsequently, by comparing the variability among these stages to the one observed in the final output, we can narrow the search space to the stage(s) exhibiting the largest variation differences.

References

American National Standards Institute and the American Society for Quality Control (1978). *Quality Systems Terminology*. ASQC, Milwaukee, WI.

Amster, S. and Tsui, K. L. (1993). Counterexamples for the component search procedure. *Quality Engineering*, 5(4):545–552.

Antony, J. (1999). Spotting the key variables using Shainin's variables search design. *Logistics Information Management*, 12(4):325–331.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society A*, 160(901):268–282.

Bhote, K. R. (1991). *World Class Quality*. American Management Association, New York, first edition.

Bhote, K. R. and Bhote, A. K. (2000). *World Class Quality*. American Management Association, New York, second edition.

Breyfogle, F. (1999). *Implementing Six Sigma-Smarter Solutions Using Statistical Methods*. Wiley, New York, NY.

Browne, R., MacKay, R. J., and Steiner, S. H. (2009a). Improved measurement system assessment for processes with 100% inspection. *Journal of Quality Technology*, 41(4):376–388.

Browne, R., MacKay, R. J., and Steiner, S. H. (2009b). Two-stage leveraged measurement system assessment. *Technometrics*, 51(3):239–249.

Browne, R., Steiner, S. H., and MacKay, R. J. (2010a). Leveraged gauge R&R studies. *Technometrics*, 52(3):294–302.

Browne, R., Steiner, S. H., and MacKay, R. J. (2010b). Optimal two-stage reliability studies. *Statistics in Medicine*, 29(2):229–235.

Chittaro, L. and Ranon, R. (2004). Hierarchical model-based diagnosis based on structural abstraction. *Artificial Intelligence*, 155(1–2):147–82.

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey, third edition.

Cox, S. (2011). Concise process improvement methods. Master thesis, Durham University. Available at Durham E-Theses Online: http://etheses.dur.ac.uk/3275/.

Dasgupta, T., Adiga, N., and Wu, C. (2011). Another look at Dorian Shainin's variable search technique. *Journal of Quality Technology*, 43(4):273–287.

De Mast, J. (2004). A methodological comparison of three strategies for quality improvement. *International Journal of Quality and Reliability Management*, 21(2):198–213.

De Mast, J. (2011). The tactical use of constraints and structure in diagnostic problem solving. *Omega*, 39(6):702–709.

De Mast, J. (2013). Diagnostic quality problem solving: A conceptual framework and six strategies. *Quality Management Journal*, 20(4):21–36.

De Mast, J. and Lokkerbol, J. (2012). An analysis of the six sigma DMAIC method from the perspective of problem solving. *International Journal of Production Economics*, 139(2):604–614.

De Mast, J., Roes, K. C., and Does, R. (2001). The multi-vari chart: A systematic approach. *Quality Engineering*, 13(3):437–447.

De Mast, J., Schippers, W., Does, R., and Van den Heuvel, E. (2000). Steps and strategies in process improvement. *Quality and Reliability Engineering International*, 16(4):301–311.

De Mast, J., Steiner, S. H., Kuijten, R., and Funken, E. (2019). Statistical reasoning in diagnostic problem solving – the case of flow-rate measurements. *Quality Engineering*, 31(3):484–498.

De Mast, J., Steiner, S. H., Nuijten, W. P. M., and Kapitan, D. (2023). Analytical problem solving based on causal, correlational and deductive models. *The American Statistician*, 77(1).

Ghurka, A. and Pawar, N. (2015). Use of Shainin design of experiments to reduce the tripping force of an air circuit breaker. *The International Journal of Engineering and Science*, 4(11):11–18.

Gryna, F. M. and Juran, J. M. (1988). *Juran's Quality Control Handbook*. McGraw-Hill, New York, fourth edition.

Hahn, G., Hill, W., Hoerl, R., and Zinkgraf, S. (1999). The impact of Six Sigma improvement - a glimpse into the future of statistics. *The American Statistician*, 53(3):208–215.

Hahn, G. J., Doganaksoy, N., and Hoerl, R. W. (2000). The evolution of Six-Sigma. *Quality Engineering*, 12(3):317–326.

Harry, M. (1997). *The Vision of Six Sigma*. Tri Star, Phoenix, AZ, fifth edition.

Hernan, M. A., Hsu, J., and Healy, B. (2019). A second chance to get causal inference right: A classification of data science tasks. *Chance*, 32(1):42–49.

Ho, J. K. K. and Sculli, D. (1997). The scientific approach to problem solving and decision support systems. *International Journal of Production Economics*, 48(3):249–257.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Juran, J. (1989). *Juran on Leadership for Quality: An Executive Handbook*. Free Press, New York.

Juran, J. M. and Gryna, F. M. J. (1980). *Quality Planning and Analysis*. McGraw-Hill Book Company, New York, second edition.

Kiatcharoenpol, K., Seeluang, R., and Klongboonjit, S. (2023). Applying Shainin's tools to process improvement for reducing cracking defect of sanitary product. *International Journal of Membrane Science and Technology*, 10(3):181–189.

Lederer, D. J., Bell, S. C., Branson, R. D., Chalmers, J. D., Marshall, R., and Maslove, D. M., e. a. (2019). Control of confounding and reporting of results in causal inference studies. *Annuals of the American Thoracic Society*, 16(1):22–28.

Ledolter, J. and Swersey, A. (1997). Dorian Shainin's variables search procedure: a critical assessment. *Journal of Quality Technology*, 29(3):237–247.

Levene, H. (1960). *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press, Palo Alto.

Linderman, K., Schroeder, R. G., Zaheer, S., and Choo, A. S. (2003). Six Sigma: a goal-theoretic perspective. *Journal of Operations Management*, 21:193–203.

Logothetis, N. (1990). A perspective on Shainin's approach to experimental design for quality improvement. *Quality and Reliability Engineering International*, 6(3):195–202.

MacDuffie, J. P. (1997). The road to "root cause": Shop-floor problem-solving at three auto assembly plants. *Management Science*, 43(4):479–502.

Montgomery, D. C. (2017). *Design and Analysis of Experiments*. John Wiley & Sons, New Jersey.

Mooren, J., de Mast, J., and Does, R. J. M. M. (2012). Quality quandaries*: The case of premature drill wear out. *Quality Engineering*, 24(2):354–359.

Nair, V. N. (1992). Taguchi's parameter design: a panel discussion. *Technometrics*, 34(2):127–161.

Panahi, M., De Mast, J., and Steiner, S. H. (2021). Identifying dominant causes using leveraged study designs. *Quality Engineering*, 33(4):581–593.

Panahi, M., Steiner, S. H., and De Mast, J. (2023). Verifying a dominant cause of output variation. *Quality Engineering*. In press. doi: 10.1080/08982112.2023.2253303.

Panchal, D., Patel, B., and Shah, D. (2020). Application of Shainin component search approach as root cause analysis tool for drop in two-wheeler fuel economy. *International Journal of Six Sigma and Competitive Advantage*, 12(2-3):187–208.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, second edition.

Pearl, J. and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic books, New York, second edition.

Pietraszek, J., Krawczyk, M., Sobczyk, A., and Skrzypczak, E. (2016). The dominant factor detection in the Shainin's approach. *Technical Transactions*, 113(4-M):95–100.

Prashar, A. (2016). A conceptual hybrid framework for industrial process improvement: Integrating Taguchi methods, Shainin system and six sigma. *Production Planning & Control*, 27(16):1389–1404.

Pyzdek, T. (2001). *The Six Sigma Handbook*. McGraw-Hill, London.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. John Wiley & Sons, New York, second edition.

Ross, P. (1988). *Taguchi Techniques for Quality Engineering*. McGraw-Hill, London.

Rubin, D. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331.

Ryan, T. (1989). *Statistical Methods for Quality Improvement*. John Wiley & Sons, New York.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components*. John Wiley & Sons, New York.

Shainin, P. (1993a). Managing quality improvement. In *47th Annual Quality Congress Transactions*, ASQC, pages 554–560, Milwaukee, WI.

Shainin, P. D., Shainin, R. D., and Nelson, M. T. (1997). Managing statistical engineering. *51st Annual Quality Congress Proceeding, ASQC*, pages 818–532.

Shainin, R. D. (1993b). Strategies for technical problem solving. *Quality Engineering*, 5(3):433–448.

Shainin, R. D. and Shainin, P. D. (1988). Better that Taguchi orthogonal tables. *Quality and Reliability Engineering International*, 4(2):143–149.

Smith, G. F. (1988). Towards a heuristic theory of problem structuring. *Management Science*, 34(12):1489–1506.

Steiner, S. H. and MacKay, R. J. (2005). *Statistical Engineering: An Algorithm for Reducing Variation in Manufacturing Processes*. Quality Press, Milwaukee.

Steiner, S. H. and MacKay, R. J. (June, 2006). Statistical engineering: A case study. *Quality Progress*, pages 33–39.

Steiner, S. H., MacKay, R. J., and Ramberg, J. S. (2008a). An overview of the Shainin System$^{\text{TM}}$ for quality improvement. *Quality Engineering*, 20(1):6–19.

Steiner, S. H., MacKay, R. J., and Ramberg, J. S. (2008b). Rejoinder to the discussion of "An overview of the Shainin System™ for quality improvement". *Quality Engineering*, 20(1):42–45.

Taguchi, G. (1986). *Introduction to Quality Engineering: Designing Quality into Products and Processes*. Asian Productivity Organisation, Tokyo.

Tukey, J. W. (1959). A quick, compact, two-sample test to Duckworth's specifications. *Technometrics*, 1(1):31–48.

Vining, G. and Meyers, R. (1990). Combining Taguchi and response surface philosophies: a dual response surface approach. *Journal for Quality Technology*, 22(1):38–45.

Wagner, C. (1993). Problem solving and diagnosis. *Omega*, 21(6):645–656.

Xu, L., Gotwalt, C., Hong, Y., King, C. B., and Meeker, W. Q. (2020). Applications of the fractional-random-weight bootstrap. *The American Statistician*, 74(4):345–358.

APPENDICES

Appendix A

Group Comparison - Details and Additional Findings

A.1 The Derivation of ρ^2 Formula

In the following, we show how to obtain Equation 2.2. Assuming Model 2.1 and the independence between X and ϵ, we have

$$\begin{aligned} E(XY) &= E\big(X\,(\alpha+\beta X+\epsilon)\big) = \alpha\,E(X)+\beta\,E(X^2)+E(X)\,E(\epsilon) \\ &= \alpha\,E(X)+\beta\big(Var(X)+E^2(X)\big)+E(X)\times 0 \\ &= \alpha\,E(X)+\beta\,\sigma_X^2+\beta\,E^2(X) \end{aligned}$$

Therefore, the covariance between X and Y is given by

$$Cov(X,Y) = E(XY)-E(X)\,E(Y) = \alpha\,E(X)+\beta\,\sigma_X^2+\beta\,E^2(X)-E(X)\times\big(\alpha+\beta\,E(X)\big) = \beta\,\sigma_X^2$$

As a result, the squared correlation between X and Y is given by

$$\rho^2 = \frac{Cov^2(X,Y)}{\sigma_X^2\,\sigma_Y^2} = \frac{(\beta\,\sigma_X^2)^2}{\sigma_X^2\,(\beta^2\,\sigma_X^2+\sigma_\epsilon^2)} = \frac{\beta^2\,\sigma_X^2}{\beta^2\,\sigma_X^2+\sigma_\epsilon^2}$$

which is Equation 2.2.

A.2 The Conditional Distribution of $(X^*|Y = y)$ for Binary X^*

The conditional distributions of $P(X_i^* = -1|Y_i = y_i)$ and $P(X_i^* = 1|Y_i = y_i)$, which were discussed in Section 2.4.1, are $\frac{(q)P(Y_i|X_i^*=-1)}{P(Y_i=y_i)}$ and $\frac{(1-q)P(Y_i|X_i^*=+1)}{P(Y_i=y_i)}$, respectively. As a result,

$$P(X_i^* = x_i^*|Y_i = y_i) = \frac{(\frac{1-x_i^*}{2})\, q\, P(Y_i|X_i^* = -1) + (\frac{1+x_i^*}{2})(1-q)P(Y_i|X_i^* = +1)}{P(Y_i = y_i)},$$

where $(Y_i|X_i^* = \pm 1) \sim N(\alpha \pm \beta, \sigma_\epsilon^2)$ and the denominator is given by Equation 2.5. Then, for binary inputs, the log-likelihood can be derived by using Equation 2.5 and plugging the above equation into Equation 2.3.

Appendix B

Component Swapping - Details and Additional Findings

B.1 An Overview of Variants of Component-Swapping Procedures in the Literature

The existing literature presents slightly different component-swapping procedures. Table B.1 compares the data collection and analysis procedure used in Phase I. Note that all these procedures select two extreme products for disassembling and reassembling.

Table B.1 reveals some inconsistencies within the component-swapping literature. Note that only Shainin and Shainin (1988) explicitly explained the origin of the threshold value d, while the others failed to provide a statistical reference. Therefore, this paper focuses solely on examining Shainin's procedure, and thus, $d = 1.07$.

Table B.2 is devoted to Phase II and compares two of the most important component-swapping procedures, which are proposed by Shainin and Shainin (1988) and Bhote and Bhote (2000). Although both procedures employ the same decision intervals, their conclusions and subsequent steps do not always align.

Literature	Recommended number of disassembling and reassembling (r)	D is the difference between	Threshold value d in the decision criterion $D > d \times \bar{R}$
Shainin and Shainin (1988) Ledolter and Swersey (1997) Dasgupta et al. (2011)	$r = 2$	Medians	$d = 1.07$
Logothetis (1990) Bhote (1991) Amster and Tsui (1993)	$r = 1$	Means	$d = 5$
Antony (1999) Bhote and Bhote (2000) Ghurka and Pawar (2015) Pietraszek et al. (2016) Prashar (2016)	$r = 2$	Medians	$d = 1.25$
Cox (2011) [a]	$r = 2$	Medians	$d = 5$
Steiner and MacKay (2005)	$r \geq 3$	The analysis is based on graphical tools	

Table B.1: Comparison of *Phase I* component-swapping procedures in some literature where $\bar{R}$ represents the average of the ranges.

Literature	Situation	Conclusion	Consequent step
Shainin and Shainin (1988)	Both swapped results lie within the corresponding DI's	No significant change (insignificant cause) Eliminate the swapped component from consideration	Swap the next-ranked component
	One or both of the swapped results fall outside the corresponding DI's	Significant change (significant cause) Keep the swapped component under consideration	Swap the next-ranked component.
Bhote and Bhote (2000)	Both of the swapped results lie within the corresponding DI's	Minor change (unimportant cause) Eliminate the swapped component from consideration	Swap the next-ranked component
	Both swapped results fall outside the corresponding DI's	Complete change (dominant cause)	Stop the procedure
	Only one of the swapped results falls outside the corresponding DI's	Partial change (part of the dominant causes) Keep the swapped component under consideration	Swap the next-ranked component.

Table B.2: Comparison of *Phase II* component-swapping procedures in some literature.

[a] Cox (2011) cites Bhote and Bhote (2000) and Antony (1999), but they used $d = 1.25$.

As Table B.2 demonstrates, Shainin and Shainin (1988) continue the search after identifying a dominant cause. They only eliminate non-significant components and continue swapping the next-ranked components. Therefore, their procedure keeps any significant components in the analysis and later conducts a capping run. Consequently, they tend to identify the combined effect of two or more components as the dominant cause, even when the dominant cause is a single component. On the other hand, Bhote and Bhote (2000) distinguish between complete, partial, and minor changes. Therefore, their procedure will stop after identifying the dominant cause and only retain components under consideration if there is a partial change.

Figures B.1 and B.2 provide flowcharts of the Phase II procedures for Shainin and Shainin (1988) and Bhote and Bhote (2000), respectively. In these figures, we assume that only two components C_1 and C_2 are swappable, C_1 is the top-ranked component, and C_R denotes the remaining components.

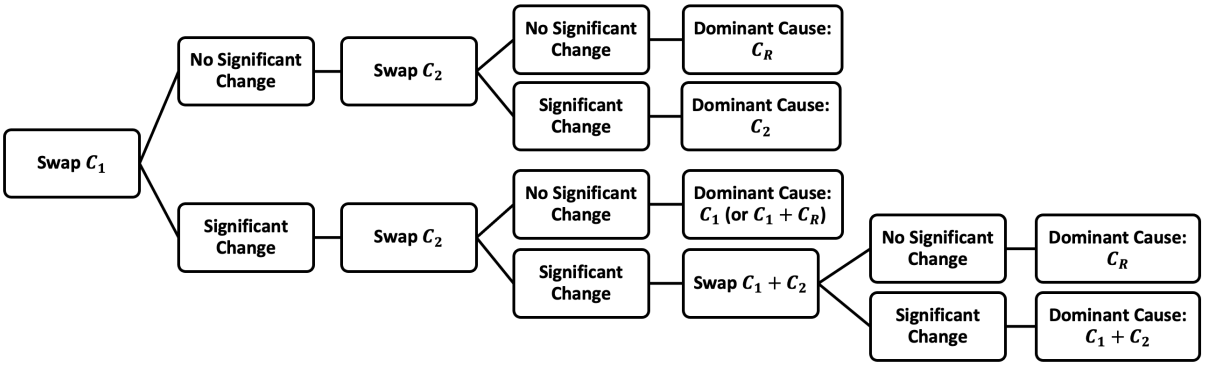

Figure B.1: Flowchart of the Phase II component-swapping procedure proposed by Shainin and Shainin (1988).

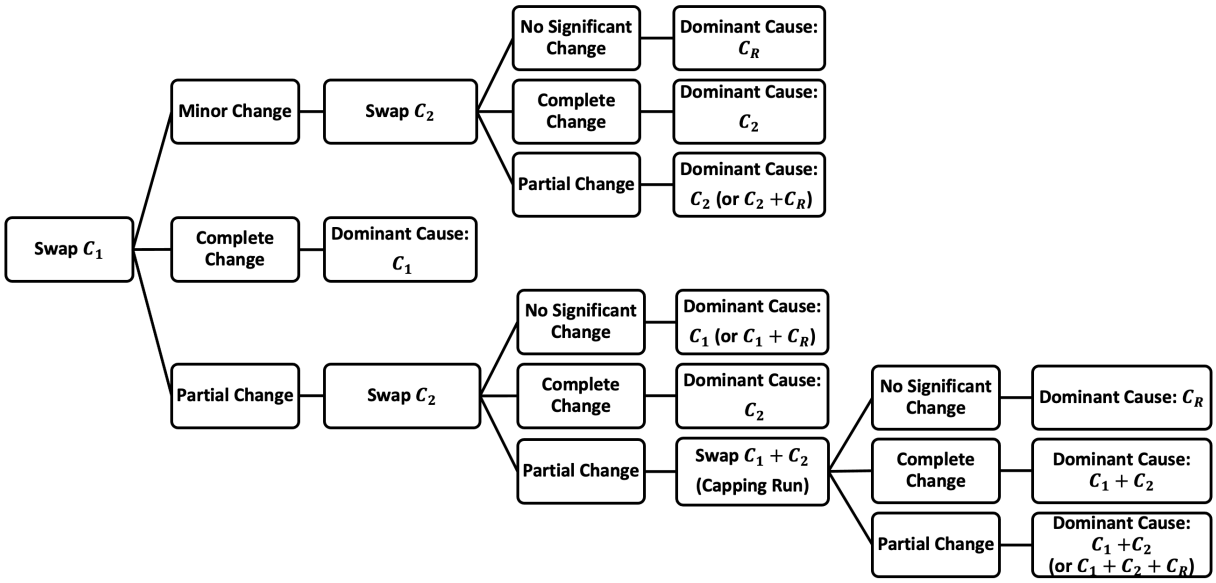

Figure B.2: Flowchart of the Phase II component-swapping procedure proposed by Bhote and Bhote (2000).

B.2 Evaluation of Shainin's Procedure

Here, we present our data generation model and critiques of Shainin's procedure for Phases I and II.

B.2.1 Shainin's Phase I Procedure

To assess the reliability of Shainin's Phase I, procedure we conducted a simulation study and estimated the probability of identifying assembly as the dominant cause. In our model, we denote the properties of all components by C with an effect of β_C on Y, properties of the assembly-related causes by A with an effect of β_A on Y, and let β_{CA} be their interaction effects. Assuming that their effects on the output characteristic Y are approximately linear with intercept α, we have

$$Y = \alpha + \beta_C\, C + \beta_A\, A + \beta_{CA}\, C\, A + M, \tag{B.1}$$

where the error, denoted M, models the measurement variation. Note that Model B.1 is unsuitable for analyzing the Phase I data because we cannot observe the values of A and

C in component swapping. However, we use this model to generate simulation data.

In the simulations, for simplicity, we assume $\beta_{CA} = 0$, $C \sim N(\mu_C, \sigma_C^2)$, $A \sim N(\mu_A, \sigma_A^2)$, and $M \sim N(0, \sigma_M^2)$ in Model B.1. Moreover, without loss of generality, we set $\alpha = 0$, $\beta_A = \beta_C = 1$, $\mu_A = \mu_C = 0$, and $\sigma_Y^2 = 1$. Note that we would not conduct a component-swapping investigation without first confirming that measurement variation is small. Therefore, for illustration, we assume there is no measurement variation, i.e., $\sigma_M^2 = 0$. Our results do not markedly change for small measurement variation. We also set σ_C^2 and σ_A^2 such that $\rho_A^2 \in \{0.0, 0.1, \ldots, 1.0\}$ (as defined by Equation 3.1), consider a baseline of $n_b = 1000$ samples, select $k = 2$ extreme products, and disassemble and reassemble them $r = 2$ times. Then, in each run, we investigate whether Shainin's procedure identifies assembly as the dominant cause or not. Note that Shainin's procedure eliminates assembly as the dominant cause when $Max(y_0^L, y_1^L, y_2^L) < Min(y_0^H, y_1^H, y_2^H)$ and $D > 1.07 \times \bar{R}$. The first criterion is an application of the Tukey end-count test and is equivalent to having an end-count of six (Tukey (1959)). However, we show below that the former criterion is not very helpful. We run the simulation with and without the end-count criterion. Figure B.3 demonstrates the result for 5000 simulation runs for different ρ_A^2 values.

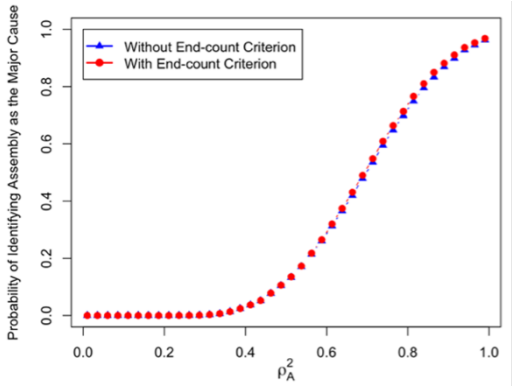

Figure B.3: Probability of identifying assembly as the dominant cause using Shainin's Phase I procedure vs. ρ_A^2 with and without the end-count criterion when there is no measurement error, $r = 2$, and $n_b = 1000$.

Figure B.3 reveals that Shainin's procedure leads to very similar results whether we consider the end-count criterion or not. Thus, the end-count criterion is not very helpful. It also reveals that with their recommendation of $r = 2$, Shainin's procedure is unreliable in correctly identifying assembly as the dominant cause when it is indeed the dominant cause. For instance, Shainin's procedure identifies assembly as the dominant cause only 20% of the time, when its actual contribution to the overall variation in Y is 50%.

B.2.2 Shainin's Phase II Procedure

To investigate the reliability of Shainin's procedure in identifying the dominant cause among components, we conducted a simulation study and estimated the probability of identifying each component as the dominant cause. In the simulation, we add the simplifying assumption that we can only swap C_1, C_2, or both simultaneously and C_R represents the effect of the remaining components. The conclusions, however, can be extended to more general cases. Note that with our notation, C_i could be a single or a group of components. Assuming an approximately linear relationship between the effect of components and the output characteristic Y with intercept α, we have

$$Y = \alpha^* + \beta_{C_1} C_1 + \beta_{C_2} C_2 + \beta_{C_1C_2} C_1 C_2 + \beta_{C_R} C_R + \beta_A A + \sum_{i \in \{C_1, C_2, C_1C_2, C_R\}} \beta_{iA}\, i\, A + M^*, \quad \text{(B.2)}$$

where M^* reflects the measurement effects.

Note that Model B.2 is unsuitable for analyzing the results of Phase II data because we cannot observe the characteristics of any components. However, we use this model to generate simulation data.

To simulate data in Phase II, for simplicity, we first assume there is no interaction between assembly and components, i.e., $\beta_{iA} = 0$ for $i \in \{C_1, C_2, (C_1, C_2), C_R\}$ in Model B.2. Later in Appendix B.3.2, we assess the situation where there is an interaction between assembly and component(s). Moreover, for simplicity, we assume C_1 and C_2 are the only components, i.e., $\beta_{C_R} = 0$. We also assume $C_1 \sim N(\mu_{C_1}, \sigma^2_{C_1})$, $C_2 \sim N(\mu_{C_2}, \sigma^2_{C_2})$, $A \sim N(\mu_A, \sigma^2_A)$, and $M^* \sim N(0, \sigma^2_{M^*})$. Without loss of generality, we set $\alpha^* = 0$, $\mu_{C_1} = \mu_{C_2} = \mu_A = 0$ and $\sigma^2_{C_1} = \sigma^2_{C_2} = \sigma^2_A = 1$ in Model B.2. We also assume no measurement variation, i.e., $\sigma^2_{M^*} = 0$. Since we have completed Phase I, we assume only a small proportion of the total variation in Y is due to the assembly process, e.g., $\rho^2_A = 0.05$.

Here, for simplicity, we also add the simplifying assumption of having no interaction between components (i.e., $\beta_{C_1C_2} = 0$). However, later in Appendix B.4.2, we consider cases with such interactions. Therefore, we set different sets of values to β_{C_1}, β_{C_2}, and β_A such that $\rho^2_A = 0.05$, $\rho^2_{C_1} = \{0.01, \ldots, 0.94\}$, and $\rho^2_{C_2} = 1 - \rho^2_{C_1} - \rho^2_A$. In each simulation run, we follow Shainin's recommendation of selecting two extreme products and $r = 2$. Figure B.4 demonstrates the probability of identifying C_1 vs. the combined effect of C_1 and C_2 as the dominant causes by Shainin's procedure using 5000 simulation runs for different sizes of $\rho^2_{C_1}$ when $n_b = 1000$. Note that by concluding the combined effect as the dominant cause, the dominant cause could be the sum of the two main effects $C_1 + C_2$ or their interaction effect C_1C_2. However, we use the notation $C_1 + C_2$ here because Shainin's procedure does

not distinguish between the two cases. The remaining times, Shainin's procedure identifies C_2 as the dominant cause.

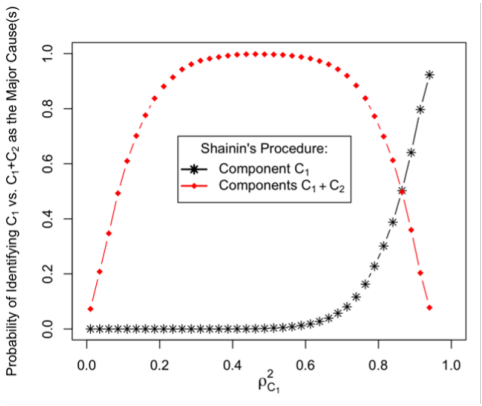

Figure B.4: Probability of identifying C_1 vs. $C_1 + C_2$ as the dominant cause by Shainin's procedure for different $\rho^2_{C_1}$ values when $\rho^2_A = 0.05$, $r = 2$, and $n_b = 1000$.

Figure B.4 reveals that unless $\rho^2_{C_1}$ or $\rho^2_{C_2}$ are extreme, e.g., higher than 0.85, Shainin's procedure tends to identify $C_1 + C_2$ as the dominant cause. For instance, when $\rho^2_{C_1} = 0.75$ and there is no interaction between C_1 and C_2, Shainin's procedure identifies $C_1 + C_2$ as the dominant cause almost 85% of the time.

B.3 Proposed Phase I Setup and Analysis

Here, our focus is on Phase I of our proposed component-swap procedure. We allocate Appendix B.3.1 to present different approaches to estimate ρ^2_A and propose our recommended method among them. Appendix B.3.2 is devoted to identifying assembly by component interaction effects. We provide our recommended values for Phase I design parameters in Appendix B.3.3.

B.3.1 Estimating ρ^2_A

Phase I of component swapping, in which we disassemble and reassemble extreme assemblies, is to some extent like the measurement assessment study proposed by Browne et al. (2009b, 2010a). They suggested leveraging to select several extreme parts from a baseline and then remeasuring the parts several times each. From this data, they estimated the

proportion of variation due to the measurement system. We propose to adopt the analysis proposed by Browne et al. (2009b, 2010a) to our situation. They showed the benefits of selecting extreme parts from a baseline for a measurement assessment investigation.

Here, we consider four approaches to estimate ρ_A^2. The first method is based on Maximum Likelihood. The other three methods estimate ρ_A^2 using Phase I data conditional on the values of the output of the selected assemblies from the baseline. The second approach uses regression because the conditional mean of the Phase I data depends on ρ_A^2. The third is based on the analysis of variance (ANOVA), which uses the variation within the Phase I data for each product. The fourth estimator combines the regression and ANOVA estimators.

B.3.1.1 Maximum Likelihood Estimator for ρ_A^2

There are two types of available data for the Phase I analysis: the baseline and the measured output values after we disassemble and reassemble the selected products r times. Therefore, the log-likelihood is the summation of the following pieces:

- l_0: the log-likelihood contribution for μ_Y and σ_Y^2 using the n_b randomly selected Y values for the baseline data, where $Y \sim N(\mu_Y, \sigma_Y^2)$.

- $l_1, \ldots, l_k$: the k log-likelihood contributions for μ_Y, σ_Y^2, ρ_A^2, where l_i corresponds to the output values from disassembling and reassembling of the i^{th}, $i = 1, \ldots, k$ selected products r times. For the $i \in \{1, \ldots, k\}$ selected independent products, the r repeated disassembling and reassembling given the observed extreme y from the baseline has the following multivariate normal (MVN) distribution:

$$\begin{pmatrix} Y_1^i \\ \vdots \\ Y_r^i \end{pmatrix} \Bigg| \; Y_0^i = y_0^i \Bigg) \sim MVN \left(\left[\mu_Y + (1-\rho_A^2)(y_0^i - \mu_Y)\right] \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \sum_r \right),$$

  where $\sum_r$ is the $r \times r$ covariance matrix given by

$$\sum_r = \sigma_Y^2 \rho_A^2 \begin{bmatrix} 2-\rho_A^2 & & 1-\rho_A^2 \\ & \ddots & \\ 1-\rho_A^2 & & 2-\rho_A^2 \end{bmatrix}.$$

Therefore, the overall log-likelihood is given by

$$-\frac{n_b}{2}\,log\,\sigma_Y^2-\frac{1}{2\,\sigma_Y^2}\Big((n_b-1)s_b^2+n_b(\bar{y}_b-\mu_Y)^2\Big)-\frac{rk}{2}\,log\,(\sigma_Y^2\rho_A^2)-\frac{k}{2}\,log\,(1+r-r\,\rho_A^2)$$
$$-\frac{\sum_{i=1}^{k}\sum_{j=1}^{r}(y_j^i-\bar{y}_{PhI}^i)^2}{2\sigma_Y^2\rho_A^2}-\frac{r}{2\,\sigma_Y^2\rho_A^2(1+r-r\rho_A^2)}\sum_{i=1}^{k}\Big(\bar{y}_{PhI}^i-\mu_Y-(1-\rho_A^2)(y_0^i-\mu_Y)\Big)^2,$$

where $\bar{y}_b$ and s_b^2 represent the sample mean and variance of the baseline data, and $\bar{y}_{PhI}^i = (y_1^i+\cdots+y_r^i)/r$ is the average of Phase I data for the i^{th} product.

Although maximum likelihood is an efficient estimator, we need to use numerical iteration to find the maximum likelihood estimates (MLEs) as there is no closed-form solution. Therefore, in the following, we propose alternate estimators with closed forms.

B.3.1.2 Regression Estimator for ρ_A^2

Since $Y\sim N(\mu_Y,\,\sigma_Y^2)$, for $i\in\{1,\ldots,k\}$, the distribution of the average of the r measurements arising from disassembling and reassembling (i.e., $\bar{Y}_{PhI}^i$) given the observed output value in the baseline (i.e., y_0^i) follows a normal distribution, namely

$$\bar{Y}_{PhI}^i|(Y_0^i=y_0^i)\sim N\Big(\mu_Y+(1-\rho_A^2)(y_0^i-\mu_Y),\,\sigma_Y^2\rho_A^2(1-\rho_A^2+r^{-1})\Big).$$

We can use regression to estimate ρ_A^2, because $\bar{Y}_{PhI}^i$ for $i\in\{1,\ldots,k\}$ are mutually independent, the mean depends on ρ_A^2 linearly (and μ_Y that can be estimated by baseline sample mean $\bar{y}_b$), and the variance is the same for all k products. Following Browne et al. (2009b, 2010a), we have

$$\hat{\rho}_{A_{Reg}}^2=1-\frac{\sum_{i=1}^{k}(\bar{y}_{PhI}^i-\bar{y}_b)(y_0^i-\bar{y}_b)}{\sum_{i=1}^{k}(y_0^i-\bar{y}_b)^2}.$$

B.3.1.3 ANOVA Estimator for ρ_A^2

Note that to obtain $\hat{\rho}_{A_{Reg}}^2$, we only used the average of the r disassembling and reassembling output values from Phase I, and not their variability. To use the variation information within Phase I data for each selected product, we propose an ANOVA estimator. Following Browne et al. (2009b, 2010a), its closed form is

$$\hat{\rho}_{A_{ANV}}^2=\frac{\sum_{i=1}^{k}\sum_{j=1}^{r}\left(y_j^i-\bar{y}_{PhI}^i\right)^2}{k\,(r-1)\,s_b^2},$$

where s_b^2 represents the sample variance of the baseline data. Note that to have a non-zero denominator, we require $r \geq 2$.

B.3.1.4 Combined Estimator for ρ_A^2

Although regression and ANOVA estimators have closed forms, they are not as good as the MLE. We can obtain an estimator with a closed form having similar properties as the MLE, if we combine $\widetilde{\rho}^2_{Reg}$ and $\widetilde{\rho}^2_{ANV}$ as they are uncorrelated (Browne et al. (2009b, 2010a)).

Suppose $\widetilde{\rho}^2_{A_{Reg}}$ and $\widetilde{\rho}^2_{A_{ANV}}$ have known variances σ^2_{Reg} and σ^2_{ANV}, respectively. Then, the minimum variance linear combination is given by

$$\frac{\sigma^2_{ANV}}{\sigma^2_{Reg}+\sigma^2_{ANV}}\,\widetilde{\rho}^2_{Reg} + \frac{\sigma^2_{Reg}}{\sigma^2_{Reg}+\sigma^2_{ANV}}\widetilde{\rho}^2_{ANV}.$$

Substituting the quantities, Browne et al. (2009b, 2010a) demonstrated that $1-\hat{\rho}^2_{A_{Com}}$ is the smaller root of

$$\begin{aligned}(v_F-q)\,(1-\rho^2_{A_{Com}})^2 + \Big(q(1-\hat{\rho}^2_{A_{ANV}}-r^{-1}) - v_F(2-\hat{\rho}^2_{A_{Reg}})\Big)\,(1-\rho^2_{A_{Com}}) \\ + \Big(v_F\,(1-\hat{\rho}^2_{A_{Reg}}) + q\,r^{-1}(1-\hat{\rho}^2_{A_{ANV}})\Big) = 0,\end{aligned}$$

where $\rho^2_{A_{Com}}$ is the the combined estimate of ρ^2_A, $v_F = \frac{2(n_b-1)^2(k(r-1)+n_b-3)}{k(r-1)(n_b-3)^2(n_b-5)}$, $q^{-1} = \sum_{i=1}^{k}\frac{\left(y_0^i-\bar{y}_b\right)^2}{s_b^2}$, and s_b^2 is the sample variance of the baseline data. Note that we choose the smaller root since the other one gives estimates of ρ^2_A less than zero, while $0 \leq \rho^2_A \leq 1$. Therefore,

$$\begin{aligned}\hat{\rho}^2_{A_{Com}} = 1 + \frac{q(1-\hat{\rho}^2_{A_{ANV}}-r^{-1}) - v_F(2-\hat{\rho}^2_{A_{Reg}})}{2\,(v_F-q)} \\ + \frac{\sqrt{\Big(q(1-\hat{\rho}^2_{A_{ANV}}-r^{-1})-v_F(2-\hat{\rho}^2_{A_{Reg}})\Big)^2 - 4(v_F-q)\Big(v_F(1-\hat{\rho}^2_{A_{Reg}})+qr^{-1}(1-\hat{\rho}^2_{A_{ANV}})\Big)}}{2\,(v_F-q)}.\end{aligned}$$

To compare the four different estimators of ρ^2_A, we use 1000 simulation runs, and in each run, we estimate the bias and standard deviation of $\hat{\rho}^2_A$ with the four estimators for different ρ^2_A values and $r \in \{2, 10\}$.

We assume the same simulation setup as in Appendix B.2.1 to generate the data. Figures B.5 and B.6 provide the results for $Bias\left(\hat{\rho}_A^2\right) = \rho_A^2 - \hat{\rho}_A^2$ and $SD\left(\hat{\rho}_A^2\right)$ of Phase I estimators for different ρ_A^2 values.

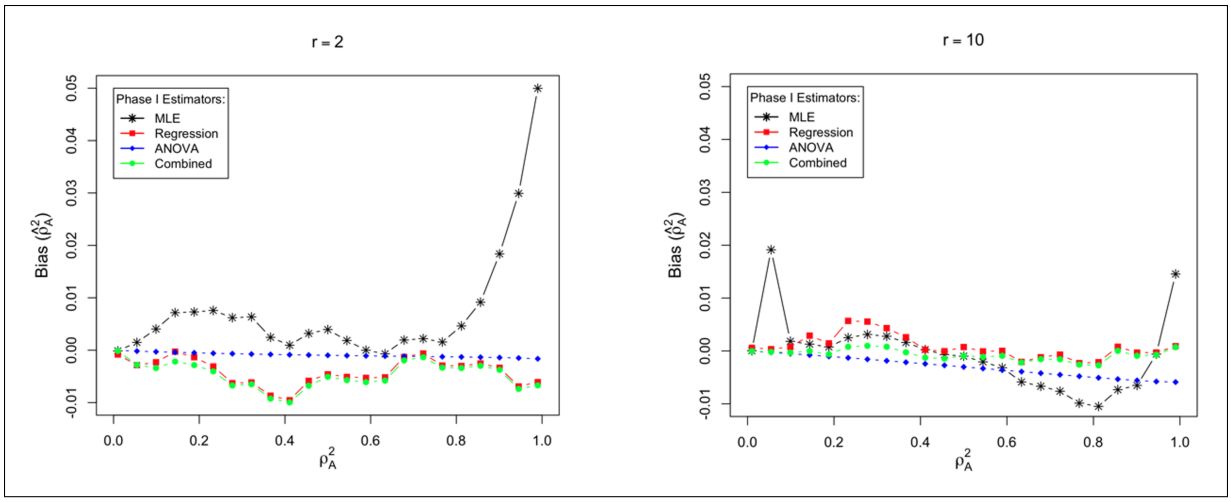

Figure B.5: $Bias(\hat{\rho}_A^2)$ vs. ρ_A^2 for different estimation methods with $r = 2$ (left panel) and $r = 10$ (right panel) when $n_b = 1000$ and $k = 2$.

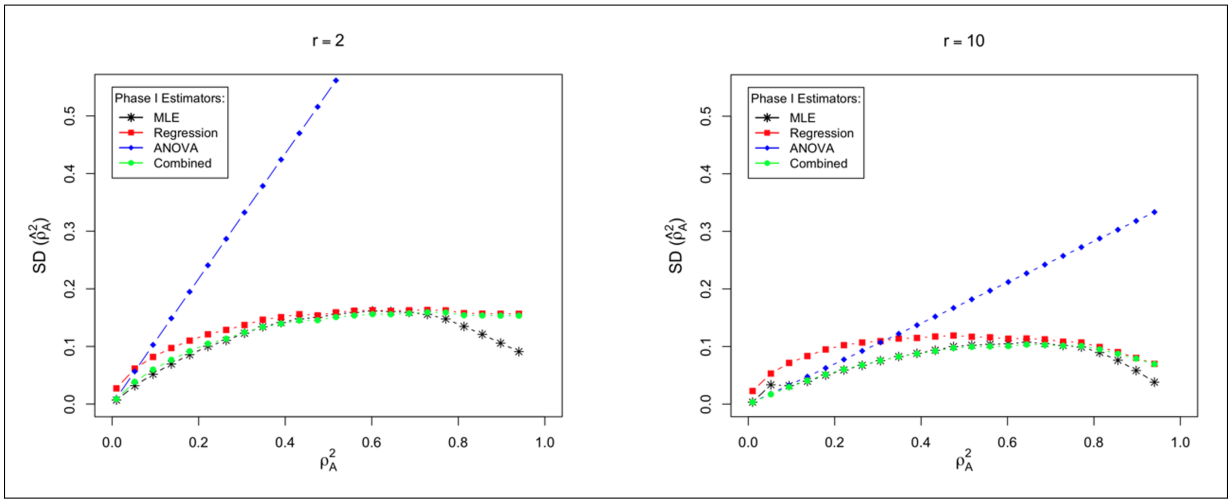

Figure B.6: $SD(\hat{\rho}_A^2)$ vs. ρ_A^2 for different estimation methods with $r = 2$ (left panel) and $r = 10$ (right panel) when $n_b = 1000$ and $k = 2$.

Figures B.5 and B.6 reveal that the MLE has the lowest standard deviation. However, it has a high bias for low r and high ρ_A^2 values. ANOVA is the best estimator in terms of bias, but it has a dramatically high standard deviation, particularly for low r values and high ρ_A^2. Moreover, the regression and combined estimators exhibit similar bias behaviour, but the combined estimator slightly outperforms the regression estimator with a smaller $SD(\hat{\rho}_A^2)$. Therefore, we recommend using the combined estimator, which has a closed form and provides a low standard deviation of $\hat{\rho}_A^2$ with a negligible bias.

B.3.2 Identifying Interaction between Assembly and Component(s)

To evaluate the effectiveness of our proposed "checks for irregularities" in Section 3.3.1.2 in signalling possible important interactions, we use simulation to calculate the p-value of Bartlett's and Levene's variance equality tests. To do so, we conduct simulations with and without selecting the median product (in addition to the two extreme products) and summarize the results of 1000 simulations in Figure B.7. In the simulation, we assume $\rho_{AC}^2 = 0.80$ and $\rho_A^2 = \rho_C^2 = 0.10$. In the left panel of Figure B.7, we disassemble and reassemble $k = 2$ extreme products $r = 8$ times each, and in the right panel, we disassemble and reassemble $k = 3$ products including one median and two extreme products $r = 5$ times each. Note that with these settings, we have more data in the left panel ($rk = 16$ vs. $rk = 15$).

Figure B.7 confirms that when there is strong interaction, despite the left panel benefiting from more data, using only the extreme products usually does not suggest unequal variances. Therefore, using only extreme products, the irregularity check is unlikely to provide evidence of interaction even though there is a strong interaction between assembly and component(s). On the other hand, we more often reject the variance equality assumption when we also disassemble and reassemble the median product (the right panel), particularly when using Bartlett's test. Therefore, to ensure the model assumption is not violated, we recommend using the median as well as the extreme products in Phase I. We also simulated other scenarios in which assembly or components are the dominant cause, and there is no interaction (results not shown here). In both cases, the rejection rate from both Bartlett's and Levene's tests are close to 0.05, indicating that the tests reject the variance equality assumption only when the assembly process and components interact. Figure B.7 reveals Bartlett's test superiority in identifying evidence of interaction between assembly and component(s) in case such interaction exists. However, Bartlett's test is more sensitive to the normality assumption (Levene (1960)). Therefore, if the normality assumption is a concern, we recommend Levene's test.

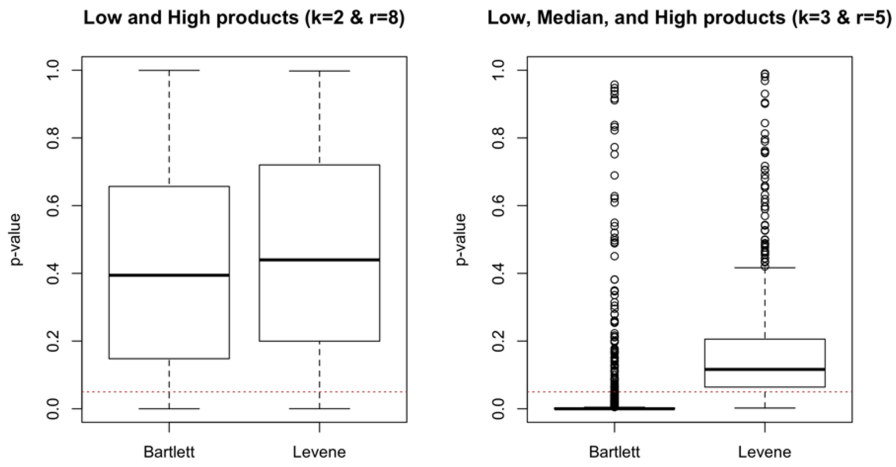

Figure B.7: Boxplots of p-values of Bartlett's and Levene's tests when there is extreme interaction between the assembly process and components without (the left panel with $k = 2$ and $r = 8$) and with (the right panel with $k = 3$ and $r = 5$) the median product where dashed lines indicate the critical value of 0.05.

B.3.3 Recommended Parameters for Phase I

Selecting k products from the n_b baseline data and disassembling and reassembling each of them r times leads to a total number of n_b+rk measurements. However, collecting baseline data is usually much cheaper than collecting the disassembling and reassembling data since the baseline data often have already been collected for another purpose. Therefore, to compare different designs, we fix n_b and investigate how different r and k combinations with rk fixed affect the standard deviation of $\widetilde{\rho}_A^2$ using the combined estimator introduced in Appendix B.3.1. We also investigate whether selecting the median and the $k-1$ extreme products (we call this the "Leveraged" plan) is more beneficial than selecting k random products from the baseline (we call this the "Random" plan).

To examine the trade-off between r and k, we consider $n_b = 400$. Note that these results are not sensitive to the baseline size unless it is small (e.g., less than 50). We examine two situations: $rk = 30$ and $rk = 15$. While assuming $r \geq 2$ (to avoid a zero denominator in the ρ_A^2 estimator's formula) and $k \geq 2$ (to include at least two random/extreme products), there are six and two possible combinations of k and r when $rk = 30$ and $rk = 15$, respectively. If $k = 2$, we select two extreme products and no median product; however, for $k > 2$, we select one median and $k - 1$ extreme products. Through 5000 simulation runs, we estimate $SD(\rho_A^2)$ for each design and summarize the results in Figure B.8.

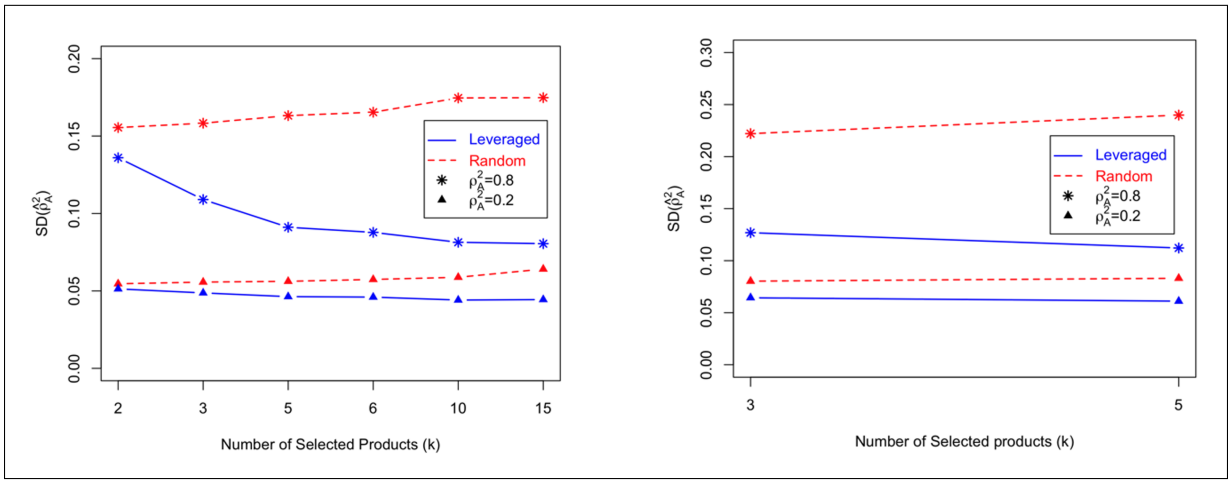

Figure B.8: $SD(\hat{\rho}_A^2)$ vs. k with $r = 30/k$ (the left panel) and $r = 15/k$ (the right panel) for the "Leveraged" and "Random" plans when $n_b = 400$ and $\rho_A^2 \in \{0.2, 0.8\}$.

Figure B.8 demonstrates that leveraging is a valuable idea, as the "Leveraged" plan consistently yields lower $SD\left(\hat{\rho}_A^2\right)$ compared to the "Random" plan, particularly when assembly is the dominant cause (e.g., $\rho_A^2 = 0.8$). The results also indicate that, for the "Leveraged" plan, selecting more products and performing a small number of disassemblies and reassemblies leads to slightly more precise conclusions than disassembling and reassembling fewer products more times, especially when assembly is the dominant cause and rk is high.

While the relationship between r and k is more evident in the left panel of Figure B.8 and $rk = 30$ allows us to explore a wider range of r and k combinations, collecting this amount of disassembly and reassembly data would often be too expensive. Instead, we believe $rk = 15$ is a more economical sample size. As illustrated in the right panel of Figure B.8, $rk = 15$ yields very similar performance in terms of $SD\left(\hat{\rho}_A^2\right)$ for both combinations of r and k. Consequently, to propose a Phase I design, we also consider other criteria, including obtaining more precise estimates in Phase II and being better able to identify irregularities. Compared to $k = 5$ and $r = 3$, the setting $k = 3$ and $r = 5$ allows us to collect more disassembly and reassembly measurements for the two extreme products used in Phase II. This leads to more precise estimates for $\bar{Y}_{PhI}^L$ and $\bar{Y}_{PhI}^H$. Moreover, with this combination, Bartlett's and Levene's tests are more effective in identifying evidence of potential assembly by component(s) interactions. Therefore, we recommend setting $k = 3$ and $r = 5$ for the Phase I study design, i.e., selecting a low, median, and high product from

the baseline and disassemble and reassemble them $r = 5$ times. Selecting a low, median, and high product from the baseline, i.e., $k = 3$ also aligns with the recommendation of Steiner and MacKay (2005) for measurement assessment studies.

B.4 Proposed Phase II Analysis

Here, the simulation setup is like Appendix B.2.2. We devote Appendix B.4.1 to compare the ANOVA and LVR estimators for a few scenarios when there is no interaction between components. In Appendix B.4.2, we consider some scenarios where there are component-by-component interaction effects.

B.4.1 Comparison between the ANOVA and LVR Estimators

We conducted a comprehensive analysis between ANOVA and LVR, under the simplifying assumption of having only two non-interacting components. Through 1000 simulation runs, we estimate $\rho^2_{C_1}$ for $\rho^2_{C_1} \in \{0.0, 0.1., \ldots, 1.0\}$, where $\rho^2_{C_2} = 1 - \rho^2_{C_1} - \rho^2_A$. The simulation setting is the same as Appendix B.2.2 (involving two extreme products), but with our recommendation of $r = 5$. Figure B.9 represents the results of $Bias\left(\hat{\rho}^2_{C_1}\right) = \rho^2_{C_1} - \hat{\rho}^2_{C_1}$, $SD(\hat{\rho}^2_{C_1})$, and $MSE\left(\hat{\rho}^2_{C_1}\right) = Bias^2\left(\hat{\rho}^2_{C_1}\right) + SD^2(\hat{\rho}^2_{C_1})$ when we only have two components, and we first swap C_1. Figure B.9 also compares the estimators for $\rho^2_A = 0.05$ and $\rho^2_A = 0.35$.

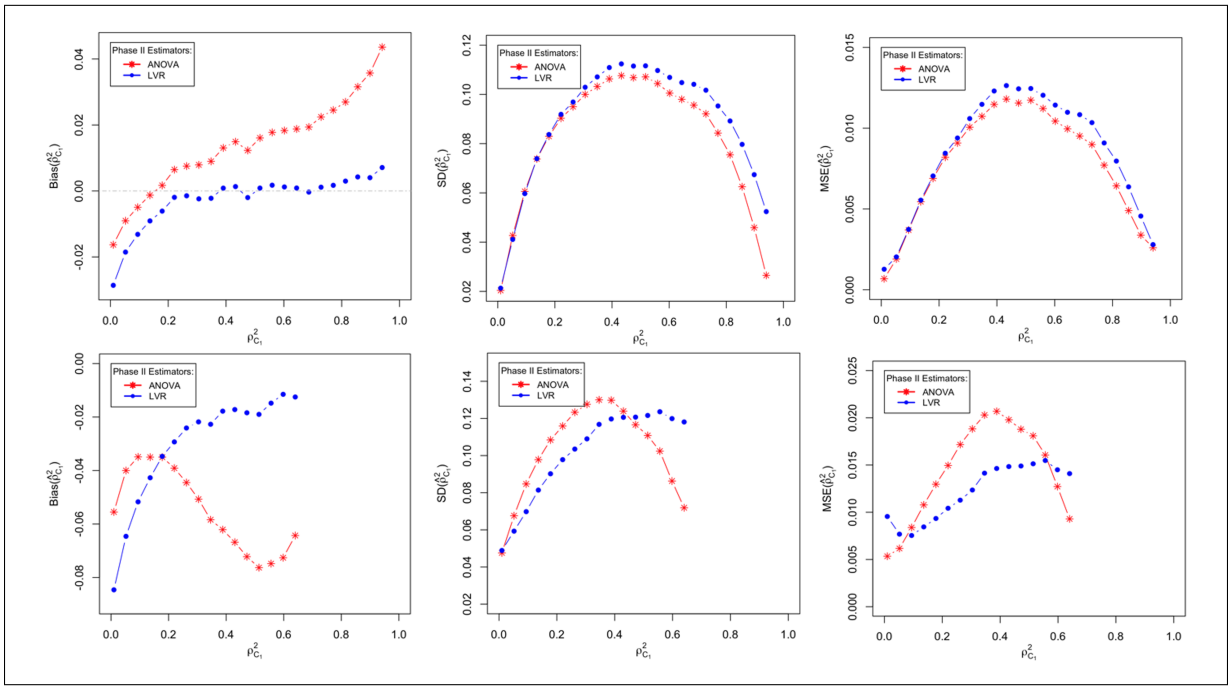

Figure B.9: $Bias(\hat{\rho}^2_{C_1})$ (left panel), $SD(\hat{\rho}^2_{C_1})$ (middle panel), and $MSE(\hat{\rho}^2_{C_1})$ (right panel) vs. $\rho^2_{C_1}$ for ANOVA and LVR estimators when $\rho^2_A = 0.05$ (first row) and $\rho^2_A = 0.35$ (second row) for $n_b = 1000$ and $r = 5$.

Figure B.9 reveals that the LVR method is more robust against large assembly variations compared to the ANOVA estimator.

B.4.2 Identifying Interaction between Two or More Components

To investigate how well we can identify interactions between components, we conducted a simulation study and estimated the probability of identifying each component as the dominant cause using ANOVA and LVR estimators. Considering the sequential nature of the ANOVA and LVR methods, different simulation runs may correspond to a different number of swaps due to different stopping times. We use similar assumptions and simulation settings as Appendix B.2.2, however, we allow interaction between components (i.e., $\beta_{C_1C_2} \neq 0$). Therefore, we set different sets of values for β_{C_1}, β_{C_2}, $\beta_{C_1C_2}$, β_{C_R}, and β_A (and thus $\rho^2_{C_1}$, $\rho^2_{C_2}$, $\rho^2_{C_1C_2}$, $\rho^2_{C_R}$, and ρ^2_A) to match various scenarios. In each simulation run, we select two extreme products from a baseline of $n_b = 1000$ products and disassemble and

reassemble them $r = 5$ times. Through 1000 simulation runs, we investigate the probability of identifying C_1, C_2, C_1C_2, and C_R as the dominant causes using the ANOVA and LVR methods. In the simulation, we make decision by comparing $\hat{\rho}^2_{C_i}$ to the threshold 0.5. Then, we investigate the effect of considering our proposed Partial and Extreme criteria on identifying an interaction between components. We limited the simulation results to only a few representative scenarios in the following subsections.

B.4.2.1 Scenario: No Interaction between Components

To check that our proposed criterion for identifying interactions does not lead to many false alarms, we consider a scenario with no interaction between components. An example of this scenario is when we set $\rho^2_{C_1} = 0.20$, $\rho^2_{C_2} = 0.75$, $\rho^2_{C_1C_2} = 0$, $\rho^2_{C_R} = 0$, and $\rho^2_A = 0.05$. To illustrate, Figure B.10 provides an example of the interaction and data collection plot for one simulation run. This figure denotes the data related to the low and high products in red and blue colors, respectively. In the left panel of Figure B.10, the red and blue crosses demonstrate the true (but unknown) values of C_1 and C_2 for the high and low products, respectively. In the right panel of Figure B.10, the circles represent the disassembly and reassembly results, and the crosses show the results when components are swapped.

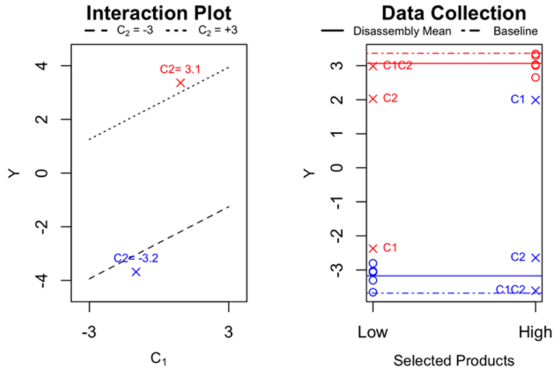

Figure B.10: An example of interaction and data collection plot when there is no interaction between components.

Figure B.11 and Table B.3 compare the ANOVA and LVR methods using 1000 simulation runs when we apply the component-swapping procedure and stopping rules. In this scenario, C_2 is the dominant cause and the blue and red horizontal dashed lines show the true values for $\rho^2_{C_1}$ and $\rho^2_{C_2}$, respectively. The results demonstrate that LVR leads to less biased estimates and slightly more often identifies the actual dominant cause.

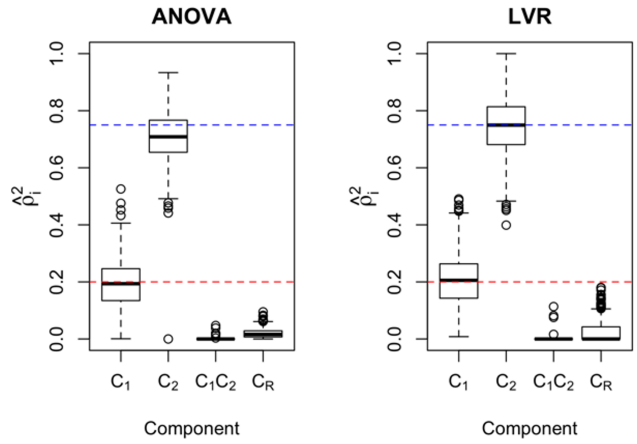

Figure B.11: Boxplots of $\hat{\rho}_i^2$ for $i \in \{C_1, C_2, C_1C_2, C_R\}$ for ANOVA and LRV estimators when $\left(\rho^2_{C_1}, \rho^2_{C_2}, \rho^2_{C_1C_2}, \rho^2_{C_R}, \rho^2_A\right) = (0.20, 0.75, 0, 0, 0.05)$.

	C_1	C_2	C_1C_2	C_R	$C_1 + C_2$
ANOVA	0.001	0.992	0.000	0.000	0.007
LVR	0.000	0.993	0.000	0.000	0.007

Table B.3: Probability of the dominant cause identifications when $\left(\rho^2_{C_1}, \rho^2_{C_2}, \rho^2_{C_1C_2}, \rho^2_{C_R}, \rho^2_A\right) = (0.20, 0.75, 0, 0, 0.05)$.

Table B.4 reveals how adding the Partial and Extreme criteria affects our ability to identify interactions. In this scenario, only 1.3% of the time we identify evidence of C_1C_2 interaction using the "Either of Partial or Extreme" criterion.

Interaction Identification Methods	Partial	Extreme	Either of Partial or Extreme
Probability of Identifying C_1C_2 as the Dominant Cause	0.000	0.013	0.013

Table B.4: The effect of adding Partial and Extreme criteria when $\left(\rho^2_{C_1}, \rho^2_{C_2}, \rho^2_{C_1C_2}, \rho^2_{C_R}, \rho^2_A\right) = (0.20, 0.75, 0, 0, 0.05)$.

B.4.2.2 Scenario: Mild Interaction between Components

An example of this scenario is when we set $\rho^2_{C_1} = 0.65$, $\rho^2_{C_2} = 0.29$, $\rho^2_{C_1C_2} = 0.01$, $\rho^2_{C_R} = 0$, and $\rho^2_A = 0.05$. Recall that when there is an interaction, we can no longer interpret the

ρ^2's as the proportion of variability in the output explained by that source. To illustrate the scenario, Figure B.12 provides an example of the interaction and data collection plot for one simulation run.

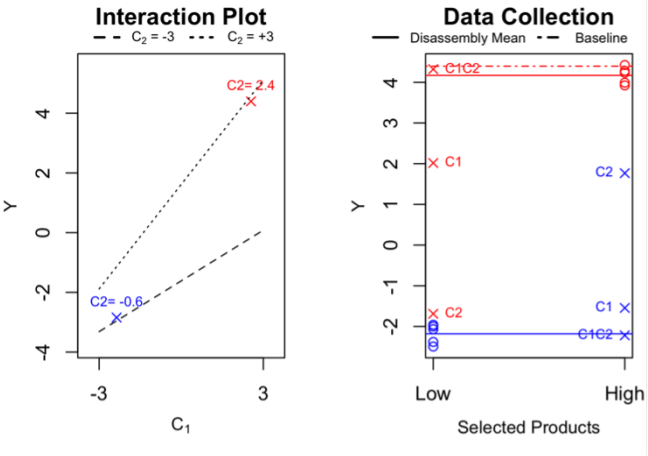

Figure B.12: An example of interaction and data collection plot when there is a mild interaction between components.

Figure B.13 and Table B.5 compare the ANOVA and LVR estimators using 1000 simulation runs. In this scenario, C_1 is the dominant cause, but there is a mild interaction between C_1 and C_2. The results demonstrate that the ANOVA method leads to less biased estimates and more often identifies the actual dominant causes. However, both the LVR and ANOVA estimators fail to identify any evidence of the mild C_1C_2 interaction.

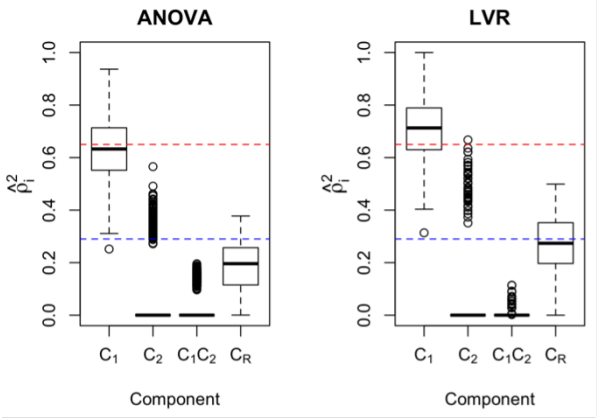

Figure B.13: Boxplots of $\hat{\rho}_i^2$ for $i \in \{C_1, C_2, C_1C_2, C_R\}$ for ANOVA and LRV estimators when $\left(\rho^2_{C_1}, \rho^2_{C_2}, \rho^2_{C_1C_2}, \rho^2_{C_R}, \rho^2_A\right) = (0.65, 0.29, 0.01, 0, 0.05)$.

	C_1	C_2	C_1C_2	C_R	$C_1 + C_2$
ANOVA	0.888	0.002	0.000	0.000	0.110
LVR	0.966	0.020	0.000	0.000	0.014

Table B.5: Probability of the dominant cause identifications when $\left(\rho^2_{C_1}, \rho^2_{C_2}, \rho^2_{C_1C_2}, \rho^2_{C_R}, \rho^2_A\right) = (0.65, 0.29, 0.01, 0, 0.05)$.

Table B.6 reveals how adding the Partial and Extreme criteria affects the interaction identifications. In this scenario, we often identify evidence of C_1C_2 interaction using the "Either of Partial or Extreme" criterion.

Interaction Identification Methods	Partial	Extreme	Either of Partial or Extreme
Probability of Identifying C_1C_2 as the Dominant Cause	0.870	0.011	0.873

Table B.6: The effect of adding Partial and Extreme criteria when $\left(\rho^2_{C_1}, \rho^2_{C_2}, \rho^2_{C_1C_2}, \rho^2_{C_R}, \rho^2_A\right) = (0.65, 0.29, 0.01, 0, 0.05)$.

B.4.2.3 Scenario: Large Interaction between Components

An example of this scenario is when we set $\rho^2_{C_1} = 0.40$, $\rho^2_{C_2} = 0.50$, $\rho^2_{C_1C_2} = 0.05$, $\rho^2_{C_R}$, and $\rho^2_A = 0.05$. To illustrate, Figure B.14 provides an example of the interaction and data collection plot for one simulation run.

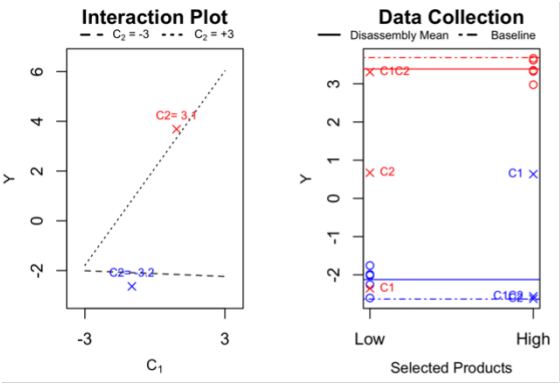

Figure B.14: An example of interaction and data collection plot when there is a large interaction between components.

Figure B.15 and Table B.7 compare ANOVA and LVR estimators using 1000 simulation runs. In this scenario, C_2 is the dominant cause, but C_1 is also a large cause and there is a large C_1C_2 interaction. The results demonstrate that LVR identifies the main effect dominant causes with higher probability, but neither analysis approach identifies the large interaction.

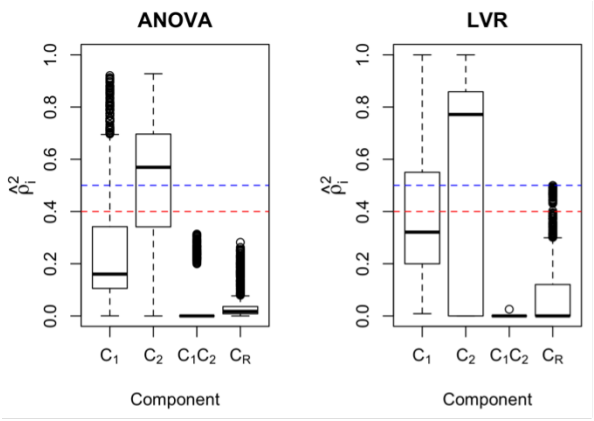

Figure B.15: Boxplots of $\hat{\rho}_i^2$ for $i \in \{C_1, C_2, C_1C_2, C_R\}$ for ANOVA and LRV estimators when $\left(\rho^2_{C_1}, \rho^2_{C_2}, \rho^2_{C_1C_2}, \rho^2_{C_R}, \rho^2_A\right) = (0.40, 0.50, 0.05, 0, 0.05)$.

	C_1	C_2	C_1C_2	C_R	$C_1 + C_2$
ANOVA	0.231	0.593	0.000	0.000	0.176
LVR	0.307	0.691	0.000	0.000	0.002

Table B.7: Probability of the dominant cause identifications when $\left(\rho^2_{C_1}, \rho^2_{C_2}, \rho^2_{C_1C_2}, \rho^2_{C_R}, \rho^2_A\right) = (0.40, 0.50, 0.05, 0, 0.05)$.

Table B.8 reveals how adding the Partial and Extreme criteria affects the interaction identifications. In this scenario, we often identify evidence of C_1C_2 interaction using the "Either of Partial or Extreme" criterion.

Interaction Identification Methods	Partial	Extreme	Either of Partial or Extreme
Probability of Identifying C_1C_2 as the Dominant Cause	0.836	0.361	0.893

Table B.8: The effect of adding Partial and Extreme criteria when $\left(\rho^2_{C_1}, \rho^2_{C_2}, \rho^2_{C_1C_2}, \rho^2_{C_R}, \rho^2_A\right) = (0.40, 0.50, 0.05, 0, 0.05)$.

B.4.2.4 Scenario: Pure Interaction between Components

An example of this scenario is when we set $\rho^2_{C_1} = 0.10$, $\rho^2_{C_2} = 0.10$, $\rho^2_{C_1C_2} = 0.75$, $\rho^2_{C_R} = 0$, and $\rho^2_A = 0.05$. To illustrate, Figure B.16 provides an example of the interaction and data collection plot for one simulation run.

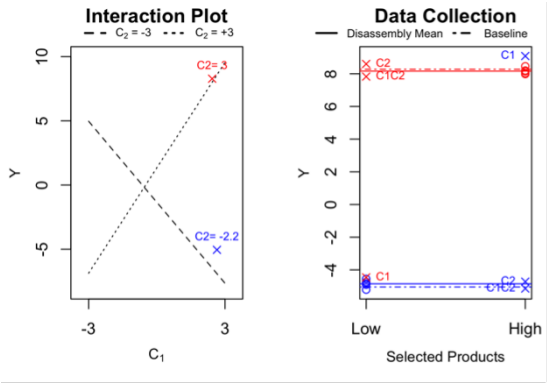

Figure B.16: An example of interaction and data collection plot when there is a pure interaction between components.

Figure B.17 and Table B.9 compare ANOVA and LVR estimators using 1000 simulation runs. In this scenario, C_1C_2 is the dominant cause. The results demonstrate that both estimators identify C_1 and C_2 with almost equal probabilities and fail to identify evidence of C_1C_2 interaction. Section 3.5 provides an explanation for these results, demonstrating how the selection of extreme products affects the conclusion in this scenario.

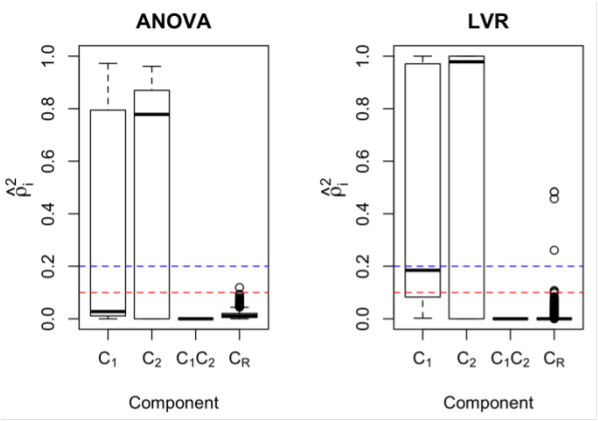

Figure B.17: Boxplots of $\hat{\rho}_i^2$ for $i \in \{C_1, C_2, C_1C_2, C_R\}$ for ANOVA and LRV estimators when $\left(\rho^2_{C_1}, \rho^2_{C_2}, \rho^2_{C_1C_2}, \rho^2_{C_R}, \rho^2_A\right) = (0.10, 0.10, 0.75, 0, 0.05)$.

	C_1	C_2	C_1C_2	C_R	$C_1 + C_2$
ANOVA	0.489	0.510	0.000	0.000	0.001
LVR	0.493	0.507	0.000	0.000	0.000

Table B.9: Probability of the dominant cause identifications when $\left(\rho^2_{C_1}, \rho^2_{C_2}, \rho^2_{C_1C_2}, \rho^2_{C_R}, \rho^2_A\right) = (0.10, 0.10, 0.75, 0, 0.05)$.

Table B.10 reveals how adding the Partial and Extreme criteria affects the interaction identifications. In this scenario, we almost always identify evidence of C_1C_2 interaction using the "Either of Partial or Extreme" criterion.

Interaction Identification Methods	Partial	Extreme	Either of Partial or Extreme
Probability of Identifying C_1C_2 as the Dominant Cause	0.658	0.906	0.911

Table B.10: The effect of adding Partial and Extreme criteria when $\left(\rho^2_{C_1}, \rho^2_{C_2}, \rho^2_{C_1C_2}, \rho^2_{C_R}, \rho^2_A\right) = (0.10, 0.10, 0.75, 0, 0.05)$.

B.4.3 Interaction Identification as the Dominant Cause

In the following, we elaborate on why data from a component swapping study analyzed as a factorial experiment never concludes that an interaction is the dominant cause. While here we added the simplifying assumption of having only two components, the argument can be extended to experiments involving more than two components. This is important because, in classical factorial experiments, we can identify the interaction effect as the dominant cause. However, as we explain later, the interaction effect cannot exceed 33% of the total in component swapping due to the selection of extreme assemblies for the experiment.

Using the notation used by Montgomery (2017) notation, consider a 2^2 factorial experiment with factors A and B, each having the low and high settings of $-$ and $+$. Then, as shown in Table B.11, we have four treatment combinations.

Factor A	Factor B	Name of the Combination
$-$	$-$	(1)
$+$	$-$	a
$-$	$+$	b
$+$	$+$	ab

Table B.11: Four treatment combinations and their names in a 2^2 factorial experiment.

Note that, in the component-swapping investigation, we use the two selected extreme assemblies to label the levels of factors A and B. Therefore, the run with A and B at the high setting yields y_0^L, and the run with A and B at the low setting yields y_0^H. In other words, $(1) \approx y_0^L$ and $ab \approx y_0^H$.

Following the notation used by Montgomery (2017) and the preceding explanation, we can write the effects as follows:

$$\begin{aligned} A &\approx ab + a - b - (1) \approx (y_0^H - y_0^L) + a - b, \\ B &\approx ab + b - a - (1) \approx (y_0^H - y_0^L) + b - a, \\ AB &\approx ab + (1) - a - b \approx (y_0^H + y_0^L) - a - b. \end{aligned}$$

When we swap components, the results almost always fall between y_0^L and y_0^H. Otherwise, we may not have selected the two extreme assemblies from the baseline. Therefore, $y_0^L \leq a \leq y_0^H$ and $y_0^L \leq b \leq y_0^H$. The consequence of these constraints is that $y_0^L - y_0^H \leq AB \leq y_0^H - y_0^L$, and the size of the interaction effect AB is the largest when either of the following satisfied:

(i) $a = b = y_0^L$, resulting in $AB = A = B = y_0^H - y_0^L$, or

(ii) $a = b = y_0^H$, resulting in $AB = A = B = y_0^L - y_0^H$.

Therefore, in a component-swapping investigation using extreme assemblies, the three effects should have equal magnitude to maximize the interaction effect. In other words, estimating an interaction effect larger than 33% of the total is impossible, and we never conclude that the interaction effect is the dominant cause.

Appendix C

Verification Study - Details and Additional Findings

C.1 Analytical Power Calculation for the Test $H_0 : \beta = 0$ vs. $H_A : \beta \neq 0$

Denote the low and high settings of X by x^- and x^+, respectively. By using the experimental data, the above hypothesis test can be written as $H_0 : \mu_{Y|x^+} = \mu_{Y|x^-}$ vs. $H_A : \mu_{Y|x^+} \neq \mu_{Y|x^-}$, where $\mu_{Y|X^+} = E(Y|X = x^+)$ and $\mu_{Y|x^-} = E(Y|X = x^-)$. Using the measured mean values $\bar{y}|x^+$ and $\bar{y}|x^-$ from $\frac{n_E}{2}$ experimental observations in each group, we estimate the parameters of interest. Then, considering Model 4.1 and the normality assumption for the residuals, the corresponding estimators are normally distributed with means $\mu_{Y|x^+}$ and $\mu_{Y|x^-}$, and variance $\frac{\sigma_\epsilon^2}{n_E/2}$. Then, the test is based on the statistic

$$T = \frac{(\bar{y}|x^+) - (\bar{y}|x^-)}{\frac{2\, s_\epsilon}{\sqrt{n_E}}},$$

where s_ϵ represents the pooled sample standard deviation of $y|x^+$ and $y|x^-$. Under the alternative hypothesis, T has a noncentral t-distribution on $n_E - 2$ degrees of freedom with noncentrality parameter

$$\theta = \frac{\mu_{Y|x^+} - \mu_{Y|x^-}}{\frac{2\,\sigma_\epsilon}{\sqrt{n_E}}}.$$

The power to detect a difference of $\delta = \mu_{Y|x^+} - \mu_{Y|x^-}$ with two-sided significant level 0.05 is given by

$$\text{Power} = 1 - T_{n_E-2}\Big(t_{(0.975;n_E-2)}|\frac{\delta}{\frac{2\,\sigma_\epsilon}{\sqrt{n_E}}}\Big) + T_{n_E-2}\Big(-t_{(0.975;n_E-2)}|\frac{\delta}{\frac{2\,\sigma_\epsilon}{\sqrt{n_E}}}\Big), \tag{C.1}$$

where $T_{df}(.|\theta)$ is the cumulative distribution function of the noncentral t-distribution with df degrees of freedom and noncentrality parameter θ, and $t_{(0.975;n_E-2)}$ is the probability of the t-distribution with $n_E - 2$ degrees of freedom being less than 0.975.

From Model 4.1, we have $\mu_{Y|x^\pm} = \alpha + \beta\, x^\pm$. Also, Figure 4.2 considers $x^\pm = \mu_X \pm 2\sigma_X$. Therefore,

$$\delta = (\alpha + \beta\, x^+) - (\alpha - \beta\, x^-) = 4\,\beta\,\sigma_X.$$

So, the power formula provided in Equation C.1 simplifies to

$$\text{Power} = 1 - T_{n_E-2}\Big(t_{(0.975;n_E-2)}|\frac{4\,\beta\,\sigma_X}{\frac{2\,\sigma_\epsilon}{\sqrt{n_E}}}\Big) + T_{n_E-2}\Big(-t_{(0.975;n_E-2)}|\frac{4\,\beta\,\sigma_X}{\frac{2\,\sigma_\epsilon}{\sqrt{n_E}}}\Big).$$

Considering $\rho^2 = \frac{\beta^2\sigma_X^2}{\beta^2\sigma_X^2+\sigma_\epsilon^2} = 0.5$ in Figure 4.2, we have $\sigma_\epsilon = \beta\,\sigma_X$. As a result, the power of the test simplifies to

$$\text{Power} = 1 - T_{n_E-2}\Big(t_{(0.975;n_E-2)}|2\sqrt{n_E}\Big) + T_{n_E-2}\Big(-t_{(0.975;n_E-2)}|2\sqrt{n_E}\Big). \tag{C.2}$$

From Equation C.2, it is clear that the power of the test depends only on the experimental sample size, n_E. Figure 4.2 represents the results when $n_E \in \{6, 8, 10, \ldots, 32\}$.

C.2 Study Designs and Models for Binary Causes

In practice, suspect dominant causes may not be continuous variables and instead may be binary. For example, we may find that the suspect dominant cause is the process stream number when a process consists of two streams running in parallel.

For simplicity, consider the situation where we have only a single binary suspect, denoted X^* to verify. However, we discuss a verification study with multiple suspect dominant causes in Section 4.7. Also, we model the effect of X^* on the output Y as

$$Y = \alpha + \beta\, X^* + \epsilon, \tag{C.3}$$

where X^* and ϵ are independent, and the distribution of X^* is

$$X^* = \begin{cases} -1 & \text{with probability } q, \\ +1 & \text{with probability } 1-q, \end{cases}$$

with $0 < q < 1$. Note that in some applications, we may assume q is known. For example, suppose X^* represents two parallel processing streams. Then, if there are approximately equal volumes in the two streams, we know $q \approx 0.5$.

To make our results more general, we continue with the assumption that q is unknown. Given this setup, $Var\,(X^*) = 4\,q\,(1-q)$ and $Var\,(Y) = \beta^2(4\,q)(1-q) + \sigma_\epsilon^2$. Therefore, X^* is strictly a dominant cause of variation if $\beta^2(4q)(1-q) > \sigma_\epsilon^2$. Similar to the continuous case, it is convenient to consider the squared correlation between Y and X^* given by

$$\rho^{*2} = \frac{\beta^2(4q)(1-q)}{\beta^2(4q)(1-q) + \ \sigma_\epsilon^2}, \tag{C.4}$$

where $0 \leq \rho^{*2} \leq 1$. Using this parameterization, X^* is strictly a dominant cause if $\rho^{*2} > 0.5$.

To verify X^* as a dominant cause in Model C.3, the goal is to determine whether:

> The *causal* contribution of X^* to the output variation (i.e., $\beta^2(4\,q)(1-q)$) is large compared to the variation due to noise and other causes (i.e., σ_ϵ^2).

Meeting this goal requires estimating three parameters: β, σ_ϵ, and q, or two of these three along with σ_Y, given that $\sigma_Y^2 = \beta^2\,(4\,q)\,(1-q) + \sigma_\epsilon^2$. Note that besides these parameters, we also need to estimate the nuisance parameter α to fit Model C.3. Plugging in the estimated parameters into Equation C.4, we obtain an estimate for ρ^{*2}.

Here, we consider the same study designs as Section 4.4. Assuming $\epsilon \sim N(0, \sigma_\epsilon^2)$, the log-likelihood when we have experimental $(x^*, y)_E$ data is given by Equation 4.3. This is true because the distribution of $(Y|x^*)$ does not depend on the distribution X^*. However, the log-likelihoods are different for the various types of observational data and are given below.

1. With observational paired $(x^*, y)_O$ data, we have

$$P(X_i^* = x_i^*,\ Y_i = y_i) = (\frac{1-x_i^*}{2})\ q\ P(Y_i|X_i^* = -1) + (\frac{1+x_i^*}{2})(1-q)\ P(Y_i|X_i^* = +1),$$

where $(Y_i|X_i^* = \pm 1) \sim N(\alpha \pm \beta,\ \sigma_\epsilon^2)$. So, the log-likelihood is

$$l_O^* = \sum_{i=1}^{n_O} \log \left(\frac{q\ (\frac{1-x_i^*}{2})}{\sqrt{2\pi\sigma_\epsilon^2}}\ e^{-\frac{(y_i-\alpha+\beta)^2}{2\sigma_\epsilon^2}} \right) + \left(\frac{(1-q)(\frac{1+x_i^*}{2})}{\sqrt{2\pi\sigma_\epsilon^2}}\ e^{-\frac{(y_i-\alpha-\beta)^2}{2\sigma_\epsilon^2}} \right). \tag{C.5}$$

2. With only observational $(x^*)_O$ data, we have

$$P\left(X_i^* = x_i^*\right) = q(\frac{1-x_i^*}{2}) + (1-q)(\frac{1+x_i^*}{2}),$$

where $x_i^* \in \{-1, +1\}$. So, the log-likelihood is

$$l_{Ox}^* = \sum_{i=1}^{n_{Ox}} \log \left(q(\frac{1-x_i^*}{2}) + (1-q)\,(\frac{1+x_i^*}{2}) \right). \tag{C.6}$$

3. With only observational $(y)_O$ data, we have

$$P\left(Y_i = y_i\right) = q\ P\left(Y_i = y_i|X_i^* = -1\right) + (1-q)\ \ P\left(Y_i = y_i|X_i^* = +1\right),$$

where $(Y_i|X_i^* = \pm 1) \sim N(\alpha \pm \beta,\ \sigma_\epsilon^2)$. So, the log-likelihood is

$$l_{Oy}^* = \sum_{i=1}^{n_{Oy}} \log \left(\frac{q}{\sqrt{2\pi\sigma_\epsilon^2}}\ e^{-\frac{(y_i-\alpha+\beta)^2}{2\sigma_\epsilon^2}} + \frac{1-q}{\sqrt{2\pi\sigma_\epsilon^2}}\ e^{-\frac{(y_i-\alpha-\beta)^2}{2\sigma_\epsilon^2}} \right). \tag{C.7}$$

In the following, we investigate different experimental and observational sample sizes to verify the dominant cause. Similar to Section 4.4, to first assess the effect of different experimental sample sizes (i.e., n_E), we calculate the power of the hypothesis test $H_0 : \beta = 0$ vs. $H_A : \beta \neq 0$ from $(x, y)_E$ data using Equation C.1. However, here X^* is binary and $x^{*\pm} = \pm 1$. Therefore,

$$\delta = \left(\alpha + \beta x^{*+}\right) - \left(\alpha - \beta x^{*-}\right) = 2\beta.$$

So, the power formula provided in Equation C.1 simplifies to

$$Power = 1 - T_{n_E-2}\left(t_{(0.975;n_E-2)} \mid \frac{\sqrt{n_E}\ \beta}{\sigma_\epsilon} \right) + T_{n_E-2}\left(-t_{(0.975;n_E-2)} \mid \frac{\sqrt{n_E}\ \beta}{\sigma_\epsilon} \right).$$

As in Section 4.4, our focus is on the case where $\rho^{*2} = 0.5$. Also, we focus on q=0.5 because it is likely the most common case and, more importantly, has the lowest power.

Considering $\rho^{*2} = \frac{\beta^2(4\,q)(1-q)}{\beta^2(4\,q)(1-q)+\ \sigma_\epsilon^2} = 0.5$ and $q = 0.5$, we have $\sigma_\epsilon = \beta$. So, the power of the test simplifies to

$$Power = 1 - T_{n_E-2}\left(t_{(0.975;n_E-2)} \mid \sqrt{n_E}\right) + T_{n_E-2}\left(-t_{(0.975;n_E-2)} \mid \sqrt{n_E}\right). \tag{C.8}$$

From Equation C.8, the power of the test depends only on n_E. Figure C.1 represents the results of the analytical power calculation for $n_E \in \{6, 8, 10, \ldots, 32\}$ when $\rho^{*2} = 0.5$ and $q = 0.5$. For other ρ^{*2} values, although the results are slightly different, the overall conclusions remain the same.

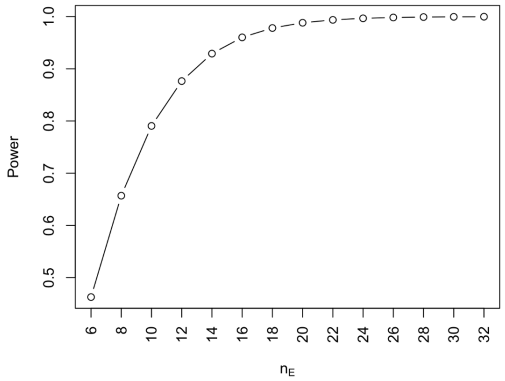

Figure C.1: Power of the test $H_0 : \beta = 0$ vs. $H_A : \beta \neq 0$ from $(x, y)_E$ data for $n_E \in \{6,\ 8,\ 10, \ldots,\ 32\}$ when $\rho^{*2} = 0.5$ and X^* is binary with $q = 0.5$.

Figure C.1 reveals that to achieve a high power for the hypothesis test with binary X^*, we need larger n_E values compared to the continuous suspect case. We leave it up to the practitioner to decide on the appropriate size of n_E in a way that makes sense in their context, but our recommendation is to have at least twelve experimental samples to obtain a power close to 0.90. If doing the experiment is not too expensive, $n_E = 16$ gives the power of almost 0.95, which is highly reliable. Similar to Section 4.4, having a high power is important because mistakenly eliminating X^* as the cause of variation in Y when it is the actual cause can waste considerable time and effort.

Through simulation, we investigate how well we can estimate ρ^{*2} for each viable composite study design. In Model C.3, we have four parameters to estimate, namely, α, β, σ_ϵ^2, and q, but ρ^{*2} depends only on β, σ_ϵ, and q. In the simulation, without loss of generality, we generate data with $\alpha = 0$. We also generate X^* with $q = 0.5$ as it is likely to be the most common case. Similar to Section 4.4, we fix $\beta = 1$ and determine the corresponding

σ_ϵ so that $\rho^{*2} = 0.5$, i.e., we consider $\sigma_\epsilon = 1$. In each simulation run, we first estimate all the four parameters, and then we obtain $\widehat{\rho^*}^2$ by plugging the estimates into Equation C.4. We summarize the results of 2000 simulation runs using the bias and standard deviation of $\widehat{\rho^*}^2$, denoted by $Bias(\widehat{\rho^*}^2)$ and $SD(\widehat{\rho^*}^2)$, respectively.

To estimate the parameters in Model C.3 when we have $(x^*, y)_O$, $(x^*)_O$ & $(y)_O$, or $(y)_O$ data, we use maximum likelihood estimation. Assuming the independence of different parts used in each study, the overall log-likelihood can be written as the sum of the log-likelihoods with non-zero corresponding sample sizes provided by Equations 4.3, C.5, C.6, and C.7.

When we only have $(x^*)_O$ as the observational data, only $(x^*, y)_E$ data provide information about α, β, and σ_ϵ^2. In this case, if we use maximum likelihood to estimate these parameters, since n_E is typically relatively small, the obtained $\hat{\sigma}_\epsilon^2$ will be biased. For this case, as in Section 4.4, to correct the bias, we suggest instead estimating σ_ϵ^2 with a $n_E - 2$ divisor. Note that in this case, q is estimated by the proportion of $(x^*)_O$ data that equals -1.

First, we investigate the relative merits of the viable composite study designs for $n_E = 16$ (recall that this experimental sample size gives very high power for the hypothesis test $H_0 : \beta = 0$ vs. $H_A : \beta \neq 0$ from $(x^*, y)_E$ data), where the observational sample sizes are $\{50, 100, 150, \ldots, 1000\}$. Figure C.2 presents the results for $Bias(\widehat{\rho^*}^2) = \widehat{\rho^*}^2 - \rho^{*2}$ and $SD(\widehat{\rho^*}^2)$. Note that in Figure C.2, similar to Section 4.4, when the observational sample size is n, $n_O = n$, $n_{Ox} = n$, and $n_{Oy} = n$ for the $(x^*, y)_O$, $(x^*)_O$, and $(y)_O$ cases, and $n_{Ox} = n_{Oy} = n$ for the $(x^*)_O$ & $(y)_O$ case.

The left panel of Figure C.2 demonstrates that the bias for all combinations is fairly small. The right panel reveals that, as in Section 4.4, the most precise estimates arise when we supplement the experiment with $(x^*, y)_O$ data. However, recall that to use such data, we must assume no confounder with a large inference exists. Figure C.2 also reveals that unlike in Section 4.4, only having $(x^*)_O$ data does not provide very valuable information in terms of $SD(\widehat{\rho^*}^2)$, whereas $(y)_O$ data help considerably more. Having $(x^*)_O$ & $(y)_O$ data is only slightly better than $(y)_O$ data because $(x^*)_O$ data are not very informative, even though $(x^*)_O$ & $(y)_O$ data represent twice as many observational data.

Second, we investigate the relative merits of the viable composite study designs for n_O, n_{Ox}, or $n_{Oy} = 200$, when $n_E \in \{12, 14, \ldots, 32\}$. Figure C.3 presents the results for $Bias(\widehat{\rho^*}^2) = \widehat{\rho^*}^2 - \rho^{*2}$ and $SD(\widehat{\rho^*}^2)$.

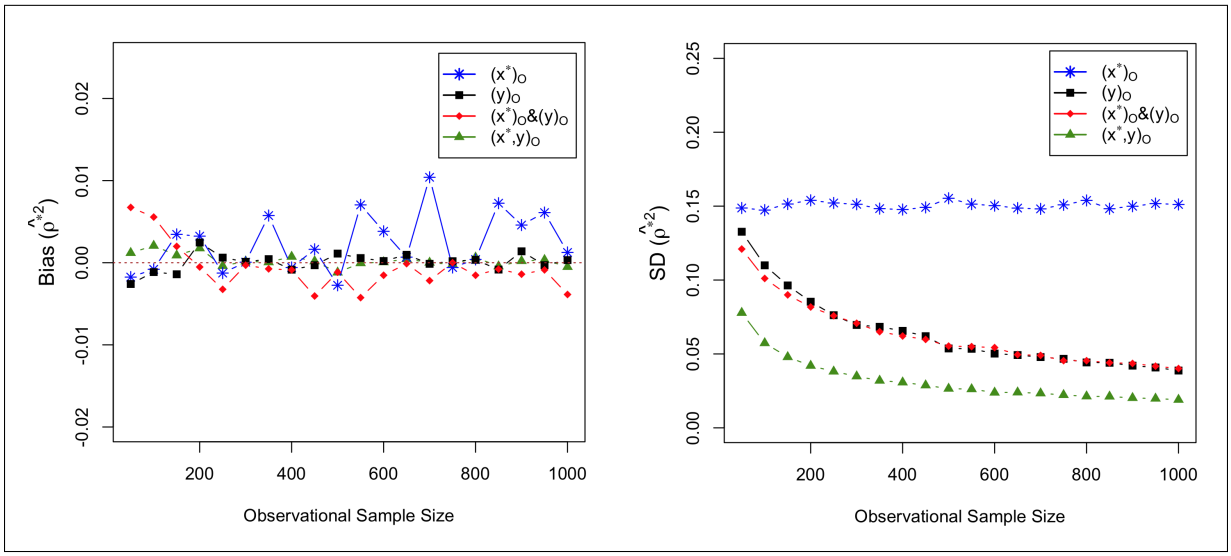

Figure C.2: $Bias(\widehat{\rho^*}^2)$ (left panel) and $SD(\widehat{\rho^*}^2)$ (right panel) for different viable combinations of data when $\rho^{*2} = 0.5$ and $n_E = 16$.

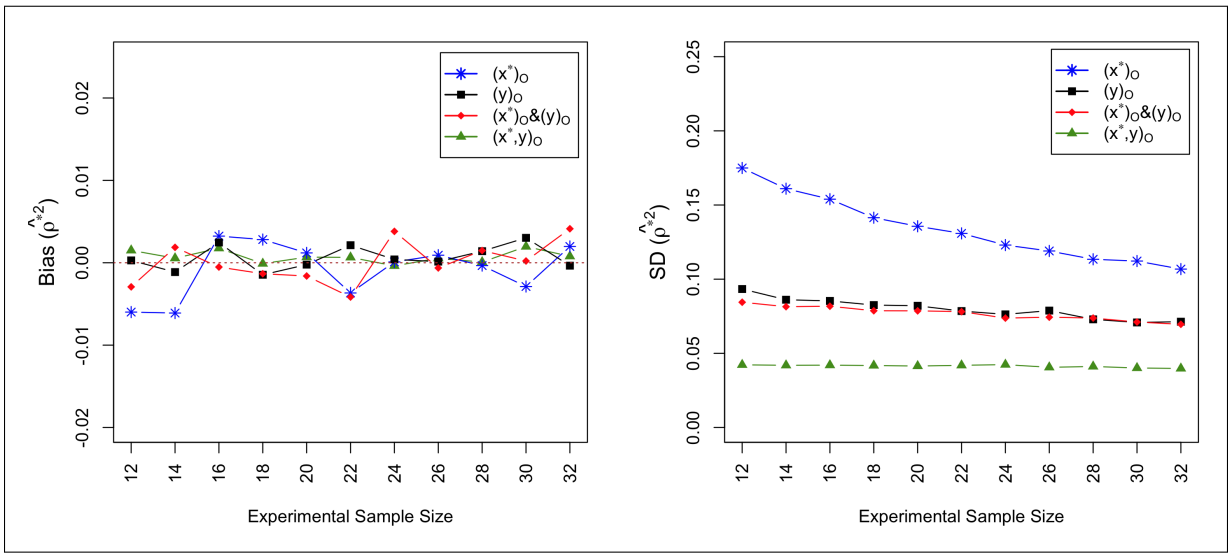

Figure C.3: $Bias(\widehat{\rho^*}^2)$ (left panel) and $SD(\widehat{\rho^*}^2)$ (right panel) for different viable combinations of data when $\rho^{*2} = 0.5$ and there are 200 observational data.

The left panel of Figure C.3 shows that the bias for all combinations is fairly small. The right panel reveals that when we have $(x^*)_O$ data, as n_E increases, $SD(\widehat{\rho^*}^2)$ substantially decreases, but $SD(\widehat{\rho^*}^2)$ is still worse than the other data combinations. However, when we have $(x^*, y)_O$, only $(y)_O$, or $(x^*)_O$ & $(y)_O$ data, having more experimental data does not help much in terms of $SD(\widehat{\rho^*}^2)$.

C.3 The Derivation of ρ_C^2 Formula for $(x, y, c)_O$ Data

Consider Model 4.7 when the mean and standard deviation of X are denoted by μ_X and σ_X, and the corresponding values for C are μ_C and σ_C, respectively. Then, $E(Y) = \alpha + \beta\,\mu_X + \gamma\,\mu_C$, and $Var(Y) = \sigma_Y^2 = \beta^2\sigma_X^2 + \gamma^2\sigma_C^2 + 2\beta\gamma\sigma_{XC} + \sigma_\epsilon^2$. So,

$$\begin{aligned} E(XY) =& E\big(X(\alpha + \beta X + \gamma\, C + \epsilon)\big) \\ =& \alpha E(X) + \beta E(X^2) + \gamma E(XC) + E(X)E(\epsilon) \\ =& \alpha\mu_X + \beta\big(Var(X) + E^2(X)\big) + \gamma\big(\sigma_{XC} + E(X)E(C)\big) + 0 \\ =& \alpha\mu_X + \beta\sigma_X^2 + \beta\mu_X^2 + \gamma\big(\sigma_{XC} + \mu_X\mu_C\big). \end{aligned}$$

As a result,

$$\begin{aligned} \sigma_{XY} =& E(XY) - E(X)E(Y) \\ =& \Big(\alpha\mu_X + \beta\sigma_X^2 + \beta\mu_X^2 + \gamma(\sigma_{XC} + \mu_X\mu_C)\Big) - \mu_X(\alpha + \beta\mu_X + \gamma\mu_C) \\ =& \beta\sigma_X^2 + \gamma\sigma_{XC}. \end{aligned}$$

Therefore,

$$\begin{aligned} \rho_C^2 =& \frac{\sigma_{XY}^2}{\sigma_X^2\sigma_Y^2} = \frac{(\beta\sigma_X^2 + \gamma\sigma_{XC})^2}{\sigma_X^2(\beta^2\sigma_X^2 + \gamma^2\sigma_C^2 + 2\beta\gamma\sigma_{XC} + \sigma_\epsilon^2)} \\ =& \frac{(\beta\sigma_X + \gamma\sigma_{XC}/\sigma_X)^2}{\beta^2\sigma_X^2 + \gamma^2\sigma_C^2 + 2\beta\gamma\sigma_{XC} + \sigma_\epsilon^2} = \frac{\beta^2\sigma_X^2 + \gamma^2\sigma_{XC}^2/\sigma_X^2 + 2\beta\gamma\sigma_{XC}}{\beta^2\sigma_X^2 + \gamma^2\sigma_C^2 + 2\beta\gamma\sigma_{XC} + \sigma_\epsilon^2}, \end{aligned}$$

which is Equation 4.8.