Randomization: A Core Principle of DOE

Authored by:
Emily Divis
Sarah Burke, PhD
Melissa Key, PhD
Steven Thorsen, PhD

31 August 2020

The goal of the STAT COE is to assist in developing rigorous, defensible test strategies to more effectively quantify and characterize system performance and provide information that reduces risk. This and other COE products are available at www.afit.edu/STAT.

STAT Center of Excellence
2950 Hobson Way – Wright-Patterson AFB, OH 45433

Table of Contents

Executive Summary 2
Introduction 2
Why Randomize? 3
Statistical Assumptions and Generalizability 3
Example 3
Safeguarding Against "Unknown Unknowns" 4
When Complete Randomization is Infeasible 6
Hard-to-Change Factors and Cost/Scheduling Constraints 6
Uncontrollable Factors 8
Stability of the Test Conditions 8
Change in Test Plan 8
Impossible Randomization 9
Justification to Leadership 11
Conclusion 12
References 13

Executive Summary

The practice of implementing test run randomization techniques in the control of tests and experiments is often overlooked or downplayed significantly as a matter of course during test planning and design. Programs that do not pay attention to this key principle of design of experiments (DOE) (the key principles being 1) randomization, 2) replication, and 3) local control or blocking) run unknown risks associated with not adequately understanding the importance of assumptions within statistical tests and methods. This best practice addresses in detail why we randomize and the STAT COE approach to mitigating the randomization requirement when it is not feasible within a test or experiment. As a reference, you can find the more general STAT considerations enumerated within the document, "Guide to Developing an Effective STAT Test Strategy", found on the STAT COE best practices webpage (https://www.afit.edu/STAT/statdocs.cfm?page=1126), which outlines the general STAT Process.

The STAT COE recommends this best practice for every test planner and test manager. The material here reviews the principle of randomization, advocates the "why" of randomizing test runs, and considers the constraints to randomization that may lead a test planner/manager to seek a mitigating alternative to complete randomization that does not sacrifice rigor in the results.

This paper also provides critical rationale to justify test plan changes to run randomization for communication to more senior test leaders within your organization when that is necessary, especially when your leadership is expecting randomization, but the test planning process indicates constraints restricting full implementation of the principle.

Keywords: constraints, DOE, generalizability, randomization, rigor, statistics, split plot

Introduction

Randomization is the practice of using chance methods (such as flipping a coin) to assign treatments to experimental units in a manner that protects against unintended influences on the assignments (Stat Trek). For our purposes, a **treatment** is a one specific combination of conditions (several **factors** set at specific levels) controlled by the test team. These treatments are applied to a set of **experimental units** – a generic term which can refer to samples, test runs, individuals, etc. Randomization forms one of the core principles of DOE [Burke et al. (2019)]. It is a necessary step when planning a test to ensure valid statistical analysis is possible. Randomization safeguards experimenters against unforeseen and/or uncontrollable variables which might otherwise mask relationships between the factors and the response. While randomization may appear counterintuitive, difficult, and expensive, the costs of not randomizing are far greater. With proper planning, these perceived problems can be mitigated effectively without sacrificing rigor.

Why Randomize?

Randomization is standard practice in DOE and in the Scientific Test and Analysis Techniques (STAT) process because it ensures that the statistical assumptions required for generalizable results are met. This safeguards against potential setbacks and produces designs which hold up under scrutiny.

Statistical Assumptions and Generalizability

A critical assumption of most statistical tests is that the experimental units being analyzed are a representative subset of the population from which they were drawn. When this assumption holds, statistical test results can be generalized to the population. If the assumption is violated (i.e., not actually true), the conclusions from the statistical test may not be true. **Bias** occurs when the selection process produces a non-representative subset. A randomized assignment strategy counters bias by assuming that (on average) individual differences in experimental units are balanced out. Alternatively, a methodical approach counters bias by explicitly representing all relevant aspects of the population proportionally. This generally requires a thorough understanding of the population under test and large sample sizes to represent all (relevant) features of the population. Random sampling is not the only way to prevent bias, but it is very often the simplest and most efficient. Thus, randomization is an integral step in DOE because it ensures that the experimental units are representative of the entire population.

Another critical assumption of many statistical tests is that experimental observations are **independent** of each other; in other words, the response of one observation does not influence the responses of others. When the independence assumption does not hold, parameter estimates, specifically the variance, become inaccurate. Consequently, every common statistical test becomes suspect, including t-tests, analysis of variance, and linear regression (NIST). In an experiment, the simplest way to meet the independence assumption is to run the experiment such that the treatments are assigned in a random order and to "reset" all factor levels between runs (e.g., set the switch to the "off" position between runs, even if the test design requires it to be set to "on" for two consecutive runs). This **complete randomization** minimizes the influence any single test run can have on the outcomes of future test runs.

Example

Consider a shipment consisting of 20 boxes with 100 items per box. A quality control analysis is to be performed on 40 items to determine whether or not to accept or reject the shipment. A **convenience sample** is just that: it involves sampling the most accessible items under the assumption that all items are effectively identical (at least with respect to the metrics of interest). In this case, this could produce 40 items from the top of one easy-to-reach box.

There is no way to be certain, however, that all the items are the same based on the convenience sample. Even if we ignore the possibility of an unethical supplier filling less-accessible positions with substandard items, small variations will likely exist across items due to minute changes in machine calibration, weather, material composition, and many other components of the production line process. Since these changes build over time, the convenience sample is not representative of the entire shipment; the items in the sample are more likely to be similar to one another, and less likely to be

similar to other items elsewhere in the shipment, than if a random sampling strategy was used. This is evidence of a lack of independence.

Safeguarding Against "Unknown Unknowns"

In most situations, even the most carefully constructed and monitored testing environment is subject to random fluctuations. These fluctuations – called **noise –** can heavily confound an experiment. Another common hazard is **lurking variables**, which are factors that are unaccounted for (and therefore uncontrolled) but have an impact on responses. Lurking variables are often correlated with time trends (for example, a new process can appear to become more efficient over time as the operator grows accustomed to it). If the runs of a designed experiment are executed in a systematic order as opposed to a random order, the effect of a factor of interest could be masked by potential noise and lurking variables. As if that weren't frustrating enough, noise and lurking variables are often difficult to track or even notice. Even when we are aware that we are missing information, it is impossible to always know the extent of the information gap; this is the idea of "unknown unknowns."

Figures 1a, 1b, and 1c below provide an illustration of a lurking variable trend. Suppose we have an experiment with a factor set at two levels, +1 and -1 (high and low). The test team does not know what effect this factor has on the response – determining this effect is the purpose of the experiment, after all. Figure 1a shows the true effect of the factor.

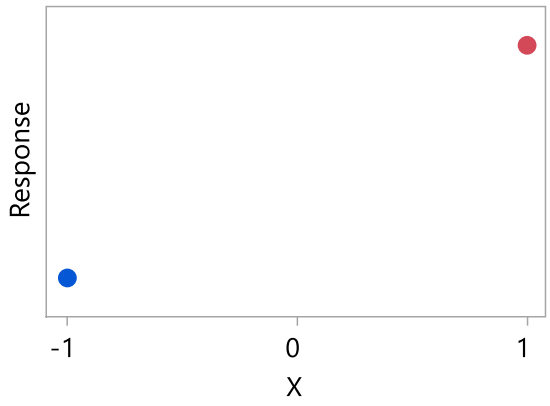


Figure 1a. True effect of example factor

The response increases when the factor is set to +1 and decreases when the factor is set to -1.

Unbeknownst to the test team, there is another factor unaccounted for that causes the response to increase over time (e.g., an increase in temperature over time or an operator's efficiency increasing due to repeating the test procedure), represented by the dotted black line in Figures 1b and 1c. Suppose we run the experiment in a systematic order, with the factor held at the high level for the first half of the experiment and held at the low level for the second half.

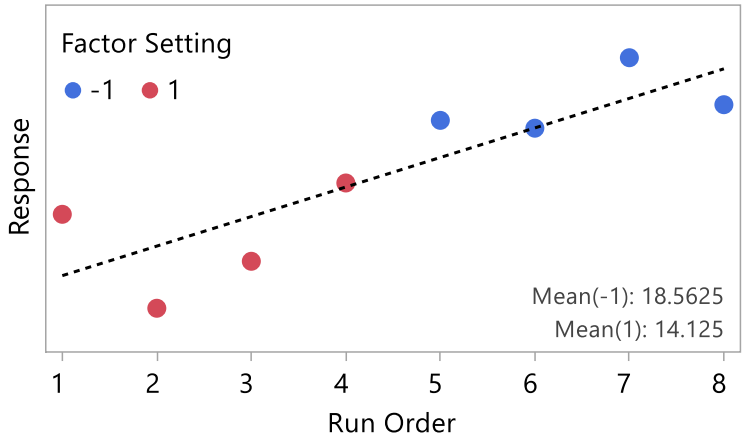


Figure 1b. Non-randomized example experiment (+1 settings followed by -1)

From these results, we see that the response is higher on average when the factor is set to -1. The obvious but erroneous conclusion here is that as the factor level increases, the response decreases. Due to the lurking variable's influence, we reach a wrong and possibly dangerous conclusion.

By contrast, suppose the same experiment is run in a randomized order:

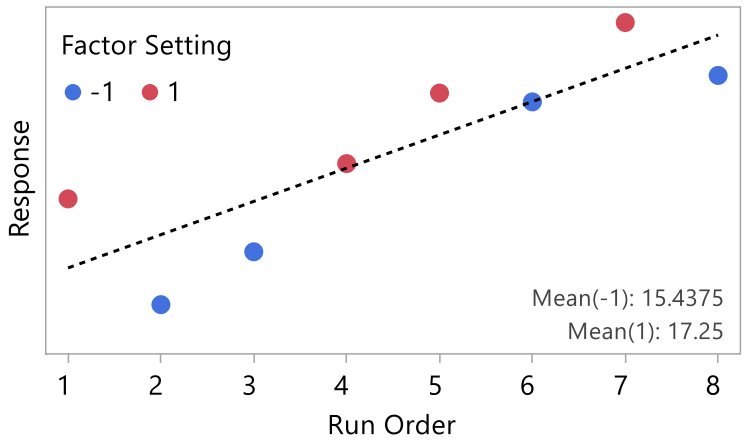


Figure 1c. Randomized example experiment

This time, the effect of the lurking variable is averaged out across the test runs, allowing us to observe the true effect of the factor. The response increases over time regardless of the factor's levels, but the runs with the factor set at -1 yield a lower average response than the runs at +1. By randomizing this experiment, we can correctly conclude that the response increases as the factor increases.

The best defense against the potential effects of any variable we cannot control in the test (or did not think to control) is to randomize the order of the test runs. Doing so will spread any unknown effects of

those uncontrolled variables evenly across the test points. Randomizing allows us to still draw the correct conclusions regarding the system; we are able to determine the factors of interest that have an effect on the response.

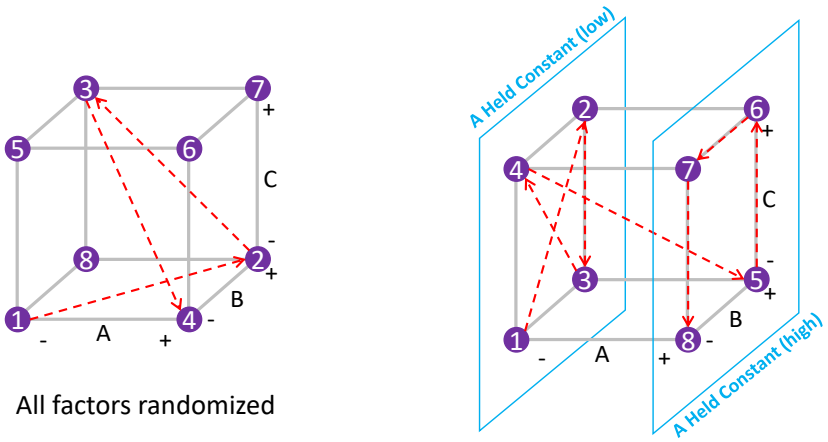
When Complete Randomization is Infeasible

If complete randomization is easy and inexpensive, then experimenters should always randomize the runs in the test. However, it is not always possible or practical to randomize all test runs. Below are some common constraints to randomization and methods to mitigate them without sacrificing rigor.

Hard-to-Change Factors and Cost/Scheduling Constraints

A common roadblock to a fully randomized test design is hard-to-change factors. These are factors whose levels take a lot of time, effort, money, or resources to change, resulting in conflicts with cost and scheduling constraints when the test run order is fully randomized. A split-plot design can account for hard-to-change factors by restricting the randomization in the experiment. The process for this takes two steps. First, randomize with respect to the hard-to-change factors only. Use **replicates** – that is, run some or all of the hard-to-change factor combinations more than once. Second, for each combination of hard-to-change factors, hold that combination fixed and create a set of test runs randomized with respect to the remaining factors. This greatly reduces the number of times hard-to-change factor levels must be adjusted, potentially allowing an otherwise infeasible experiment to be completed within budget. Statistical software should be used to help choose the best combination of hard-to-change and easy-to-change factors in the design.

Consider Figure 2 below, which illustrates this concept. Suppose there are three factors A, B, and C, to be tested for a total of eight runs. Suppose that factors B and C are reasonably easy-to-change, but factor A is hard-to-change. A traditional, fully randomized factorial experiment (shown on the left) would require factor A's level to be changed randomly (i.e., often), which is undesirable. A split-plot design (shown on the right) divides the test runs into two groups: four runs where factor A is held at the low level, and then four runs where factor A is held at the high level. Within each half of the total test runs, factors B and C are still randomized.



All factors randomized

B/C randomized at a fixed A level

Figure 2. Randomized and split-plot designs

Note that this split-plot design leaves factor A vulnerable to "unknown unknowns." Any lurking variables that are influenced by the passing of time will most likely be confounded with factor A, since all of factor A's low level runs take place before its high level runs. In order to estimate the effect of Factor A, this experiment requires another one or two sets of runs (replicates) of Factor A. Note that we still benefit in terms of feasibility; even with the added runs, the split-plot design will reduce the number of changes Factor A requires overall.

Split-plot designs come at the cost of lower **power** – the likelihood of discovering that a significant factor has a significant effect on the response – for the hard-to-change factors. Split-plot designs are intentionally not fully randomized, so they will always yield less information than their fully randomized counterparts. Therefore, a split-plot design should only be considered when complete randomization is truly infeasible. Consider whether the following actions can be taken to reduce the time or cost of a fully randomized design:

- Adjust the way hard-to-change factors are controlled or measured in order to make them easy-to-change.
- Decide that a hard-to-change factor is not a factor of interest and hold it constant throughout the experiment. (This option requires careful consideration – it comes at the cost of losing the ability to estimate the effects of the factor.)
- Implement a sequential testing strategy, including an initial screening. The screening will potentially rule out hard-to-change factors as significant so that they may be excluded or held constant in future tests.

If a split-plot design is necessary, a STAT expert can assist in creating one that strikes the proper balance between feasibility and rigor. Another example of a split-plot design is provided in the references section (Minitab).

Uncontrollable Factors

If a factor's levels are impossible to control, then it is certainly impossible to fully randomize test runs that involve that factor. Simply record this factor's levels instead of controlling them. Run the experiment, randomized with respect to the remaining factors. The recorded factor can still be analyzed as a covariate, which will yield information about correlation but not causation; consider consulting with an expert to accomplish this (Anderson & Whitcomb, 2010).

If most or all of the factors of interest are uncontrollable, DOE may not be appropriate or even possible. It may be necessary to perform an observational study instead. For guidance on this, see Stone (2018), "Test Planning for Observational Studies".

Stability of the Test Conditions

Even in a typical lab setting, variability and drift can be difficult to control with sufficient precision to justify ignoring their potential impact on experimental results. Lurking variables, time trends, and other unknown sources of variation (so-called "unknown unknowns") can impact the system under test and the testing environment in ways that are unpredictable. Randomization makes it safe to assume that the net effect of these potential influencers balances out. Consequently, even when there is a history of stability for a test, randomization is still recommended as standard practice. However, in the rare overlap between cases where the system and environment are highly stable and randomization is *prohibitively* expensive or difficult, it is acceptable to perform the experiment without randomization (Box, 1990). Evidence of these circumstances should be reported for credibility.

Change in Test Plan

There will be times when an experiment that is well-designed on paper is impractical or impossible to execute. For example, a factor that was thought to be easy-to-change may prove to be hard-to-change, or the number of allowed test runs may be significantly reduced after testing has begun. In such a situation, deviating from the original test design may be necessary. If a STAT expert assisted in designing the test, seek their input before proceeding with further testing.

If the experiment was performed differently from the recommended design – the test runs were reordered or reduced, for instance – then this difference, along with analysis of the potential negative effects of it, must be reported. Explain the motivation for deviating from the test design. Analyze and present the experiment as it was actually performed. Consult with a STAT expert for analysis on the impact of incomplete randomization on the experiment's conclusion. If the experiment has been weakened as a result of the deviation, acknowledge and take ownership of that fact. There could still be some value in analyzing a real experiment, even if it ultimately fails to answer the question it was designed to answer. There is no value in analyzing a perfect experiment that never took place.

For further reading on analysis of an experiment that deviated from the recommended test design, see Harman (2018), "Lessons Learned from an Incompletely Randomized Test Design," a case study which describes a test team's effort to investigate potential causes of wide shot dispersion in vehicle weapon

testing. The original design was not executed in the intended random test order; the case study explores the resulting impact on analysis of the experiment's findings.

Impossible Randomization

If randomization is extremely difficult or impossible, and test conditions are not sufficiently stable to justify not randomizing, then a designed experiment cannot be performed at this time. There are now three options moving forward: find a way to make randomization feasible, stabilize the test environment, or perform an observational study (Box, 1990).

Figure 3 below summarizes the common recommendations for randomization when there may be uncontrollable or hard-to-change factors present.

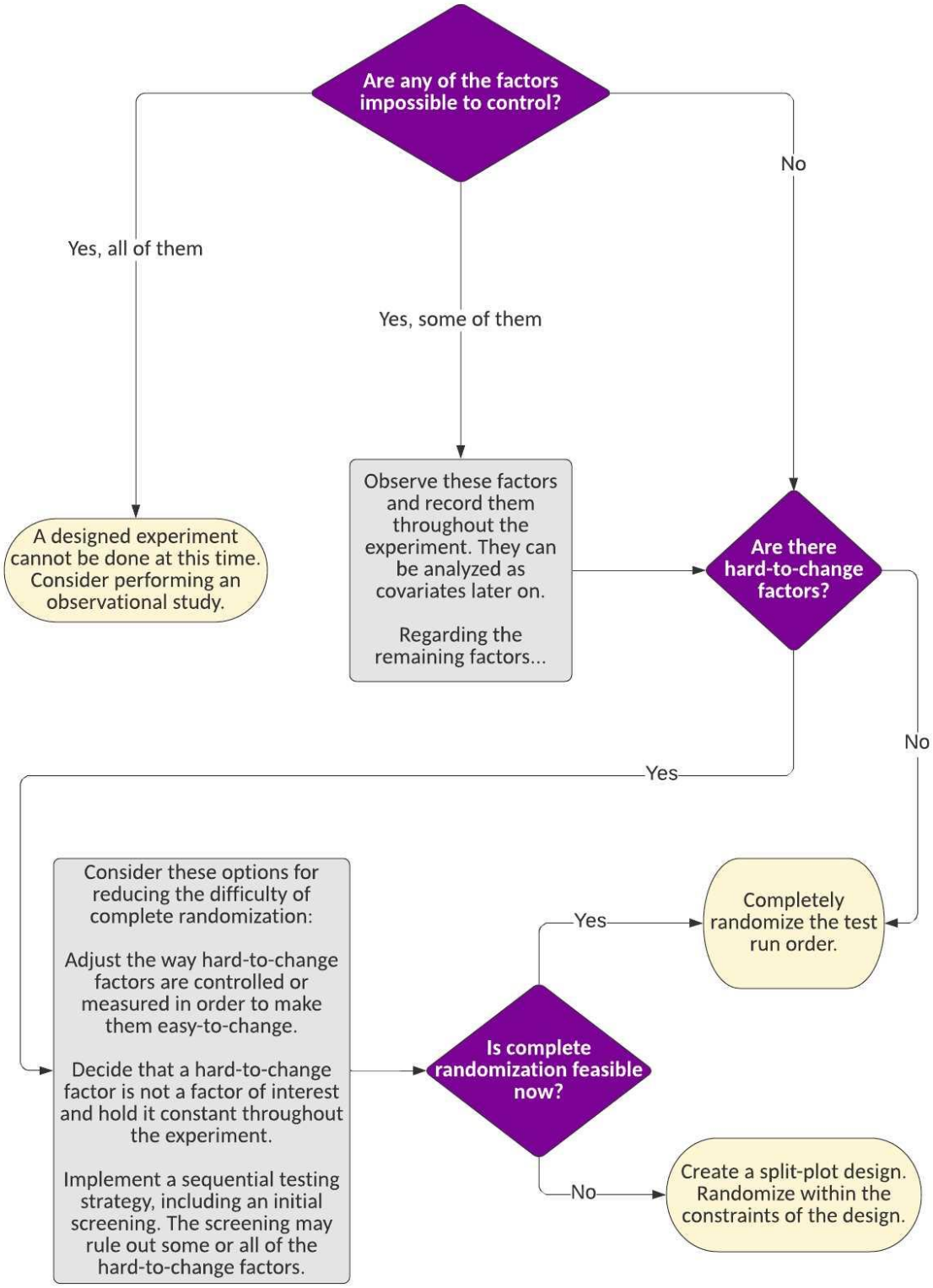


Figure 3. Flowchart of randomization recommendations

Justification to Leadership

Those in DoD leadership positions are likely to expect and advocate for rigorous test methods (Kensler et al., 2015). As a result, they should also expect randomized test designs. This section addresses how to justify the choice to randomize in the case that leadership does not see the merit in that choice; then, it discusses the alternative problem of how to justify the choice *not to randomize* when leadership expects randomization.

The case for randomization has already been made in this paper. A similar (though much briefer) case can be made here:

- The random order of test runs offers protection from the confounding effects of any lurking variables – any effects not accounted for or that are outside of the test team's control. This is especially important because even the most thorough of non-randomized experiments can fall victim to "unknown unknowns."
- Statistical analysis performed on the data from this experiment depends on the assumption that the results of each test run are independent of the results of all other test runs. The best way to meet this assumption is to randomize the order in which the tests are run. Independence allows us to generalize the results of the experiment and make correct conclusions, and therefore, correct decisions.
- If there are replicates, the random order of test runs will allow analysis to determine the amount of noise present in the system, which in turn will provide more information about the success of the current test and guide future testing.

In the case where leadership expects randomization and the test run order has not been randomized, or not been completely randomized, some potential justifications for this follow:

- Some of the factors in this experiment are difficult to change, making complete randomization infeasible. Therefore, we will partially randomize by using a split-plot design.
- Many of the factors in this experiment are difficult to change, making both complete randomization and a split-plot design infeasible. We will be implementing a sequential test design approach, with some of these factors held constant during the first test to allow the rest to be completely or partially randomized. It is likely that most of the factors we are considering will be proven insignificant and can therefore be excluded from future testing; we will reexamine the ability to randomize after we have ruled those factors out.
- Some of the factors in this experiment are impossible to control. We will observe these factors, randomize with respect to the remaining factors, and analyze the uncontrollable factors as covariates after the fact.
- Many of the factors in this experiment are impossible to control. Therefore, we cannot perform an experiment at this time. We will perform an observational study instead. (Note: in

observational studies, unlike experiments, factors are not controlled. However, analysis of an observational study will not allow us to show causative effects of factors on the responses.)

- Randomization, even with a split-plot design, is prohibitively expensive or time-consuming. However, we have compelling evidence (such as subject matter expertise and historical data) that test conditions are sufficiently stable to minimize the risk of non-randomization. (Note: this option is rarely viable and should be used with utmost caution.)

Whenever randomization is not implemented, the issues of lurking variables, "unknown unknowns," independence between test runs, and potentially weaker statistical analysis must be addressed. Consider consulting with a STAT expert for analysis of the consequences of non-randomization in a particular experiment.

Conclusion

The practice of randomization is crucial to DOE because it facilitates statistical analysis, accounts for the bias introduced by "unknown unknowns," and allows conclusions derived from test data to stand up under scrutiny. Test runs should always be completely randomized unless doing so makes the experiment too difficult or expensive to perform. If it does, alternative methods, such as split-plot designs, offer a compromise between rigor and feasibility. Without some degree of randomization, and without a strong reason to believe that the test environment is stable, a valid experiment cannot be performed.

References

Anderson, Mark J. & Whitcomb, Patrick J. *DOE Simplified: Practical Tools for Effective Experimentation.* 3rd ed., Taylor & Francis Group, 2015.

Box, George. "George's Column." *Quality Engineering*, vol. 2, no. 4, 1990, pp. 497-502., doi:10.1080/08982119008962743

Burke, Sarah et al. "Guide to Developing an Effective Test Strategy." Scientific Test and Analysis Techniques Center of Excellence (STAT COE), 31 Dec. 2019.

Harman, Michael. "Lessons Learned from an Incompletely Randomized Test Design." Scientific Test and Analysis Techniques Center of Excellence (STAT COE), 25 Jan. 2018.

Kensler, Jennifer et al. "Critical STAT Questions DOD Leadership Should Ask." Scientific Test and Analysis Techniques Center of Excellence (STAT COE), 18 Mar. 2015.

Minitab. "DOE: Handling Hard-to-Change Factors with Split-Plot Designs in Minitab." www.minitab.com/en-us/Published-Articles/DOE--Handling-Hard-to-Change-Factors-with-Split-Plot-Designs-in-Minitab/. Accesses 16 July 2020.

National Institute of Standards and Technology (NIST). "Consequences of Non-Randomness." *NIST/SEMATECH e-Handbook of Statistical Methods.* *https://www.itl.nist.gov/div898/handbook/eda/section2/eda251.htm*. Accessed 30 Jan. 2020.

Stat Trek. "Statistics Dictionary." stattrek.com/statistics/dictionary.aspx?definition=randomization. Accessed 24 July 2020.

Stone, Brian. "Test Planning for Observational Studies." *DATAWorks 2018.* HQ Air Force Operational Test & Evaluation Center (AFOTEC), 12 Mar. 2018. *dataworks2018.testscience.org/wp-content/uploads/sites/8/2018/03/AFOTEC_Observational-Study-Test-Planning_Stone_Final_12MAR2018.pdf.* Accessed 27 Aug. 2020.