Case Study: Effects of Incomplete Randomization and Test Dependence in Flight Helmet Testing

Gina Sigler
Nick Jones
Dr. Steven Thorsen

June 2020

The objective 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. 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

This case study provides a logical thought process behind challenges to an inadequately planned test in support of flight helmet acquisitions. The Scientific Test and Analysis Techniques Center of Excellence (STAT COE) provided consultation on design and analysis for three tests, all aimed at characterizing the service suitability of prototype flight helmets as compared with helmets currently in use in the Department of Defense (DoD). The program found itself under significant test constraints of limited budget, availability of test articles, and access to test facilities. Such limitations are often dealt with using design of experiments (DOE) approaches which plans designs according to three principles: 1) Randomization, 2) Replication, and 3) local control (blocking). These principles are more thoroughly explained in best practices for test planning found on the STAT COE webpage: https://www.afit.edu/STAT/statdocs.cfm?page=1126 . The STAT COE became a stakeholder in the test only after much planning consideration was already documented and tests were underway. The limits imposed caused the STAT COE to propose a set of minimal designs with high degrees of inter-dependence between runs. Even when optimized, such designs are not ideal and require strict adherence by the test team to 1) the principle of randomization and 2) recording variations to ensure the results can be interpreted correctly. This case study focuses on improving test plans and designs using experiences from an experiment that did not initially adhere to the principle of randomization. The overall lesson learned presented in this case study is that dependencies created by not adhering to DOE principles can't be undone, though some mitigation is possible. Randomization and having a necessary and sufficient number of independent runs is what will properly characterize the system under test.

Background

A DoD program responsible to advise through testing on procurement of new flight helmets contacted the Scientific Test and Analysis Techniques Center of Excellence (STAT COE) to consult on test designs and analysis in two separate instances. The first instance was in regards to testing across multiple prototype helmets to determine if the probability of neck injury was consistent as well as determine the statistical and operational significance of the factor "nape tightness". The prototype helmets allow the chin and nape straps to be tightened independently from one another, and this differs from the legacy helmet which has only a single chin/nape strap. The second instance featured questions on test designs and sizing for sled and impact testing to find if there were statistical and operationally significant differences for specific visor configurations. In this case, the team had a very limited number of test runs and was challenged to maximize information. In both cases, the STAT COE suggested a number of potential test designs based on statistical measures of merit and emphasized the need for randomization and designed experiments to maximize information gains.

Helmet Strap Testing

The objectives of the prototype helmet strap testing were two-fold. The first objective was to determine if the probability of neck injury is consistent across the types of helmets. The second objective was to decide if nape tightness has any impact on head and neck loads. Neck loading was assessed in terms of the magnitudes of forces applied to the neck. The impacts of shear forces, in the X-Y plane of the head as shown in Figure 1 were of particular interest. Injury likelihood was assessed using a set of calculated responses which compare the applied force to a critical force value. Testing occurred on four types of helmets: Legacy, Prototype 1, Prototype 2, and Prototype 3. Based on requirements, the team requested a test matrix that had two acceleration levels, differing levels of nape strap tightness, and differing levels of chin strap tightness. These factors can be seen in Table 1. Some previous test data was available with all four helmet types at the high acceleration level. The value of data from one of the prototype helmets was called into question because it could have been possibly broken during previous testing, and only one helmet of each type was available to test.

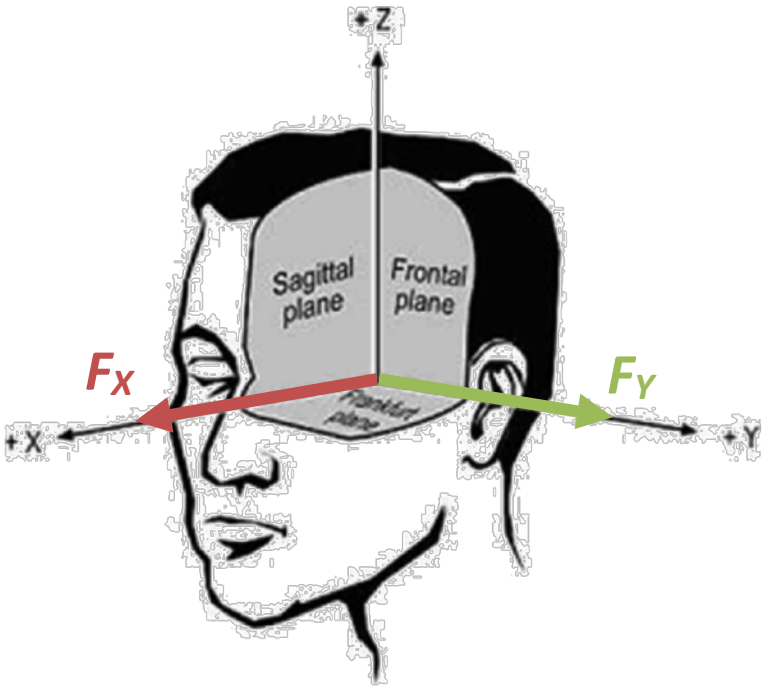

Figure 1: Shear forces in the X-Y plane

Table 1: Factor table for helmet strap testing

Factor	Levels			
Acceleration Level	0	1		
Nape strap tightness	Loose	Medium	Tight	
Chin strap tightness	Loose	Medium	Tight	
Helmet	Legacy	Prototype 1	Prototype 2	Prototype 3

The STAT COE explained the value of using continuous values for both nape and chin strap tightness. However, the test team decided that it was not possible to measure nape strap tightness as a continuous variable, and it would be set to loose, medium, and tight in the designs. The STAT COE proposed a series of designs with differing factors and levels to the team that would provide information on the magnitude of force loading on the neck due to applied forces in different directions, which can be used to assess the likelihood of neck injury.

Design

STAT COE proposed 12 designs, each with different properties. Designs varied from 16 to 20 runs, the inclusion or exclusion of chin strap tightness as a factor, categorization of helmets as legacy versus prototype, and whether or not prototype 3 was included. All designs had power calculated using a signal to noise ratio of 2 and an alpha value of 0.05. The models initially focused on main effects only, but could be adjusted to include specific two-factor interactions of interest. Some designs had also been augmented by previous data, as there were a set of runs conducted in an earlier test that gathered the same information. All presented designs were D-optimal and created using JMP.

The team selected a 16 run design as seen in Table 2 with the factors of acceleration/pulse, helmet, and nape tightness. This particular design didn't use the potentially broken prototype 3 helmet. Since there is not a separate nape strap and chin strap used in the Legacy helmet, chin strap tightness was not included as a factor. The first three runs were previously gathered and would be used to supplement the information from the new test. These points were not intended to be re-run in the new design.

Table 2: Design matrix for 16 run randomized design

Design 9				
16 Run	Accel. (G)	Helmet	Nape Tightness	
1	1	Legacy	Medium	*previously run
2	1	Proto1	Tight	*previously run
3	1	Proto2	Tight	*previously run
4	0	Legacy	Loose	
5	0	Proto2	Medium	
6	1	Proto1	Loose	
7	0	Proto1	Tight	
8	0	Proto2	Loose	
9	1	Proto2	Loose	
10	0	Legacy	Medium	
11	1	Legacy	Tight	
12	0	Proto1	Loose	
13	1	Proto2	Medium	
14	0	Proto1	Medium	
15	0	Proto2	Tight	

16	1	Legacy	Loose	
17	0	Legacy	Tight	
18	0	Legacy	Loose	
19	1	Proto1	Medium	

Execution

During the execution phase, several changes were made to the initial design matrix. It was determined during testing that it was not possible to run the legacy helmet at different nape/chin tightness levels. The legacy helmet was run at a single tightness setting, which was later determined to be at the tight level. Another major change was the run order of the design. The matrix was reordered to account for possible acceleration impacts on durability and the need to set a tight level on each helmet type. Please note that the run order seen in Table 3 doesn't match the run order seen Table 2, the design matrix.

Table 3 shows the test matrix that was executed in the lab. The previous data is not included in this table, but it does show all successfully executed runs. This includes runs that didn't produce useful results for at least one metric. These are denoted by the *** in the Run Order column of the table. Testing was conducted over the course of 4 days, and a total of 22 usable runs were collected.

Table 3: Run order matrix

Run Order	Date Tested	Accel. (G)	Helmet	Nape Tightness
0* **	Day 1	0	Legacy	Tight
1	Day 1	0	Legacy	Tight
2	Day 1	0	Legacy	Tight
3	Day 1	0	Proto1	Tight
4	Day 1	0	Proto1	Medium
5* **	Day 2	1	Proto1	Medium
6	Day 2	0	Proto1	Loose
7	Day 2	1	Proto1	Medium
8	Day 2	0	Proto2	Tight
9	Day 2	0	Proto2	Medium
10	Day 3	0	Proto2	Loose
11	Day 3	1	Proto2	Medium
12	Day 3	1	Legacy	Tight
13	Day 3	1	Proto1	Tight
14	Day 3	1	Proto2	Tight
15	Day 3	1	Proto1	Loose
16	Day 3	1	Proto2	Loose
17	Day 3	1	Legacy	Tight
18	Day 4	0	Proto1	Loose
19	Day 4	1	Proto1	Medium
20	Day 4	1	Proto1	Tight

21	Day 4	1	Proto2	Medium
22	Day 4	1	Proto2	Tight
23	Day 4	1	Proto1	Loose

Analysis

As the number of runs in the "as run" test matrix exceeded the number of runs in the original test matrix, the two-factor interactions of acceleration and helmet as well as acceleration and nape tightness were added to the planned model. Unfortunately, it's still not possible to estimate the interaction between helmet and tightness level due to the categorical factors and the disallowed combination of tightness levels and the legacy helmet. The analysis began by conducting exploratory data analysis on the collected data. Figures 2 & 3 show some preliminary graphs of the data. Figure 2 plots shear force magnitude on an arbitrary scale against all of the other factors. The injury likelihood responses have been omitted for simplicity, and because they are derived from the measured shear forces and similarly affected by the factors. The main pattern that emerges is that there are lower shear values for those set at the lower acceleration levels than at the higher levels. This is an expected result. However, there do not appear to be any clear patterns between shear and either helmet or nape tightness. This plot also highlights the lack of balance between points run at the low and high acceleration levels. Finally, the plot shows a possible lurking variable. For the repeated points of the Prototype 1 and Prototype 2 helmets run on Days 3 and 4, the runs on Day 3 have responses consistently higher than the runs on Day 4. Figure 3 shows the repeat runs as colored plus signs. Again, there is a pattern of higher values of shear on Day 3 compared to Day 4. This finding can be confirmed with further analysis.

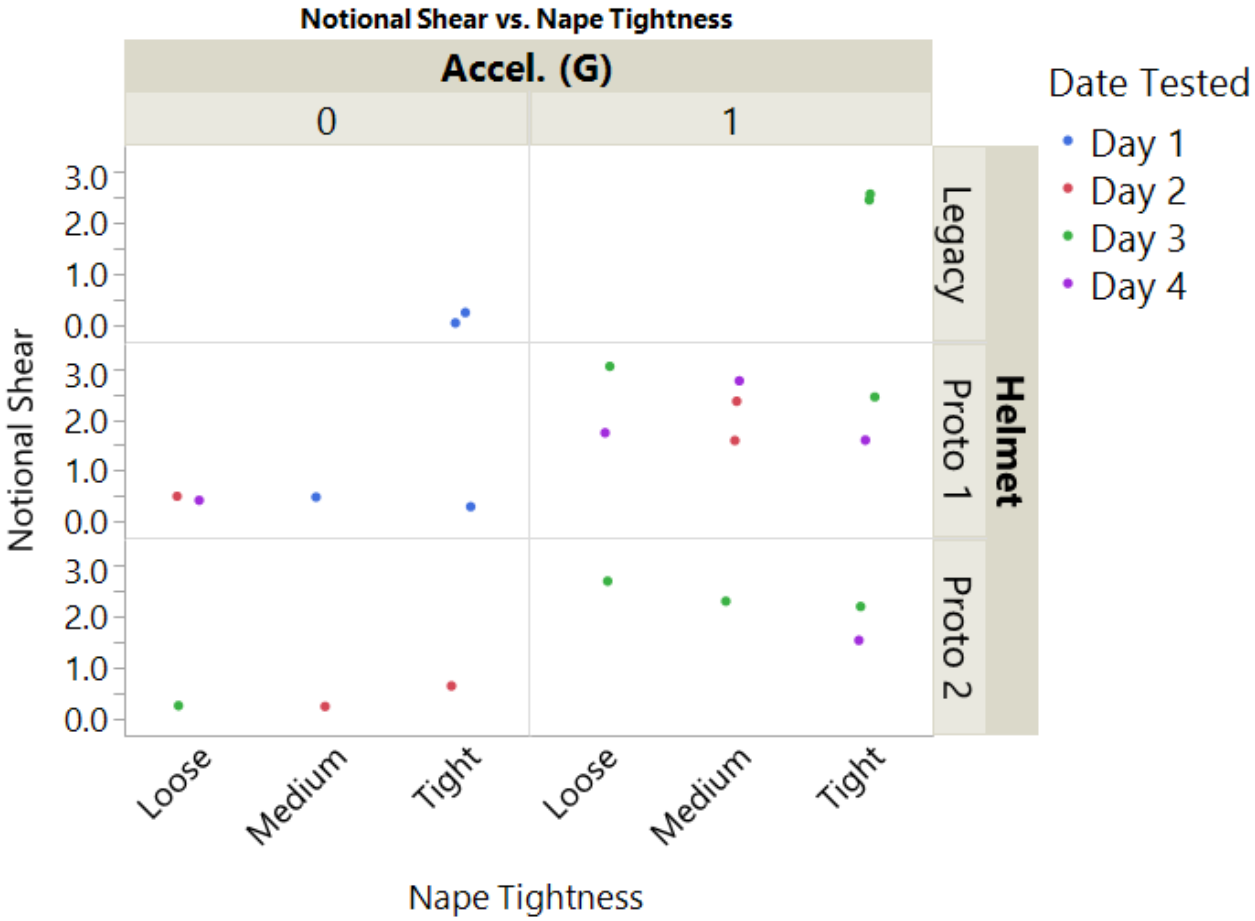

Figure 2: Shear vs. all factors (nape tightness, helmet, & acceleration)

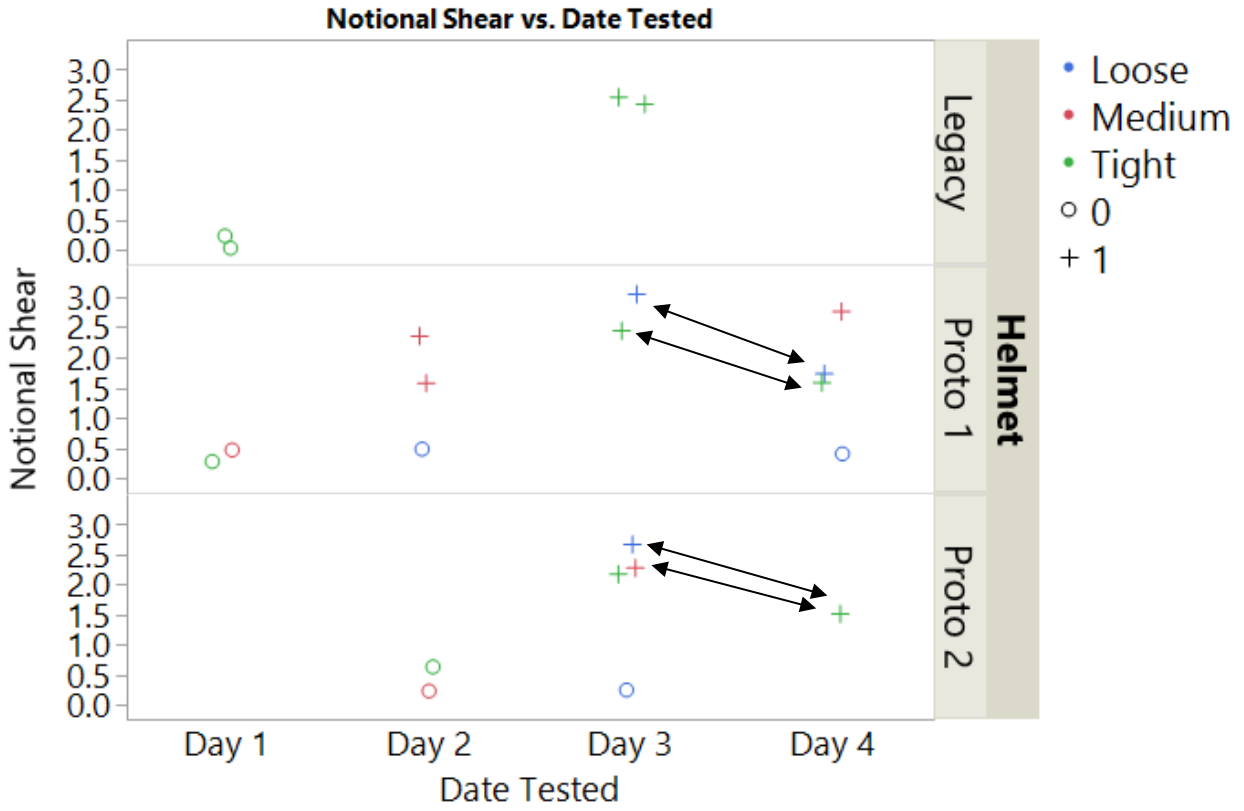

Figure 3: Shear vs. date broken up by helmet, nape tightness, and acceleration

To characterize helmet performance, generalized linear regression models were created using two versions of the data set. The first data set excludes run 0, which was a setup run and not part of the designed experiment. The second data set excludes run 0 and run 21, which are statistically different than the rest of the data points (using Cook's Distance). After discussions with test execution personnel, it's unclear whether these were trusted data points, or whether or not the helmet itself could be broken from being run at the high acceleration levels. In all cases, the models showed that lower values of acceleration corresponded to lower values of the shear response. All of the models should be very informative if using acceleration to predict any of the responses, however the effects of all other variables are overshadowed by the acceleration impact. For certain responses, helmet and nape tightness were shown to be statistically significant, but no practically significant when compared to the impact of acceleration. For other responses, all other terms except acceleration were dropped from the model.

One of the initial objectives of testing was to see if any of the helmets could be considered different from another. Pairwise t-tests were conducted to examine if there were any differences between the types of helmets or different tightness levels. In both cases, none of the categories appeared to be statistically different from one another for any of the responses. This could be due to the very large impact that acceleration has on each of the response variables, which would inflate the standard errors. Or, it could be that there is truly no difference between helmets and tightness levels.

Graphs and models show a strong relationship between the factor of acceleration and the measured response in shear forces. This was expected by the test team. However, the experiment showed larger than expected noise levels, which makes the power of the test much lower than originally anticipated. The larger than expected deviation would require a design with more samples. Additionally, the presence of a lurking variable can be seen across the days and introduces bias to the results. The design was randomized in order to avoid a potential problem such as this one, but the executed test didn't take this into account. These challenges combined with the few runs performed on categorical variables leads to inconclusive results in regards to the impact of nape tightness and helmet type.

Due to the inconclusive results of this test, follow-on testing is recommended. If at all possible, nape tightness should be measured as a quantitative response. This could be anything from measuring the length of the strap at each setting, measuring how far a click pulls the strap in, or the amount of force caused by the nape tightness. In addition, this might allow for the interaction between nape tightness and helmet type to be measured. Also, the STAT COE would suggest fully randomizing the runs to protect against lurking variables and hidden bias, or understand the constraints of testing and design an appropriate split-plot design.

Sled and Impact Testing

Following the completion of the helmet strap testing, the team contacted the STAT COE to discuss more efficient test designs for both sled and impact testing. The objectives of the sled test were to characterize the effects of speed, visor configuration, seat angle/type, and body type on a series of responses. The objective of the impact test was to characterize the effects of temperature, size, and impact location on another series of continuous responses. Again, the STAT COE proposed a series of designs with differing factors and levels that would allow for different budgets and design metrics. In addition to their own matrices, the STAT COE also looked at design metrics for some designs proposed by some of the engineers. While most of these designs captured the important factors, they had extra constraints and often lacked the randomization necessary to properly analyze the factors after testing. Due to lab and leadership hesitation to use randomized runs, the STAT COE took on the task of creating a solid explanation for the team as to why one particular design was better than the others proposed.

Sled Testing

For the sled testing effort, the STAT COE chose the design in Table 4 as this was the most efficient test in terms of the numbers of runs and the most effective at answering the objectives of the test. The effectiveness of the test was based on the calculated metrics seen in Table 5. These included signal-to-noise ratios of 2, type I error rates (alpha) of 0.1, total number of runs, assorted power values for each factor, fraction of design space (FDS) prediction error values, and aliasing. For more details on these metrics and their definitions see the STAT COE's best practice *Test Design Comparison and Selection*. For ease, the metrics have been color coded from red to green. Red indicates a metric has an undesirable value, and green indicates a desirable value.

Table 4: Suggested design for sled testing

Sled Design						
10 Test Runs	Sled Run	Manikin	Speed	Visor Config	Seat Angle	Seat Position
1	1	Large	Low	Down	High	Forward
2	1	Small	Low	Up	High	Aft
3	2	Small	High	Down	Low	Forward
4	2	Large	High	Up	Low	Aft
5	3	Large	Low	Up	Low	Forward
6	3	Small	Low	Down	Low	Aft
7	4	Small	High	Up	High	Forward
8	4	Large	High	Down	High	Aft
9	5	Small	Low	Down	Low	Forward
10	5	Large	Low	Up	Low	Aft

Table 5: Calculated metrics for selected designs

Design #	**1**	**2**	**4**	**6**
Software Package	JMP	JMP	JMP	JMP
Signal to Noise Ratio	2.0	2.0	2.0	2.0
Alpha	0.1	0.1	0.1	0.1
Total Runs	14	12	10	8
Power for Speed @ S/N	0.539	0.428	0.33	0.207
Power for Visor Config @ S/N	0.961	0.928	0.842	0.694
Power for Manikin @ S/N	0.961	0.928	0.842	0.694
Power for Seat Angle @ S/N	0.539	0.428	0.33	0.207
FDS Pred Err @50%	0.45	0.52	0.65	0.79
FDS Pred Err @95%	0.70	0.80	1.00	1.20
Aliasing	low	low	low	medium

The specifics of the design recommendations stem from the details of the test matrix. The matrix begins with a slow run that can be used to check for initial set-up issues. This design also attempts to minimize seat angle changes as this may take extreme time or effort to change multiple times. The STAT COE has stressed however, that this would have to be switched at least twice in order to see the differences between seat angles in the analysis. Finally, the STAT COE attempted to emphasize the ability of this design to determine a "bad run."

It's important to have assurance that the data gathered is representative of the truth. The variation in this design allows for identifying the cause of a bad run. For instance, say the helmet in run 4 is ineffective for some reason. The analysis would be able to conclude that it was likely not the manikin (which was used in run 1), not the speed (since run 7 went fine), not the angle (because runs 5 and 10

were also good), and not the sled run (as the behavior of the helmet in run 3 was expected). If this variation doesn't exist, these types of conclusions cannot be made.

In comparison to other designs, any un-randomized design will increase the risk that some unseen variable will impact the data (as seen in the prototype helmet testing with day). The risk here is that with so few runs, the variable would likely remain unknown. Risk of loss of information is another issue with such designs. At this point, un-randomized data should be considered compromised data. The designs proposed by the STAT COE have been optimized to have the best possible power (for each factor), lowest prediction variances, and lowest aliasing. While other proposed designs might have greater sensitivity to one particular area, say higher power for speed, this doesn't mean they are better designs overall. By definition, these optimized designs are "best" for the specified criterion.

The STAT COE was also asked to create designs that included a constraint of no runs at high speed, with the visor down. Since this was not an issue of safety, the STAT COE pushed for a design matrix with no constraints. If runs are only done at the low speed with the visor down, there can be no information gleaned of what performance would be at other speeds. The unbalanced design cannot separate the effect of visor configuration from that of manikin size, speed, or seat angle. By adding the additional runs, it would become possible to discuss what might happen at speeds in between high and low. While it may seem counterintuitive to run these, the amount of information lost in terms of power, prediction, and separation of effects makes these runs statistically relevant (if not operationally relevant). A constraint on testing should only be considered when runs are either unsafe or operationally impossible. In this case, STAT COE recommends that a randomized balanced design be used to maximize information with the most efficient test.

Impact Testing

STAT COE also proposed a series of designs for impact testing that involved changing the temperature, hit location, and size of helmets. Table 6 shows the suggested test matrix for impact testing. There was a question as to whether or not multiple drops in different hit locations on the same helmet could be considered independent. Based on previous evidence, the STAT COE has concluded that independent drops on a single helmet was not a safe assumption. Instead, the COE suggested conducting no less than 30 independent drops. Additional drops would allow further analysis of independence.

Table 6: Impact test matrix with 30 runs

Run	Temperature	Hit Location	Size	Run	Temperature	Hit Location	Size
1	Low	Crown	Medium	**16**	High	Rear	Large
2	High	Right	Medium	**17**	Low	Right	Extra Large
3	Mid	Front	Medium	**18**	Low	Rear	Medium
4	Mid	Crown	Extra Large	**19**	Mid	Front	Medium
5	High	Rear	Large	**20**	Mid	Right	Small
6	High	Front	Extra Large	**21**	Mid	Right	Large
7	Mid	Crown	Large	**22**	High	Right	Extra Large
8	Mid	Rear	Small	**23**	Mid	Left	Extra Large
9	High	Crown	Large	**24**	Low	Left	Large
10	Low	Crown	Small	**25**	Low	Left	Small
11	High	Front	Small	**26**	High	Crown	Small
12	High	Left	Medium	**27**	Low	Rear	Extra Large
13	Mid	Left	Small	**28**	Mid	Rear	Medium
14	High	Left	Extra Large	**29**	Low	Front	Large
15	Low	Front	Extra Large	**30**	Low	Right	Large

Again, from a randomization prospective, it's better to test a larger number of helmets with a randomized order. For a simple example as to the impact the number of helmets has, assume that each helmet can be dropped 3 times without loss of independence. The case could then be made to drop 10 helmets instead of the original 30. This creates a split plot design, which automatically violates the assumption of independence. Helmet size and temperature were now considered hard-to-change factors, and the power for each of these factors will decrease from the original design due to this lack of randomization. In addition, there will not be enough runs to see the variability across helmets. As these two factors are of high interest, it's not recommended to run this as a split-plot design.

As another example of why randomization and number of helmets are important, consider the hypothetical scenario with previous test data. For simplicity, the amount of needed information is shrunk to four different combinations: high temperature with a small helmet, high temperature with a large helmet, low temperature with a small helmet, and low temperature with a small helmet. Based on previous data, varied responses for repetitions of the same combinations are grouped together for ease of interpretability in Table 7. The table shows that the combination of low and small have a much larger variability in the responses than the other combinations. This indicates a potential variability of like helmets (helmets of the same size being impacted under the same conditions), meaning that responses may not be similar each time the experiment is performed.

Table 7: Historical hypothetical response data

Temperature	Helmet Size	Response
High	Large	14.0
High	Large	13.0
High	Large	12.8
High	Small	12.9
High	Small	11.7
High	Small	12.4
Low	Large	14.2
Low	Large	14.7
Low	Large	15.1
Low	Small	17.2
Low	Small	15.3
Low	Small	12.1

If this historical data is extended back to the original example, an argument can be made as to why testing 30 helmets is better than testing 10. Assume that the requirement states that the response must be below 15. In this case, there are 12 combinations of conditions to run (3 temperatures, and 4 helmet sizes). With only 10 helmets, the same condition with multiple helmets could not be run. This means there will only be one result for a small helmet at a low temperature. This leaves no context to determine if the measured response of the one helmet is at the high or low end of the response value range. There is no way of knowing if the data has been properly represented. The greater the variability of the measurements, the more difficult it becomes to decide how representative the data is of the truth.

The proposed DOE design allows for some hidden replication, but only if all of the runs are independent. The split plot violates this independence. It also still assumes that the drops on a single helmet can be considered independent of one another. If this is violated there will not be enough runs to draw useful conclusions. If multiple drops are done on each helmet, it's necessary to block for the variation of each helmet in the analysis. No matter what is done, the analysis will likely be complex due to the nature of the multiple blocks from day or other nuisance factors. Violating the independence of the runs can lead to unusable data. STAT COE recommends that every possible effort be made to ensure runs remain randomized and independent.

Conclusion

When designing a set of tests to characterize a system, rigorous planning of the experimental design and test execution are important to ensure an accurate characterization and to make efficient use of scarce test resources. To ensure such designs are effective, those executing the test must understand and adhere to the test procedure as planned. This is particularly true when executing minimal test plans to economize limited test resources and run budgets, since deviations in such tests may make it impossible

to powerfully attribute all observed effects to their causes. The importance planning and executing the DOE principle of run randomization to the fullest extent practical is easily seen in this case study. It's equally important to properly analyzing the test results.