

Design of Experiments for Ruggedness Testing: Case Study on Paraffin Therapy Bath

Mark J. Anderson, Paul J. Anderson

Design of experiments (DOE) has become an essential tool for validation of medical manufacturing processes. Kim and Kalb provided a good overview of the techniques¹. They say “the process should be challenged to discover how outputs change as variables fluctuate within allowable limits.” This article shows how we challenged our durable medical device – a paraffin heat-therapy bath. We first ran a highly-fractionated 2-level factorial design. This minimal-run DOE showed a significant degree of sensitivity amongst a panel of users. Therefore, we then ran a follow-up design called a “foldover” to find out what really caused the effects. By combining data from both phases of the DOE we revealed some surprising interactions. This ultimately led to redesign of the product for better performance at lower cost.

Background

The Therabath® paraffin therapy bath² (pictured below) is a durable medical device that holds one gallon of molten paraffin wax. Sufferers of osteoarthritis use it for physical therapy. They dip their hands repeatedly in the heated bath, which helps loosen their joints. The wax then slowly solidifies as a glove, producing further therapeutic benefits via the heat of fusion. Oils reduce the overall melt point to a comfortable level, facilitate removal of the glove and provide moisturizing for skin. To enhance the perceived benefit to skin, vitamin E and various scents and colors were added for this application.



Paraffin Therapy Bath (Test Unit)

DESIGNING THE RUGGEDNESS TEST

The experiment involved varying the following 6 factors at low and high levels:

- A. Ratio of two component waxes (W_1 to W_2)
- B. Ratio of total wax to oil
- C. Supplier of wax
- D. Amount of dye
- E. Amount of perfume
- F. Amount of vitamin E

The amounts of vitamin E, dye and perfume are very small in relation to the wax and oil.

For response measures, an “expert” panel of 10 employees provided sensory evaluation of color, scent, heating, oiliness and quality of the wax glove. They rated the paraffin on a hedonic scale from 1 ☹ (worst) to 5 😐 (so-so) to 9 ☺ (best). The results were analyzed by individual (block) and then averaged, thus providing a fairly powerful tool for discriminating changes in performance of the device.

Ideally, the variations in factor levels we tested would not affect the response. This would prove that the system is rugged, thus fulfilling the purpose of validation. On the other hand, a significant outcome would require further work to fine-tune the system.

The full factorial design for 6 factors requires 64 runs (2^6). To keep the testing at a manageable level, we chose a $1/8^{\text{th}}$ fraction of all the combinations (2^{6-3}) for a total of only 8 runs. Cutting the design back to so few runs causes main effects to be aliased with two-factor interactions (see Table 1 below).

$\begin{aligned} [\text{Intercept}] &= \text{Intercept} + \text{ABD} + \text{ACE} + \text{BCF} + \text{DEF} \\ [\text{A}] &= \text{A} + \text{BD} + \text{CE} + \text{BEF} + \text{CDF} \\ [\text{B}] &= \text{B} + \text{AD} + \text{CF} + \text{AEF} + \text{CDE} \\ [\text{C}] &= \text{C} + \text{AE} + \text{BF} + \text{ADF} + \text{BDE} \\ [\text{D}] &= \text{D} + \text{AB} + \text{EF} + \text{ACF} + \text{BCE} \\ [\text{E}] &= \text{E} + \text{AC} + \text{DF} + \text{ABF} + \text{BCD} \\ [\text{F}] &= \text{F} + \text{BC} + \text{DE} + \text{ABE} + \text{ACD} \\ [\text{AF}] &= \text{AF} + \text{BE} + \text{CD} + \text{ABC} + \text{ADE} + \text{BDF} + \text{CEF} \end{aligned}$
--

Table 1: Alias Structure for Ruggedness Test

An alias occurs when inputs are perfectly correlated. Then you cannot distinguish one from the other. For example, the effect of factor A cannot be separated from those of interactions BD, CE or two other higher-order interactions. However, the aliases become moot if nothing is significant – the hoped-for (rugged) outcome. If any factors do appear significant, the low resolution generally precludes definitive conclusions, so you will then need to do further experimentation to pin down the cause(s) for failure.

RESULTS FROM INITIAL STUDY

Analysis of the DOE with a commercially-available statistics package³ revealed significant impacts on perceptions of users. Therefore, the device did not pass the ruggedness test. For example, Figure 1 shows a half-normal probability plot for color. Half-normal plots show the absolute value of the effect on the x-axis. The effects are shown as square points. Estimates of error are displayed as triangles. The biggest effects, those to the right, are most likely to be real. However, many of the estimated effects may occur due to chance. These will be grouped near the zero effect level. The y-axis is constructed to be linear in the normal scale, so the near-zero (insignificant) effects fall on the line emanating from the origin (0,0). One effect stands out: D – the amount of dye.

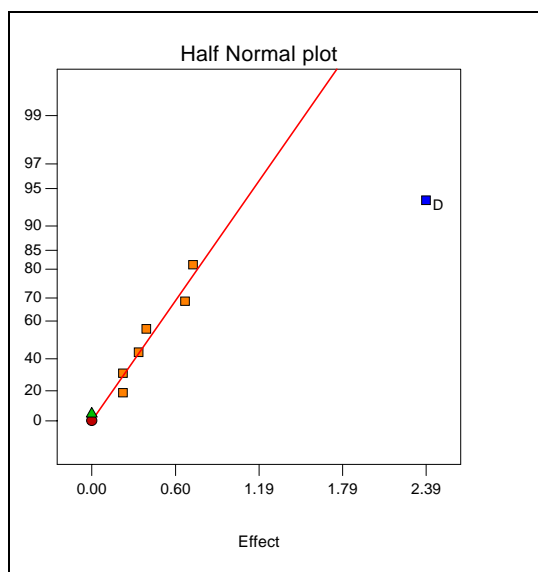


Figure 1: Probability Plot of Color

Standard statistical procedures for analysis of variance (ANOVA) reveal a probability of less than 0.1% that this big of an effect could have been caused by chance. According to the alias table, this highly significant effect could also be caused by interactions AB and/or EF. (The graph arbitrarily displays the main effect (D), since this is most likely.) However, we made the assumption that color would be affected only by amount of dye (D). The panel preferred higher levels dye/color.

For scent, the biggest effect was factor E – the amount of perfume (see Figure 2), but it did not stand out as clearly as color. However, an outlier was detected in run (bath) #1 (see Figure 3). The y-axis on this chart shows the “T”value – a statistic that indicates how many standard deviations a result differs from expected.

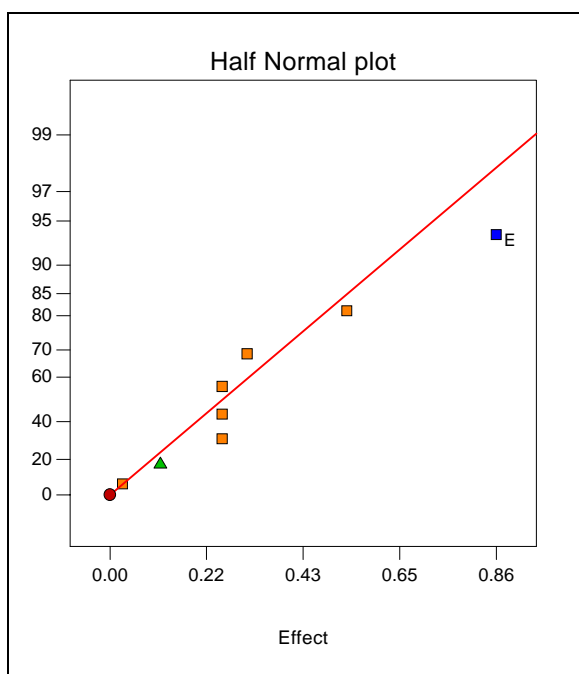


Figure 2: Probability Plot of Scent (All data)

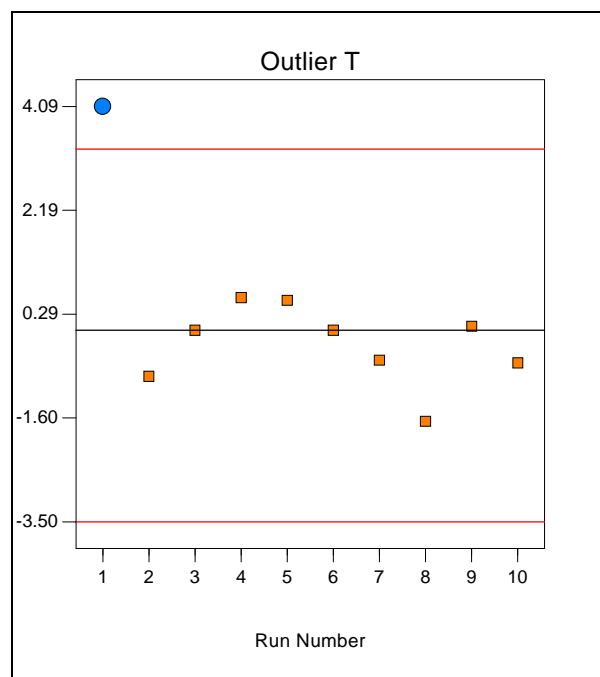


Figure 3: Outlier Detected (Run #1)

Anything outside of plus or minus 3.5 standard deviations can be considered as a possible outlier, since it's unlikely to have occurred due to chance. Upon further investigation, it was found that the temperature of this bath ran significantly high, thus generating more than the expected amount of scent.

After removing the outlier, the perfume (E) stood out even more clearly as the most likely cause of the effect on scent. ANOVA shows the probability of this happening by chance to be less than 1 percent. Again it seems reasonable to make a leap of faith that factor E is the cause and not its aliased interactions, in this case BC and DE. The panel preferred higher levels of perfume for the attribute of scent.

The statistical analysis revealed nothing significant for perception of heat (see Figure 4).

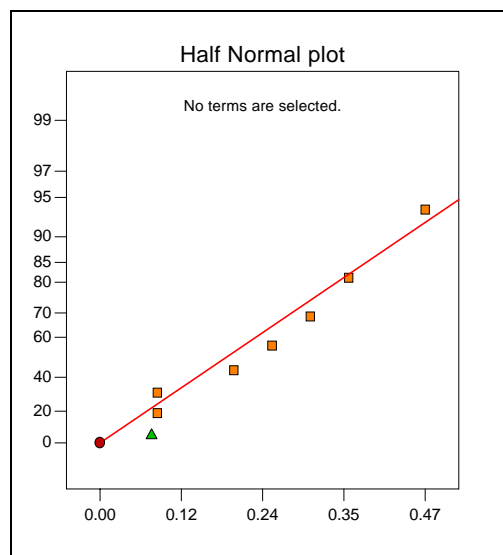


Figure 4: No Significant Effects for Perception of Heat

Perceptions of oiliness were significantly affected (Figure 5), but the aliasing of main effects with interactions in the resolution III design made it impossible to draw any definite conclusions. For example, it makes no sense that factor E, the dye, would affect oiliness. This factor must really be one of the aliased effects: AC and/or DF. Also, the other significant effect, AF, could be BE and/or CD. It's very confusing. Unlike the results for color and scent, there was no obvious explanation for the effects, especially considering the possibility of aliasing. At this stage, we exhausted the capability of the low-resolution design. We needed more experimentation to uncover the true causes for the failure of the ruggedness test.

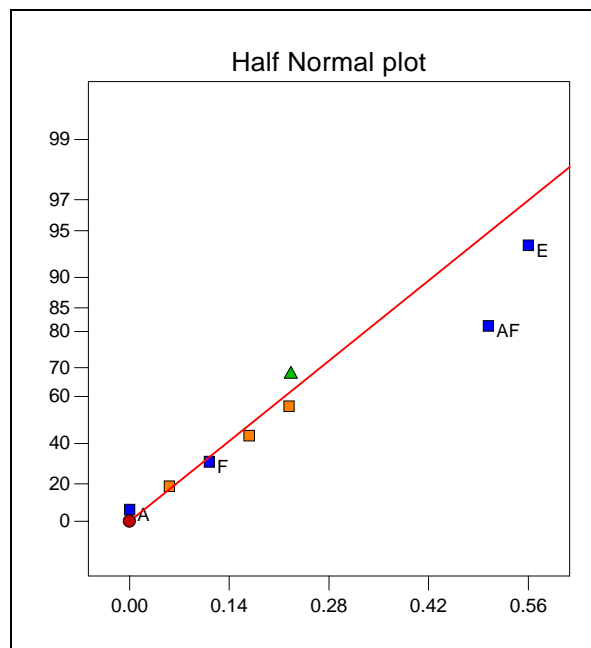


Figure 5: Plot of Effects for Oiliness

FOLLOW-UP EXPERIMENT REVEALS TRUE CAUSES FOR VALIDATION FAILURE

By adding a second block of experiments with all levels reversed, you can eliminate aliasing of main effects with 2-factor interactions.⁴ This is called a “foldover”. The combined results normally remain somewhat aliased. However, before doing the foldover, we eliminated dye (D) and perfume (E) as factors - on the assumption that they affected only color and scent respectively. We set dye and perfume at their mid-point level, and dropped color and scent from further consideration. The addition of the 8 foldover runs then resulted in a full (no aliases) 16-run factorial for the remaining four factors.

Analysis of the combined data continued to show no significant impact on perceptions of heat. This is an important finding. Prior to doing the DOE, the manufacturing people were concerned that users would be sensitive to variations in melt point caused by changes in ratios of wax and oil. For this attribute the process passed the challenge of validation: it’s robust to expected variations.

In regard to perception of the glove, the first experiment seemed to indicate some effects (graph not shown), but after reviewing data for the entire series of runs, including the foldover, it’s now believed that none of the factors affected user perceptions (see Figure 6). Therefore, this is another attribute that passed the ruggedness test.

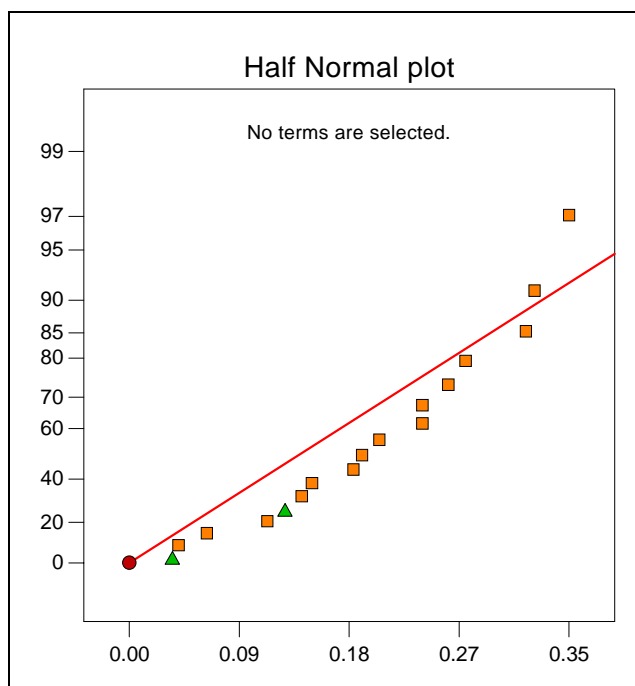


Figure 6: Plot of Effects for Glove (combined results)

The final results on perception of oiliness (see Figure 7) indicate dependence on the combination of three factors: ratio of W_1 wax to W_2 wax (A), higher ratio of total wax to oil (B) and amount of vitamin E (now labeled D).

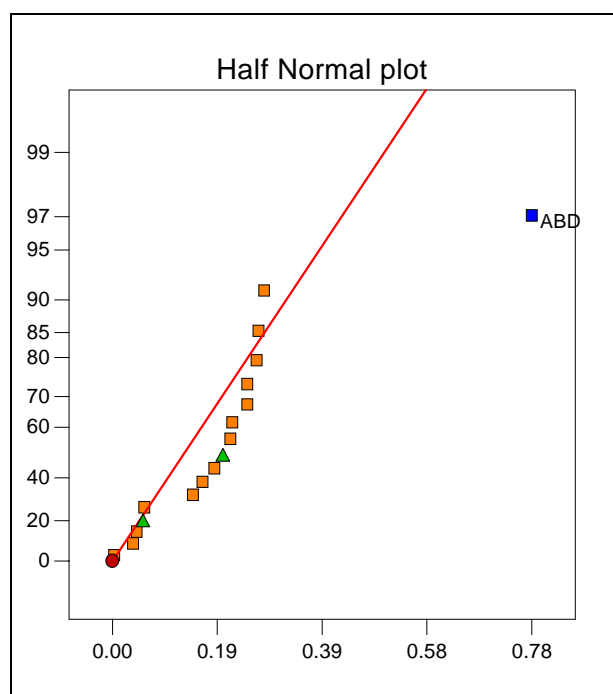
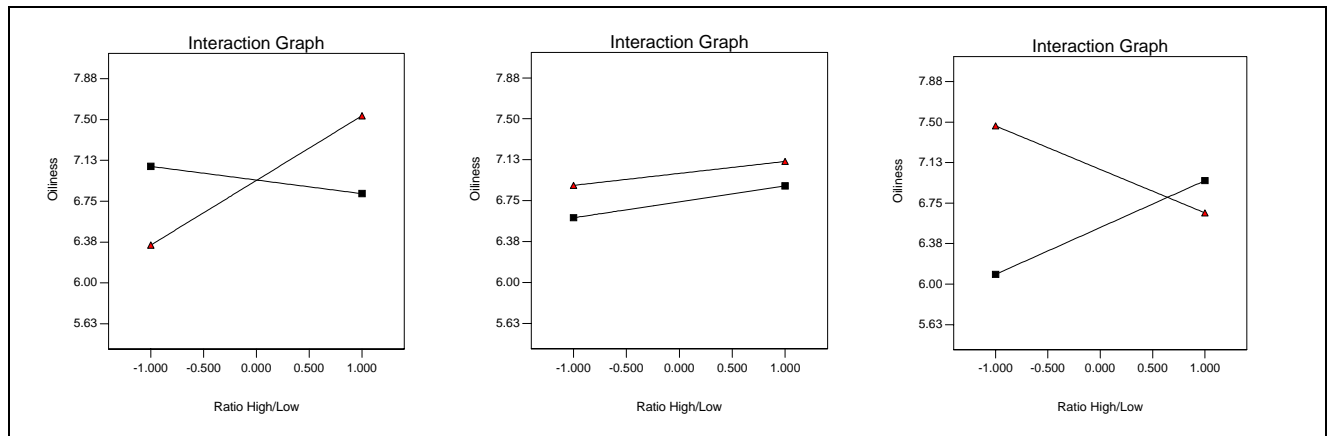


Figure 7: Plot of Effects for Oiliness (combined data)

Three-factor interactions such as this are very unusual, but more likely in experiments that involve mixtures. The series of interaction graphs shown on Figures 8 a, b and c show the complex behavior governing perception of oiliness.



Figures 8a, b and c: Interaction AB at Low level of Vitamin E, Middle level of Vitamin E, Highest level of Vitamin E (Triangles are positive levels of B (wax to oil), squares are negative levels)

To pick the winning combination of the three factors (highest rated), it's easier to work with a cube plot (Figure 9).

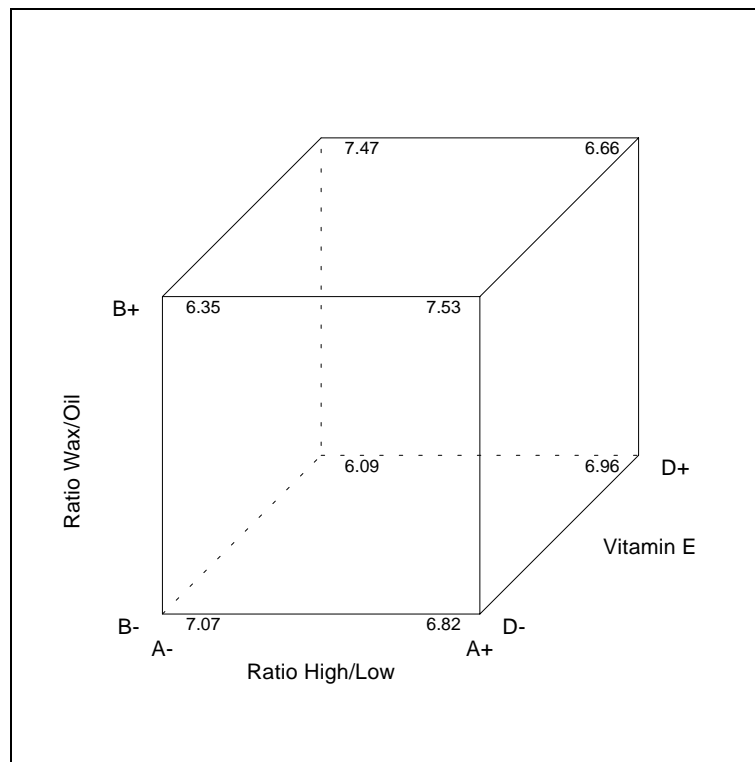


Figure 9: Cube Plot Shows Best Combination (upper right front) for 3 Factors Affecting Oiliness

CONCLUSION

Based on the results from the two-step DOE, we recommended some changes to the product:

- Cheapest supply of raw material wax - factor C, which did not significantly affect any of the tested perceptions.
- Add more color and scent, which may also mask variability of native colors and scents.
- Reduce the vitamin E and increase the ratio of W₁ wax to W₂ wax and the ratio of wax to oil.

The end result is a better and cheaper paraffin blend.

This application provides a good example of how to apply the power of two-level factorial DOE to validation testing. It demonstrates the flexibility of the approach should the validation fail. In this situation, the use of foldover runs provides a profound knowledge of how variations in factors can affect your process or product.

DOE is just one of the statistical tools used in validation. It challenges the system and identifies which factors to control. But this is not the end. Other tools, such as statistical process control (SPC) must be employed to show that the system can produce consistent outputs over time, and meet specifications with a high level of confidence and reliability.

ACKNOWLEDGEMENTS

Dave Sletten of WR Medical did all the experimental work. Patrick Whitcomb of Stat-Ease provided valuable advice on the setup and analysis of the DOEs.

REFERENCES

1. Kim JS, Kalb JW, *Design of Experiments: An Overview and Application Example*, MDDI, March 1996, pp. 78-88.
2. WR Medical Electronics Company, 123 N. 2nd St., Stillwater, MN 55082
3. Design-Ease® software, Version 5 for Windows, Stat-Ease, Inc., Minneapolis (\$395).
4. Montgomery DC, *Design and Analysis of Experiments*, 4th ed, Wiley, New York, 1997, p. 413.

Mark J. Anderson is a principal of Stat-Ease, Inc. and WR Medical Electronics Company. Paul J. Anderson is Vice-President of R&D and a principal of WR Medical Electronics Company.