Statistical analysis of heat loss data for traditional versus particulate-blocking firefighting hoods

Researcher:

Madilynn Smith
Fiber and Polymer Science
Thermal Protection and Comfort

Student Consultants:

Adam Weimerskirch: Introduction, Methods: Calculating sample size, Results: Power and sample size
Clayton Ramsey: Methods: testing assumptions, Methods: testing fixed and random effects, Results: initial study results, Summary

Introduction

Background

Firefighters rely on several types of personal protective equipment (PPE) to help mitigate health and safety risks associated with their profession. One important component of their PPE is their hood, which is carefully designed to protect against burns, harmful particles, and the elements, all while limiting risk of heat illness and improving comfort for the wearer by allowing excess body heat to escape. Heat loss performance of the hoods is measured as predicted total heat loss. Higher values are desirable in this context, as higher heat loss indicates that the hood will more effectively allow excess heat to escape. One component of this research focuses on evaluating heat loss performance of hoods made of two different materials: a traditional knit material made from 100% Nomex and a new material with an improved Gore particulate-blocking layer.

The research question of primary interest was whether the hoods made from the new material had similar heat loss performance as the hoods made from the traditional material. Prior experience of the subject matter experts engaged in this research indicate that differences of $\pm 10\ W/m^2$ are of practical significance; this difference is large enough to make a noticeable

difference in comfort for the firefighters wearing the hoods. While better heat loss performance for the hoods of the new material would be ideal, equal performance is also a desirable outcome because the primary purpose of the new material is to improve particle blocking performance. Additional research questions of interest were whether multiple hoods with no obvious differences (of the same material, design, manufacturing process, etc.) had consistent heat loss performance, and what sample size was recommended to assess heat loss performance of different materials in future studies. Each of these questions are addressed in this paper.

Study Design

A traditional approach to assessing the difference in heat loss between materials is to test flat swatches of each type of material three times for each sample. This method is highly standardized and repeatable, but does not closely reflect how the materials are actually used and leaves open the possibility that the materials will perform differently in specific applications. In this case, the researchers are interested in how the materials perform in a specific application - firefighting hoods, so they opted for a new test method that more closely replicates real-world conditions by placing full hoods on a “sweating manikin headform”, shown below.

Figure 1: Sweating manikin headform, dressed in a Nomex hood for testing

The headform was dressed in the hood and placed in a room with an air temperature of $T_a = 20°C$ and is heated to $T_s = 35°C$ while salt water was pumped through nine pores to simulate sweat. Thermal resistance R_t and evaporative resistance R_{et^A} were measured directly and used to calculate predicted total heat loss Q_t using the equation below.

$$Q_t = \frac{T_s - T_a}{R_t} + \frac{3.57kPa}{R_{et^A}} = C + E$$

$Q_t =$ *predicted total heat loss,* W/m^2
$T_s =$ *temperature at the manikin surface,* ℃
$T_a =$ *temperature of the air surrounding the manikin,* ℃

R_t = *total thermal resistance of the test ensemble and surface air layer*, $°C \times m^2/W$

R_{et^A} = *total evaporative resistance of the test ensemble and surface air layer*, $kPa \times m^2/W$

C = *predicted conductive heat loss*, W/m^2

E = *predicted evaporative heat loss*, W/m^2

Using this test method, the researchers conducted a study to assess the heat loss performance of the two materials. Five hoods of each material were tested three times each. For each test, the manikin was undressed and re-dressed with the hood, and test order was randomized all hoods. This study design corresponds to the nested mixed effects model depicted below. It includes a fixed effect for material $(i = 1, 2)$, a random effect for hood nested within material $(j = 1, 2, 3, 4, 5)$, and three replicates for each hood $(k = 1, 2, 3)$.

$$Y = \alpha_i + B_{j(i)} + \varepsilon_{ijk}$$

Y = *predicted heat loss*, $Q\ (W/m^2)$

α_i = *material type*, $i = 1, 2$

$B_{j(i)}$ = *hood*, $j = 1, 2, 3, 4, 5$

ε = *error term*, $k = 1, 2, 3$

Given this study design and model choice, there are a few statistical questions that can help assess each research question. They are:

1. Is the fixed effect associated with material type significant at a level of $\alpha = 0.05$? Does the difference in mean predicted heat loss between the two materials exceed the practically significant difference of $\pm 10\ W/m^2$?
2. Is the random effect associated with hood number significant?
3. What sample size is required in future studies to detect a difference of $\pm 10\ W/m^2$ with a significance level of $\alpha = 0.05$ and a power level of $1 - \beta = 0.8$?

Statistical Methods

Testing model assumptions

We performed the data analysis using RStudio version 1.1.463 running R version 3.5.3. References to functions refer to R functions. The model fit we had in mind based on the study design was an ANOVA type, so we wanted to check for the normality of the response variable. The sample sizes were too small to assume the tests would be robust to non-normality. Looking at the mean predicted heat loss values of each hood under study showed two groups of data with separate means. The difference between the means appeared large, so we checked the response variable for normality after separating the data by material type.

Table 1: average predicted heat loss for individual hoods

Material	Hood	averageQ
Traditional	1	215.2667
Traditional	2	236.4000
Traditional	3	218.8667
Traditional	4	243.8667
Traditional	5	221.6333
Particulate-blocking	6	167.3667
Particulate-blocking	7	169.2000
Particulate-blocking	8	170.6333
Particulate-blocking	9	159.1667
Particulate-blocking	10	186.1333

In table 1, the variable “averageQ” is the mean of the three values of “Q predicted” for the given hood. Each material contains 15 observations. This relatively small sample size gave us plots which did not provide convincing evidence for normality.

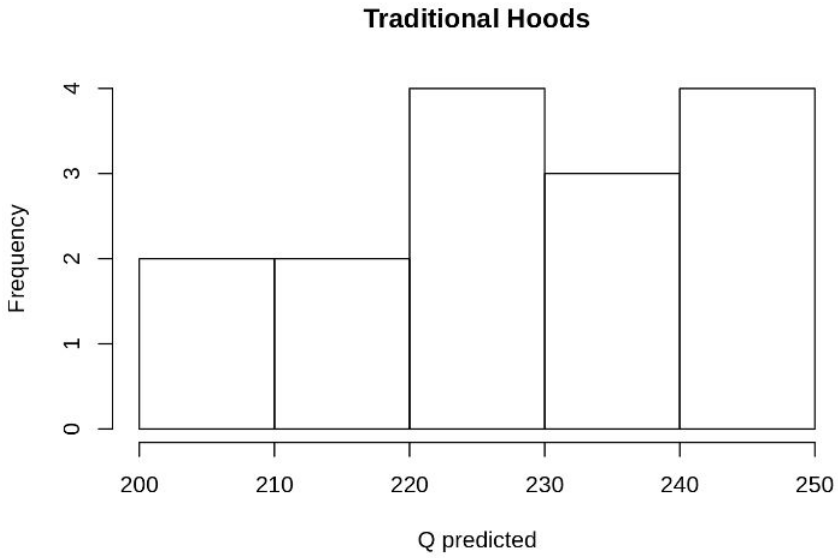


Figure 2: Histogram of traditional hood data

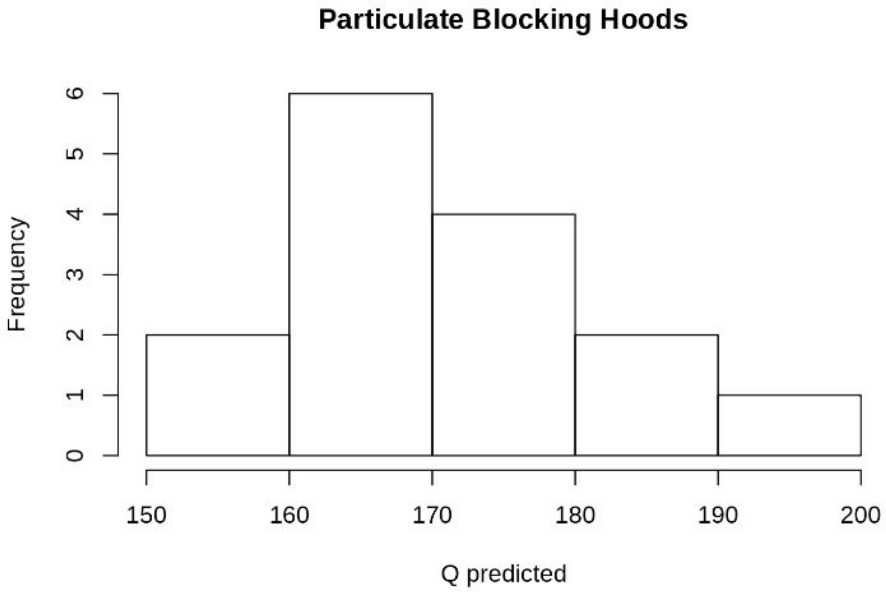


Figure 3: Histogram of particulate-blocking hood data

Because the plots did not provide evidence of normality, we decided to perform the Shapiro-Wilk test of normality on the separated data. We chose this test because it is a high-powered normality test that stands up to small sample sizes. When we tested the data from each material individually, Shapiro-Wilk did not reject normality at the 95% level either for the traditional material $(p = 0.2583)$ or for the particulate-blocking material $(p = 0.2511)$. Since the plots were not obviously non-normal, and the tests failed to reject normality, we did not reject the normality of the average heat loss data for either material.

After the model was fit, the residuals it produced showed acceptable normality. Figures 4 and 5 show that the residuals appear normal and follow the Q-Q line reasonably well for the sample size. Figure 6 shows no particular heteroskedasticity.

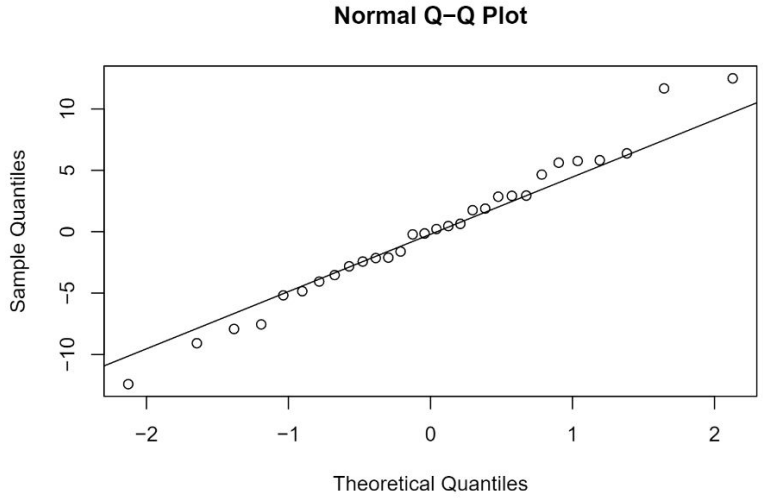


Figure 4: Q-Q plot of fit residuals

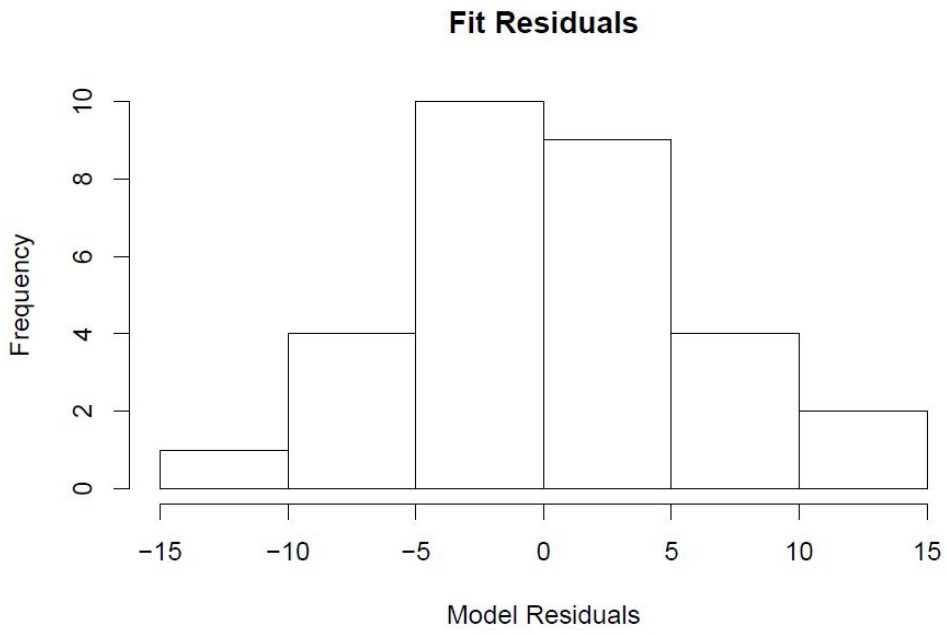


Figure 5: Histogram of residuals

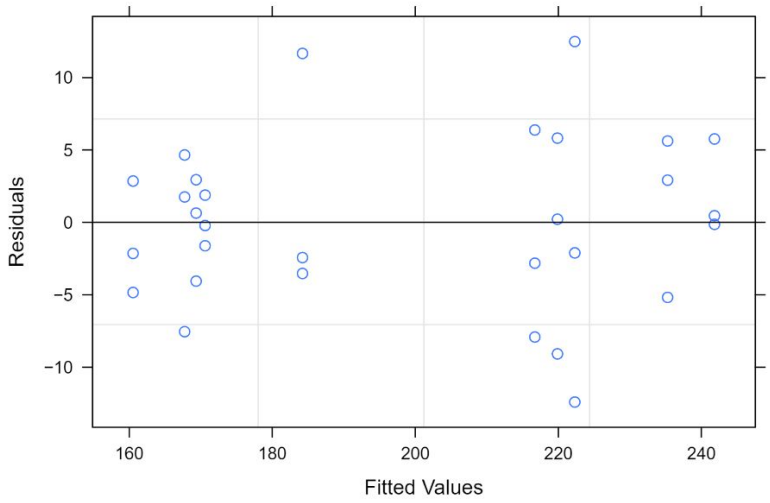


Figure 6: Residuals plotted against the model fitted values

Checking for random and fixed effects

We fit the model using the linear mixed model function *lmer()* from the package *lme4*. The function was used with the default REML fit method. The fixed effect was from the variable “Material” and the random effect was from the variable “Hood”. To evaluate the significance of the random effect, we used the *ranova()* function from the package *lmerTest* to calculate an “ANOVA-like table for random effects.” This function calculates the log-likelihoods of the given model including the random effect term and the reduced model without the random effect. It

then performs a likelihood ratio test (LRT). The null hypothesis for this test is that the model containing the extra parameter is not significantly more likely than the reduced model.

Calculating power and sample size

Prior to this study, the sample size required for a study to consistently detect a difference of $\pm 10\ W/m^2$ in mean predicted heat loss of the two materials $(\mu_1 - \mu_2)$ was not known. One of the key objectives of this research was to determine what sample size should be used in future studies to achieve a power of $1 - \beta = 0.8$, which requires an understanding of the amount of variation present in the hoods relative to the desired detection limit. A simplified model that takes the mean of the three replicates from each hood allows estimation of the required sample size for future studies. That model is:

$$Y = \alpha_i + \varepsilon_{ij}$$

$Y = predicted\ heat\ loss,\ Q_t\ (W/m^2)$
$\alpha_i = material\ type,\ i = 1, 2$
$\varepsilon = error\ term,\ j = 1, ..., n$

where n represents the number of hoods of each material type. The mean values of predicted heat loss of each hood are assumed to be independent and normally distributed, so a two-sample t-test is appropriate to assess whether the material type has a statistically significant effect on heat loss. Additionally, as the researchers are interested in understanding whether the new materials have equal or higher heat loss than the traditional Nomex material, a one-sided test is appropriate with $H_a: \mu_1 > \mu_2$. For the two-sample t-test, there is a relationship among effect size d, significance level α, power $(1 - \beta)$, and sample size n, such that choosing any three of those quantities allows calculation of the fourth. The effect size is calculated as the ratio of the difference in mean predicted heat loss of hoods of each material to the sample standard deviation of the hoods of each material.

$$d = \frac{\mu_1 - \mu_2}{s}$$

This calculation assumes that the standard deviations of the two samples are equal. That appears not to be a valid assumption in this case, so the sample standard deviation is calculated using a modification from Cohen (1988).

$$s = \sqrt{\frac{s_1^2 + s_2^2}{2}}$$

The *pwr.t.test()* function from the *pwr* package was used to complete these calculations.

Because each heat loss test with the manikin takes more than an hour to complete, another scenario of interest involves a study design with only one observation from each hood,

rather than three. The same method for estimating sample size applies, except that a different value must be used for the sample standard deviation because we are no longer averaging the results of three replicates for each hood. To estimate the sample standard deviation for this scenario, a bootstrap simulation with the following procedure was used.

1. Take a random draw of one predicted heat loss observation from each hood
2. Calculate the sample standard deviation for each material from these draws
3. Take Cohen's modification for the case of unequal variances
4. Repeat $B = 10,000$ times to estimate the distribution of sample standard deviations

The estimated distribution of the sample standard deviation in the case of one observation per hood is shown below, in figure 7. The average sample standard deviation from this simulation was $s = 12.27\ W/m^2$. This is the sample standard deviation value used in calculations for the scenario of one observation per hood.

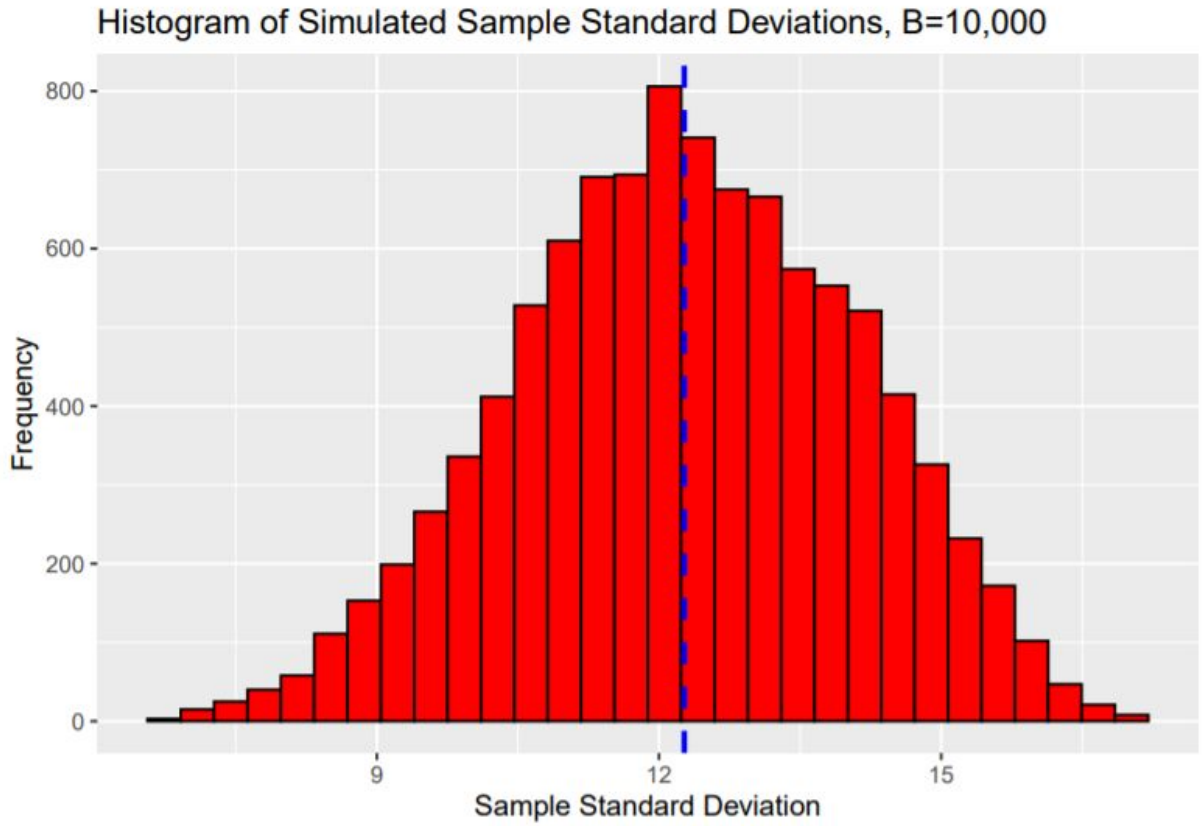


Figure 7: Histogram of sample standard deviations from the bootstrap simulation of one observation per hood

Results

Initial study results

The linear mixed model fit produced an intercept term of 227.207 with a standard error of 4.974. This parameter corresponds to the estimated heat loss for the traditional hoods. The statistic corresponding to the difference in means was calculated as -56.707 with standard error of 7.034. We used this output to construct confidence intervals for both the means. These results are shown in Table 2.

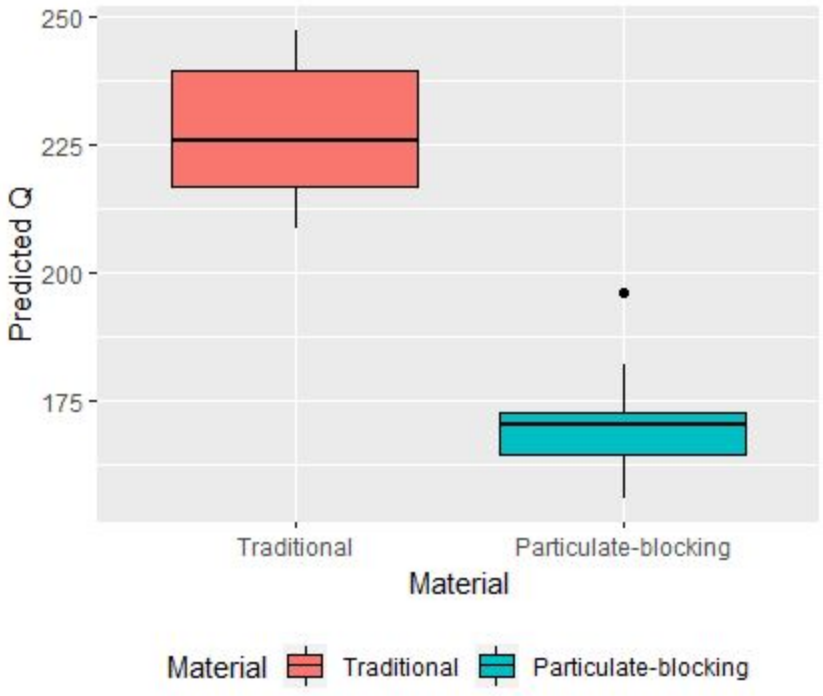
Table 2: Means and confidence intervals for both hood materials

Material	Mean of Q_Pred	CI Lower bound	CI Upper bound	CI Width
Traditional	227.2067	217.5786	236.8347	19.25616
Particulate_Blocking	170.5000	156.8838	184.1162	27.23232

The *lmer()* function calculated the difference between the mean predicted heat loss of the particulate-blocking hoods compared to the mean predicted heat loss of the traditional hoods as the model fixed effect parameter estimate of $-56.7\ W/m^2$ $(p < 0.0001)$ with a 95% confidence interval ranging from -70.3 to -41.3. This supports the hypothesis of a statistically significant material effect. Additionally, the predicted heat loss of the particulate hoods was smaller by a practically significant amount because the magnitude of the smallest practical difference was defined as $\pm 10\ W/m^2$ and the parameter estimate was larger in magnitude than that number.

For the test of the random effect done by the *ranova()* function, the LRT was 14.491 $(p < 0.0005)$. Therefore we concluded that the random effect associated with the hoods was significant.

During the data analysis we visualized the data to show the fixed and random effects. In figure 8 we see the predicted heat loss of the traditional hoods is larger than for the new hoods, and it has a higher variance.



Material	Mean	SD
Traditional	227.2067	13.17858
Particulate Blocking	170.5000	10.14270

Figure 8: Predicted heat loss of the two materials

Figure 9 shows the individual data values separated by hood number. In this plot the random hood effect is pronounced enough to be visible. Hoods two and four appear to have more heat loss than the other traditional hoods, whereas hood ten has more than the other new hoods, and hood nine has somewhat less.

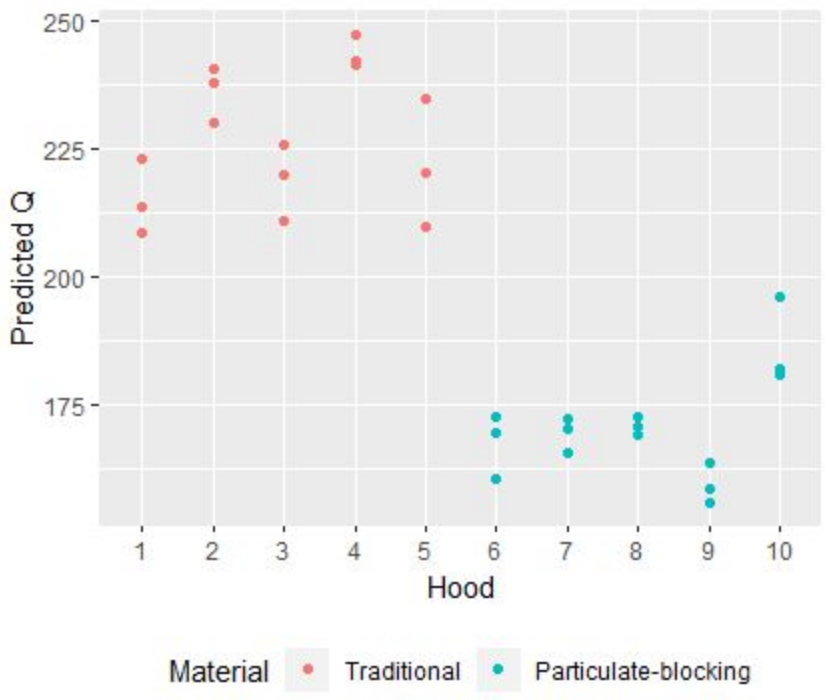


Figure 9: Predicted heat loss for individual hoods

Power and sample size

The *pwr.t.test()* function was used first to analyze the power of the study as-designed. The results are shown below, where μ_1 represents mean predicted heat loss of the Nomex hoods and μ_2 represents mean predicted heat loss of the new particle-blocking hoods, and the alternative hypothesis is $H_a: \mu_1 > \mu_2$. The statistical power resulting from the designed study was 0.37, which is significantly below the target power of 0.80. This study design required $N = 30$ total tests ($i = 2$ materials $*$ $n = 5$ hoods $*$ $k = 3$ replicates), and is represented by the blue dashed lines in figure 10.

Table 3: Power of the study as-designed

Parameter / statistic	**Specified or unspecified**	**Value**
Effect size $d = \frac{\mu 1 - \mu 2}{s}$	Specified	0.899
Significance level α	Specified	0.05
Power level $(1 - \beta)$	Unspecified	0.37
Group sample size n	Specified	5

The next logical question is: what sample size would be required to achieve the desired statistical power? The *pwr.t.test()* function was again used to analyze the power of the study as-designed. The results are shown below. The minimum sample size required to achieve the targeted power of 0.80 was $n = 17$ hoods of each material type. This would correspond to $N = 102$ total tests, and is represented by the red dashed lines in figure 10.

Table 4: Sample size required for the study to achieve desired statistical power of $1 - \beta = 0.8$

Parameter / statistic	**Specified or unspecified**	**Value**
Effect size $d = \frac{\mu 1 - \mu 2}{s}$	Specified	0.899
Significance level α	Specified	0.05
Power level $(1 - \beta)$	Specified	0.80
Group sample size n	Unspecified	16.02

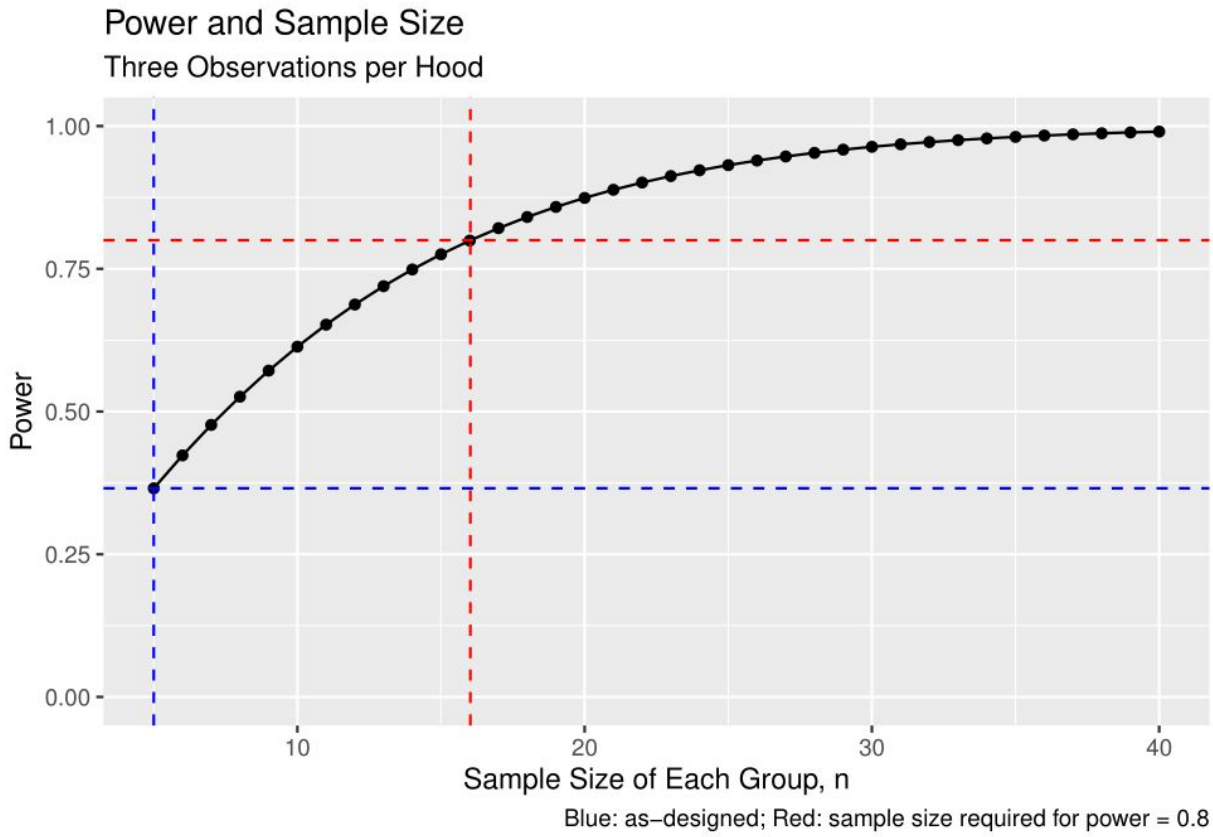


Figure 10: Power and sample size curve for a study with three observations per hood

The final sample size scenario under consideration is: what sample size would be required to achieve the desired statistical power if each hood were tested only once instead of

three times? Using the sample standard deviation calculated from the bootstrap simulation yields a result of $n = 20$ hoods of each material type, for $N = 40$ total tests. This scenario is represented by the red dashed line in figure 11.

Typically, testing each sample multiple times is a good approach to reduce the overall number of samples required. Were the performance of each hood consistent with other hoods of the same material, one would expect a reduction in required samples by a factor of $\sqrt{k} = \sqrt{3} = 1.73$. However, because the random effect of each hood is significant, testing each hood multiple times is not as effective in reducing sample size requirements; in this case the required sample size decreased only by a factor of $20/17 = 1.18$.

Table 5: Sample size required for the study to achieve the desired statistical power of $1 - \beta = 0.8$ with only $k = 1$ replicate per hood

Parameter / statistic	**Specified or unspecified**	**Value**
Effect size $d = \frac{\mu 1 - \mu 2}{s}$	Specified	0.821
Significance level α	Specified	0.05
Power level $(1 - \beta)$	Specified	0.80
Group sample size n	Unspecified	19.1

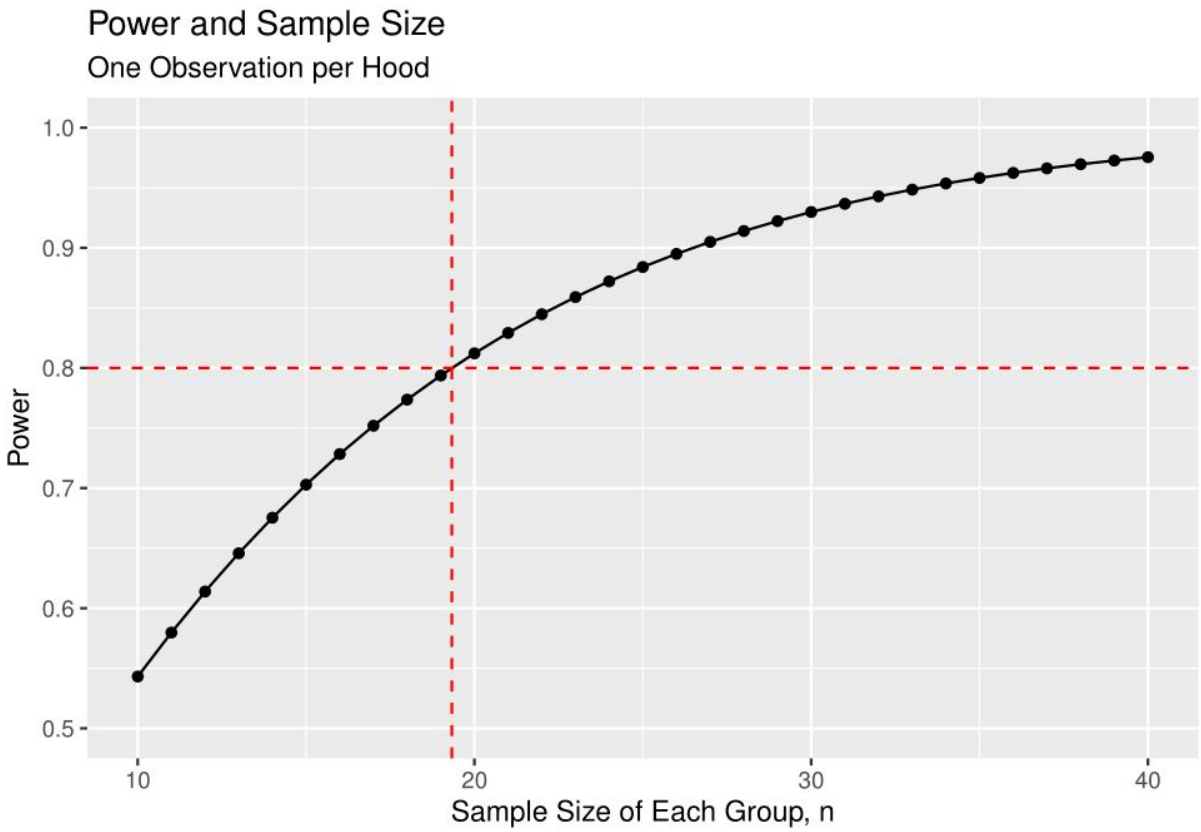


Figure 11: Power and sample size curve for a study with $k = 1$ replicate per hood

Summary

Discussion

The data analysis shows that the new hoods had a predicted heat loss value with a statistically significant difference that was smaller than the value for the traditional hoods by a practically significant amount. The new hoods having a smaller heat loss is the opposite of the ideal outcome for the material performance from the safety and comfort point of view. These results tell us the particulate-blocking hoods do not come with a pure benefit, but rather force a tradeoff between comfort and particulate-blocking performance.

The power analysis shows us that the current sample size of $n = 5$ hoods per material with $k = 3$ replicates per hood has a power of 0.37 compared to a target power of 0.8. It is underpowered for the effect size and significance level because of the amount of variation within hoods of the same material, which prevented the study from reliably detecting differences as small as 10 W/m^2 between the materials. To gain the necessary power, we had to increase the sample size.

Recommendations and Conclusion

Based on the conclusions from the power analysis, it would be necessary to use a sample size of $n = 17$ hoods in each material to achieve the desired power, effect size, and significance level with $k = 3$ replicates while also being able to detect differences between the means of $10\ W/m^2$. Under this scenario, $n = 17$ hoods from $i = 2$ different materials would each be tested $k = 3$ times, resulting in $N = 102$ hours of testing. Using $k = 1$ replicate per hood would raise the sample size per hood to $n = 20$. Therefore, the testing time for this procedure is $N = 40$ hours.

This procedure requires a tradeoff between cost and time. The data set we received from the researcher included links to representative hoods available for purchase. We used these numbers to provide an example of what this tradeoff looks like for the study data we analyzed. If different kinds of hoods are tested in the future, the numbers will not be exactly the same, but the same tradeoff will be necessary because the time expenditure derives from the test procedure. The particulate blocking hood had a listed price of $110.50 and the traditional Nomex hood is listed for $35.00. Therefore using the first scenario, where each of the $n = 17$ hoods is tested $k = 3$ times, would cost $2473.50. Testing $n = 20$ hoods once each would cost $2910.00 and reduce test time by a factor of 2.55.

In our discussions with the researcher, we learned that testing the hoods three times is the standard procedure to use. Assuming the time expenditure is less of a concern than expense, we recommend the sample size of $n = 17$ hoods per material with $k = 3$ replicates per hood. This will allow the testing procedure to detect a practically significant difference of $10\ W/m^2$ with a power of 0.8 and significance level of $\alpha = 0.05$. If the procedure can be modified in light of the time savings offered by the slightly larger sample size, we can also recommend using $n = 20$ hoods per material with $k = 1$ replicate per hood. This will also produce the desired power and significance level.

Another possibility mentioned by the researcher was to increase the detection limit to $20\ W/m^2$. This detection limit may be acceptable to the researcher and the standards committee in light of the fact that the difference in average heat loss between materials was nearly $57\ W/m^2$. The power analysis described above gives a sample size of $n = 5$ for the $k = 3$ replicate design and a sample size of $n = 6$ for the $k = 1$ replicate design. The sample sizes and numbers of tests are presented in Table 6 below. Using the $k = 3$ replicate design would take $N = 30$ hours and cost $727.50. The $k = 1$ replicate design would cost $873.00 and take $N = 12$ hours. Therefore, if the larger detection limit is acceptable, we can recommend using it to bring down the overall monetary and time expenditure.

Table 6: Potential future study designs for $i = 2$ materials, each with power $1 - \beta = 0.8$

	Detection limit $\pm 10\ W/m^2$	Detection limit $\pm 20\ W/m^2$

Three replicate design ($k = 3$)	$n = 17$ hoods $N = 102$ total tests	$n = 5$ hoods $N = 30$ total tests
One replicate design ($k = 1$)	$n = 20$ hoods $N = 40$ total tests	$n = 6$ hoods $N = 12$ total tests