Research Article

Quality and Reliability Engineering International

(wileyonlinelibrary.com) DOI: 10.1002/qre.2047

Published online 13 July 2016 in Wiley Online Library

Planning and Analyzing Experiments with Models that Distinguish Between Replicates and Repeats

Michael S. Hamada,[a*†] Stefan H. Steiner,[b] R. Jock MacKay[b] and C. Shane Reese[c]

A commonly used model to analyze experiments with normal responses does not distinguish between replicates and repeats. The same problem arises with binary and count responses where we can use a generalized linear model. In this article, we propose using models that explicitly allow for two sources of variation, that due to replicates and that due to repeats. In addition, for experiments carried out on high-volume, existing processes, there are often large amounts of data, collected in different ways, that are available to aid in the planning and analysis of the experiment. We demonstrate the value of using these available data with two detailed examples. We finish with a brief summary and raise some further issues. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: available data; Bayesian analysis; designed experiments; repeats and replicates

Introduction

In planning an experiment to investigate a manufacturing process, we must specify the definition of a run as a number of parts or a period of time with a fixed treatment. There are replicates in the experiment when there are two or more runs with the same treatment. There are repeats (also called repetitions) when we set up a run and measure the response on two or more parts within the run. Freeman and Vining[1] refer to repeats as sub-samples in their analysis of a life-testing experiment due to Zelen.[2] If a run is not replicated or if there is a single observation within a run, we are left with the awkward language that there is a single replicate or repeat.

Suppose we have an experiment in which there are two or more repeats within two or more replicates for each treatment. We expect the variation in the response among replicates to be different than its variation among repeats within the same run. In the analysis of data from factorial experiments, many practitioners use a model that does not distinguish between these two sources of variation. In this article, we propose a model that makes this distinction and consider its implications in analyzing normal and non-normal data.

In our experience, in the context of experiments on existing processes, replicates are rare but repeats are common. One exception is the use of replicated center points[3] in designs with quantitative factors. Typically, adding an extra repeat to each run is much easier and cheaper than adding a replicate for one or more treatments. We consider two examples in which there is a single replicate for each treatment but several repeats.

Experiments on existing processes are seldom run in a vacuum. We may have available statistical process control data or the results from 100% inspection on the process at the current factor settings. In a variation reduction context, Steiner and MacKay[4] recommend starting the project with a baseline investigation such as a multi-vari study[5,6] to quantify the variability and look at its behavior over differing time scales. In this paper, we propose to use these available data from the process to help define a run and specify the number of repeats and replicates in the experiment. We can then combine the existing and experimental data in the analysis to increase efficiency and sometimes avoid untestable assumptions about negligible interactions.

Mistakenly analyzing the repeats as replicates is an example of what the ecology and fisheries literatures have called pseudo-replication; see Hurlbert[7] and Millar and Anderson.[8] Millar and Anderson[8] propose mixed-effect models to handle the fisheries examples that they consider. Jones and Nachtsheim[9] and Lucas and his colleagues (e.g., Ju and Lucas[10]) warn about the dangers

[a]*Los Alamos National Laboratory, Los Alamos, NM, USA*
[b]*University of Waterloo, Waterloo, ON, Canada*
[c]*Brigham Young University, Provo, UT, USA*
**Correspondence to: Michael S. Hamada, Los Alamos National Laboratory, Los Alamos, NM, USA.*
[†]*E-mail: hamada@lanl.gov*

of inadvertent split plotting. Treating repeats as replicates can be viewed as one such case. Practitioners do not seem to be aware of this problem as demonstrated by a number of such analyses reported in the Taguchi Symposia proceedings (e.g., American Supplier Institute[11] and those that followed; there were at least 12 of these symposia).

In the next section, we consider the implications of not using the correct model in analyzing normal responses in the context of factorial or fractional factorial designs with replicates and repeats. We then look at a model for repeats and replicates for non-normal data. In the planning stage, we discuss how the available data may be used to help to specify the definition of a run and to choose the number of repeats. In the analysis stage, we propose a Bayesian approach to combine the available and experimental data. We demonstrate the proposed ideas with two examples. We finish with a summary and discussion.

Modeling normal responses with replicates and repeats

Suppose we have a balanced factorial experiment with T treatments, J replicates per treatment, and K repeats per run. For simplicity, we consider only two-level factors. In the analysis of normal data from such an experiment, many practitioners use the model

$$y_{ijk} = \alpha + x_i\beta + r_{ijk}, \tag{1}$$

where y_{ijk} is the response for the ith treatment, jth replicate, and kth repeat. The vector β, the primary parameters of interest, captures the main effects and interactions (perhaps confounded), and the row vector x_i ($x_i = \pm 1$) gives the settings for the factors on the ith treatment and the corresponding interactions. The term r_{ijk} captures the residual variation. We assume the residuals are independent and have a common standard deviation. This model does not distinguish between replicates and repeats.

Alternately, we can add a second random effect to separate the variation among replicates and repeats,

$$y_{ijk} = \alpha + \beta x_i + R_{ij} + r_{ijk}. \tag{2}$$

For normal data, we assume the variation among repeats r_{ijk} is normal with 0 mean and standard deviation σ_r and the variation among replicates R_{ij} are normal with mean 0 and standard deviation σ_R. As well, we assume r_{ijk} and R_{ij} are independent for all i, j, and k. One consequence of this model is that the response variates for repeats within a run are positively correlated. The correlation is

$$\mathrm{corr}\left(y_{ij1}, y_{ij2}\right) = \frac{\sigma_R^2}{\sigma_R^2 + \sigma_r^2}$$

and is large when the variation among replicates is large relative to the variation among repeats. Note that (2) is a split-plot model with no sub-plot factors and the repeats are the sub-plot observations; see Robinson *et al.*[12] and Jones and Nachtsheim.[9]

If there is a single repeat for each run, then models (2) and (1) are equivalent and lead to the same analysis and conclusions. When there is more than one repeat for each run, then using model (1) when model (2) applies may produce misleading conclusions depending on the relative sizes of σ_R and σ_r. On the other hand, using model (2) when $\sigma_R = 0$ leads to a loss of power. To examine these issues more closely, suppose each factor has two levels so the estimate $\widehat{\theta}$ of any main or interaction effect derived from either model is the difference of the response averages at the high and low levels of the effect. Under model (2), we have

$$\mathrm{Var}\left[\widehat{\theta}\right] = \frac{4\left(\sigma_R^2 + \sigma_r^2/K\right)}{TJ} = \frac{4\sigma^2}{TJ}, \tag{3}$$

where $\sigma^2 = \sigma_R^2 + \sigma_r^2/K$. Under model (1), we set $\sigma_R = 0$ so that $\mathrm{Var}\left[\widehat{\theta}\right] = 4\sigma_r^2/TJK$.

To test a hypothesis such as $\theta = 0$, we need the standard error of $\widehat{\theta}$, an estimate of the square root of Equation (3). Under either model, the estimated residuals $\widehat{r}_{ijk}$ are the same, and we have the three-term decomposition of their sum of squares: within runs, among replicates, and among treatments, as shown in (4),

$$\sum_{ijk}\widehat{r}_{ijk}^2 = \sum_{ijk}\left(\widehat{r}_{ijk} - \widehat{r}_{ij+}\right)^2 + K\sum_{ij}\left(\widehat{r}_{ij+} - \widehat{r}_{i++}\right)^2 + KJ\sum_{i}\widehat{r}_{i++}^2. \tag{4}$$

Note that $\widehat{r}_{ij+} = \sum_{k=1}^{K}\widehat{r}_{ijk}/K$ is the average estimated residual within the jth replicate of treatment i, $\widehat{r}_{i++} = \sum_{j,k}\widehat{r}_{ijk}/JK$ is the average residual within treatment i, and the overall average of the estimated residuals is 0.

Suppose that model (1) is appropriate (i.e., $\sigma_R = 0$) with two or more replicates per treatment, but we carry out an analysis using model (2). We would then use the second term in the decomposition (4) to estimate σ^2. Then, we have

$$\widehat{\sigma}^2 = \frac{K\sum_{ij}\left(\widehat{r}_{ij+} - \widehat{r}_{i++}\right)^2}{T(J-1)}, \tag{5}$$

an unbiased estimate of σ^2 with $T(J-1)$ degrees of freedom. So substituting the estimate (5) for σ^2 in (3) leads to a t-test with the appropriate size. However, there is a loss of power because of the loss of degrees of freedom. Using model (1) in the analysis produces

a *t*-test with $T(JK-1)$ degrees of freedom. If we have no replicates (i.e., $J=1$), then the second term of Equation (4) is 0. To estimate σ^2, we must assume certain effects are negligible and further decompose the third term on the right side of (4). Even if the negligibility assumption is correct, we have fewer than T degrees of freedom in the *t*-test, so the loss of power can be substantial.

Now suppose instead that model (2) is appropriate and we carry out an analysis based on model (1). Here, we estimate σ^2 (actually σ_r^2) in Equation (3) using the first two terms of Equation (4),

$$\widehat{\sigma}^2 = \frac{\sum_{ijk}\left(\widehat{r}_{ijk}-\widehat{r}_{i++}\right)^2}{T(JK-1)} = \frac{\sum_{ijk}\left(\widehat{r}_{ijk}-\widehat{r}_{ij+}\right)^2 + K\sum_{ij}\left(\widehat{r}_{ij+}-\widehat{r}_{i++}\right)^2}{T(JK-1)}.$$

Under model (2),

$$E\left[\widehat{\sigma}^2\right] = \frac{J(K-1)}{JK-1}\sigma_r^2 + \frac{K(J-1)}{JK-1}\left(\sigma_R^2+\sigma_r^2/K\right) = \sigma_r^2 + \frac{K(J-1)}{JK-1}\sigma_R^2,$$

and so we have a biased estimate of σ^2 for use in Equation (3). Furthermore, under model (2), $T(JK-1)\widehat{\sigma}^2$ is a weighted sum of χ^2 distributions,

$$\widehat{\sigma}^2 \sim \frac{1}{T(JK-1)}\left(\sigma_r^2\chi^2_{TJ(K-1)} + K\sigma^2\chi^2_{T(J-1)}\right),$$

and hence, the *t* statistic for $\widehat{\theta}$

$$t = \frac{\widehat{\theta}-\theta}{2\widehat{\sigma}/\sqrt{JT}} = \frac{Z}{\widehat{\sigma}}$$

no longer follows a *t* distribution. Both the size and power calculated from the assumed *t* distribution will be incorrect. We can examine the size problem quantitatively by simulation. For example, suppose we have a 2^3 design with $J=2$ replicates and $K=5$ repeats per run. Let $f=\sigma_R^2/\left(\sigma_R^2+\sigma_r^2\right)$ be the fraction of the total variation due to the random replication effect. Table I gives the actual size of the two-sided test for $\theta=0$ with nominal size 0.05 as f varies. If σ_R^2 is a significant component of the total variation, we are at increased risk of deciding that an effect is significant when it is, in fact, negligible.

When we have two or more replicates and repeats for each treatment, the run averages $\overline{y}_{ij+}$ are sufficient for β and σ^2 in model (2). The within-run estimated residuals are independent of the run averages and can be used to estimate σ_r^2 but provide no information about the treatment effects or σ^2. In the design, increasing the number of repeats K reduces σ^2 and can lead to a large increase in power if σ_r^2 dominates σ_R^2. We can use the experimental data to estimate and compare σ_r^2 and σ_R^2, which may be helpful in planning further experiments; see Example 1: normal data.

Non-normal responses for experiments with repeats and replicates

The response in many experiments cannot be described by a model such as Equation (1) or (2) based on normal residuals. Some examples are given in Chapter 14 of Wu and Hamada.[13] We can determine whether or not an individual unit is defective or, less often, count the number of defects per unit. We might also measure breaking strength or time to failure where a normal model may be inappropriate. The generalized linear models for a response y_{ijk} (McCullagh and Nelder[14]), linear on the link function scale, corresponding to models (1) and (2) are

$$\text{link}\left(\eta_{ijk}\right) = \alpha + \beta x_i, \tag{6}$$

with y_{ijk} assumed independent and

$$\text{link}\left(\eta_{ijk}\right) = \alpha + \beta x_i + R_{ij}, \tag{7}$$

where, for convenience, we suppose $R_{ij} \sim N(0,\sigma_R)$ and, given R_{ij}, y_{ijk} are independent, and $E(y_{ijk})=\eta_{ijk}$. Model (6) does not distinguish between replicates and repeats, whereas model (7) accounts separately for both between-run and within-run variability. In an experiment with replicates and repeats, if we use model (6) when model (7) is appropriate, we cannot expect that the estimates based on model (6) will behave as predicted.

Table I. Actual size of the *t*-test assuming model (1) when model (2) is correct

f	0.1	0.3	0.5	0.7	0.9
Size	0.055	0.069	0.087	0.113	0.149

In the normal case, we saw that if model (2) is appropriate, using model (1), we can analyze the within-run averages (or sums) without any loss of efficiency. Suppose that y_{ijk} is binary and we have a logistic link in model (7). Then, given the random effect R_{ij}, the within-run sums $\sum_k y_{ijk}$ are binomial, but marginally, they are not. That is, unlike the normal case, using model (6) and the within-run sums with a logistic link leads to an incorrect analysis. We see the same problem if y_{ijk} given R_{ij} is Poisson with the log link.

It is possible to analyze experimental data using model (7) with both replicates and repeats using a frequentist approach with the lmer function in the lme4 package of R.[15] However, in this article, we consider a Bayesian approach that provides an easy way to evaluate functions of the model parameters such as the probability of meeting a specification. We plan to use available process data to estimate σ_R to aid in the planning of the experiment and to combine the available and experimental data in the analysis. A Bayesian approach with vague priors is a convenient way to proceed.

Available process data

We showed that replicates and repeats in a factorial experiment should be differentiated by the model used to describe the response. Here, we consider some planning and analysis issues for such experiments.

Suppose the goal of the experiment is to find factors that can be used to adjust the process average for the response of interest that may be continuous, binary, or a count. We consider the issue of deciding on the nature and number of repeats and replicates. We limit the discussion to experiments on existing processes where the factors under consideration are normally fixed. That is, in the current process, these factors do not change and hence do not contribute directly to the variation in the output.

We also assume that there are data available collected from the process with the experimental factors held fixed at their current levels. These historical data for the response of interest can take many different forms. Examples include a control chart with a fixed sampling plan, [16] inspection records (the results of 100% inspection are the most useful), or a multi-vari study[5,6] looking for important families of variation. These data provide some idea of the pattern of variation in the current process.

For planning purposes, we make the assumption that when the experimental factors are changed to a particular treatment combination and then left fixed, the process average may shift but the pattern of variation remains the same. Further, we assume that if the same treatment combination is replicated, re-setting the factor levels makes a negligible contribution to the variation. That is, when we execute the selected design, we assume changes in factor settings make a negligible contribution to σ_R.

Under these assumptions and depending on the nature of the available data, for any proposed run definition, we may be able to obtain some information about σ_R and, in the case of a normal response, some information about σ_r as well. We use this information to help plan the experiment. The pattern of variation in the available data may also suggest possible blocking schemes.

In the analysis, we propose to combine information from the historical data with that provided by the experiment. For convenience, we take a Bayesian approach.

Example 1: normal data

In order to reduce the proportion of leaking seals in a battery case, an improvement team decided to conduct an experiment to increase the tensile strength of the seal. The goal of the experiment was to increase the average tensile strength to at least 440 lb. We have preserved the (incorrect) vernacular for the units of tensile strength; see Steiner and MacKay[4] (page 228) for more details. As part of a preliminary analysis, the team sampled 100 parts over 1 week to establish a baseline for the process average and standard deviation. Based on these data, the estimates of the process average and standard deviation were 389.3 and 44.9, respectively. The process average strength was well below the target. Other than these summaries, the baseline data played no further part in the team's analysis or conclusions.

The team planned a 2^3 full factorial experiment. Within each of the eight runs, there were five repeats. No reasons were given for either the design or the number of repeats. The order of the runs was randomized. The data are given in Table II. Because the experiment had no replication, the analysis was based on a half-normal plot (not shown) of the seven estimable factor effects. The effect of factor A stood out over all others. After some further simple experimentation to determine an appropriate value, the level of A was changed in the process with satisfactory results. The average seal strength increased, and the leak rate decreased. The project was deemed successful.

In hindsight, it was noticed that the baseline data had been collected in 20 subgroups of five consecutive batteries over the 5-day sampling period. The data are shown in Table III. How might these available data have been used to justify the experimental plan and to augment the simple analysis?

We start with the analysis. The average and standard deviation Shewhart charts (Figure 1) show that the process is close to being stable with a single point on the S chart out of control. Given how the experiment was planned with five repeats per run, we can use the available data to gain information about both σ_R and σ_r. Assuming that model (2) applies to the baseline data with a single treatment (current conditions), we obtain estimates $\widehat{\sigma}_R = 15.9$ and $\widehat{\sigma}_r = 43.2$ from a one-way analysis of variance.

Using (3) and only the baseline data, the estimate of the standard error for any main effect or interaction is 12.5. Instead of relying on the half-normal plot, we can use this standard error to isolate significant effects. Note that the within-run variation is dominant here so the repeats are very useful. After the experiment is run, we can obtain an improved estimate of σ_r by combining the within-run residuals for both the available and experimental data, although the gain in degrees of freedom is relatively small in this case.

Table II. Seal strength experimental plan and data

Treatment	A	B	C	Tensile strengths				
1	Low	Low	Low	413	505	489	452	465
2	High	Low	Low	468	493	484	393	423
3	Low	High	Low	383	368	280	377	370
4	High	High	Low	383	365	352	389	353
5	Low	Low	High	440	415	483	395	433
6	High	Low	High	466	387	505	393	456
7	Low	High	High	399	294	317	300	337
8	High	High	High	373	379	383	385	345

Table III. Seal strength baseline data

Subgroup									
1	2	3	4	5	6	7	8	9	10
381	406	320	356	368	430	409	396	354	505
457	353	370	378	429	447	433	330	392	443
503	375	391	491	452	383	428	348	379	260
481	433	390	360	393	451	387	339	416	403
405	445	436	406	314	394	371	383	407	351
Subgroup									
11	12	13	14	15	16	17	18	19	20
425	379	319	356	329	375	464	407	365	399
406	377	330	391	392	352	343	407	397	469
361	351	367	322	375	433	403	390	288	362
383	350	335	334	373	426	377	384	372	411
368	402	450	391	331	353	442	359	404	421

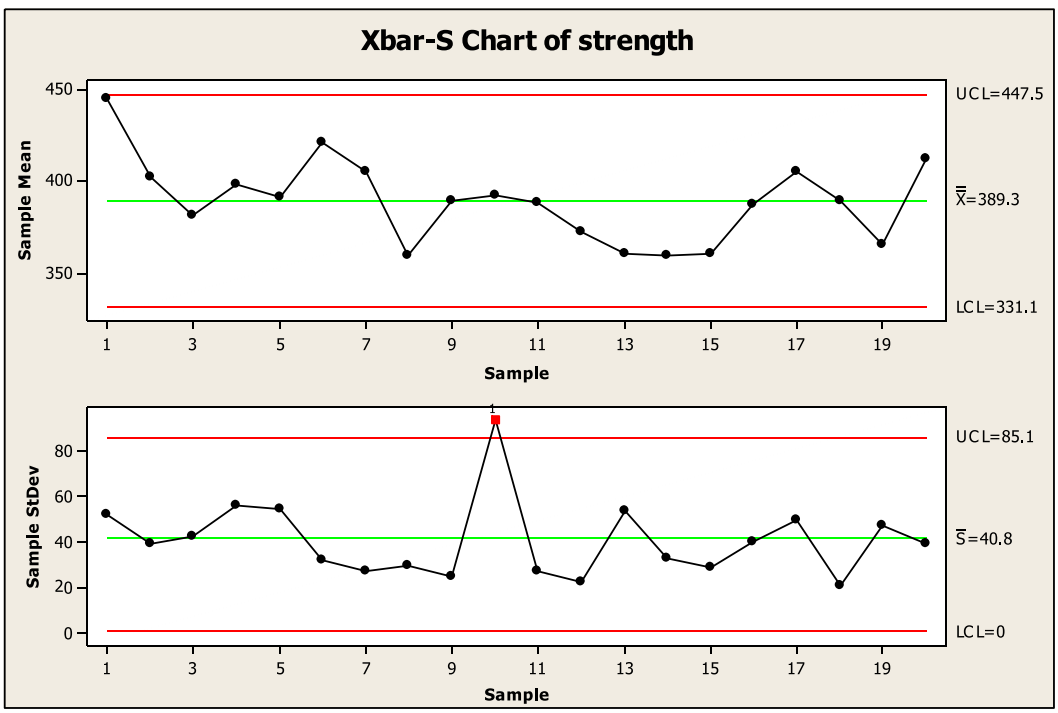

Figure 1. Control charts for the available seal strengths

A Bayesian approach is an alternate way to combine the data. We assume that model (2) is appropriate for both the available and experimental data. The model has 10 parameters: the mean μ for the baseline; the intercept for the experiment α; six factor effects (A, B, and C main effects and AB, AC, and BC two-factor interactions denoted by β_2 to β_7); and two standard deviations (σ_R and σ_r). Although not necessary, we assume that the three-factor interaction is negligible, so that the experiment gives a small amount of information about σ_R. We select diffuse priors (assuming independence) as shown in Table IV.

We used WinBUGS[17] to analyze the data that produce samples from the posterior distribution through Markov chain Monte Carlo algorithms like the Metropolis–Hastings algorithm[18]; see Appendix A for the WinBUGS code for the seal strength example. Table V displays the median and 95% probability interval of the marginal posterior distributions for the model parameters and confirms that changing the level of factor B and none of the other factors or interactions effects the seal strength.

To demonstrate the convenience of the analysis, suppose we want to estimate (i.e., find the posterior median and central 95% of the posterior distribution) for $\lambda = P(Y > 440)$. Here, Y is the seal strength of a randomly selected battery when factor B is set to its low level and the other two factors are at their original levels (A low, C high). Given $\alpha, \beta_2, \ldots, \beta_7, \sigma_r, \sigma_R$, we have $\lambda = 1 - \Phi\left(\frac{440-(\alpha-\beta_2-\beta_3+\beta_4+\beta_5-\beta_6-\beta_7)}{\sqrt{\sigma_r^2+\sigma_R^2}}\right)$, where Φ is the standard normal cdf. WinBUGS produces a large number of samples from the joint posterior distribution of $\alpha, \beta_2, \ldots, \beta_7, \sigma_r, \sigma_R$ that we can substitute into the expression for λ and hence determine its posterior distribution. The median is 0.43, and the 95% probability interval is (0.13, 0.79).

We can see the quantitative value of using the available process data by analyzing only the experimental data using the same priors as before. Table VI shows that the factor B main effect (i.e., β_3) is important but with greater uncertainty. For $\lambda = P(Y > 440)$ as defined earlier, the corresponding median and 95% probability interval are 0.43 and (0.07, 0.87). Incorporating the available data into the analysis significantly reduces the uncertainty.

The project team made no explicit use of the available data when they designed the experiment. In hindsight, looking at the available data, the team could have seen the following:

- substantial benefit in increasing the number of repeats because the within-run variation is dominant, especially if the anticipated effects are small;
- for the same reason, little benefit in introducing replication; and
- no need for blocking because, as seen in Figure 1, there are no systematic patterns in the response over time.

In this experiment, it was relatively easy to randomize the order of the runs. If randomization had not been possible or exorbitantly expensive, Figure 1 shows no systematic patterns in the variation, and so these data provide some assurance that there are no lurking variables that might confound the conclusions of the experiment.

Example 2: experiment with binary data

We received the case study report that contains this example in confidence. Hence, we have provided little detail of the process, factors, or changes based on the analysis of the data. The experimental description and data are real. The process data that were available when the experiment was conducted unfortunately no longer exist.

Table IV. Seal strength prior distributions

Parameter	Prior
μ, α	Normal(0,1000^2)
$\beta_2, \ldots, \beta_7$	Normal(0,10^2)
σ_R, σ_r	Uniform(0,100)

Table V. Seal strength posterior distribution summaries

Parameter	Median	0.95 Probability interval
μ	389.40	(379.10, 399.70)
α	402.20	(386.40, 418.90)
β_2	6.20	(−10.17, 22.52)
β_3	−45.37	(−61.83, −29.38)
β_4	−7.92	(−24.12, 8.41)
β_5	7.67	(−8.732, 23.77)
β_6	−2.53	(−13.67, 18.95)
β_7	6.27	(−10.03, 22.62)
σ_r	41.97	(37.05, 48.17)
σ_R	12.0	(0.95, 26.23)

Table VI. Seal strength posterior distribution summaries (experimental data only)

Parameter	Median	0.95 Probability interval
α	402.30	(368.10, 436.70)
β_2	6.52	(−28.11, 39.34)
β_3	−44.93	(−75.41, −8.44)
β_4	−7.79	(−40.78, 24.87)
β_5	7.44	(−27.22, 40.27)
β_6	2.55	(−31.52, 36.06)
β_7	6.22	(−28.68, 39.49)
σ_r	37.68	(29.94, 49.31)
σ_R	25.07	(1.03, 93.1)

In an injection molding process, a clear lens is inserted into the mold, and then plastic is added to seal the lens within the rest of the part. The completed parts are subject to 100% inspection. A screening experiment was planned to identify and change some normally fixed process factors with the goal of reducing crazing, a stress cracking problem. The degree of crazing was measured on a three-point ordinal scale. There was fear that the changes to reduce crazing might lead to an increase in other defect types. For illustration, we concentrate on splay defects, white streaking on the finished part. The team hoped to find factors that could be changed to reduce crazing without affecting the frequency of splay.

The experiment had 16 runs with 11 factors, labeled A–K, each at two levels. The factors did not change under normal operation. Each run consisted of 10 shots, that is, repeats. The factors were changed according to the treatment, and then the process was allowed to settle before the 10 consecutive parts were selected. There was no replication or blocking. The order of the runs was randomized. The degree of crazing and the presence or absence of splay were determined for each part in the run. The design in standard order and the data (number of parts out of 10 with splay) for each run are given in Table VII.

The original analysis for the splay data used a half-normal plot (Figure 2) of the 15 effects to see if any main effects stood out. No factor appeared to be significant.

The crazing response was analyzed separately in a similar naïve way and showed that two of the factors had a significant effect. The team was happy with these findings because it appeared that changing either of these factors would have little effect on splay and reduce the average crazing score.

Now we suppose that historical data from the 100% inspection had been considered. For our purposes, we assume that records were kept for each part in production order. As noted earlier, the data no longer exist. To provide independent information about σ_R, we look at a set of pseudo-runs of the experiment. That is, we sample groups of consecutive parts from the historical record so that each group mimics the definition of a run in the experiment. We limit the sampling to a period of production when there were no changes to the factors of interest in the experiment. Here, we demonstrate the idea using the realistic but artificial data summarized in Table VIII based on a sample of 100 pseudo-runs of 10 consecutive parts. A p-chart (not given) shows no evidence of instability.

Table VII. Experimental design and number of defects per run

Run	A	C	K	D	G	E	B	F	H	I	J	Number of parts with splay
1	−1	−1	1	−1	1	1	−1	−1	1	1	−1	0
2	−1	−1	1	−1	1	1	−1	1	−1	−1	1	5
3	−1	−1	1	1	−1	−1	1	−1	1	1	−1	0
4	−1	−1	1	1	−1	−1	1	1	−1	−1	1	2
5	−1	1	−1	−1	1	−1	1	−1	1	−1	1	1
6	−1	1	−1	−1	1	−1	1	1	−1	1	−1	1
7	−1	1	−1	1	−1	1	−1	−1	1	−1	1	1
8	−1	1	−1	1	−1	1	−1	1	−1	1	−1	0
9	1	−1	−1	−1	−1	1	1	−1	−1	1	1	1
10	1	−1	−1	−1	−1	1	1	1	1	−1	−1	2
11	1	−1	−1	1	1	−1	−1	−1	−1	1	1	2
12	1	−1	−1	1	1	−1	−1	1	1	−1	−1	3
13	1	1	1	−1	−1	−1	−1	−1	−1	−1	−1	3
14	1	1	1	−1	−1	−1	−1	1	1	1	1	3
15	1	1	1	1	1	1	1	−1	−1	−1	−1	3
16	1	1	1	1	1	1	1	1	1	1	1	5

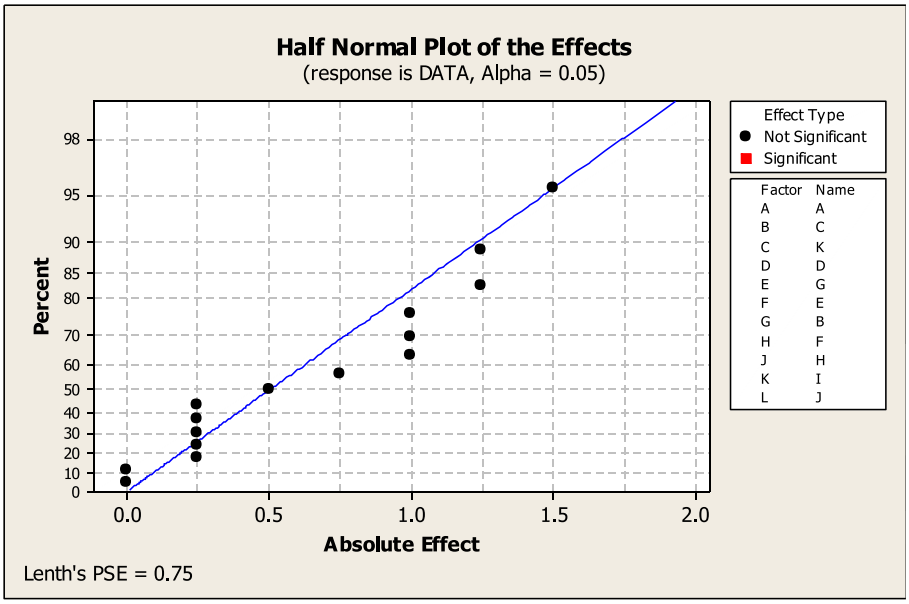

Figure 2. Half-normal plot of the splay factor effects

Table VIII. Frequency of splay defects in pseudo-runs						
Number of defects	0	1	2	3	4	≥5
Number of pseudo-runs	56	17	18	7	2	0

We carried out a Bayesian analysis for the combined data in Tables VII and VIII using model (7) and a conditional binomial distribution with logit link. For the available data in Table VII, there are no treatment effects (but a baseline intercept μ). For the experimental data, we used model (7) with 12 treatment effects (overall experimental mean α and 11 factor main effects). We assume that σ_R is the same for both data sets but that the 'mean' levels μ and α may be different. There are 14 parameters in total. We use WinBUGS[17] with the diffuse priors given in Table IX; see Appendix A for the WinBUGS code for this example.

The purpose of the experiment was to screen for large main effects among the 11 factors. We summarize the posterior distributions of these main effects in Table X. We see that the factor F main effect (β_9) is important and likely the factor I main effect (β_{11}) is as well. These effects were missed in the analysis using only the experimental data.

Now suppose we are in the planning stage for this screening experiment, and we want to make use of the available data. Consider an experiment with *J* replicates per treatment and *K* repeats per run. Select a systematic sample of *N* pseudo-runs from recent history

Table IX. Prior distributions for the splay example	
Parameter	Prior
μ	Normal(0,10^2)
α	Normal(0,100^2)
$\beta_2, \ldots, \beta_{12}$	Normal(0,10^2)
σ_R	Uniform(0,10)

Table X. Posterior distribution summaries for the splay example					
Parameter	Median	0.95 Probability interval	Parameter	Median	0.95 Probability interval
α	−2.82	(−3.32, −2.44)	β_7	−5.17	(−18.73, 8.33)
μ	33.63	(16.95, 58.53)	β_8	5.23	(−18.70, 8.08)
β_2	4.65	(−9.23, 18.95)	β_9	13.36	(0.33, 27.97)
β_3	−12.25	(−27.04, 1.73)	β_{10}	−3.16	(−17.58, 10.39)
β_4	4.47	(−9.41, 19.27)	β_{11}	13.33	(−0.14, 28.00)
β_5	−5.31	(−18.83, 7.71)	β_{12}	−2.99	(−17.45, 10.54)
β_6	−5.43	(−18.69, 8.10)	σ_R	1.028	(0.588, 1.557)

when no experimental factor was changed. The time between the pseudo-runs should be the same as the planned time between runs in the experiment. Plot the proportion of defects per run over time looking for regular patterns of variation that suggest opportunities for blocking. For example, if there is a systematic shift-to-shift effect, then blocking over shifts should be built into the experimental design. Next, we fit model (7) augmented by fixed block effects to the available data with the goal of estimating the extra between-run variation σ_R. In our example, there was no indication that blocking was necessary, and using the vague priors in Table IX and the data in Table VIII, we find that median value of the posterior for σ_R is 1.06 with 95% probability interval (0.60, 1.61).

For simplicity, suppose we decide that blocking is unnecessary. Roughly speaking, for any particular treatment, the more precisely we can estimate the probability of a defect with that treatment, the more power we will have in the experiment to detect significant effects. Let y be the number of defects in any particular run. Then, using model (7), we have

$$Y|(R=r)\sim\text{Binom}(K,p(r))\ ,\ p(r)=\frac{\exp\{\alpha+r\}}{1+\exp\{\alpha+r\}},\ R\sim N(0,\sigma_R),$$

where α depends on the particular treatment. We further assume independence among the runs. Using the conditional variance formula, we can show (Appendix B) that for any treatment with J replicates and K repeats, the corresponding variance of the proportion of defects is

$$\frac{E[p(R)](1-E[p(R)])}{JK}+\frac{K-1}{JK}\text{Var}[p(R)]. \tag{8}$$

If among replicate variation σ_R goes to zero, the second term in Equation (8) vanishes, and only the binomial variation, $p(0)$, the first term in Equation (8), remains. As the number of repeats K increases, the second term in Equation (8) is almost unchanged. Also, as the number of replicates J increases, the variance goes to zero. Thus, Equation (8) suggests that, from a statistical perspective, a replicate is always better than a repeat. On the other hand, replicates are likely to be more expensive than repeats. Using the available data, we can estimate σ_R to see which of these two terms dominates as α or the probability of a defect changes. This gives a rough idea of the relative value of repeats versus replicates.

We quantify the trade-off by comparing the case $J=K=1$ with either $K=1$ and $J=a$ (a replicates with a single repeat) or $K=a$ and $J=1$ (single replicate with a repeats). Using r replicates will reduce the variance in Equation (8) by a factor of $1/a$, while using a repeats will result in a smaller reduction. We are interested in comparing how much smaller a reduction is achieved using repeats rather than replicates. From Equation (8), the percentage reduction available using a repeats rather than a replicates is $100(1-\text{Var}(p(R))/[E(p(R)(1-p(R)))])$. Note that this result does not depend on a.

In our example, using the available data only, we have plausible values for σ_R in the range (0.60, 1.61). For these values, Table XI shows the percentage reduction from using repeats rather than replicates, that is, the relative benefit of increasing the number of repeats compared with increasing the number of replicates as $p(0)=\exp(\alpha)/(1+\exp(\alpha))$ varies.

As mentioned earlier, in all cases, adding an extra replicate rather than a repeat for each treatment will lead to a greater reduction in the variance of the proportion defective for any treatment. However, in many cases, the benefit of adding repeats is close to that of adding replicates. For instance, with $\sigma_R=1.06$ and $p(0)=0.20$, using repeats results in 84.1% of the reduction in the variance of the proportion defect we could obtain by instead using replicates.

In general, for other values of σ_R and $p(0)$ and for any given choice of J and K, we can estimate Equation (8) to examine how precisely we can estimate a treatment proportion of defects.

The aforementioned analysis breaks down if the number of repeats is large as we do not expect the between-run variation quantified by σ_R to stay the same. If we have sufficient inspection data as in the example, we can investigate different choices of K with appropriate sampling from the records.

Discussion

In this paper, we consider the issue of differentiating between repeats and replicates in factorial experiments conducted on high-volume, data-rich processes. For normal data, we look at a model with two independent sources of variation and quantify possible errors resulting from treating repeats as replicates. For more general situations when we model the response with a generalized linear model, we include an extra source of variation to distinguish between the within-run and between-run variations.

In this context, there are often considerable data available before the experiment. As we show by example, these data may be used in the planning stage to define blocking schemes and to determine the relative value of repeats and replicates. We also demonstrate

Table XI. Percentage benefit of a repeat versus a replicate (compare relative benefit of $J=K=1$ with $J=a$, $K=1$, or $J=1$, $K=a$)

			$p(0)$		
σ_R	0.05	0.10	0.20	0.30	0.50
0.60	97.9	96.3	94.3	93.1	92.3
1.06	91.3	87.7	84.1	82.3	81.1
1.61	78.2	74.2	70.7	69.2	68.3

the value of the available data when it is incorporated into the analysis with the experimental data. We can increase the sensitivity of the experiment and sometimes avoid untestable assumptions such as the negligible effects of high-order interactions.

There are many reasons for conducting experiments on existing processes. For example, suppose we want to desensitize the output to the variation in an identified noise factor z that normally varies in the process. In the experiment, we control z at two or more levels and vary a number of normally fixed factors with the goal of identifying suitable interactions between z and the control factors. It is not clear how to use available data in this case. In many experiments, as in the splay example, there are multiple outputs and data available on all of them. Again, it is unclear how to incorporate the data into the planning and analysis of the experiment.

Models (2) and (7) imply a form of stability for the process with two sources of variation within and between runs. We suggest using run or control charts to examine the behavior of the process over time. We may be able to deal with other systematic (over time) sources of variation using blocking. However, if the variation is unpredictable (e.g., say in the splay example, the process had large unpredictable spikes in the number of splay defects), it is advisable to first bring the process into some semblance of control before trying to conduct experiments.

We have adopted a pragmatic Bayesian approach to our analyses. WinBUGS[17] can easily deal with the combined available and experimental data.

Appendix A

This appendix provides WinBUGS code for the combined analysis of the available process data and the experimental data for the seal strength (normal) and crazing (binomial) examples.

Seal strength example

```
model
{
for(i in 1: NP) {
                    strength[i] ~ dnorm(muP[i],tauRepeat)
                    muP[i] < -muP0 + thetaP[subgroup[i]]
                }
                for(j in 1: NPsg){
                thetaP[j] ~ dnorm(0,tauReplicate)
                }
                for(i in 1: N) {
                    mu[i] < - beta[1] + beta[2]*x1[i] + beta[3]*x2[i] + beta[4]*x3[i] + beta[5]*x1x2[i] + beta[6]*x1x3[i] + beta[7]*x2x3
[i] + thetaE[run[i]]
                    Y[i] ~ dnorm(mu[i],tauRepeat)
                    }
                    for(j in 1:nRun){
                    thetaE[j] ~ dnorm(0,tauReplicate)
                    }
            #priors
            beta[1] ~ dnorm(0.0,1.0E-6)
            for(k in 2:7){
            beta[k] ~ dnorm(0.0,1.0E-4)
            }
            A < -100
            sigmaReplicate ~ dunif(0,A)
            tauReplicate < - pow(sigmaReplicate, -2)
            B < -100
            sigmaRepeat ~ dunif(0,B)
            tauRepeat < - pow(sigmaRepeat, -2)
            muP0 ~ dnorm(0.0,1.0E-6)
            }
Data
list(NP = 100,N = 40,nRun = 8,NPsg = 20) #5 repeats
Inits
list(beta = c(0,0,0,0,0,0,0),sigmaReplicate = 10,sigmaRepeat = 45,
                thetaE = c(0,0,0,0,0,0,0,0),muP0 = 389
                )
```

Splay example

```
model{
```

```
                        for(i in 1: NP) {
                        yy[i] ~ dbin(p[i],n[i])
                        p[i] < - exp(logitp[i])/(1 + exp(logitp[i]))
                        logitp[i] < -muP0 + thetaP[run[i]]
                    }
                    for(j in 1: NP){
                    thetaP[j] ~ dnorm(0,tauReplicate)
                    }
                    for(i in 1: NE) {
                        logitpp[i] < - beta[1] + beta[2]*A[i] + beta[3]*C[i] + beta[4]*K[i] + beta[5]*D[i
        beta[6]*G[i] + beta[7]*E[i] + beta[8]*B[i] + beta[9]*F[i]+
                        beta[10]*H[i] + beta[11]*II[i] + beta[12]*J[i]+
                        thetaE[runE[i]]
pp[i] < -exp(logitpp[i])/(1 + exp(logitpp[i]))
                        y[i] ~ dnorm(pp[i],10)
                        }
                        for(j in 1:nRun){
                        thetaE[j] ~ dnorm(0,tauReplicate)
                        }
                    #priors
                    beta[1] ~ dnorm(0.0,1.0E-4)
                    for(k in 2:np){
                    beta[k] ~ dnorm(0.0,1.0E-2)
                    }
                    AA < -10
                    sigmaReplicate ~ dunif(0,AA)
                    tauReplicate < - pow(sigmaReplicate, -2)
                    muP0 ~ dnorm(0.0,1.0E-2)
        }
Data
list(NE = 16,nRun = 16,np = 12,NP = 100)
Inits
list(beta = c(0,0,0,0,0,0,0,0,0,0,0,0),sigmaReplicate = 1,
                    thetaE = c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
                    )
```

Appendix B

In this appendix, we derive the variance of the proportion of defects for a particular treatment as given by Equation (8). Let Y_{jk} be a binary random variable corresponding to the jth replicate and kth repeat. Then, denoting $p_1, p_2, \ldots, p_J$ as the replicate effects for replicates 1, 2, …, J, and p as the generic replicate effect, we have

$$\begin{aligned}\mathrm{Var}\left(\frac{\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{jk}}{JK}\right) &= \frac{1}{J^2K^2}\left[\mathrm{Var}\left(E\left(\sum\sum Y_{jk}\right)|p_1,\ldots,p_J\right) + E\left(\mathrm{Var}\left(\left(\sum\sum Y_{jk}\right)|p_1,\ldots,p_J\right)\right)\right] \\ &= \frac{1}{J^2K^2}[\mathrm{Var}(K(p_1+\cdots+p_J)) + E(Kp_1(1-p_1)+\cdots+Kp_J(1-p_J))] \\ &= \frac{1}{JK}\left[K\mathrm{Var}(p) + E(p) - E\left(p^2\right)\right] = \frac{1}{JK}[(K-1)\mathrm{Var}(p) + E(p)(1-E(p))]\end{aligned}$$

References

1. Freeman LJ, Vining GG. Reliability data analysis for life test experiments with subsampling. *Journal of Quality Technology* 2010; **42**:233–241.
2. Zelen M. Factorial experiments in life testing. *Technometrics* 1959; **51**:249–288.
3. Montgomery DC. Design and Analysis of Experiments (8th edn). John Wiley & Sons: New York, 2013.
4. Steiner SH, MacKay RJ. Statistical Engineering: An Algorithm for Reducing Variation in Manufacturing Processes. Quality Press: Milwaukee, 2005.
5. Seder L. Diagnosis with diagrams—Part I. *Industrial Quality Control* 1950a; **7**(1):11–19.
6. Seder L. Diagnosis with diagrams—part II. *Industrial Quality Control* 1950b; **7**(2):7–11.
7. Hurlbert SH. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 1984; **54**:187–211.

8. Millar RB, Anderson MJ. Remedies for pseudoreplication. *Fisheries Research* 2004; **70**:397–407.
9. Jones B, Nachtsheim C. Split-plot designs: what, why and how. *Journal of Quality Technology* 2009; **41**:340–361.
10. Ju HL, Lucas JM. L^k factorial experiments with hard-to-change and easy-to-change factors. *Journal of Quality Technology* 2002; **34**:411–421.
11. American Supplier Institute. Supplier Symposium on Taguchi Methods. American Supplier Institute: Romulus, 1984.
12. Robinson T, Anderson-Cook CM, Hamada MS. Bayesian analysis of split-plot experiments with nonnormal responses for evaluating non-standard performance criteria. *Technometrics* 2009; **51**:56–65.
13. Wu CFJ, Hamada MS. Experiments: Planning, Analysis and Optimization (2nd edn). John Wiley and Sons, Inc.: New York, 2009.
14. McCullagh P, Nelder JA. Generalized Linear Models (2nd edn). Chapman & Hall: London, 1989.
15. R Development Core Team (2004). R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. (http://www.R-project.org) (accessed on 25 August 2015).
16. Montgomery DC. An Introduction to Statistical Quality Control (6th edn). John Wiley & Sons: New York, 2009.
17. Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2010). WinBUGS Version 1.4.3 user manual. http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml (accessed on 14 July 2015).
18. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian Data Analysis (2nd edn). Chapman and Hall: Boca Raton, 2003.

Authors' biographies

Michael S. Hamada is a Scientist at Los Alamos National Laboratory and holds a Ph.D. in Statistics from the University of Wisconsin-Madison. He is a Fellow of the American Statistical Association.

Stefan H. Steiner is a Professor and Chair of Ph.D. from McMaster University. He is a Fellow of the American Society for Quality and the American Statistical Association.

R. Jock Mackay is a Professor Emeritus (retired) from the Department of Statistics and Actuarial Science at the University of Waterloo.

C. Shane Reese is a Professor in the Department of Statistics at Brigham Young University and holds a Ph.D. in Statistics from the Texas A&M University. He is a Fellow of the American Statistical Association.