Quality Engineering
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lqen20

Statistical Engineering and Variation Reduction

Stefan H. Steiner [a] & R. Jock MacKay [a]
[a] Business and Industrial Statistics Research Group, Department of Statistics and Actuarial Sciences , University of Waterloo , Waterloo , Ontario , Canada
Published online: 11 Dec 2013.

To cite this article: Stefan H. Steiner & R. Jock MacKay (2014) Statistical Engineering and Variation Reduction, Quality Engineering, 26:1, 44-60, DOI: 10.1080/08982112.2013.846069

To link to this article: http://dx.doi.org/10.1080/08982112.2013.846069

Quality Engineering, 26:44–60, 2014

ISSN: 0898-2112 print/1532-4222 online
DOI: 10.1080/08982112.2013.846069

Taylor & Francis
Taylor & Francis Group

Statistical Engineering and Variation Reduction

Stefan H. Steiner,
R. Jock MacKay

Business and Industrial Statistics Research Group, Department of Statistics and Actuarial Sciences, University of Waterloo, Waterloo, Ontario, Canada

ABSTRACT Statistical engineering as proposed by Hoerl and Snee (2010a) aims to develop a discipline devoted to better understanding how to use statistical tools to support project goals. Existing examples abound but more work is needed. We discuss the use of statistical engineering to improve problem solving—that is, reducing variation in processes—and note that this requires a series of empirical investigations where we should use information gained to help plan subsequent investigations. The systematic use of prior/existing information, especially baseline information, in problem solving is illustrated using a crossbar dimension case study. The baseline results are used to help plan and analyze all subsequent investigations both when looking for a dominant cause of the variation and when assessing a possible solution. The effective use of prior statistical information and the consequences of its use in the variation reduction context are not commonly taught and thus opportunities for more efficient problem solving are lost.

KEYWORDS baseline, families of causes, full extent of variation, measurement system assessment, method of elimination, process improvement, sequential learning, six sigma

INTRODUCTION

Hoerl and Snee (2010a, 2010b) proposed a new discipline they termed *statistical engineering* (SE). They defined SE as "the study of how to best use statistical concepts, methods and tools, and integrate them with IT and other relevant sciences to generate improved results" (2010a, p. 52).

This broad definition suggests that students of SE should look for better (if not the best) ways to apply statistical methods. See also the panel discussion in Anderson-Cook and Lu (2012a, 2012b) for more examples and opinions about SE.

Process improvement is an important and rich context in which to think about SE and its development and consequences. Hoerl and Snee (2001) provided the foundation for process improvement through the principles of statistical thinking:

1. All work occurs in a system of interconnected processes, where a *process* is a chain of activities that turns inputs into outputs.
2. Variation, which gives rise to uncertainty, exists in all processes.

Article presented at the First Stu Hunter Research Conference in Heemskerk, Netherlands, March 2013.

Address correspondence to Stefan H. Steiner, Business and Industrial Statistics Research Group, Department of Statistics and Actuarial Sciences, University of Waterloo, 200 University Ave. West, Waterloo, ON, Canada N2L 3G1. E-mail: shsteine@uwaterloo.ca

3. Understanding and reducing variation are keys to success.

We could quibble that some processes can be improved by shifting the average (e.g., increasing the yield in a chemical process), but experience has shown that there are ample opportunities to reduce variation and hence make improvements. To highlight the importance, Neave (1990, p. 57) attributed the following quote to Deming: "If I could reduce my message to management to just a few words, I'd say it all has to do with reducing variation."

For mass-produced components and assemblies, reducing variation can simultaneously lower overall cost, improve function, and increase customer satisfaction with the product. Excess variation can have dire consequences, leading to scrap and rework, the need for added inspection, customer returns, impairment of function, and a reduction in reliability and durability.

So let us concentrate on how SE can help with variation reduction. Juran and Gryna (1980) provided the basic two-step algorithm for understanding and reducing variation (Principle 3) via the diagnostic and remedial journey shown in Figure 1.

Within Six Sigma (Breyfogle 1999), define–measure–analyze–improve–control (DMAIC) fleshes out this algorithm. There are many other such expansions. For improving a medium- to high-volume manufacturing processes, Steiner and MacKay (2005) developed the version shown in Figure 2 that we apply to the case study given later in this article.

Unfortunately, statistical engineering is a term that has been used repeatedly with different meanings. Most recently there is the general concept of SE as introduced by Hoerl and Snee (2010a, 2010b) and

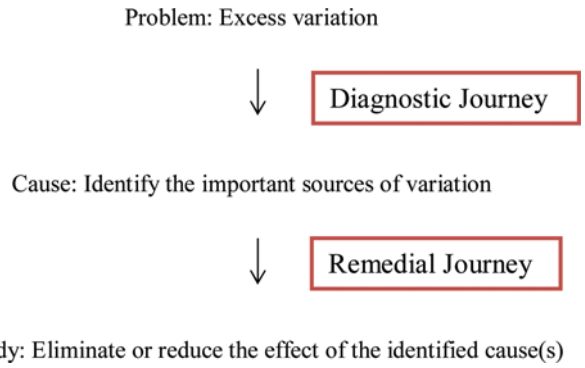


FIGURE 1 Diagnostic and remedial journey. (Color figure available online.)

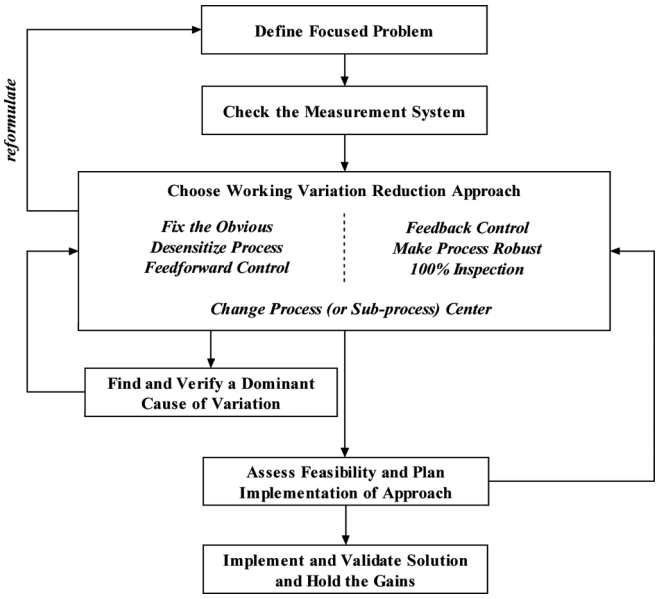


FIGURE 2 Statistical engineering (StatEng) variation reduction algorithm.

discussed earlier in this article. Steiner and MacKay (2005) also chose the name statistical engineering for their process improvement algorithm. To limit possible confusion, we use the acronym StatEng to refer to this algorithm and its application. SE as described by Hoerl and Snee is a more general concept. Statistical engineering is also a name associated with the Shainin Red X problem-solving system (Shainin 1993). The StatEng algorithm builds on some of the ideas in the Shainin system.

Where does statistics come in then? What is its purpose? We recommend that you take the broad view of the discipline of statistics that provides the concepts, methodologies, tactics, and tools for empirical learning. Note that *learning* is the key word in the previous sentence. The purpose of statistics is to learn, in either an exploratory or confirmatory sense. *Empirical* means by observation or experiment; that is, we learn about the process by watching it without intervention (observational), after changing one or more inputs (experimental) or some combination of the two. Because applying statistics is inductive, learning from an empirical investigation is often imperfect; that is, we are left with some uncertainty.

Consistent with the principles of statistical thinking, Deming purportedly said, "If you can't describe what you are doing as a process, you don't know what you're doing." We believe that there is great value in also applying process thinking to the planning and

execution of any empirical investigation (i.e., statistics). At a high level, we suggest the five step process called QPDAC. The steps of QPDAC are as follows:

Question:	Specify what you are trying to learn.
Plan:	Formulate what, when, and how you will collect the data.
Data:	Execute the plan.
Analysis:	Examine the data in light of the question and plan.
Conclusion:	Specify what has been learned with limitations (uncertainty, deviations from the plan, etc.).

See Steiner and MacKay (2005) and MacKay and Oldford (2000) for the details of each process step. QPDAC provides a process framework for carrying out (or criticizing) any empirical investigation.

Now we have the elements to discuss SE in the context of variation reduction. We will apply an algorithm such as DMAIC or StatEng. Our first conclusion is that it is better practice to use a sequence of empirical investigations rather than a single investigation. In most applications, it is a recipe for disaster to use a single investigation to try to identify the cause(s) of variation and, at the same time, to try to find a remedy that reduces or eliminates the effects of the causes. We look to George Box for support. The use of statistical tools to solve nontrivial problems requires "sequential learning" as described in Box (1999) and Box and Liu (1999). The main theme of these two papers can be summarized as "Investigations are conveniently conducted sequentially with results from previous experiments interacting with subject matter knowledge to motivate the next step"(Box, 1999, p. 27).

In Box's view, too much emphasis in statistics has been given to "one-shot" procedures, such as hypothesis testing and optimal designs that follow a mathematical paradigm. He felt instead that there should be more studies of statistics from a dynamic point of view. The focus on the mathematical paradigm can be partly explained by the relative ease of deriving mathematical results and many statisticians' mathematical training.

Box and Liu (1999) also noted that in the context of process improvement, there is often immediacy. That is, we can apply QPDAC to get the results of any investigation in a relatively short time frame so a sequence of investigations is feasible. The same may not be true in other contexts such as agriculture, medicine, engineering, and so on.

A famous historical example of the successful application of sequential empirical learning (and SE) is the Wright brothers' development of a heavier-than-air fixed-wing manned airplane in the years leading up to their maiden flight in 1903. The Wright brothers experimented extensively, looking for an effective way to control flight and design wings that provided sufficient lift. They employed many tools/methods including kites mounted on bicycles, a rudimentary wind tunnel, and gliders both tethered and manned. The brothers tested over 200 wing designs in their wind tunnel and conducted hundreds of unmanned and manned glider flights before they felt ready to tackle actual flight. Along the way they learned a lot but also suffered many setbacks. They even discovered an error in the assumed value of a physical constant, called the *Smeaton coefficient* of air pressure needed to calculate the expected lift from a particular wing design. The Wright brothers ultimately succeeded in part due to their use of empirical learning. Other better funded teams such as one led by Samuel Langley failed when following a much more theory-based approach. SE could have helped the brothers to reach their goal more quickly.

One well-developed application of SE is response surface methods (RSM) as initially proposed by Box and Wilson (1951) and summarized by Box and Draper (2007). Using RSM we conduct a series of experiments to try to optimize an objective function defined in terms of process outputs. With RSM, there is explicit use of initial experiments to drive further investigation. We may start with a highly fractionated two-level screening experiment that looks for inputs (factors) with large main effects. This is followed by further factorial experiments (possibly with center points) with higher resolution using promising factors found in the screening experiments. The results are then used to suggest promising areas of the design space (in terms of optimizing the objective function). Further experiments are conducted with new levels in the direction of steepest ascent (or descent). As needed, these designs are augmented with axial points to examine nonlinear and interaction effects.

The purpose of this article is to illustrate the important advantages and surprising consequences of using sequential empirical investigations (and learning) within a variation reduction algorithm such as DMAIC or StatEng. In our experience, Six Sigma books and

training material make few connections between and within the stages of DMAIC. There is no explicit use of information from previous stages to help complete the current stage. For instance, in the well-known Six Sigma book by Breyfogle (1999), few of the examples refer to anything learned in a previous stage of DMAIC. This is especially strange when moving from the analysis to the improvement stage; you would think that knowing the cause would be helpful when looking for a remedy.

To illustrate, we consider a case study that follows an improvement team through a project to reduce variation in the crossbar dimension of a plastic switch base. We show how the knowledge gained early in the improvement project drives and influences the choices made later on. The new knowledge impacts fundamental details such as sampling plans, when and what to measure, etc., of subsequent investigations. Here we focus on the use of information gained in a baseline investigation conducted (in part) to assess the magnitude of the problem at the start of the project.

The article is organized in the following manner. In the next section, we outline the importance and the use of the knowledge gained in the baseline investigation to help plan and analyze subsequent investigations. Next, we discuss an appropriate plan and analysis of a baseline investigation that takes into account the proposed uses. This is then followed by a series of sections that illustrate the use of the knowledge gained in the baseline investigation to assess a measurement system, find the dominant cause(s) of the variation, verify the identified dominant cause, and assess the feasibility of various variation reduction remedies. We conclude with a summary and some additional discussion.

BASELINE INVESTIGATION

Establishing a baseline is the first step in most variation reduction algorithms. For example, it is one of the necessary activities in the measure stage of DMAIC in Six Sigma (Breyfogle 1999). It is also the first stage of the StatEng algorithm (Steiner and MacKay 2005) illustrated in Figure 2.

We define the *baseline* as a numerical and graphical summary of the current process performance. In other words, the baseline quantifies the size and nature of the process variation. The baseline may come from data previously collected such as weekly scrap rates or stored values from an end-of-line 100% inspection. However, in many instances, we may decide to carry out an empirical investigation to establish the baseline.

We propose to use the baseline to help

- set the goal—i.e. determine how big a reduction in variation is required;
- validate a potential solution if and when one is found; and
- plan and analyze subsequent investigations when searching for a cause or a solution.

The first two uses are commonly acknowledged. However, it is our contention that, unlike most current practice, the information gained in the baseline can be exploited in planning and analyzing subsequent investigations designed to gain the process knowledge necessary to meet the project goal. Some may argue that this is common sense; we should always use any prior information as a guide; that is, use sequential learning when planning any investigation. However, in our experience, mistakes and oversights are common in practice. In addition, explicitly acknowledging the intended use of the baseline suggests a particular plan for the baseline investigation itself. We give our recommendations in the next section.

We use the StatEng algorithm (Figure 2) to illustrate the benefits of using the baseline information in variation reduction. We hope that in the future project teams will make more systematic use of the baseline information and achieve better results in less time.

Here are some details about the case study. In the manufacture of the injection-molded plastic base shown in Figure 3, there was excessive variation in a key crossbar dimension, measured as a difference from a nominal value. With rescaling, the target dimension was 1.0 inch and the specifications were 0 to 2.0 thousandths of an inch (thou). Note that *thou* is often referred to as *mil* in the United States. In a later assembly process, many mechanical and electronic components are inserted into spaces in the plastic base. Problems occurred due to both breakage when spaces were too small and loose assembly when spaces were too large. The crossbar dimension of the plastic base was used as a surrogate for all of

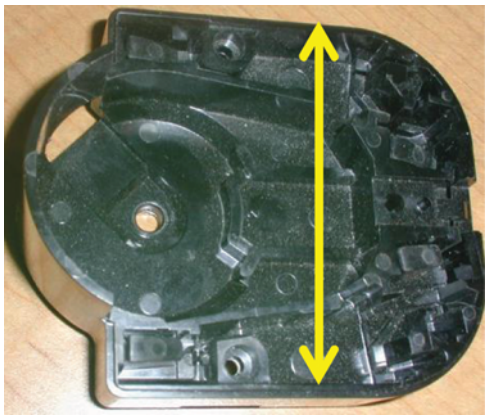


FIGURE 3 Plastic base showing crossbar dimension with arrow. (Color figure available online.)

the internal dimensions. If crossbar dimension was small (large), the spaces were generally too big (small). The project goal was to reduce variation in the crossbar dimension.

PLANNING AND ANALYZING A BASELINE INVESTIGATION

To determine the baseline, we need an empirical investigation to estimate the long-term properties (mean, standard deviation, etc.) of the critical process output(s). For the purposes of illustration, we assume a single output of interest and that a performance measure is given. There are many feasible choices for a performance measure—standard deviation, capability ratio, etc. The choice helps to define the question in QPDAC for this investigation. If possible, the process output should be a continuous rather than a binary characteristic because that provides more process information per observation. In addition, we would rather work with an output that has two-sided specification limits and is not already a measure of variation itself like out-of-roundness.

We propose a plan for the baseline investigation that is designed to help progress through the StatEng variation reduction algorithm (Figure 2). Specifically, to accomplish the goals, the baseline investigation should allow us to

- estimate the long-term performance measure,
- estimate the full extent of variation (denote FEoV) in the output, and, perhaps most critically,
- determine the nature of the output variation over time.

We define the FEoV as the range within which the vast majority of output values lie. The range (minimum to maximum) defines the FEoV when the sample size is in the hundreds and there are no wild outliers (as in the case study). More generally, for a histogram with a bellshape, the FEoV corresponds to the range of output values given by the average plus or minus three times the standard deviation. This way the FEoV covers 99.7% of output values using a Gaussian assumption. To define the FEoV we ignore rare outliers. For binary and discrete outputs the FEoV is given by all of the output values seen in normal production.

To accomplish the baseline investigation goals, the sampling scheme is critical. First we must decide over what time frame we will sample. This study period must capture the long-run performance of the process characteristic of interest. To help decide, we use any prior knowledge and/or experience about the process we have. For example, if process performance is already summarized using weekly scrap rates (but we decide we want a baseline for a continuous output characteristic), we can use the pattern in the scrap rates to help decide the time frame.

Instead of random sampling, we recommend a systematic sampling plan that includes consecutive parts and parts sampled from the process at regular time intervals. Such a systematic sampling plan is desirable because it provides information about the time nature of the output variation. This proposed plan can be thought of as a multi-vari investigation focused on the time families of variation. See Snee (2001) and De Mast et al. (2001) for more details on multi-vari investigations. In this light, our suggestion for the baseline investigation is similar to the suggestion in Shainin (1993) to start problem solving with a multi-vari investigation. An alternative is to use a random sample over the proposed time frame and keep track of when each observation is made.

We can describe the nature of the output (or any other process characteristic) variation over time using the idea of a time family of variation. If the output changes quickly (that is, over a short time frame we observe values across most of the FEoV), we say that the output variation acts in the part-to-part family. On the other hand, if the output changes slowly—for example, to see values on both ends of the FEoV we need to measure parts separated by a

long time; for example, days—we say that the output variation acts day-to-day. By sampling parts consecutively at regular intervals we are able to distinguish between situations where the output varies quickly (part-to-part) or slowly (say, day-to-day) or somewhere in between. This information is valuable both to help us choose the time frame for subsequent investigations and to give us clues about the possible major causes of variation.

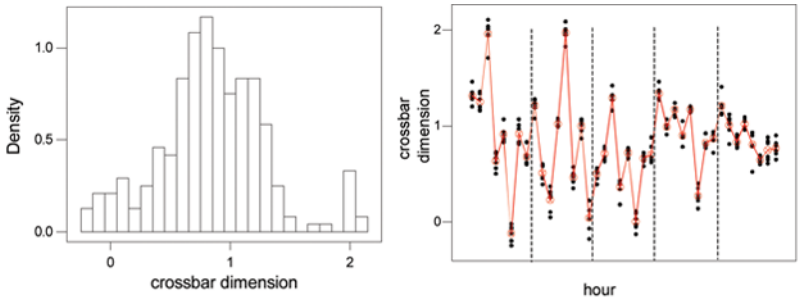
In the crossbar dimension example, the team planned and executed a baseline investigation where six consecutive parts were selected from the process each hour for 5 days. This choice was expected to provide ample time for the process output to vary over its normal range and give a large enough sample size to reasonably estimate the process variation. We provide graphical and numerical summaries of the data in Figure 4. We suggest always using both a histogram and some sort of run chart. The right panel in Figure 4 gives a multi-vari chart that illustrates how crossbar dimension varies over time. The six consecutive values each hour are plotted at the same horizontal location. The vertical dashed lines show the division into the 5 days.

From the graphical and numerical summaries of the data in Figure 4, we see that the FEoV of crossbar dimension variation is −0.25 to 2.1 thou (as indicated by the dashed lines in subsequent figures) and the major source of variation acts hour-to-hour with some evidence of day-to-day differences. More formally, we also fitted a nested analysis of variance (ANOVA) model. In agreement with the multi-vari chart, the dominant source of variation is among hours (with variance component standard deviation equal to 0.45). The variation in crossbar dimension for consecutive parts is small. The standard deviation of the baseline data is 0.45 thou. The team set the goal to reduce the standard deviation to less than 0.25 thou. There was no immediate explanation for the smaller variation in crossbar dimension observed on the fifth day. Note that had there been a large day effect—that is, had the day averages been very different—the baseline investigation was (probably) not conducted over enough days to capture the long-term performance. In that case, the team should collect data over some additional days before drawing conclusions.

One of the goals of the baseline investigation is to estimate the current process performance in terms of the output variation. Estimating a measure of variation like a standard deviation is difficult with a small sample size. In addition, because of the multitude of uses we make of the results of a baseline investigation, we favor a large baseline sample size, ideally consisting of hundreds of parts for a continuous output characteristic and thousands of parts for a binary characteristic.

Due to the time nature of the crossbar dimension variation, the team concluded that the time frame for further observational investigations should be hours and days. We expect to see the FEoV in the output over that period. Investigations conducted over a shorter time frame, say, only an hour, would not show the FEoV and thus not reflect the long-term behavior of the process.

One final point about the outcome of the baseline investigation is that we recommend that several parts with extreme values (i.e., to span the FEoV) be set aside because they can be useful in subsequent studies, such as the measurement system assessment investigation discussed in the next section.



Variable	N	Mean	SD	Minimum	Q1	Median	Q3	Maximum
Dimension	240	0.8383	0.4497	−0.2500	0.6025	0.8300	1.0900	2.1100

FIGURE 4 **Histogram, multi-vari chart, and numerical summary for crossbar dimension baseline data. (Color figure available online.)**

Next, we illustrate the use of the baseline in subsequent investigations needed at various stages of the StatEng algorithm.

USING THE BASELINE TO HELP CHECK THE MEASUREMENT SYSTEM

After establishing the baseline, the next step in the StatEng algorithm (Figure 2) is to assess the measurement system for the output. The goal of this investigation is to compare the size of the measurement variation, denoted. σ_{meas}, to the process variation, denoted $\sigma_{process}$. We want to determine whether the measurement system is a large source of variation and whether it is adequate to support further process investigations. If the measurement variation is large, improving the measurement system is necessary before proceeding with problem solving and may solve the original problem. Note that for this reason, in many other problem-solving systems, checking the measurement system is often recommended *before* we conduct a problem baseline investigation. However, we propose establishing the baseline first because we use the results from the baseline investigation to help plan and analyze a better measurement system assessment investigation. This is a small example of SE where reversing the order of the two investigations can increase efficiency.

A generic plan for measurement assessment is to measure the same parts repeatedly over a variety of conditions and times. We plan to use the baseline estimate of the overall variation (i.e., the combined effect of the process and measurement) to improve the precision of the conclusion about the relative size of the measurement variation. If we assume independence—that is, the part dimension does not affect the measurement variation—we have $\sigma_{overall} = \sqrt{\sigma^2_{process} + \sigma^2_{meas}}$. The measurement investigation will provide an estimate for σ_{meas}, and combining that with the estimate for $\sigma_{overall}$ given by the baseline allows us to solve for $\sigma_{process}$.

In the measurement system assessment investigation, we suggest selecting three parts chosen (from the baseline) to cover the FEoV for the output observed in the baseline. We select one large, one small, and one intermediate-sized part. The benefits of choosing extreme parts were explored in more detail by Browne et al. (2009, 2010), who also proposed a more complicated analysis that incorporates the measured part size from the baseline investigation used to select the parts. Note the difference from the usual suggestion in gage repeatability and reproducibility (R&R) investigations for 10 randomly selected parts (Automotive Industry Action Group 2010). The traditional gage R&R estimates both σ_{meas} and $\sigma_{process}$ using only the measurement investigation data.

In the crossbar dimension example, the three (small, medium, and large) parts were measured nine times each on two separate days. If the measurement system is a dominant source of the variation, based on what we observed in the baseline, we expect to see the FEoV within the measurements on each part over the 2 days. The results are shown graphically in Figure 5 and the one-way ANOVA numerical results are provided in Table 1.

In Figure 5 we added horizontal dashed lines to show the output FEoV(−0.3 to 2.1) seen in the baseline. Later we continue this suggestion and always include lines showing the FEoV in all plots of individual output values. This practice helps ensure that the plots are interpreted in an appropriate way when we want to try to explain the FEoV as seen in the baseline. In this investigation, because we deliberately selected extreme parts from the baseline, we will always see the FEoV. However, this is not necessarily true with other investigations. Because the measurement assessment investigation repeatedly measured parts, the error variance in the ANOVA corresponds to measurement error. Thus, from the ANOVA, we find $\hat{\sigma}_{meas} = \sqrt{0.020} = 0.14$ (given by the pooled SD

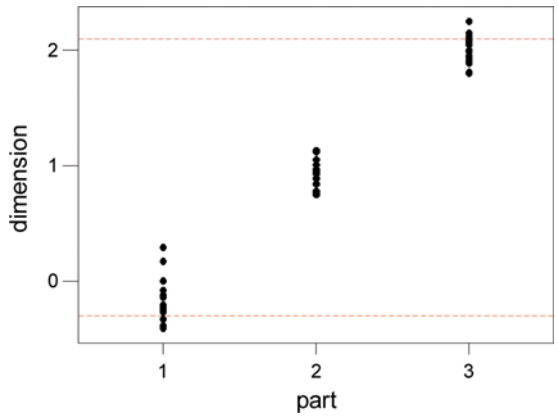


FIGURE 5 **Crossbar dimension measurement investigation results. Dashed horizontal lines show the output FEoV from the baseline. (Color figure available online.)**

TABLE 1 One-way ANOVA: Dimension Versus Part

Analysis of variance for dimension

Source	df	SS	MS	F	P
Part	2	42.5111	21.2556	1,064.87	0.000
Error	51	1.0180	0.0200		
Total	53	43.5291			

Individual 95% confidence intervals for mean based on pooled SD

Level	N	Mean	SD		
1	18	−0.1722	0.1797	(*)	
2	18	0.9222	0.1166	(*)	
3	18	2.0011	0.1183	(*)	
Pooled SD	0.1413	0.00	0.70	1.40	2.10

SS: Sum of squares. MS: meansquare.

or the square root of the mean squared error). The estimated baseline standard deviation was $\hat{\sigma}_{overall}$=0.45. Thus, we estimate $\hat{\sigma}_{process} = \sqrt{0.45^2 - 0.14^2} = 0.43$. Because the measurement variation is small relative to the process variation, we conclude that the measurement system is adequate for the project. The system can distinguish between the three parts and the measurement variation is relatively small. Some may complain that, unlike with traditional measurement assessment studies, our investigation was conducted over two separate days rather than as quickly as possible. At the time the measurement investigation was conducted, we spread the investigation out over many hours because the baseline investigation results suggested that this was needed to generate the FEoV. However, we now realize (continuous improvement) that conducting the measurement assessment investigation as quickly as possible is preferred because we do not need to worry about generating the FEoV in the measurement investigation if we select extremes parts from the baseline (as we have done). We could clarify this idea by adding the initial measurements for each part with a special symbol to a plot like Figure 5.

The proposed assessment plan is different than the traditional gage R&R investigation (Automotive Industry Action Group 2010) with 10 randomly selected parts measured four to six times each. We can use fewer parts in our investigation because we have an estimate of the overall variation from the earlier baseline investigation. The benefit of the proposal can be quantified as in Stevens et al. (2010 and 2013) using the asymptotic precision of the estimator for $\gamma = \sqrt{\sigma^2_{meas}/\sigma^2_{overall}}$ obtained as a linear approximation from the Fisher information. In Figure 8 we compare the approximate standard deviation of the estimator for γ (when the true value is 0.2) for three different plans defined in terms of (k, n) where k represents the number of randomly selected parts and n is the number of repeated measurements per part. The three selected plans all have a total of 60 measurements and correspond to a plan similar to the one used in the case study; that is, (3,20), the standard gage R&R plan (10,6), and the plan (30,2) proposed by Shainin (1993). Our proposed measurement assessment plan and analysis will provide slightly worse results than shown in Figure 8 for the (3,20) plan because we did not select the three parts at random. However, if we adopt the more complicated analysis proposed by Browne et al. (2009, 2010) that incorporates the baseline part measurements, we can do substantially better. In addition, selecting extreme parts will make it easier to assess the model assumption that measurement variation does not depend on part size.

From Figure 8 we see that when there is no baseline data (i.e., $b = 0$) the Shainin plan has the lowest standard error at about 0.035. However, as we add baseline information the proposed plan with only three parts quickly becomes the best one. With a baseline sample size of $b = 240$ (just off the right-hand edge of Figure 8) the proposed measurement investigation should have a standard error of a little more than 0.02, which is less than half as big as using the traditional gauge R&R plan with no baseline data that has a standard error of more than 0.45. In summary, Figure 8 shows the substantial benefits of the baseline information (for all plans) and how for a reasonable baseline size (say, greater than 50 parts) the proposed plan with three selected parts measured 20 times each is the best. We see similar results for other values of γ.

USING THE BASELINE TO HELP SEARCH FOR A DOMINANT CAUSE

Following the diagnostic and remedial journey (Juran and Gryna 1980), the next step in the StatEng algorithm is to identify one or more dominant causes of the variation. A dominant cause is a process input that, if held

fixed, would substantially reduce the variation in the output. Assuming independence, we can partition the variation in the process output into two parts:

$$\sigma_{\text{overall}} = \sqrt{\sigma^2_{\text{due to specific cause}} + \sigma^2_{\text{due to all other causes}}}$$

We discussed a special case of this formula in the measurement assessment section. The notion of a dominant cause uses the Pareto principle applied to causes (Juran and Gryna 1980). For a dominant cause, the residual variation—that is, $\sigma_{\text{due to all other causes}}$ — must be relatively small; that is, $\sigma_{\text{due to all other causes}} << \sigma_{\text{due to specific cause}}$. Figure 9 shows the percentage reduction in the overall variation possible if we eliminate the contribution due to a specific cause. We see that little improvement is possible unless we reduce the contribution of a cause that is dominant. For instance, suppose that we find a cause that accounts for half the overall variation (on the standard deviation scale). Then, in the unlikely event that we are able to completely eliminate the effect of this cause, we reduce the overall variation by only about 14%. Figure 9 also suggests that if the problem is defined by multiple large causes and not a dominant cause it will be more difficult to solve. In such cases we will need to address a number of large causes to make a substantial difference and it will be much more difficult to identify any large cause due to the masking effect of the other large causes.

In searching for a dominant cause, we use the baseline in several ways:

- The results of a baseline investigation can be used to eliminate many inputs as suspect dominant causes because we determined the contribution of some time families to the output variation. If a dominant cause exists, it must act in the time family that is the largest source of variation. For instance, if the output varies slowly (say, hour to hour), then any input that changes from part to part cannot be a dominant cause.
- The baseline suggests a time frame for the plan of any observational investigation designed to look for a dominant cause. We want to collect data over a long enough time period (or in such a way) to be sure that the dominant cause acts during the investigation.
- We can use the FEoV to *check* that the dominant cause has acted during the investigation. There is no sense in finding causes that explain only a small part of the output variation. If the output variation in an investigation does not closely match the FEoV seen in the baseline, we conclude that the dominant cause did not act. Then, it is not possible to generate strong clues about the identity of the dominant cause using the investigation results.

In the case study, what clues about the dominant cause are provided by the baseline investigation? We know that the dominant cause must vary the same way over time as the output crossbar dimension. The dominant cause is thus not an input that varies quickly, say, part-to-part, such as cavity or mold number. Otherwise, we would not have seen the pattern of variation in the crossbar dimension in the right panel of Figure 4.

To search for a dominant cause, the team planned an investigation where they measured five varying inputs and the crossbar dimension on 40 parts haphazardly selected over a 2-day period. The five inputs were all thought to be possible substantial causes and all varied to match the pattern observed in the baseline; that is, all five inputs were expected to vary over hours. The investigation was conducted over 2 days because the baseline investigation suggested that we should see the FEoV within that time.

The input–output investigation results are summarized using the two scatterplots of an input versus the crossbar dimension output given in Figure 6. The plots for the remaining three inputs showed no pattern; that is, they looked similar to the left panel of Figure 6. In the scatterplots, the horizontal dashed lines give the FEoV seen in the baseline. First, we conclude that the dominant cause acted in the

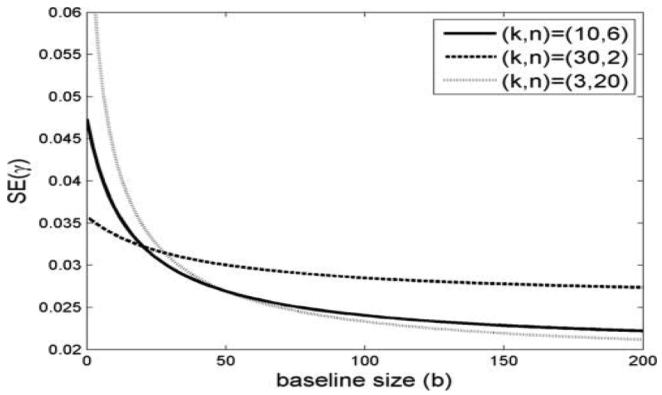


FIGURE 6 **Effect of baseline size (b) on the precision of the estimator for γ where $\gamma = \sqrt{\sigma^2_{meas}/\sigma^2_{overall}} = 0.2$.**

investigation because variation in the crossbar dimension seen in the 2 days of sampling was close to the FEoV. Second, we see that barrel temperature is a strong suspect for the dominant cause. If we could hold barrel temperature fixed, (it appears) that there would be much less variation in crossbar dimension. The other four inputs were eliminated as possible dominant causes. Note that at this point we could fail to find a dominant cause if it is measured with large measurement variation. We should ideally check the measurement systems for all inputs (suspected dominant causes) in a way similar to the measurement assessment investigation for crossbar dimension we conducted earlier.

Note the contrast between the observational studies and the typical brainstorming and screening experiment approach suggested in many implementations of Six Sigma (Breyfogle 1999). Observational plans are preferred because they are usually cheaper and easier to conduct than an experimental investigation where we must select, and deliberately set, one or more (normally varying) process inputs. We suggest using an experimental plan, as described in the next section, to verify the dominant cause only after we have generated as many clues as we can about the dominant cause with simpler and cheaper investigations.

Here we illustrated searching for a dominant cause using only an input–output investigation. There are many other types of investigations that can be useful, including disassembly–reassembly and component swap (offline) experiments, group comparison, and other simple stratification investigations (Steiner and MacKay 2005). As a rule, the aim of these investigations is to use time or location families (groups) of causes to narrow down the list of possible suspect dominant causes. Each new investigation is planned using the knowledge gained from all of the previous investigations until we (hopefully) identify a single (or small number) of remaining suspect(s). This employs the "method of elimination" popularized by Dorian Shainin (Shainin 1993; Steiner et al. 2008) that is another example of SE and should be the subject of further research. In each investigation we use the baseline knowledge in the same way as for the input–output investigation. We use the output time family to help decide on an appropriate time frame and the observed FEoV to check that the dominant cause acted in the investigation. Note that the crossbar dimension case study is not a good example of applying the method of elimination because we use only a single investigation, rather than a series of investigations, to find the dominant cause.

USING THE BASELINE TO HELP VERIFY A SUSPECT DOMINANT CAUSE

We want to be sure that the suspected dominant cause(s), here called a *suspect*, is dominant before moving to the remedial journey. We need to verify the suspect because in the search for the dominant cause using observational studies, we might have inadvertently ruled out a family of causes that contains the dominant cause or been misled by confounding. To verify that a suspect is a dominant cause, we use an experimental plan (if feasible) where the value of one or more suspects is deliberately manipulated. A verification experiment should only be considered when we have a single or only a small number of remaining suspects. That is, a verification experiment should only be used to verify clues previously attained and not to search for the dominant cause.

We also use the baseline information to help plan and analyze the verification experiment. The time nature of the output variation in the baseline helps us to

- define an experimental run,
- determine the importance of replication (i.e., choosing the number of runs), and
- determine the importance of randomization to reduce the risk of confounding in the experiment.

To draw conclusions, we compare the output variation observed in the verification experiment to the FEoV. Note that we are not primarily concerned with statistical significance. The range of values for the suspect dominant cause seen in regular production should generate (close to) the FEoV in the output if it is a dominant cause. We first illustrate these ideas using our motivating example and then draw general conclusions about how to use the baseline information when verifying a dominant cause.

In the case study, the team concluded that barrel temperature was a suspect dominant cause. They decided that verification was necessary because it

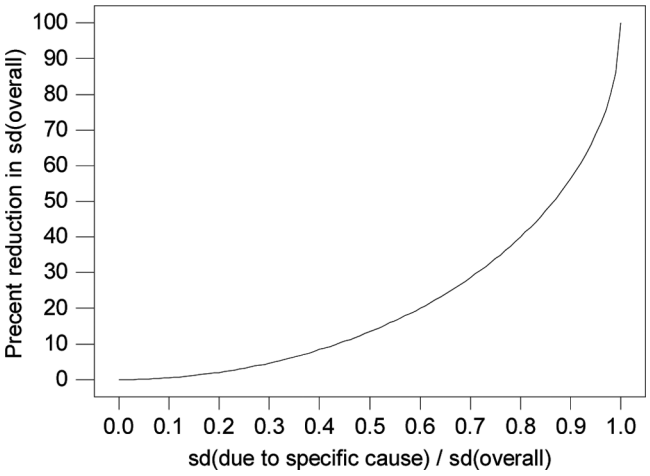


FIGURE 7 **Reduction in variation if we remove a cause contributing a given proportion of the overall variation.**

was possible that, in the earlier observational investigation, barrel temperature may have been confounded with the real dominant cause (that was not measured).

To verify barrel temperature as the dominant cause, the team planned a simple two-level experiment. They chose the low and high levels for barrel temperature as 75 and 80°C to cover the range of barrel temperatures seen in the input–output investigation (see Figure 6). Barrel temperature was difficult to hold fixed in normal production but could be controlled for an experiment. The verification experiment was conducted with only two runs, one at each of the selected barrel temperatures. For each run, the barrel temperature was set, 25 parts were made to ensure that the temperature had stabilized, and the next 10 parts were selected and measured. Then, barrel temperature was changed as quickly as possible for the second run. Using design of experiments terminology, the experiment consisted of two runs with 10 repeats per run and no replication.

We see from the experimental results in Figure 7 that barrel temperature had a large effect on crossbar dimension relative to the baseline variation. The team concluded that they had verified barrel temperature as a dominant cause of crossbar dimension variation. The small number of runs and lack of randomization was not a major concern. The earlier investigations had shown that the dominant cause acted in the hour-to-hour family and, thus, over the 30 minutes needed to conduct the verification experiment, the team felt that it was very unlikely that they would have seen the FEoV in crossbar dimension unless barrel temperature was a dominant cause. In other words, they concluded that there was insufficient time for other causes in the hour-to-hour family to change substantially during the experiment. This suggests that during the verification experiment barrel temperature could not have been confounded with any other reasonable suspect.

We now draw some general conclusions about conducting verification experiments. Assuming that the verification experiment can be conducted in a short time, if the dominant cause acts over a long time, as in the crossbar dimension example, we do not need to worry about confounding in the verification experiment. Other causes in the same time family as the suspect will not have time to vary substantially during the verification experiment. As a result, the experimental principles of replication and random assignment are not critical. On the other hand, if the dominant cause acts over a short time, we do need to worry about possible confounding between the suspect and other inputs in the same time family in the verification experiment. Then, in this case, we need a verification experiment that utilizes sufficient replication (i.e., many runs at each of

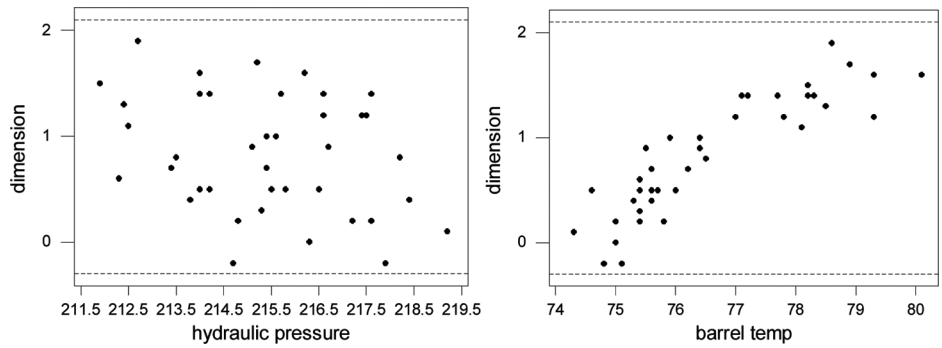


FIGURE 8 **Scatterplots of crossbar dimension by hydraulic pressure and barrel temperature. Dashed horizontal lines show the FEoV from the baseline.**

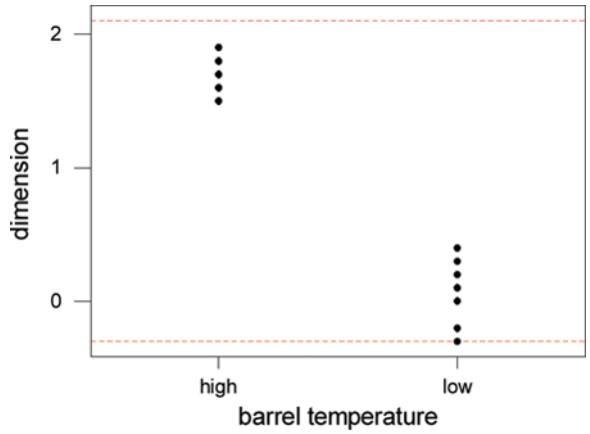


FIGURE 9 **Barrel temperature dominant cause verification experiment results. Dashed horizontal lines show the FEoV from the baseline. (Color figure available online.)**

the two levels of the suspect) and random ordering to control the risk of confounding.

USING THE BASELINE TO HELP ASSESS THE FEASIBILITY OF A VARIATION REDUCTION APPROACH

As suggested in Figure 2, there are seven possible approaches to reducing variation: fix the obvious, desensitization, feed forward control, feedback control, robustness, 100% inspection, and moving the process center. More details on these approaches are given in MacKay and Steiner (1997–1998). We can use the time nature of the output variation from the baseline to help assess the feasibility of some of these variation reduction approaches. For instance, if the output FEoV is seen over a short time, feedback control is not feasible because any observed output values provide only a poor prediction for future values.

To reduce variation, we must make some appropriate change to the process. We can add, remove, or change a processing step (inspection, controller), change some normally fixed input, or apply the StatEng algorithm upstream to reduce variation in the identified dominant cause. Here we look at improving a process by changing the level of one or more normally fixed inputs. Because the input is normally fixed, we will need an experiment to find the appropriate fixed input(s) to change and its best level. The baseline information is useful to help plan and analyze subsequent experiments designed to determine whether an approach is feasible and/or how to implement a particular approach. The time nature of the output variation seen in the baseline can help define a run. Generally, for experiments conducted to check the feasibility of a variation reduction approach, we want each run to resemble a mini baseline investigation; that is, we want each run to provide an estimate of the long-term behavior of the process with the process changes specified by the factor levels in the run. This suggests, for instance, that if the output FEoV is seen over a long time, the robustness approach (as defined in Steiner and MacKay 2005; see also the upcoming example) is likely not feasible because each run in a robustness experiment would need to be conducted over too long a time frame.

In the crossbar dimension example, the team decided that the obvious solution of reducing variation in the barrel temperature was too expensive and difficult with the existing process. Instead, they hoped that they could change the process in some other way to make it less sensitive to the variation in barrel temperature. The team then noticed the nonlinear relationship between barrel temperature and crossbar dimension in the right panel of Figure 6. As a result, they decided to raise the barrel temperature set point (average) to make the process less sensitive to barrel temperature variation. Afterwards it was straightforward to adjust the crossbar dimension average (downward to compensate for the increase that resulted from increasing the barrel temperature setpoint) by changing another normally fixed process input. However, when validating the solution, they discovered that while the crossbar dimension variation was reduced substantially, the higher barrel temperature setpoint resulted in an increase in the frequency of a mold defect called *burn*. The burn problem arises when the barrel temperature, which will still vary, is too high. The team decided to retain the crossbar dimension solution they worked hard to find and attack the burn defect as a new problem. Investigating further (details not shown here), they showed that the dominant cause of burn acted in the part-to-part family, but the specific dominant cause was not found. They suspected that the defect occurred due to variation in filling of the mold. Next the team decided to try to make the process robust to the unknown dominant cause(s) of the burn defect.

To look for a solution to the burn problem, the team planned an experiment with four factors that

TABLE 2 Factors and levels for the burn robustness experiment

Factor	Label	Low level	High level
Injection speed	A	Slow*	Fast
Injection pressure	B	1,000*	1,200
Back-pressure	C	75	100*
Screw speed	D	0.3	0.6*

*Indicates level in current process.

are normally fixed inputs: injection speed, injection pressure, back-pressure, and screw speed (rpm). These factors were selected because of their influence on fill speed and other potential dominant causes in the part-to-part family. They selected two levels for each factor as given in Table 2. Just for the experiment, the team planned to classify burn on each part into one of four categories of increasing severity. Levels 1 and 2 were acceptable, whereas levels 3 and 4 resulted in scrap. Using a single rater and boundary samples (i.e., photos of plastic bases at the agreed-upon boundaries between the burn levels), the team felt that this measurement system would add little variation. A full baseline investigation with the new burn classification system was not conducted, but because burn levels 1 through 4 had been seen in the earlier investigations, that gave the FEoV.

The team selected a fractional factorial experiment with eight runs as given in Table 3 Because there was no proper baseline investigation for the new burn problem, the team assigned the labels A, B, C, and D to the factors so that one of the treatments (treatment 5) corresponded to the current process settings.

In the resolution IV design, pairs of two-factor interactions are confounded, as given in Table 4.

The team defined a run as five consecutive parts. Because they knew from the baseline that the time family of variation containing the dominant cause (of burn) was part-to-part, they hoped that the dominant cause would act within each run in the planned robustness experiment. Deciding to use only five parts for each run was a great risk. Having more parts would have made it more likely that each run would reflect the long-term behavior of the process but would have cost more time and money. Each run was carried out once the process stabilized after changing the values of the factors. The order of the runs was randomized. The results from this robustness experiment are given in Table 3.

We plot the individual burn scores against treatment number in Figure 10. Because the data are discrete, we add jitter in the vertical direction. Examining the results, we see that treatments 2 and 3 are promising and look much better than the existing process performance as given by treatment 5. It is a bit worrisome, but not surprising given the run size, that we did not see the FEoV (scores from 1 to 4) in the treatment 5 run.

The team used average burn as the performance measure for the formal analysis and looked for process settings that made the performance measure as small as possible. We can think of this as reducing variation in the burn score about the ideal score of zero. Fitting a full model with all possible effects (four main and three two-way interactions) we get the Pareto plot of the effects for the average burn score in Figure 11. Note that in Figure 11 the factor labels arbitrarily show only the first of the pairs of aliased effects as given in Table 4. We see that only factor C (back-pressure) has a large effect. In

TABLE 3 Experimental Plan and Data for the Burn Robustness Experiment

Treatment	Order	Injection speed (A)	Injection pressure (B)	Back-pressure (C)	Screw speed (D)	Burn scores	Average burn
1	4	Slow	1,000	75	0.3	1, 2, 1, 1, 1	1.2
2	8	Fast	1,000	75	0.6	1, 1, 1, 1, 1	1.0
3	2	Slow	1,200	75	0.6	1, 1, 1, 1, 1	1.0
4	3	Fast	1,200	75	0.3	1, 2, 2, 2, 2	1.6
5*	5	Slow	1,000	100	0.6	1, 3, 2, 2, 1	2.2
6	7	Fast	1,000	100	0.3	3, 3, 2, 2, 4	3.4
7	1	Slow	1,200	100	0.3	1, 1, 1, 2, 2	2.0
8	6	Fast	1,200	100	0.6	2, 2, 4, 3, 2	3.2

*Treatment 5 uses the current process levels.

TABLE 4 Robustness Experiment Aliasing Structure

A + BCD
B + ACD
C + ABD
D + ABC
AB + CD
AC + BD
AD + BC

drawing this conclusion the team assumed that the three-input interaction (ABD) aliased with C was negligible. Checking Table 3 we see that low level of back-pressure gives less burn on average than the high level and that the results appear better than the baseline for the existing process. The team decided to address the burn defect problem by reducing the back-pressure to 75 and leaving the other fixed inputs at their original values.

We need to be careful drawing conclusions from the experiment designed to look for a remedy. We want to select new settings for one or more of the experimental factors that results in better performance than we saw in the baseline. We are not simply looking for a significantly large effect in the experiment. If other settings had been very poor, then a factor might be significant even if both levels result in a process that is worse than the current process settings. The baseline results again provide the appropriate comparison. We could have added the horizontal lines showing the baseline FEoV for burn to Figure 10, though here it does not help much because the output has only four possible values.

Suppose that the team had been able to identify the dominant cause of burn that acts in the part-to-part

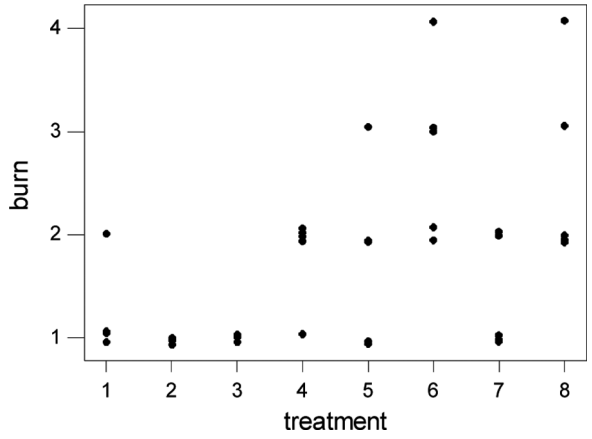


FIGURE 10 Burn by treatment plot for burn robustness experiment with added vertical jitter.

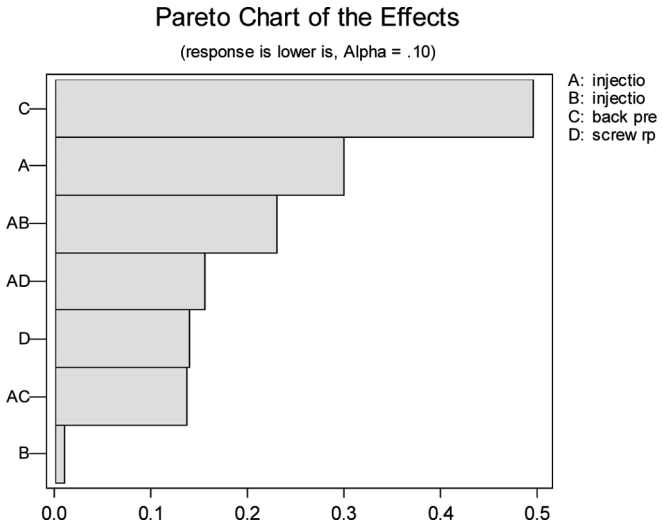


FIGURE 11 Pareto plot of input effects on average burn score.

family. Then, if that cause could be controlled in an experiment, but not easily in the regular process, it would make sense to use a desensitization rather than a robustness experiment (Steiner and MacKay 2005). The goal of the desensitization experiment is the same as in the robustness experiment, namely, we want to see whether changing the level of one or more fixed inputs can make the process less sensitive to variation in the dominant cause. However, with a desensitization experiment the team would deliberately manipulate *both* of the dominant cause(s) and the candidate fixed inputs. In this way, with the desensitization experiment, we observe the variation due to the dominant cause acting explicitly rather than implicitly as with the robustness experiment. This will make finding a better process (if one exists) easier and more reliable because we no longer have to rely on the five repeats to provide a measure of the process variability. As a side point, note that knowledge of the dominant cause may also have suggested other (better) choices for the experimental factors (fixed inputs) than given in the example. For further comparison of robustnes and desensitization experiments see Asilahijani et al. (2010).

To finish the project, the team conducted a validation investigation with the new process settings. They produced 300 parts over a number of hours and measured both the crossbar dimension and the burn defect score. The standard deviation of the crossbar dimension was 0.23 thou and only two parts were scrapped for the burn defect. The team recommended the new settings for the back-pressure and

the target barrel temperature that resulted from investigating the two problems.

SUMMARY AND DISCUSSION

In SE, we look for better ways to use statistics to achieve a specified goal. In the important context of variation reduction (using DMAIC or StatEng) of existing medium- to high-volume processes we suggest that starting with a well-designed baseline investigation is an improvement over most current practice. We showed how the results of the baseline investigation can be of great help in subsequent investigations, first looking for a dominant cause of the variation and then looking for the remedy. The key feature of the proposed baseline investigation is the recommendation to sample parts from the process systematically over time. From the baseline data we quantify the magnitude of the problem, determine the FEoV in the output, and the time nature of the output variation.

The baseline knowledge is helpful in the planning and analysis of subsequent investigations to

- assess the measurement system,
- search for and verify a dominant cause,
- assess a variation reduction approach (i.e., search for a solution), and
- validate a proposed solution.

To meet these goals, the baseline investigation should consist of a reasonably large sample size (e.g., hundreds of observations for a continuous output) so that we can well quantify the output FEoV and standard deviation. We also recommend that the baseline investigation use a systematic (rather than random) sampling plan that allows us to identify how the output variation acts over time. The time nature of the output variation is valuable information to help plan subsequent investigations. It can be used to

- choose an appropriate study population time frame;
- generate clues about the dominant cause of variation;
- help define a run, assess the risk of confounding, and determine the importance of the experimental principles of replication and random assignment in an experimental plan; and
- rule out some variation reduction approaches as not feasible.

The estimated performance measure and output FEoV can be used in planning and are useful when analyzing the results of any subsequent process investigation. We recommend adding lines showing the baseline output FEoV to all plots that show individual output values. Knowing the FEoV allows us to

- select extreme parts (as in the measurement assessment investigation) that must be generated by the action of the dominant cause(s),
- directly see whether the dominant cause has acted in an observational investigation, and
- determine how the process variation compares to the baseline variation in an experimental investigation.

In addition, the explicit use of the baseline FEoV in the analysis of subsequent investigations forces problem solvers to address the important difference between statistical and practical significance. In problem solving, practical significance is what matters. Comparing results to the baseline FEoV provides a direct way to determine whether any observed effects are large relative to the baseline variation. Small effects can be statistically significant while being unimportant. When searching for causes we want to find the dominant cause(s); that is, an input that explains a lot of the output variation, not one that is only statistically significant. The issue of practical versus statistical significance is even more critical when we use experiments to look for a solution. We want to find new process settings that are better than the current process rather than better than other treatments used in the experiment. One method to alleviate the concern about drawing inappropriate conclusions from an experiment is to always include a treatment with the current setting for each of the fixed inputs (though this costs a run).

This article has illustrated how using information gained in the baseline investigation can be effectively used to better plan and analyze future process investigations. The work has addressed a number of important issues and suggests many further questions related to how to better run variation reduction projects, including the following:

- How important is the power/generality tradeoff (De Mast and Lokkerbol 2012) in the choice of problem-solving system?

- How important is process stability, as defined by statistical process control?
- What are the consequences if there is no single dominant cause?
- When should blocking be used in the design of an experiment?
- What is the best way to train novice problem solvers to use sequential learning effectively?

Related to the last question, training problem solvers to effectively reduce variation is challenging. DMAIC and StatEng provide general roadmaps of how to proceed. But there are no specific recipes. Each application is different. There are always choices concerning what should be done next. Problem solvers must choose among the numerous available investigation plans, each with their own cost and likelihood of success. Novice practitioners of StatEng or Six Sigma will find making these choices difficult. It is clearly so much more than just applying the appropriate tool. Variation reduction involves conducting a series of investigations, and for each investigation we must choose an appropriate goal, study population, sample size, inputs to set (and their levels), and/or inputs to measure. In the appropriate circumstances the advantages of a more targeted variation reduction method are evident. However, even providing good examples is difficult because each step requires not just describing the goals of the current investigation but also background on information obtained in earlier investigations. To address this need we have, over a number of years, developed a virtual manufacturing process, called Watfactory (Steiner and MacKay 2009). Watfactory can be accessed through the website http://www.student.math.uwaterloo.ca/~watfacto/login.htm allows a wide variety of process investigations and possible remedies.

ABOUT THE AUTHORS

Stefan H. Steiner is Professor in the Department of Statistics and Actuarial Science as well as the Director of the Business and Industrial Statistics Research Group at the University of Waterloo. He holds a Ph.D. in business administration (management science/systems) from McMaster University. His primary research interests include quality improvement, statistical process control, experimental design, and measurement system assessment. He is a Fellow of the American Society for Quality.

R. Jock MacKay is a retired associate professor in the Statistics and Actuarial Science Department and past director of the Institute for Improvement of Quality and Productivity at the University of Waterloo. He is also an active consultant who has worked with organizations from a wide range of industries, including automotive, telecommunications, aerospace, government, and more.

REFERENCES

Anderson-Cook, C., Lu, L., Eds. (2012a). Statistical engineering—Forming the foundations. *Quality Engineering*, 24:110–132.

Anderson-Cook, C., Lu, L., Eds. (2012b). Statistical engineering—Roles for statisticians and the path forward. *Quality Engineering*, 24:133–152.

Asilahijani, H., Steiner, S. H., MacKay, R. J. (2010). Reducing Variation in an Existing Process With Robust Parameter Design. *Quality Engineering*, 22:30–45.

Automotive Industry Action Group. (2010). *Measurement Systems Analysis.* 4th ed. Southfield, MI: Automative Industry Action Group.

Box, G. E. P. (1999). Statistics as a catalyst to learning by scientific method part II—A discussion. *Journal of Quality Technology*, 31:16–29.

Box, G. E. P., Draper, N. (2007). *Response Surfaces, Mixtures, and Ridge Analyses.* 2nd ed. New York: John Wiley & Sons.

Box, G. E. P., Liu, P. Y. T. (1999). Statistics as a catalyst to learning by scientific method part I—An example. *Journal of Quality Technology*, 31:1–15.

Box, G. E. P., Wilson, K. B. (1951). On the experimental attainment of optimum conditions (with discussion). *Journal of the Royal Statistical Society, Series B*, 13:1–45.

Breyfogle, F. W., III. (1999). *Implementing Six Sigma: Smarter Solutions Using Statistical Methods* New York: John Wiley &Sons.

Browne, R., MacKay, R. J., Steiner, S. H. (2009). Improved measurement system assessment for processes with 100% inspection. *Journal of Quality Technology*, 41:376–388.

Browne, R., Steiner, S. H., MacKay, R. J. (2010). Leveraged gauge R&R studies. *Technometrics*, 52:294–302.

De Mast, J., Lokkerbol, J. (2012). An analysis of the Six Sigma DMAIC method from the perspective of problem solving. *International Journal of Production Economics*, 139:604–614.

De Mast, J., Roes, K. C. B., Does, R. J. M. M. (2001). The multi-vari chart: A systematic approach. *Quality Engineering*, 13:437–448.

Hoerl, R., Snee, R. (2001). *Statistical Thinking: Improving Business Performance.* Pacific Grove, CA: Duxbury.

Hoerl, R. W., Snee, R. D. (2010a). Closing the gap. *Quality Progress*, 43(5):52–53.

Hoerl, R. W., Snee, R. D. (2010b). Further explanation: Clarifying points about statistical engineering. *Quality Progress*, 43(12):68–72.

Juran, J. M., Gryna, F. M. (1980). *Quality Planning and Analysis.* 2nd ed. New York: McGraw-Hill.

MacKay, R. J., Oldford, R. W. (2000). Scientific method, statistical method and the speed of light. *Statistical Science*, 15:254–278.

MacKay, R. J., Steiner, S. H. (1997–1998). Strategies for variability reduction. *Quality Engineering*, 10:125–136.

Neave, H. R. (1990). *The Deming Dimension* Knoxville, TN: SPC Press Inc.

Shainin, R. D. (1993). Strategies for technical problem solving. *Quality Engineering*, 5(3):433–448.

Snee, R. D. (2001). My process is too variable—Now what do I do?: How to produce and use a successful multi-vari study. *Quality Progress*, December:65–68.

Snee, R. D. (2004). Six-Sigma: The evolution of 100 years of business improvement methodology. *International Journal of Six Sigma and Competitive Advantage*, 1:4–20.

Steiner, S. H., MacKay, R. J. (2005). *Statistical Engineering: An Algorithm for Reducing Variation in Manufacturing Processes* Milwaukee, WI: ASQ Quality Press.

Steiner, S. H., MacKay, R. J. (2009). Teaching process improvement using a virtual manufacturing environment. *American Statistician*, 63(4): 361–365.

Steiner, S. H., MacKay, R. J., Ramberg, J. S. (2008). An overview of the Shainin System™ for quality improvement (with discussion). *Quality Engineering*, 20(1):6–19.

Stevens, N. T., Steiner, S. H., Browne, R., MacKay, R. J. (2012). Gauge R&R studies that incorporate baseline information. *IIE Transactions*, 45:1166–1175.

Stevens, N., Browne, R., Steiner, S. H., MacKay, R. J. (2010). Augmented measurement system assessment. *Journal of Quality Technology*, 42:388–399.