A DATA ANALYSIS STRATEGY FOR QUALITY ENGINEERING EXPERIMENTS

Vijay N. Nair and Daryl Pregibon

Vijay N. Nair *and* ***Daryl Pregibon*** *are members of technical staff in the Mathematical Sciences Research Center at AT&T Bell Laboratories in Murray Hill, New Jersey. Mr. Nair holds a B.Econ. from the University of Malaya and a Ph.D. in statistics from the University of California at Berkeley. He joined AT&T in 1978. Mr. Pregibon holds a B.S. in mathematics from Youngstown State University and a Ph.D. in statistics from the University of Toronto. He joined AT&T in 1981.*

This paper deals with techniques for analyzing data from quality engineering experiments for optimizing a process with fixed target. We propose a structured data-analytic approach with three phases of analysis: an exploratory phase, a modeling phase, and an optimization phase. We emphasize the use of graphical methods in all three phases to guide the analysis and facilitate interpretation of results. We discuss the role of data transformations and the relationship between an analysis of transformed data and Taguchi's signal-to-noise ratio analysis. Our strategy is presented in a form that can be easily followed by users, and the various steps are illustrated by an example.

A Structured-Data Approach

Experimental design methods have traditionally focused on identifying the factors that affect the level of a production/manufacturing process. The Japanese, in particular Genichi Taguchi,[1] have demonstrated that for quality improvement, we also need to identify factors that affect the variability of a process. By setting these factors at their "optimal" levels, the product can be made robust to changes in operating and environmental conditions in the production line. Thus both the location and the dispersion effects of the design factors are of interest. The present paper deals with techniques for analyzing data from quality engineering experiments for optimizing a process with fixed target.

We propose a structured-data analytic approach with three phases of analysis: an exploratory phase, a modeling phase, and an optimization phase. The focus in the exploratory phase is on determining the need for transforming the data. Graphical methods are used to determine the type of transformation, to assess location and dispersion effects, and to detect possible irregularities in the data. In the modeling phase, standard analysis-of-variance techniques, supplemented with probability plots of estimated effects, are used to identify the important design factors. In the optimization phase, the model is interpreted, the optimal levels of the factors are determined, and a feasible factor level

combination is determined to optimize process quality.

Taguchi recommends analyzing the signal-to-noise ratio to determine the important dispersion effects. This is nearly equivalent to applying a logarithmic transformation to the data and modeling the variance of the transformed data. We propose a more general approach where data-analytic methods can be used to infer the appropriate transformation.

This paper is organized as follows. It provides an overview of our three-phase data analysis strategy and it describes an experiment conducted at AT&T Bell Laboratories to optimize the process of forming contact windows in complementary metal-oxide semiconductor circuits.[2] It then discusses each of the three phases of the data analysis strategy in detail, and uses the data from the contact window example to demonstrate the techniques. It includes some remarks on maximum likelihood techniques, the analysis of nonreplicated data, using other measures besides mean and variance, analysis of ordinal data, and software availability.

Our Strategy: Rationale and Overview

The Parameter Design Problem. Consider the *parameter design* problem for optimizing a process with a fixed target.[1,3] We are interested in designing a process whose output Y is a function of two sets of factors: design factors—factors that can be controlled and manipulated by the process engineer, and noise factors—all the uncontrollable factors including variability in the operating and environmental conditions. We use the notation $\mathbf{d}$ and $\mathbf{n}$ to denote the settings of the design and noise factors respectively. The target value for this process is fixed at t_0. If we can specify the cost when the output Y deviates from the target formally in terms of a loss function $L(Y, t_0)$, then the goal of the experiment is to determine the settings of the design factors $\mathbf{d}$ — parameter design — to minimize the average loss $E_{\mathbf{n}}\{L(Y, t_0)\}$ where $Y = f(\mathbf{d}, \mathbf{n})$.

Typically, the loss function cannot be specified precisely. Moreover, the form of the loss function could depend on the unknown transfer function $f(\cdot)$. Less formally then, the parameter design problem is to select the settings of the design factors to make the output Y as close as possible to the target t_0. In this paper, we interpret this as choosing $\mathbf{d}$ to minimize the variance of Y, $v_Y(\mathbf{d})$, subject to the constraint that the mean of Y, $m_Y(\mathbf{d})$, is as close as possible to t_0. See the discussion section for the use of other measures of location and dispersion.

The Role of Data Transformations. Before analyzing the data to determine the important design factors, we must first determine the extent to which the variance depends upon the mean. For if $v_Y(\mathbf{d}) = \gamma(m_Y(\mathbf{d}))$ for some function $\gamma(\cdot)$, then the variability of the process is completely determined by the constraint that the mean should be close to t_0. Nothing can be done to minimize the variability further.

Typically, the design factors $\mathbf{d}$ can be separated into two sets $\mathbf{d}_1$ and $\mathbf{d}_2$ so that the mean m_Y is a function of possibly both $\mathbf{d}_1$ and $\mathbf{d}_2$ while the variance v_Y can be factored as

$$v_Y(\mathbf{d}) = \gamma(m_Y(\mathbf{d}_1, \mathbf{d}_2))\sigma^2(\mathbf{d}_2) \qquad (1)$$

To determine the effects of the design factors on variability, we should restrict our attention to the component $\sigma^2(\mathbf{d}_2)$ in equation (1). The remaining contribution to the variability is determined by the constraint that the mean has to be on target. In fact, if one or more of the factors $\mathbf{d}_1$ (called adjustment factors[3]) are known a priori, we need to model only the factors $\mathbf{d}_2$ to minimize $\sigma^2(\mathbf{d}_2)$. The mean can then be brought on target by manipulating the adjustment factors.

Suppose the function $\gamma(\cdot)$ is known. Then one approach to analyzing the dispersion effects is to estimate $\sigma^2 = v_Y/\gamma(m_Y)$ and model it as a function of the design parameters. Taguchi's signal-to-noise ratio analysis is a special case of this with $\gamma(m) = m^2$. There is no reason, however, that this special case should always hold.

An alternative approach for analyzing the dispersion effects in equation (1), and that which we recommend, is to *transform* Y to $Z = \tau(Y)$ where $\tau'(Y) =$

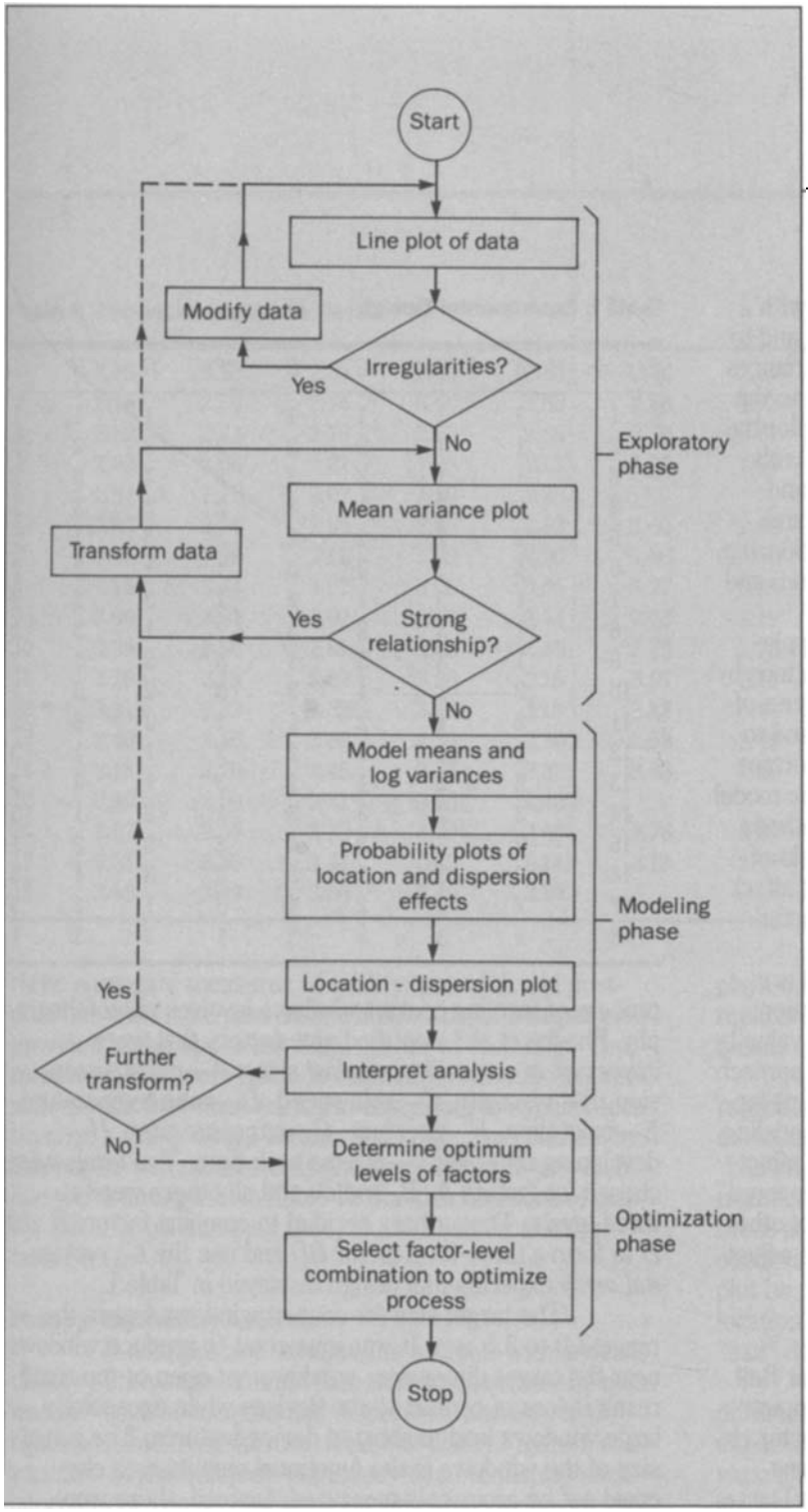

Figure 1. Flow diagram describing data analysis strategy. Arrowheads indicate the direction of flow. Diamonds represent decision points. Dashed lines indicate optional paths recommended for a thorough analysis.

$1/[\gamma(Y)]^{1/2}$. Then it can be shown by a Taylor series argument that $v_Z \approx \sigma^2$. We can then estimate the variance of Z and model it as a function of the design factors. Such transformations are known in the statistical literature as variance-stabilizing transformations.[4] Empirical evidence also suggests that often these transformations have other side benefits such as enhancing both the symmetry of the underlying (noise) distribution and the additivity of the mean as a function of the design factors. For this reason, when $\gamma(m) = m^2$ so that Taguchi's signal-to-noise analysis is appropriate, we recommend that a logarithmic transformation be applied to the data. The mean and the variance of the transformed data can then be analyzed separately to determine the location and dispersion effects.

The real difficulty lies in diagnosing the form of the unknown function $\gamma(\cdot)$ in equation (1) from the data. When there are a sufficient number of replications so that both the mean and variance of Y can be estimated with a reasonable degree of accuracy, it is possible to use data-analytic methods to determine the form of $\gamma(\cdot)$.

Overview of Strategy. Figure 1 gives a schematic representation of our recommended

analysis strategy. We begin the exploratory phase with a plot of the data to check for possible irregularities, and to visually assess the differences in the means and variances from the different factor-level settings. To determine the need for a variance-stabilizing transformation, we plot the variances versus the means. If no strong functional relationship is apparent, we do not transform the data and proceed to the next phase to model the dispersion and location effects separately. If there is a strong relationship, we use the plot to guide the choice of transformation, and repeat the process on the transformed data.

In the modeling phase, we recommend supplementing the standard analysis-of-variance computations by decomposing the effects into meaningful single-degree-of-freedom contrasts. Probability plots can then be used to determine the important design factors. If it is important to obtain good estimates of the effects, a parametric model can be fitted to the data by maximum likelihood methods.

In the optimization phase, the optimal levels of the factors that affect the dispersion and those that affect the location are determined. From this a feasible factor-level combination that optimizes process quality is obtained. When there are adjustment factors[1,3] that are known a priori, one needs to model only the dispersion effects. The mean can be made close to the target value by fine-tuning the adjustment factors. This two-step approach to optimization can also be used when the adjustment factors are not known a priori. Typically, during the modeling phase, we would discover some design factors that affect the location but not the dispersion. Given the "discovered" adjustment factors, we can first choose the levels of other factors to minimize the dispersion and then use the adjustment factors to bring the mean close to target.

An Example

The following experiment was conducted at Bell Laboratories to optimize the process of forming contact windows in complementary metal-oxide semiconductor circuits. The contact windows facilitate interconnections between the gates, sources, and drains in a circuit. The process of forming contact windows involves photolithography. Phadke et al.[2] identified nine factors that were important in controlling window sizes: *A*—mask dimension, *B*—viscosity, *C*—spin speed, *D*—bake temperature, *E*—bake time, *F*—aperture, *G*—exposure time, *H*—developing time, and *I*—plasma etch time. Two levels were chosen for factors *A*, *B*, and *D*, and all others were at three levels. The authors decided to combine factors *B* and *D* to form a three-level factor *BD* and use the L_{18} *orthogonal array* experimental design displayed in Table I.

Table I. Experimental Design

Exp. no.	*A*	*BD*	*C*	*E*	*F*	*G*	*H*	*I*
1	1	1	1	1	1	1	1	1
2	1	1	2	2	2	2	2	2
3	1	1	3	3	3	3	3	3
4	1	2	1	1	2	2	3	3
5	1	2	2	2	3	3	1	1
6	1	2	3	3	1	1	2	2
7	1	3	1	2	1	3	2	3
8	1	3	2	3	2	1	3	1
9	1	3	3	1	3	2	1	2
10	2	1	1	3	3	2	2	1
11	2	1	2	1	1	3	3	2
12	2	1	3	2	2	1	1	3
13	2	2	1	2	3	1	3	2
14	2	2	2	3	1	2	1	3
15	2	2	3	1	2	3	2	1
16	2	3	1	3	2	3	1	2
17	2	3	2	1	3	1	2	3
18	2	3	3	2	1	2	3	1

The target size for contact windows was in the range 3.0 to 3.5 μm. It was important to produce windows near the target dimension; windows not open or too small result in loss of contact to the devices while excessively large windows lead to shorted device features. The actual size of the windows in the functional circuits on a chip could not be accurately measured. Instead, there were

Table II. Pre-etch Line Width Data

1	2.43	2.52	2.63	2.52	2.50	2.36	2.50	2.62	2.43	2.49
2	2.76	2.66	2.74	2.60	2.53	2.66	2.73	2.95	2.57	2.64
3	2.82	2.71	2.78	2.55	2.36	2.76	2.67	2.90	2.62	2.43
4	2.02	2.06	2.21	1.98	2.13	1.85	1.66	2.07	1.81	1.83
5	1.87	1.78	2.07	1.80	1.83					
6	2.51	2.56	2.55	2.45	2.53	2.68	2.60	2.85	2.55	2.56
7	1.99	1.99	2.11	1.99	2.00	1.96	2.20	2.04	2.01	2.03
8	3.15	3.44	3.67	3.09	3.06	3.27	3.29	3.49	3.02	3.19
9	3.00	2.91	3.07	2.66	2.74	2.73	2.79	3.00	2.69	2.70
10	2.69	2.50	2.51	2.46	2.40	2.75	2.73	2.75	2.78	3.03
11	3.20	3.19	3.32	3.20	3.15	3.07	3.14	3.14	3.13	3.12
12	3.21	3.32	3.33	3.23	3.10	3.48	3.44	3.49	3.25	3.38
13	2.60	2.56	2.62	2.55	2.56	2.53	2.49	2.79	2.50	2.56
14	2.18	2.20	2.45	2.22	2.32	2.33	2.20	2.41	2.37	2.38
15	2.45	2.50	2.51	2.43	2.43					
16	2.67	2.53	2.72	2.70	2.60	2.76	2.67	2.73	2.69	2.60
17	3.31	3.30	3.44	3.12	3.14	3.12	2.97	3.18	3.03	2.95
18	3.46	3.49	3.50	3.45	3.57					

three surrogate measures of quality: pre-etch and post-etch line width and post-etch window size of test patterns provided in the upper left-hand corner of each chip. Ten measurements were made at each experimental run: two wafers with five chips each corresponding to specific locations on a wafer—top, bottom, left, right, and center. In this paper, we consider only the analysis of the pre-etch line width data given in Table II. For this measure of quality, factor I, plasma etch time, is not a relevant design factor.

Phase I: Exploratory Analysis

It is important to begin the analysis with a visual display of the experimental data that summarizes the main features such as location and dispersion, and also highlights possible irregularities. The *line plot* in Figure 2 is a variation of the box plot[5] commonly used to display such information. For each of the experiments, the line plot displays the individual mean and standard deviation, and all replicates which fall outside a $\pm\sigma$ interval about the mean.

The line plot is useful for identifying individual replicates within an experiment which are several standard deviations away from the norm. These could possibly be "bad" data points, in which case they should be omitted from further analysis. For the pre-etch line width data, there are no apparent irregularities in the data. If some observations are discarded, we recommend that the line plot be redone. This ensures that apparent differences in location and dispersion are due to real effects and not "bad" data.

The line plot can also be used to visually assess differences in location and dispersion across the different experiments. Figure 2 shows that the between-experiment variability is large compared to the within-experiment variability. Thus we can expect to find large location effects in

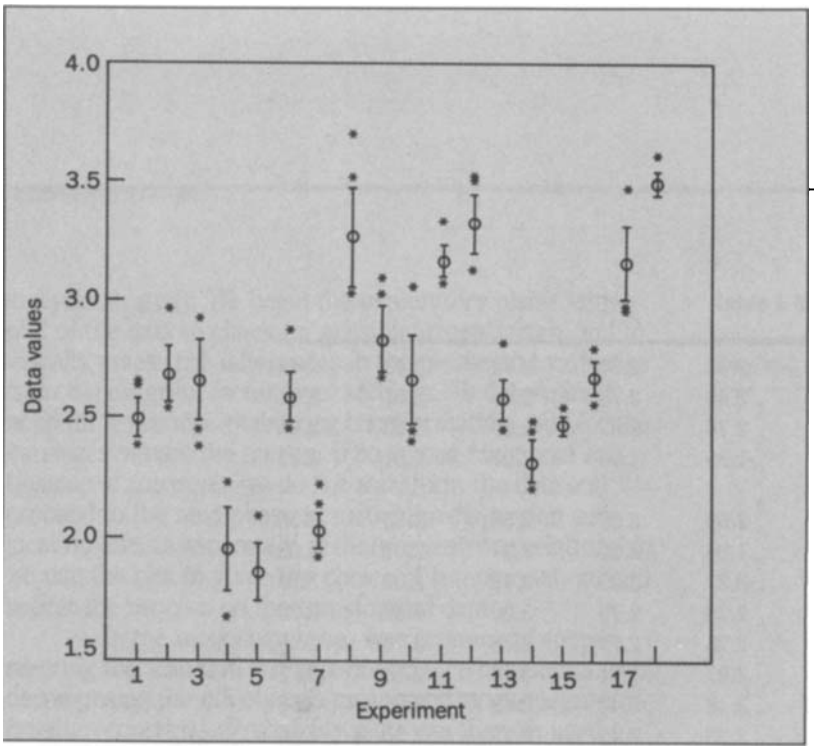

Figure 2. Line plot of the pre-etch data. For each of the eighteen individual experiments, the symbol "○" denotes the sample mean, the line extends ±1 s.d. about the mean, and the symbol "*" denotes individual values which lie outside this interval. There are obvious location effects in the data, and possibly some dispersion effects.

the modeling phase. The $\pm\sigma$ intervals appear to vary across experiments, but not dramatically so. This suggests the presence of only moderate dispersion effects.

The next step is the *mean-variance plot* shown in Figure 3. The means and the variances from the different runs are plotted on a log-log scale. If $\sigma^2(\mathbf{d}_2)$ in equation (1) is constant, i.e., it does not depend on the design factors, the mean-variance plot will reveal the form of $\gamma(\cdot)$. If $m_Y(\mathbf{d})$ is constant, the plot will show no relationship between mean and variance. The log-log scale is used since the plot will be approximately linear with slope k for the common situation where $\gamma(m) = m^k$.

Typically, however, both σ and m_Y will depend on the design factors to some extent. It may not be possible, in general, to separate $\gamma(\cdot)$ and $\sigma(\cdot)$ in equation (1). But if $\gamma(m_Y(\mathbf{d}))$ is the dominant component, the mean-variance plot will still be useful in detecting the relationship.

If the plot suggests a strong relationship, we can fit a line through the plot (say by eye) and estimate k from the slope of the line. If $k \approx 0$, there is no strong relationship and no transformation is necessary. Values of k near 1, 2, and 4 correspond respectively to the square-root, logarithmic, and reciprocal transformations. These are the three common transformations found to be useful in practice. From Figure 3, we see that there is no strong relationship and so no transformation is necessary. We can now proceed to the next phase and model the means and variances. If we had found the need for transformation, we should repeat the above cycle with the transformed data (see Figure 1).

Phase II: Modeling

Dispersion Effects. To determine the important dispersion effects, we model the logarithms of the variances of the (possibly transformed) data as an additive function of the design factors. Additivity is more likely to be satisfied on the logarithmic scale, and moreover, allows estimation of dispersion effects by unconstrained least squares.

Standard analysis of variance techniques[6] and the resulting F tests can be used to determine the significant factors when an estimate of error is available. Otherwise, one must resort to less formal probability plotting techniques. However, we recommend that the probability plotting methods be used to supplement the formal ANOVA computations and the F tests in all cases. If we perform many F tests, as would be the case in screening experiments with many factors, we would find some of these to be significant even when none of the factors has any effect.

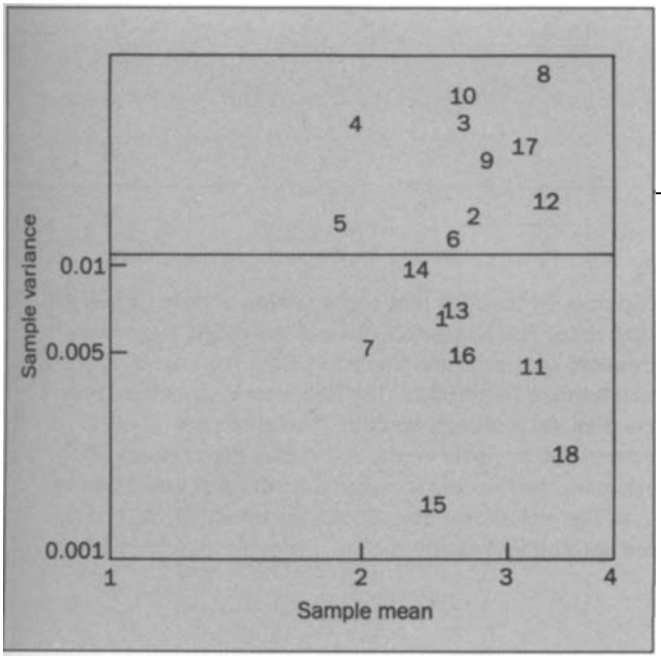

Figure 3. Mean-variance plot of the pre-etch data. On a log-log scale, the sample variance from each experiment is plotted versus the corresponding sample mean. The experiment number is used as the plotting character. The line superimposed on the scatter plot was fitted by eye. There appears to be no significant association between variance and mean, linear or otherwise.

Table III. ANOVA Table for Dispersion Effects

	df	*SS*	*MS*	*F*
A	1	0.493	0.493	4.059
BD	2	0.121	0.060	0.498
C	2	0.150	0.075	0.615
E	2	0.380	0.190	1.563
F	2	0.831	0.415	3.421
G	2	0.496	0.248	2.042
H	2	0.011	0.006	0.045
Residual	4	0.486	0.121	

The probability plotting method compensates for this appropriately.

To do a probability plot, we first decompose factors at more than two levels into meaningful single-degree-of-freedom contrasts. For quantitative factors, this would typically be the components that are linear, quadratic, and so forth. The half-normal probability plot[7] is then obtained by plotting the ordered absolute values of the contrasts against the half-normal quantiles. A linear configuration through the origin suggests that there are no significant contrasts. If there are a few important effects, they will tend to fall toward the top right-hand corner above the linear configuration determined by the rest of the contrasts. Thus the probability plot uses the variability among the different contrasts to detect the few important ones.

The results from the ANOVA computations for the pre-etch line width data are given in Table III. The F tests show that none of the factors is really significant. Factors *A*, *F*, and *G* have the largest observed effects. Figure 4 is the half-normal probability plot of the single-degree-of-freedom dispersion contrasts. The linear component of *BD* measures the effect of *D*. The quadratic component measures the effect of *B* provided *D* has no effect. Since this appears to be the case, we have denoted these terms as *B* and *D* in Figure 4. For the other three-level factors, the components are the usual linear and quadratic terms. The overall linear appearance of the plot suggests that there are no strong dispersion effects. But the effects of factors *F*, *A*, and *G* are somewhat separated from the others. Notice also that by decomposing the factors into linear and quadratic terms, we have not diffused the linear effects of *F* and *G*.

Although the dispersion effects of *F*, *A*, and *G* appear to be only marginally important, we will include these factors in the optimization phase. The possible error involved in making this decision is less costly than in wrongly concluding that the factors have no effect.

Location Effects. The important location effects can be determined by modeling the means of the (possibly transformed) data as an additive function of the design fac-

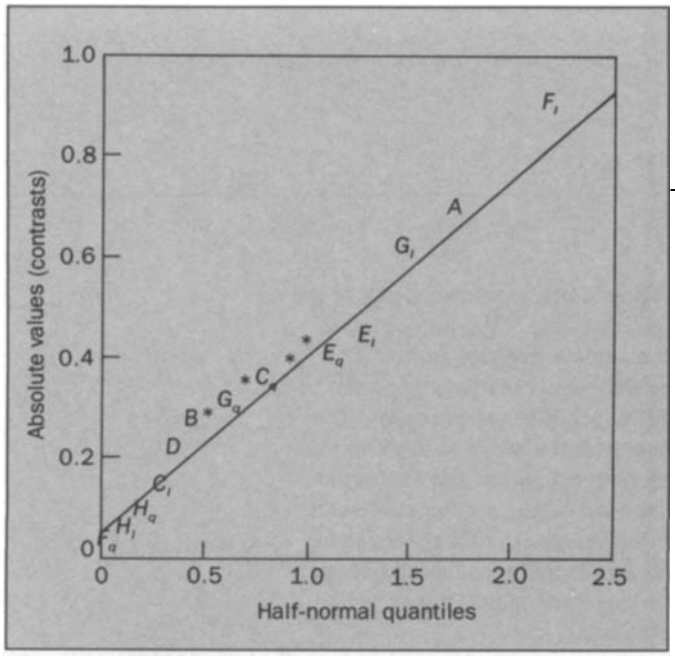

Figure 4. Probability plot of dispersion effects for the pre-etch data. The absolute values of the single degree-of-freedom contrasts are plotted against the quantiles of the half-normal distribution. The factor level combination is used as the plotting character. The contrasts labeled "*" correspond to "error contrasts" rather than effects of interest. The line superimposed on the plot was fitted by eye. The plot shows that dispersion effects F_1, A, and G, are marginally important.

Table IV. ANOVA Table for Location Effects

	df	*SS*	*MS*	*F*
A	1	0.651	0.651	22.459**
BD	2	1.345	0.672	23.186**
C	2	0.765	0.383	13.193*
E	2	0.002	0.001	0.038
F	2	0.032	0.016	0.545
G	2	0.545	0.273	9.397*
H	2	0.281	0.140	4.838
Residual	4	0.116	0.029	
Within reps	147	2.481	0.017	

** $p < 0.01$
* $p < 0.05$

tors. It is possible to do a careful and efficient weighted least-squares analysis that takes into account the results from the previous section on the differences in variances. For the sake of simplicity, we will restrict attention to a least-squares analysis which would be adequate in detecting the really important effects. See, however, the "Maximum Likelihood Analysis" section for a more efficient analysis. The F tests and the probability plots in this section should also be considered as approximate procedures.

The ANOVA table for the location effects for the pre-etch line width data are given in Table IV. In addition to the residual sum of squares, we have provided the within-replications sum of squares which also provides an "error" estimate. We use the term "error" here as some average measure of the underlying variability since the variances are possibly different. The within-replications mean square in Table IV is smaller than the residual mean square error. To be conservative, the F statistics in Table IV are computed with the residual mean square error. They suggest that the location effects associated with factors *A*, *BD*, and, to a lesser extent, *C* and *G* are significant. The next largest observed effect is due to factor *H*.

Figure 5 shows the half-normal probability plot of the single degree-of-freedom contrasts. We see that *A*, *B*, and the linear components of *C*, *G*, and *H* are the important location effects.

Interpreting the Analysis. The results from the modeling phase can be summarized in the *location-dispersion plot* shown in Figure 6. The single-degree-of-freedom dispersion contrasts are plotted against the corresponding location contrasts, and the shaded regions, obtained from the probability plots, indicate effects that are not important. Thus the factors in the intersected area of Figure 6 exhibit neither location nor dispersion effects; those lying in the horizontal (vertical) band exhibit location (dispersion) effects only. Those in the unshaded areas have both

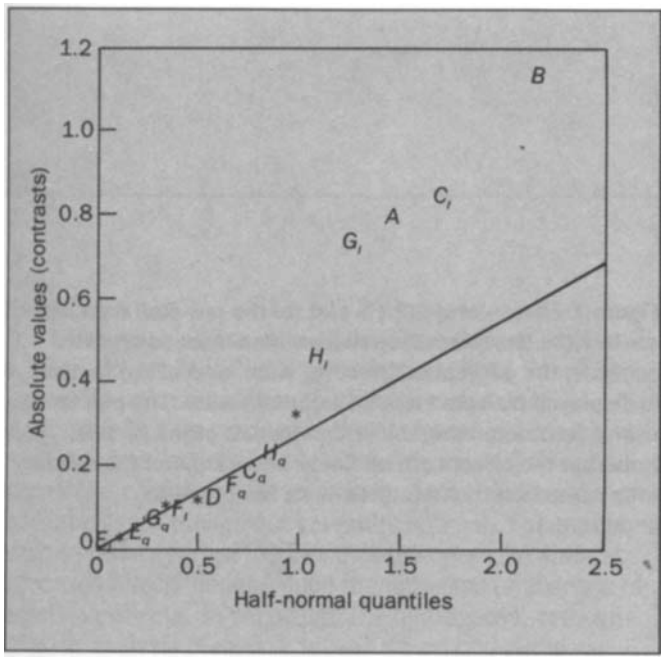

Figure 5. Probability plot of location effects for the pre-etch data. The absolute values of the single-degree-of-freedom contrasts are plotted against the quantiles of the half-normal distribution. The factor level combination is used as the plotting character. The contrasts labeled "*" correspond to "error contrasts" rather than effects of interest. The line superimposed on the plot was fitted (to the nonsignificant contrasts) by eye. The plot shows that location effects *B*, C_l, *A*, G_l, and to a lesser extent *H*, are important.

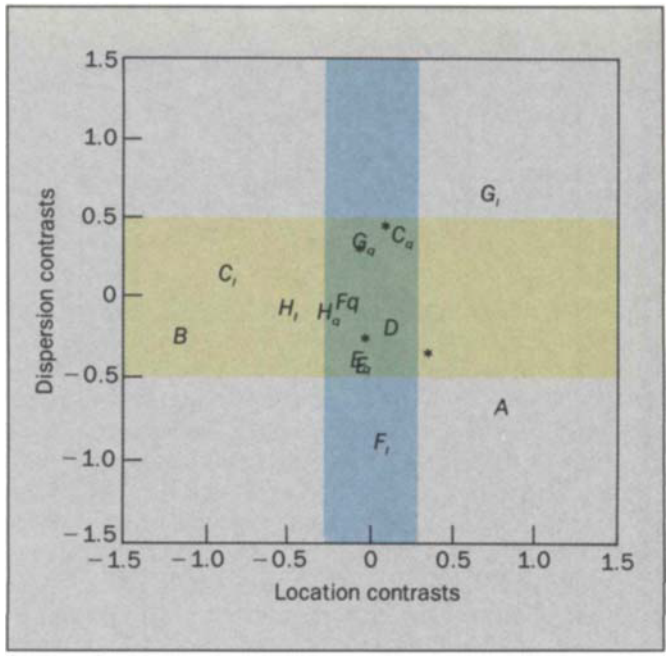

Figure 6. Location-dispersion plot of effects for the pre-etch data. The single-degree-of-freedom contrasts for dispersion are plotted against those for location. The factor-level combination is used as the plotting character. The contrasts labeled "*" correspond to "error contrasts" rather than effects of interest. The shaded bands are derived from the individual probability plots of location and dispersion effects. Contrasts appearing in the intersection of these bands are those which exhibit neither location nor dispersion effects. Those lying in the horizontal (vertical) band exhibit location (dispersion) effects only. Those contrasts appearing in unshaded areas exhibit both location and dispersion effects.

location and dispersion effects. For the pre-etch data, *B*, *C*, and to a lesser extent *H* are the adjustment factors, i.e., factors with only location effects. *F* is a pure dispersion factor while *A* and *G* have both location and dispersion effects. Further, only the linear terms of the three-level factors are important; the quadratic terms appear to be insignificant.

The location-dispersion plot can also be used to assess the effectiveness of the data transformations in the exploratory phase. If most of the important factors have both location and dispersion effects and they fall in either the top right-hand or the lower left-hand unshaded areas in Figure 6, we should suspect that the variance of our (possibly transformed) data is an increasing function of the mean. Thus a further transformation may be necessary (see Figure 1). Even if none of the factors is important, we should examine the location-dispersion plot to see if the location and dispersion estimates appear to be positively (negatively) correlated. Strong positive (negative) correlation suggests that the underlying noise distribution is very

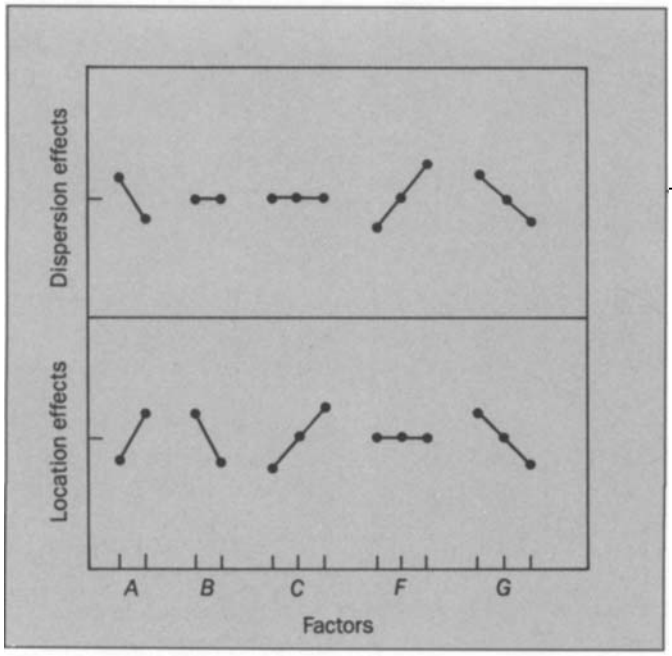

Figure 7. Factor-level effects plot for the pre-etch data. For each of the important factors identified by our suggested analysis, the estimated effect for each level of the factor is displayed for both location and dispersion. The plot is useful for determining the optimal levels of the factors. Note that the effects are all linear since none of the quadratic components of the three-level factors were important.

skewed to the right (left). It may then be worthwhile to transform the data and see if there are any important effects in the transformed scale. For the pre-etch data, Figure 6 does not not indicate the need for a transformation.

Phase III: Optimization

We can use a two-step approach to optimization if there are adjustment factors, i.e., factors with only location effects, that are either known a priori or are discovered during the modeling phase. In the first step, we can set the factors with dispersion effects at the optimal levels to minimize variability. The mean can then be brought on target by fine-tuning the adjustment factors.

For the pre-etch line width data, as seen in the preceding section, B, viscosity, and C, spin speed, are the adjustment factors while F, aperture, is a pure dispersion factor. A, mask dimension, and G, exposure time, have both location and dispersion effects. Figure 7, the *factor effects plot*, provides both the magnitude and the direction of these effects. This information can also be gleaned from the location-dispersion plot if the factors have only linear effects. We see from Figure 7 that the optimal levels for minimizing dispersion are: 1 for F, 2 for A, and 3 for G. At these optimal levels, the location effects of A and G effectively cancel each other. For the adjustment factors, the location decreases with viscosity, B, while it increases with spin speed, C.

We can now choose between F, A, or G (or even a combination of these) to minimize dispersion. Note also that G is a quantitative variable so that we may be able to reduce variability even further by linearly extrapolating beyond level 3. However, the final choice of the factor levels should be guided by engineering as well as statistical considerations.

The standard operating levels for the factors A, B, C, F, and G were, respectively, 1, 1, 2, 2, and 2. Purely on the basis of our statistical analysis, we would have recommended that the level of A be changed to 2, C be changed to 3, F be changed to 1 and, possibly, G be changed to 3. Phadke et al. decided not to change the level of F because of engineering considerations. In addition, they did not recommend changing the level of C. Their conclusions were based on a signal-to-noise ratio analysis of not only the data we consider here (pre-etch line width), but also the other two measures of quality (post-etch line width and window size).

In cases where there are no adjustment factors or where the adjustment factors cannot be easily manipulated, the optimization process is more complicated. The optimal levels of the dispersion factors and the location-dispersion factors have to be determined simultaneously. It is possible to formulate this formally as a mathematical problem and use a constrained optimization technique to solve it. However, a simple iterative search procedure over the desirable factor space is likely to be more useful.

Discussion

Maximum Likelihood Analysis. The analyses discussed in "Phase II: Modeling" are geared primarily to the identification of important effects, both for location and dispersion. As a by-product of the computations used in these analyses, estimates of the effects are available. By appropriate choice of data transformation, these estimates should be good, though not generally optimal. The problem derives primarily from the fact that the optimal estimates of location effects depend upon the presence or absence of dispersion effects. In particular, the unweighted, orthogonal array analysis of sample means we carried out is formally incorrect, since we are weighting each sample mean equally, though the presence of dispersion effects dictates otherwise.

A more formal method of estimation can be carried out.[8] What results is an iterative method of estimation, alternating between estimating location effects by weighted least-squares, with weights depending on the current estimates of dispersion effects, and estimating dispersion effects by least-squares (as suggested), where now dispersion is measured about current location effects rather than sample means. This more detailed analysis requires using all the data, not just the sample means and variances of the individual experiments. We don't recommend this procedure in general, but if the results of the analysis are to be used for producing forecasts and balance sheets, then the extra computational effort is recommended.

Analysis of Nonreplicated Data. The data analytic approach developed in this paper assumes that there is a sufficient number of replications at each experimental setting. Some of our suggestions, particularly those in the exploratory phase, either have to be modified or cannot be used when there are no replications. Little if anything can be inferred from the data about transformations. The appropriate scale of analysis must be determined by the experimenter on the basis of similar previous analyses and/or insight into the physical mechanism giving rise to the observed data. To estimate dispersion effects, one has to first identify the important location effects. Assuming all other location effects to be zero induces pseudo-replication in the data, allowing the methodology discussed in the present paper to be applied. For more discussion on this topic, see References 9 and 10.

Other Measures of Location and Dispersion. We have limited our discussion to the mean and variance as measures of location and dispersion respectively, mainly due to their familiarity among engineers. It should be noted however that similar analyses can be carried out with other measures, and it may even be desirable to do so. For example, it may be more meaningful to choose the design factors so that the *median*, rather than the *mean*, is close to the target t_0. Similarly, the *interquartile range*, the range of the central 50 percent of the data, may describe dispersion better than *variance*. Apart from such subject matter considerations, an important property of measures such as the median and interquartile range is that their sample-based estimates are more resistant to "bad" data points than those corresponding to the mean and variance.

Analysis of Ordered Categorical Data. In some cases, the response variable in a quality improvement study consists of categorical data with an implied ordering in the categories. Taguchi has proposed a method called "accumulation analysis" for analyzing such data. See Reference 11 and the discussion following that paper for the properties of this method and for some methods of analysis. McCullagh[12] discusses fitting parametric models to ordered categorical data by maximum likelihood techniques.

Software. The analysis reported in this paper was carried out entirely in the S system and language for data analysis.[13] S is an environment for statistical computing and graphics, but as such, has no special facilities for the analysis of designed experiments. A special-purpose macro package provided the calculations we required.

The flow diagram in Figure 1 explicitly describes how we feel an analysis should proceed. This expertise can be coded into software, forming the "knowledge base" of so-called *expert systems*. We have limited experience with this type of software, though a prototype system for linear regression analysis, REX, has been developed.[14] Of particular relevance is the fact the knowledge coded into REX is

exactly of the sort dispayed in Figure 1 and described in more detail in previous sections. We thus foresee the possibility of providing an intelligent interface to S to help engineers model their data and optimize process quality.

Conclusion

In quality engineering experiments, there is typically little interest in detailed structural modeling and analysis of the data. The user's primary goal is to identify the really important factors and determine the levels to optimize process quality. For these reasons, we have tried to keep our recommended strategy fairly simple. It can be supplemented with more sophisticated techniques depending on the needs and statistical training of the user. The availability of software would also reduce the need for simplicity. Finally, we note that we have not considered the analysis of data from the confirmation experiment in our data analysis strategy since this analysis is routine. It should be emphasized, however, that the confirmation experiment to determine if the new factor levels do improve process quality is an integral part of quality engineering experiments.

References

1. K. Taguchi and Y. Wu, *Introduction to Off-Line Quality Control*, Central Japan Quality Control Association, 1980.
2. M. S. Phadke, R. N. Kackar, D. V. Speeney, and M. J. Grieco, "Off-Line Quality Control in Integrated Circuit Fabrication Using Experimental Design," *The Bell System Technical Journal*, Vol. 62, No. 5, pp. 1273-1310, 1983.
3. R. N. Kackar, R. V. Leon, and A. C. Shoemaker, "Performance Measures Independent of Adjustment," *AT&T Technical Journal*, Vol. 65, Issue 3, 1986.
4. M. S. Bartlett, "The Use of Transformations," *Biometrics*, Vol. 3, pp. 39-52, 1947.
5. J. W. Tukey, *Exploratory Data Analysis*, Addison Wesley, Reading, Mass., 1977.
6. G. E. P. Box, W. G. Hunter, and J. S. Hunter, *Statistics for Experimenters*, John Wiley and Sons, New York, 1978.
7. C. Daniel, "Use of Half-Normal Plots in Interpreting Factorial Two-Level Experiments," *Technometrics*, Vol. 1, No. 4, pp. 311-341, November 1959.
8. J. A. Nelder and D. Pregibon, "An Extended Quasi-Likelihood Function," *Biometrika* [to appear].
9. G. E. P. Box and D. Meyer, "Dispersion Effects from Fractional Designs," *Technometrics*, Vol. 28, No. 1, pp. 19-28, February 1986.
10. V. N. Nair and D. Pregibon, "Analysis of Dispersion Effects: When to Log?," *Technometrics* [submitted for publication].
11. V. N. Nair, "Testing in Industrial Experiments with Ordered Categorical Data," *Technometrics* [to appear].
12. P. McCullagh, "Regression Models for Ordinal Data (with Discussion)," *Journal of the Royal Statistical Society*, Series B, Vol. 42, pp. 109-142, 1980.
13. R. A. Becker and J. M. Chambers, *S: An Interactive Environment for Data Analysis and Graphics*, Wadsworth, Belmont, Calif., 1984.
14. W. A. Gale and D. Pregibon, "REX: An Expert System for Regression Analysis," *Proceedings*, COMPSTAT 84, Prague, September 1984, pp. 242-248, Physica-Verlag.

(Manuscript received December 11, 1985)