Effective Use of Ordinal Scales

Bill Ross,
Sigma Science Inc.

Abstract

This paper is intended to highlight some of the issues and provide some advice regarding the appropriate use of ordinal data sets used to quantify variation in samples[1] being measured. Ordinal scales are most useful in assessing variation when the measurement is the result of sensory perception (i.e., visual, taste, smell, audible, feel). Companies often want to understand how their customers perceive their product, what factors affect this perception and then optimize the product design or process to improve customer satisfaction. Important clues may be derived from the collection and analysis of such data, however it is typically more efficient and effective to model quantitative data.

Questions stimulating the discussion include: How are customer preferences quantified? How can judgment or sensory perception be quantified? How are respondent[2] biases handled? How should such data sets be analyzed? How should data be collected? What are the limitations of such data sets?

Figure 1 shows a simplified comparison of types of data. Ordinal data is a set of data whose values/observations can be ranked (i.e., put in rational order). Typically these categories have a numerical rating scale assigned. The distance between the categories may not be equal or known. Ordinal data can be counted or ordered, but not measured. The categories for an ordinal set of data have a natural order, for example, suppose a group of people were asked to taste the flavor of cookies and classify each cookie on a rating scale of 1 to 5, representing strongly dislike, dislike, neutral, like, strongly like[3]. A rating of 5 indicates more enjoyment than a rating of 4, for example, so such data are ordinal.

[1] Specimens to be measured (e.g., products, experimental units)
[2] Respondent is synonymous with inspector, appraiser, scorer, evaluator or survey participant.
[3] A special ordinal data rating scale based on the work of Rensis Likert.

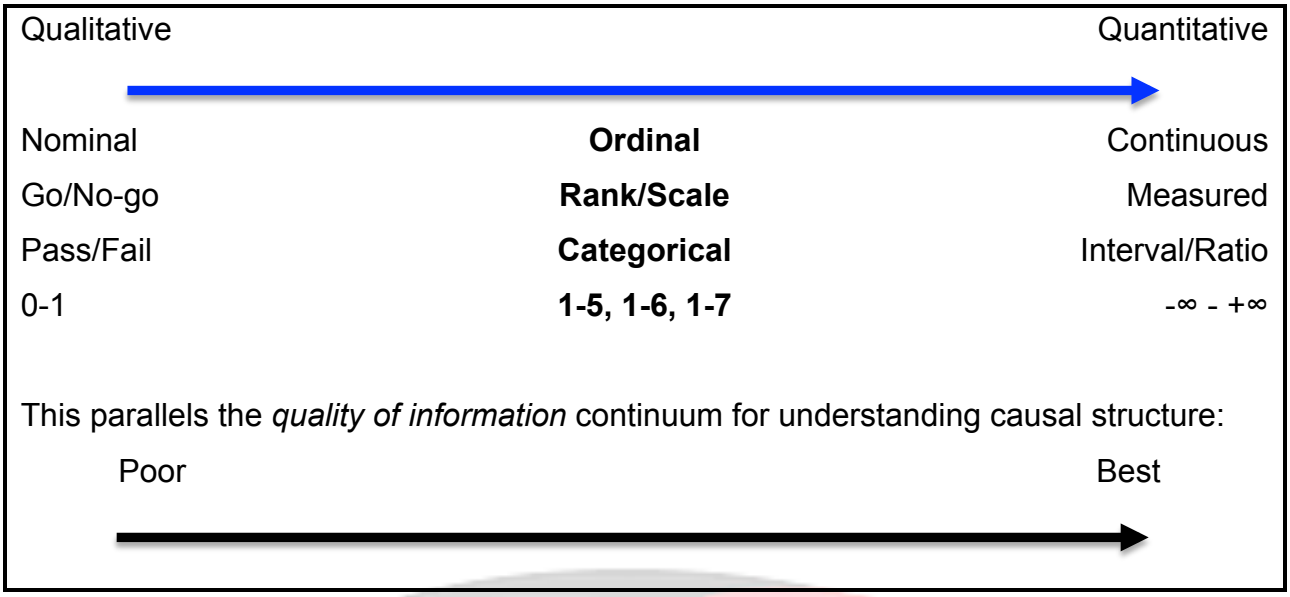

Figure 1: Summary of Data Types

Issues

The issues associated with ordinal scales and their usefulness in evaluating the mean and variation of samples include:

1. First and foremost, ordinal scales have limited effective resolution. In essence they lack enough measurement units to provide effective discrimination in the samples being measured.
2. They are particularly susceptible to respondent bias. Both within and between respondent. The respondent may or may not be cognizant of the bias.
3. It is challenging to have the proper precision and consistency due to human nature and subsequent subjectivity.
4. Accuracy can seldom be assessed. The scales are likely only applicable to the specific situation for which they are used.

Recommendations

The following is a list of considerations regarding the use of such qualitative data when evaluating samples (e.g., from a sampling plan or designed experiment). There are a number of actions that can be taken to improve the use ordinal response variables:

1. Ordinal scales rely on comparisons. Develop physical specimen as a means of direct comparison. Make sure the specimen is similar to the samples to be evaluated. For example, if you are assessing the cleanliness of the inside of an

oven, make the specimen from similar material, in a similar shape, with similar lighting conditions, etc. to represent typical ovens. The creation of comparison specimen is also useful for reducing bias amongst the respondents. The question posed to the respondents is something like "Which test specimen does the sample most match", not whether or not you like the sample.

2. Objectively describe each scale category. Use descriptions (operational definitions[4]) that create universal understanding. Do not use ambiguous words that are subject to varying interpretation.
3. Wisely consider the selection of respondents. The selection of individuals to take part in the evaluation must be representative of the population you wish to draw inference over. Involve the appropriate people in the evaluation process (e.g., the customer). I recommend using hypotheses to define the target audience. Without some rationale as to who the target audience would be, we are left with enumerative guidance: large sample size, randomly selected.
4. ALWAYS have more than one respondent and assess the consistency.
   - The variability, consistency and amount of respondent-to-respondent variation can be assessed using variability and range charts[5]. Then, if that variation is consistent, summary statistics (e.g., means and ranges) are used to evaluate the samples in the study. Two response variables will be used for analyzing the experiment. Averages will both reduce the respondent measurement error, $\hat{\sigma}_{avg} = \left(\frac{\hat{\sigma}_{ind}}{\sqrt{n}} \right)$ and expand the ordinal scale resolution while ranges will identify if any effects influence variability respondent-to-respondent.
   - In addition to the respondent-to-respondent variation, have each respondent evaluate each sample more than once. This will enable assessment of the within respondent variation for consistency and again use averages to reduce within respondent variation, increase the resolution of the scale and increase the inference space. If performing a designed experiment, incorporate nested layers for within and between respondent.
5. Describe the measurement process. Use process maps to identify factors that may effect measurement system variation. For example if the inspection is visual,

[4] Deming, W. Edwards (1986) "*Out of the Crisis*" MIT Press (ISBN 0-911379-01-0)
[5] Intraclass Correlation Coefficient (ICC) and Kappa statistics may be used to assess the reliability of such data

evaluate the effects of light intensity, lighting source, proximity to sample, magnification, angle, etc. on measurement variation. If the inspection is a taste test, consider: environmental conditions, items used to cleanse the palette, appetite, etc. Perhaps design and run an experiment to determine what factors effect measurement variation and subsequently choose levels for factors to reduce that variation and create a consistent evaluation process.

6. The entire scale must be used. For a 1-5 scale, all 5 of those categories must be represented in the data set. If not, perhaps the scale can be adjusted to accomplish the full extent of the scale. For data acquisition strategies:
   - Sampling: sample over a large inference space perhaps including customers as respondents
   - Experimenting: include a large number of factors, manipulated at bold/extreme level settings
7. The scales must be created á priori the evaluation. While they may be created after the samples are obtained, they should be in place before any assessment is done.
8. Training the respondents may be helpful. While training likely has a transitory effect, it may be useful for short periods of time. Communicating the purpose, how the data will be used and what the predicted actions are as a result of the study is recommended.
9. Develop alternative quantitative measures. Creating quantitative (interval or ratio) measures that correlate with the ordinal Y will ultimately be more effective and efficient for understanding causality. Use multivariate analysis to determine how correlated the quantitative Y's are.

Planning

Incorporating multiple layers[6] of components of the ordinal measurement system into the sampling plan will help to understand the consistency of those layers. Averaging those layers will reduce the variability. Figure 2 is an example sampling plan incorporating nested layers of respondents (scorers) and repeats (reps) within scorers. The consistency within respondent and between respondent can be evaluated to detect bias.

[6] Hierarchical study

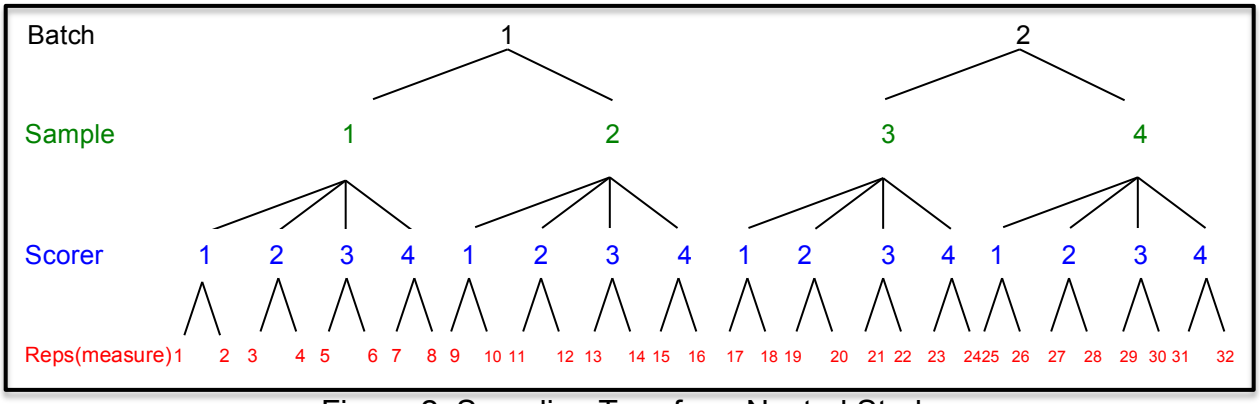

Figure 2: Sampling Tree for a Nested Study

Figure 3 shows a Factor Relationship Diagram (FRD)[7] with nested layers of the measurement system (inspector and repeats within inspector) within treatments.

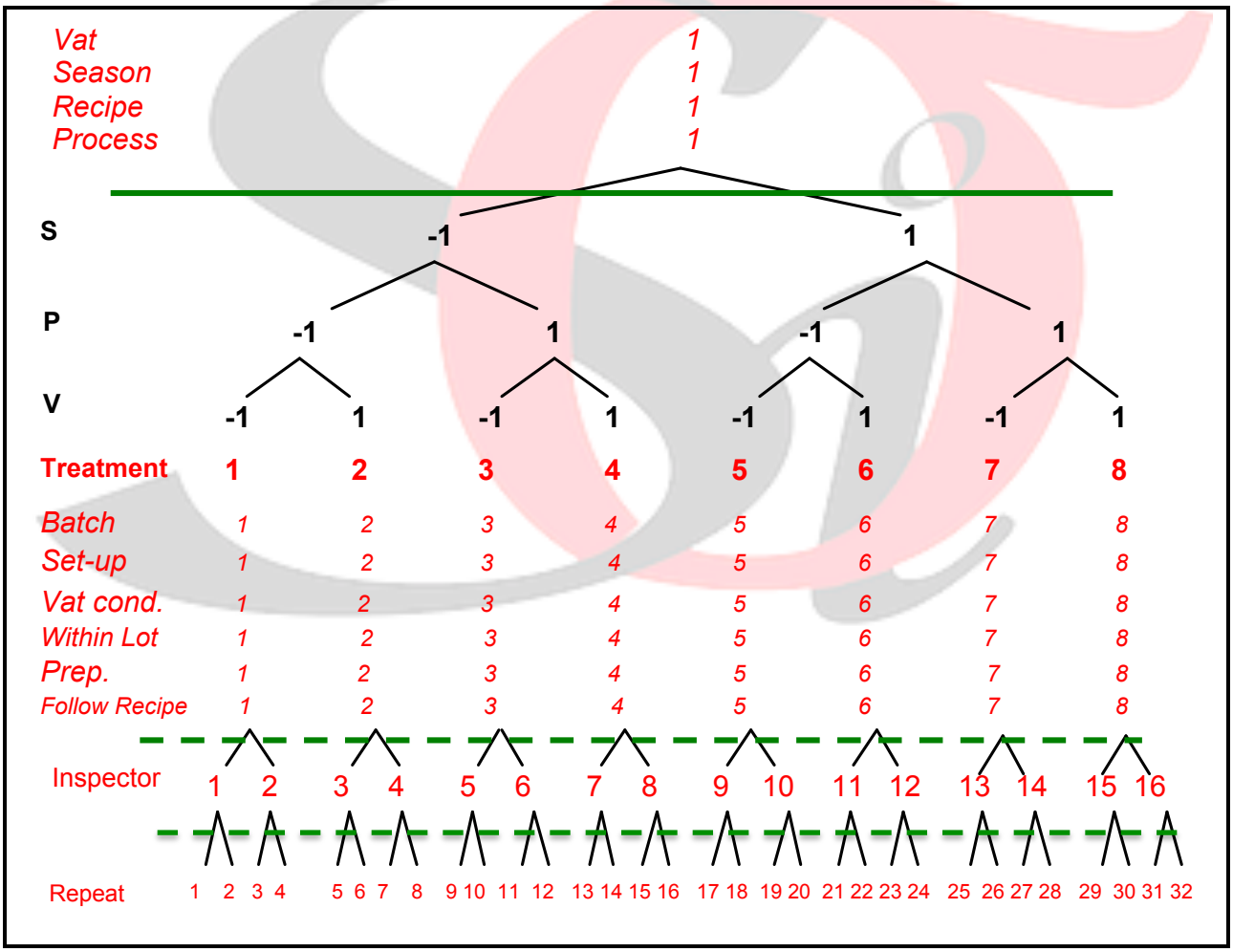

Figure 3: FRD for an experiment on factors: S, P & V

[7] Sanders, Doug and Jim Coleman (1999), "Considerations Associated with Restrictions on Randomization in Industrial Experimentation", *Quality Engineering*, Volume 12, No. 1

Execution

Both plans (figure 2 & 3) require each sample or experimental unit to be evaluated twice by multiple respondents. This should be carried out in "blind" fashion where the respondents do not know which sample or experiment unit is being evaluated to prevent bias[8].

Analysis

One of the initial steps in the analysis is to look at the data via a variability plot[9]. This will aid in recognition of obvious patterns in the data. Figure 4 shows an example of a variability plot for the sampling plan shown in figure 2. It is obvious there are potential issues with scorers 2 & 3 as they rate each sample the same value.

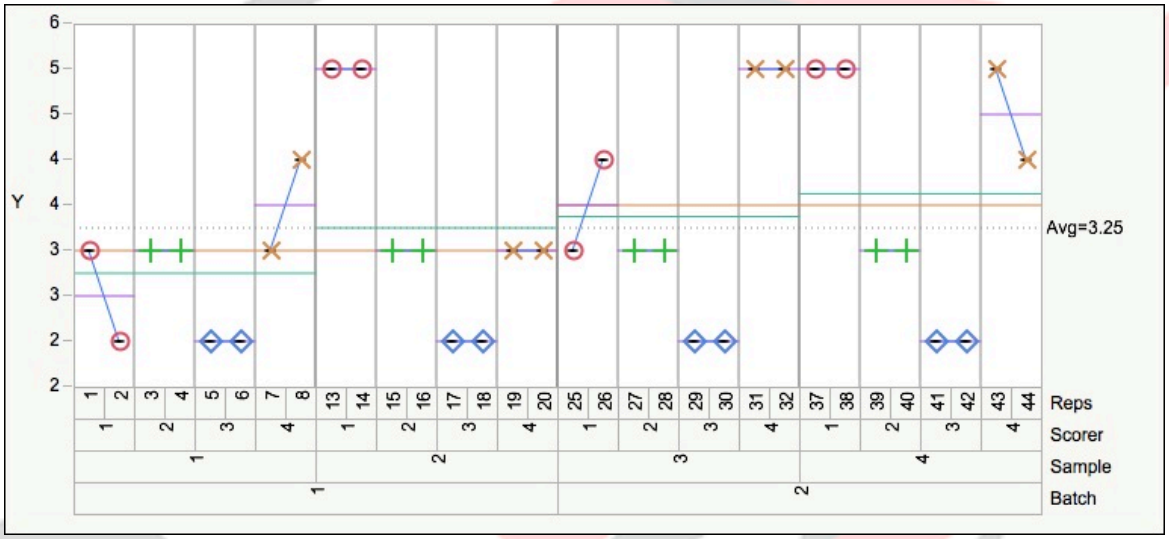

Figure 4: Variability Plot Color Coded by Scorer

The consistency of the variation between repeats (within respondent variation) can be assessed using range charts (although with the limitations of measurement discrimination, outliers can be easily detected on variability charts). The range charts can also provide insight into the effective resolution. Figure 5 shows the range chart for the sampling plan in figure 2. While the chart is out-of-control, the maximum range is 1, which seems reasonable. The reason for the out-of-control condition is the lack of variation reported by scorers 2 & 3.

[8] The bias may or may not be intentional.
[9] Plots of the data with no summarization, variously known as Box plots, individual value plots, multi-vari studies or dot plots.

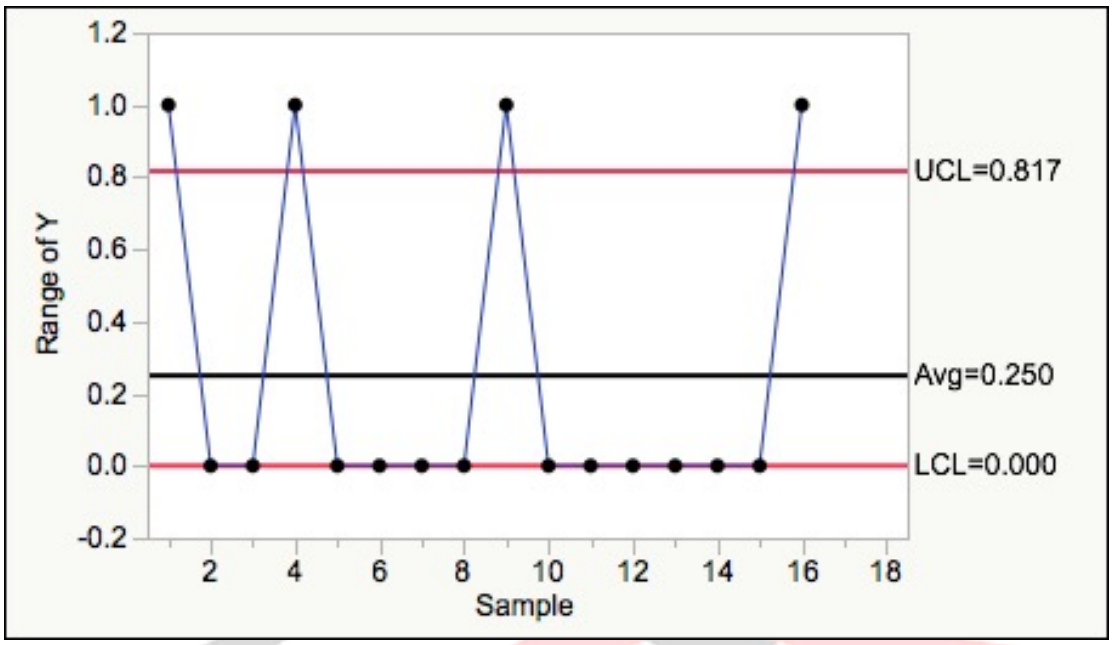

Figure 5: Range Chart

If there are multiple nested hierarchical layers, the data can be summarized and evaluated for consistency at each layer using range charts. Averaging the data has the effect of increasing the resolution of the scale making it more discriminant while reducing the variability. Quantitative assessment using analysis of variation (ANOVA) may also be performed to assign the mean square components of variation and assess significance.

Application Example

This case study discusses the application of ordinal scales as a response variable in an experiment run to determine what factors affect the quality (i.e., legibility, visual perception) of a screen printing operation. Although the screening operation was quite mature, the quality of the product coming from the operation seemed more a result of "black magic" than from optimum process understanding control. Operator skill, tweaking & “luck” were required to make acceptable product.

A cross-functional team was formed to improve the operation. The team developed thought maps[10] to graphically display hypotheses explaining why potential sources of variation might effect the screen print quality and process maps[11] to document factors potentially related to the qualitative characteristics of the resultant print. A scale for the

[10] Hild, Cheryl, D. Sanders (2000) “The Thought Map”, *Quality Engineering*, Vol. 12, No. 1.
[11] Doug Sanders, W. Ross, and J. Coleman (2000), “The Process Map”, *Quality Engineering*, Vol. 11, No. 4.

response variable was set-up using ordinal measurement categories 1-5 (Y1). The factors being manipulated in the experiment are listed in table 1.

Table 1: Listing of factors

A.	Screen tension
B.	Mesh count
C.	Screen height
D.	Squeegee skew
E.	Squeegee angle
F.	Squeegee speed
G.	Squeegee hardness
H.	Ink viscosity
I.	Ink type
J.	Squeegee pressure

In addition to the ordinal scale response variable, a quantitative measure of the width of the printed text was also made (Y3). It was hoped the quantitative measure could be used as a replacement for the more subjective qualitative measurement.

The FRD for the designed experiment is shown in figure 6. A large number of factors set at bold levels were included to insure the full extent of the scale is created in the experiment. The experiment was conducted in two incomplete blocks to include the effect of noise on the output. Note for this example, no within inspector measures were taken.

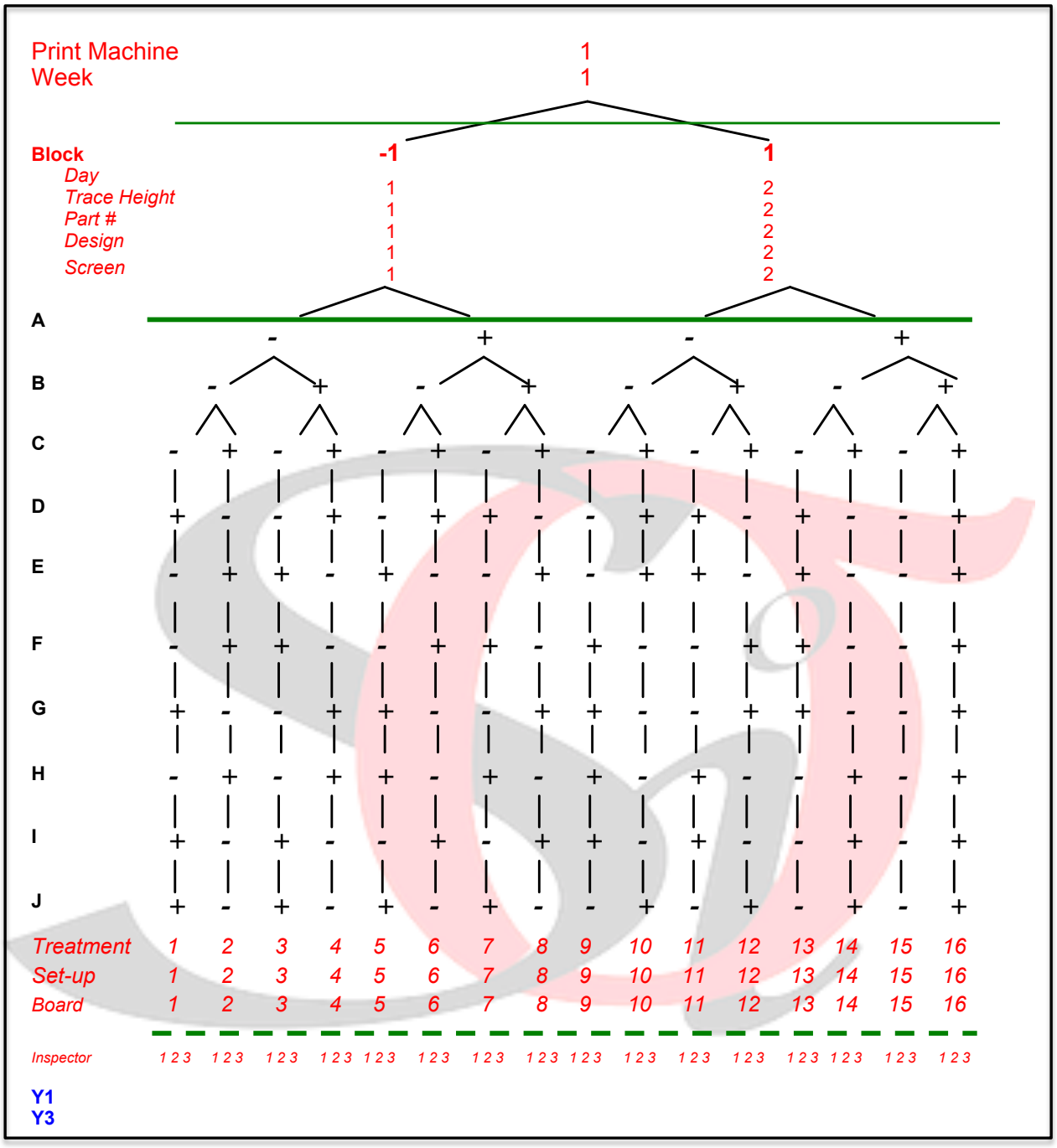

Figure 6: FRD for Screening DOE
(A resolution III fractional factorial in two incomplete blocks)

The range chart (figure 7) shows consistency between inspectors. Therefore the averages and ranges are calculated for analysis including the correlation with Y3 and ultimately the effect factors have on those response variables. Again, this has the effect of reducing the inspector-to-inspector variation, increasing the effective resolution of the scale, and increasing the inference space.

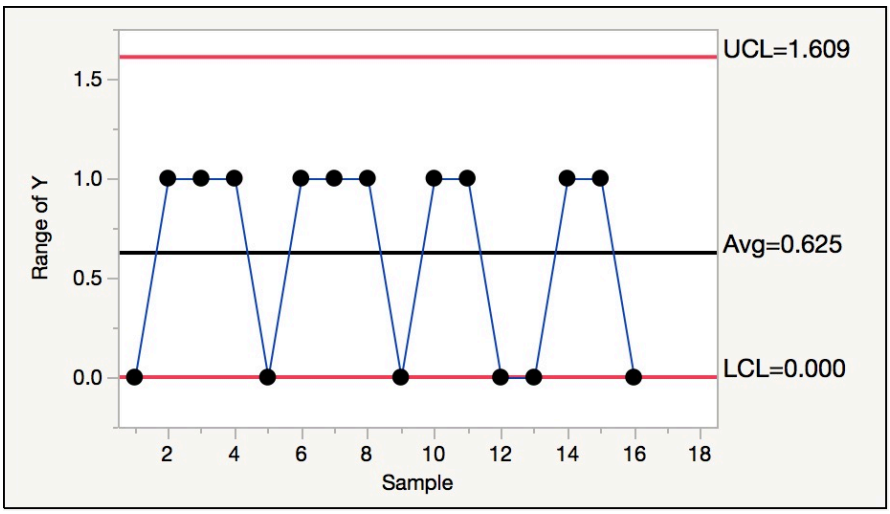

Figure 7: Range Chart for Inspector-to-Inspector Variation (within treatment)

Figure 8 shows the correlation matrix for Y1 and Y3. While the overall correlation coefficient (R) is .91, the correlation varies throughout the range of the response variables. The correlation between Y1 and Y3 is quite strong when Y1 ≤ 3. However the correlation when Y > 3 is very weak. Therefore Y3 will not make a good surrogate response variable.

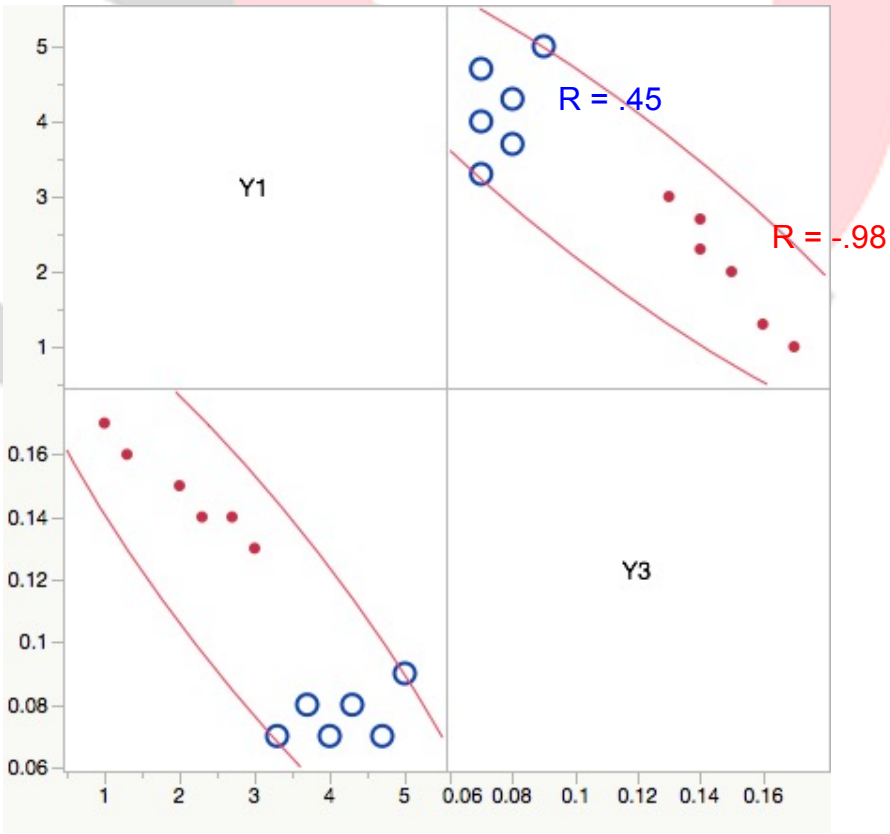

Figure 8: Scatter Plot for Y1 vs. Y3

Figure 9 shows a normal probability plot[12] of the effects on the average estimated in the experiment. Figure 10 shows a Pareto plot of the same effects (practical significance = 0.3). The ranges may provide some clues for reducing the measurement variation between inspectors (Figure 11).

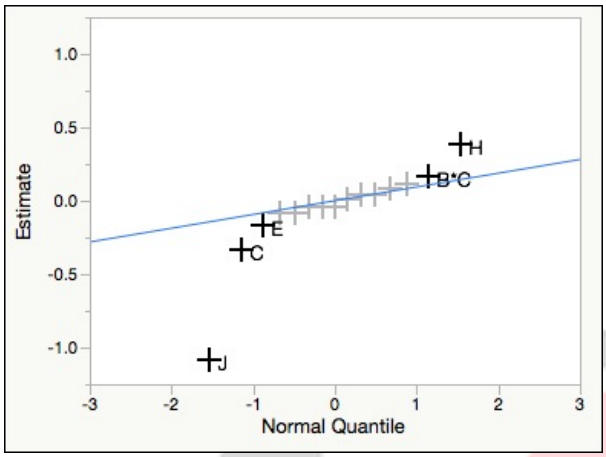

Figure 9: Normal Plot of Effects (Y1)

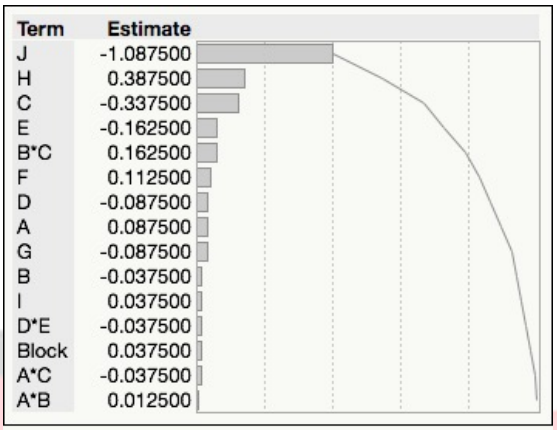

Figure 10: Pareto Plot of Effects (Y1)

Figure 11: Analysis of the Range of Y1: Shaded Areas = Smallest Ranges

Using the ordinal scale response variable, factors affecting the qualitative characteristics of the screen printing process were identified: Squeegee pressure (J), Screen height (C) & Ink Viscosity (H) in particular. The thought map was updated and process controls for those factors were implemented resulting in a significantly improved product.

[12] Daniel plot see Daniel, Cuthbert (1976) "*Applications of Statistics to Industrial Experiments*" Wiley

Definitions

Analytical Statistics (aka. Inferential statistics): the application of statistical thinking and methods together with principles and laws of the sciences to explain and predict phenomena by understanding the causal structure, thus increasing the confidence in the extrapolation of results.

Bias: Condition where the central tendency (e.g., mean) of an estimate deviates from the true value.

Control Charts: a graphical technique used to study a process over time. Control charts are used in pairs. One is the Range chart used to determine if the ranges (within subgroup) are consistent. The other is a Y-bar or X-bar chart. These charts are charts used to compare the sources within subgroup and between subgroup to the within subgroup sources to determine which source has more leverage.

Discrimination: Measure of the smallest unit of measurement effectively reported by the device.

Factor Relationship Diagram (FRD): A graphical depiction of an experiment consisting of design structure, unit structure and line(s) of restriction depicting partitioning of the unit structure.

Nominal Data: A set of data is said to be nominal if the values/ observations belonging to it can be assigned a code in the form of a number where the numbers are simply labels. You can count but not order or measure nominal data. For example, in a data set males could be coded as 0, females as 1; marital status of an individual could be coded as Y if married, N if single.

Operational Definitions (see Deming[13]):

1. All persons have the same understanding of the criteria on which decisions are to be made.
2. There exists a consistent, agreed upon method for the evaluation or measurement of the response metric.
3. The decision (e.g., Good/Bad, category 1-5) made from the evaluation is the same; irrespective of the person making the decision.

[13]Deming, W. Edwards (1986) "*Out of the Crisis*" MIT Press

Precision: The variation between successive measurements obtained under stipulated conditions.

1. Repeatability: Precision where the conditions are the same characteristic on the same part by the same person using the same instrument.
2. Reproducibility: Precision where conditions include using different operators on different instruments (or labs).

Sampling Plan (Sampling Tree): Graphical depiction of the procedure to acquire the units and the relationship of layers to hypotheses (thought map) and x's (process/product map).

Scientific Method: the iterative process of induction and deduction.

Statistics: the science of extracting information from data. This *science* includes the collection, analysis, interpretation and communication of information based on data.

References

1. Likert, Rensis (1961), *New Patterns of Management*, New York: McGraw-Hill Book Co.
2. Likert, Rensis (1932), "A Technique for the Measurement of Attitudes", *Archives of Psychology*, New York, Vol. 22, No. 140
3. Uebersax, John (2006) "Likert Scales: Dispelling the Confusion", *Statistical Methods for Rater Agreement* website
4. Hild, Cheryl, D. Sanders (2000) "The Thought Map", *Quality Engineering*, Vol. 12, No. 1.
5. Sanders, Doug, W. Ross, and J. Coleman (2000), "The Process Map", Quality Engineering, Vol. 11, No. 4, 2000.
6. Hild, Cheryl and Doug Sanders (2008), "Factor Relationship Diagrams: A Tool for Experimenters", *Wiley's Encyclopedia of Reliability and Statistics.*
7. Sanders, Doug and Jim Coleman (1999), "Considerations Associated with Restrictions on Randomization in Industrial Experimentation", *Quality Engineering*, Volume 12, No. 1
8. Daniel, Cuthbert (1976) *Applications of Statistics to Industrial Experiments*, Wiley
9. Deming, W. E. (1982), *Out of The Crisis,* Cambridge, MA: MIT Center for Advanced Engineering Study
10. Wheeler, Donald and Lyday, Richard (1984) *Evaluating the Measurement Process*, SPC Press, Inc., Knoxville, TN

JMP Pro by SAS statistical software used for analysis.