A CRITICAL LOOK AT ACCUMULATION ANALYSIS AND RELATED METHODS

M.S. Hamada and C.F.J. Wu

IIQP Research Report
RR-88-03

November 1988

(Revised November 1989)

A CRITICAL LOOK AT ACCUMULATION ANALYSIS AND RELATED METHODS*

M. Hamada and C.F.J. Wu

Department of Statistics and Actuarial Science
University of Waterloo
April 1989

ABSTRACT

Industrial quality characteristics are often measured categorically rather than numerically, such as recording a response as "slight", "moderate", or "extreme." Accumulation analysis, a method proposed by Taguchi (1974) for analyzing ordered categorical data from industrial experiments, is used in Japanese industry and is becoming popular in the United States. Nair (1986) proposed using the first two components of the accumulation analysis statistic separately as well as simpler alternatives to detect location and dispersion effects, respectively. We expose some problems with accumulation analysis in the multifactor setting, the usual industrial setting since it is more efficient to simultaneously investigate many factors. Our results show that accumulation analysis detects spurious factor effects and reverses the order of factor importance. Furthermore, reanalysis of data from two real experiments reveals that these problems with accumulation analysis are realized in practice. We demonstrate an inherent problem with detecting dispersion effects from ordered categorical data. Even in the absence of this problem, we show that the dispersion tests still detect spurious effects and reverse the order of factor importance more seriously than does accumulation analysis. On the other hand, the location tests are generally useful, especially Nair's simple alternative which happens to be the Kruskal-Wallis test. Moreover, we provide an explanation of why location tests provide a particularly sensible method of analysis in the industrial context. We also consider other alternatives: the method of scoring categories, a mean response model, and a proportional odds model. The method of scoring the categories is simple and particularly effective if the scores are reasonably chosen.

Key Words: quality improvement, ordered categorical data, multifactor experiments, location effects, dispersion effects, Kruskal-Wallis test, method of scoring categories, mean response model, proportional odds model.

* This report will appear with discussion in the May 1990 issue of *Technometrics*.

1. Introduction

Accumulation analysis (henceforth abbreviated as AA) is a method proposed by Taguchi (1974) for analyzing ordered categorical data from industrial experiments. Taguchi advocates AA as a superior alternative to Pearson's chi-squared test because "it accounts for the ordered nature of the categories." He also criticized the chi-squared test for *reversing the order of importance among factors and detecting spurious significant factors* (Taguchi 1987, p. 105). AA has been used in many Japanese industrial studies. In North America it is becoming increasingly popular. For example, in the *Fourth Symposium on Taguchi Methods,* almost a quarter of the case studies used accumulation analysis. Nair (1986) and Box and Jones (1986) pointed out several unnecessary complications of Taguchi's AA statistic and proposed using its numerator sum of squares instead. Throughout the paper we refer to this modified version as the AA statistic. Nair (1986) showed that in the single factor setting the AA statistic can be expressed as a weighted sum of score statistics. Since the scores are linear, quadratic, etc., the first two components have been interpreted as tests for location and dispersion effects. Then he proposed using these two components and simpler alternatives to test for location and dispersion effects.

The main purpose of our paper is to study the properties and particular shortcomings of AA and related methods and to consider methods more useful for analyzing ordered categorical data from industrial experiments. Previous authors considered these methods primarily for the single factor setting. This paper concentrates on the problems with these methods in the multifactor setting which is common in industrial experiments. For ANOVA on continuous data, no problems are encountered for orthogonal experiments. For ordered categorical data, however, the design's orthogonality does not insure independence of test statistics for different factors. This deficiency of ordered categorical data leads to serious consequences.

We begin with a brief review of AA and its components in Section 2. This includes AA's equivalence to analyzing collapsed tables in the multifactor setting and an interpretation of AA useful for explaining AA's problems. We reanalyze data from two real experiments in Section 3 to demonstrate that serious problems with AA and the dispersion tests are realized in practice.

In Sections 4 and 5, we study these problems with AA and related methods. In Section 4.1 we review some previous work on problems caused by analyzing collapsed tables. Results from a simulation study based on two models demonstrate that AA and the dispersion tests can often detect spurious effects. If a formal significance testing is not adopted, one can still identify important factors by ranking the factor effects. In Section 4.2 it is shown via two simulation studies that order reversal of factor importance is most serious for the dispersion tests, somewhat serious for AA, and not of much concern for the location tests. Recall that AA is a compromise between location and dispersion tests. In Section 4.3, we demonstrate that AA and the related tests for a factor depend on the other factors. This dependency is a main source of the problems observed in the multifactor setting.

Recall that Taguchi criticized Pearson's chi-squared test for the very same reasons, order reversal of factor importance, and spurious detection of effects. Detection of spurious effects and misidentification of important factors have serious practical implications. A more complicated situation may be perceived than really exists, unnecessary experimental effort may be expended to resolve ambiguities, and factor levels may be chosen that unnecessarily result in increased material costs and processing time.

In Section 5, we study the location and dispersion tests in the multifactor setting. We discuss and summarize the problems with the dispersion tests in Section 5.1. Unless some very stringent and difficult-to-verify conditions are met, they cannot be used to detect genuine dispersion effects. In some situations there is an inherent problem with detecting dispersion effects from ordered categorical data. But even in the absence of this problem, we show that the dispersion tests still detect spurious effects and reverse the order of factor importance more seriously than does accumulation analysis. In Section 5.2, we demonstrate how problems with these methods are also caused by a categorization effect; locations of observed distributions can appear different from those of underlying distributions. However, the location tests are shown to be generally useful and thus are recommended for identifying legitimate effects. Moreover, we provide an explanation of why location tests provide a particularly sensible method of analysis in the industrial context.

In Section 6 we consider some advantages and disadvantages of other alternatives by reanalyzing the first real data set from Section 3. The methods considered are the method of scoring categories, a mean response model, and a proportional odds model. The method of scoring the categories is simple and particularly effective if the scores are reasonably chosen. The paper concludes with a summary and discussion in Section 7.

2. Review of Accumulation Analysis and Related Methods

We begin by describing AA and related methods in the single factor setting and then consider their extensions to the multifactor setting.

Consider a single factor A with I levels and n observations taken at each level. Let $\mathbf{y_i} = (y_{i1}, \ldots, y_{iK})^T$ denote the frequencies of K ordered categories observed at the ith level of the factor; the data can be viewed as an I$\times$K contingency table. Let C_{ik} be the cumulative frequencies of the first k categories at the ith level and let $C_{\bullet k}$ be the average of the C_{ik} across the I levels. AA is so named because it analyzes these cumulative frequencies.

AA is an ANOVA-like procedure which can be seen as follows. AA builds K-1 I$\times$2 tables where the ith row of the kth table is $(C_{ik}, n-C_{ik})$. Standard ANOVA is performed on each table by analyzing it as if C_{ik} ones and $n-C_{ik}$ zeros had been observed at the ith level of the factor. AA then weights the sums of squares from the kth table ANOVA by $(d_k(1-d_k))^{-1}$, where $d_k = C_{\bullet k}/n$, and adds the corresponding weighted sums of squares over the K-1 tables. This yields the AA sums of squares:

$$SS_A = n\sum_{k=1}^{K-1}\sum_{i=1}^{I}(C_{ik}-C_{\bullet k})^2 \;/\; (C_{\bullet k}(n-C_{\bullet k})), \tag{2.1}$$

$SS_{tot} = nI(K-1)$, and

$$SS_e = SS_{tot} - SS_A = n\sum_{k=1}^{K-1}\sum_{i=1}^{I} C_{ik}(n-C_{ik}) \,/\, (C_{\bullet k}(n-C_{\bullet k})).$$

To obtain mean squares (MS), AA uses (K-1)(I-1) and (K–1)I(n–1) as degrees of freedom for SS_A and SS_e, respectively. Finally, AA calculates an F-like statistic given by $F_A = MS_A/MS_e$. Thus, AA is an ANOVA-like procedure.

AA's simplicity and similarity to ANOVA is appealing. Unfortunately, it does not possess ANOVA's property of independent sums of squares. Noticing that SS_e = constant – SS_A, Nair (1986) and Box and Jones (1986) pointed out the undesirable property that SS_e depends on the effect of factor A. Consequently, they proposed using only the numerator sum of squares SS_A from (2.1). In the following, we also consider this modified statistic and refer to it as the AA statistic T (T_A for factor A). Furthermore, Nair (1986) noted that in the multifactor setting, the modified AA statistic eliminates one way in which the AA statistic for a factor depends on the other factors. For example, this can be easily seen in the two-factor main effects setting, where the distribution of the original AA statistic for factor A depends on factor B since SS_e = constant – SS_A – SS_B.

Nair (1986) and Box and Jones (1986) decomposed the AA statistic by solving an eigenvalue problem as

$$T = \sum_{j=1}^{K-1} \lambda_j Z_j^2 .$$

Nair interpreted the first two components, Z_1^2 and Z_2^2, as tests for location and dispersion effects, respectively. Because the weights λ_j decrease rapidly in j, Nair interpreted AA as a test primarily for location.

Nair suggested using the first two components of the AA statistic separately to increase the power of detecting location and dispersion effects. He also proposed using simpler alternatives SS(**l**) and SS(**d**) with data-based scores **l** and **d** (see (5.1)-(5.5) of Nair(1986)), where

$$SS(\mathbf{l}) = \sum_{i=1}^{I} (\mathbf{l}^T(\mathbf{y}_i - \mathbf{y}_\bullet))^2 /n ,$$

and

$$SS(\mathbf{d}) = \sum_{i=1}^{I} (\mathbf{d}^T(\mathbf{y}_i - \mathbf{y}_\bullet))^2 /n ,$$

and $\mathbf{y}_\bullet = \sum_{i=1}^{I} \mathbf{y}_i / I$. Interestingly, SS(**l**) can be shown to be equivalent to the Kruskal-Wallis statistic. Note that the scores **l** and **d** depend only on the marginal frequencies of the ordered categories.

The components Z_j^2 can also be shown to have the same form as SS(**l**) and SS(**d**):

$$Z_j^2 = \sum_{i=1}^{I} (\mathbf{s}_j^T(\mathbf{y}_i - \mathbf{y}_\bullet))^2 /n , \qquad (2.2)$$

where the $\mathbf{s}_j$ also depend only on the marginal frequencies of the ordered categories. Since $(\mathbf{y}_i - \mathbf{y}_\bullet)$ can be interpreted as a comparison of the distribution at the ith level of the factor with a reference distribution (the mixture of all the distributions at the I levels), Z_j^2 can be viewed as a comparison of the distributions at the I levels of the factor with a reference distribution. Since Z_j^2 tests whether the factor has an effect with respect to the scores $\mathbf{s}_j$, AA can be viewed as a weighted combination of K-1 different tests to detect a factor effect.

Next we describe AA in the multifactor setting. Consider a multifactor experiment using a fractional factorial design (or more generally an orthogonal array) with r runs and n observations taken at each run on a K ordered categorical response. Denote the frequency of the jth category for the ith run by n_{ij}. For the extension of AA to the multifactor setting, K-1 multiway ANOVA's (over all the factors) are performed. The data for the kth multiway ANOVA consists of $\sum_{j=1}^{k} n_{ij}$ ones and $n - \sum_{j=1}^{k} n_{ij}$ zeros for the ith run. The AA sums of squares (i.e., the AA statistics T) are then formed by adding the corresponding weighted ANOVA sums of squares (using the same weights as in the single factor setting).

Recall that in multiway ANOVA, the sum of squares for a particular factor is the same as that for a one-way ANOVA treating the multifactor experiment as a single factor experiment. Consequently, the AA statistic for a factor main effect, say factor A with I levels, is equivalently obtained up to a constant by *collapsing* the design onto the factor (producing an I×K contingency table) and by calculating the AA statistic for the single factor as was described above. It follows that the extension of the data-based scoring methods to the multifactor setting which includes AA's components can be explained as applying the single factor procedure to this collapsed table.

3. Problems Encountered in Analyzing Real Data

In this section we reanalyze data from two real experiments to demonstrate some problems with AA and the dispersion tests. These problems are studied in Sections 4 and 5. In Section 6 we consider some alternative methods and apply them to the first experiment.

3.1. An Arc Welding Experiment

This experiment was performed by the National Railway Corporation of Japan and reported in Taguchi and Wu (1980). One facet of the experiment was to find the important factors which affect the workability of an arc welded section between two steel plates. Workability is the degree of difficulty in welding the two steel plates together which was classified into three categories: easy, normal, and difficult. The experimenters were initially interested in nine factors (A-I) and four two-factor interactions (AG, AH, AC, GH). An experiment using a 2^{9-5} fractional factorial design was performed with one observation per run. The experimental design and workability data appear in Appendix 1 (x-ray response data are also included which Koch, Tangen, Tudor and Stokes will use in their discussion).

Table 3.1 presents the results of AA, its first two components, Z_1^2 and Z_2^2, SS($\mathbf{l}$), and SS($\mathbf{d}$). For this data, the AA statistic $T = 1.28\, Z_1^2 + .72\, Z_2^2$. Using AA, the original analysis concluded that the main effects for factors D, F, and G were significant. However, Z_1^2 or SS($\mathbf{l}$) identify only the main effects for factors D and F as important location effects, while Z_2^2 or SS($\mathbf{d}$) identify main effect G as an important dispersion effect.

The results in Section 4 suggest that for this experiment the factor G main effect that AA detected was spurious; they demonstrate that AA detects spurious interactions when the corresponding main effects are significant. For this experiment, this very situation holds since the factor G main effect is confounded with the DF interaction effect, where D and F main effects are significant. Furthermore, related results in Section 5 show that if the dispersion tests had been used, they would have identified a spurious factor G dispersion effect.

Table 3.1: Results of AA, Z_1^2, Z_2^2, SS(**l**), and SS(**d**) for the Welding Experiment

Factor	AA	Z_1^2	Z_2^2	SS(**l**)	SS(**d**)
A	1.74	.10	2.23	.18	2.15
B	.41	.28	.16	.13	.31
C	1.74	1.26	.18	1.32	.12
D	5.03	3.70	.41	3.52	.59
E	.41	.16	.28	.13	.31
F	9.03	7.01	.10	7.09	.02
G	5.03	.23	6.55	.09	6.69
H	.41	.16	.28	.13	.31
I	1.74	1.26	.18	1.32	.12
AG	.41	.16	.28	.13	.31
AH	1.74	.10	2.23	.18	2.15
AC	.41	.16	.28	.13	.31
GH	1.74	.10	2.23	.18	2.15

Taguchi uses a simple way to choose the "optimal" levels of the important factors (Taguchi and Wu 1980). He simply looks at the marginal frequency table for each factor and chooses the level with the most desirable distribution (i.e. the distribution with the most in the most desirable category or with the least in the least desirable category). For this data, this yields recommendations of setting factor D at the low level and factor F at the high level. Since AA erroneously identified factor G in the original analysis, it was recommended to set factor G at the low level. It turns out that for this experiment, the low level was in fact more costly to use.

3.2. A Contact Stain Experiment

A company produces a rubber product which must meet a contact stain specification; it should not stain or mar the painted panel to which it is attached. The data and original analysis are found in Lear and Stanton (1985). A 2^{4-1} fractional factorial experiment (4=123) was performed to determine the important factors affecting the contact stain characteristic of the product. The four factors were chemical compounds used in making the rubber product. For each run, one product was attached to a painted metal panel and subjected to high temperature for three days. The panel was then inspected and its contact stain characteristic was classified as one of the following: none to very slight, slight to moderate, and moderately severe to severe. The experimental design and stain data appear in Appendix 2. Table 3.2 presents the results of AA, its first two components, Z_1^2 and Z_2^2, SS(**l**), and SS(**d**). For this data, the AA statistic $T = 1.45\,Z_1^2 + .55\,Z_2^2$.

Table 3.2: Results of AA, Z_1^2, Z_2^2, SS(**l**), and SS(**d**) for the Stain Experiment

Factor	AA	Z_1^2	Z_2^2	SS(**l**)	SS(**d**)
A	3.20	.28	5.05	.11	5.23
B	.53	.18	.48	.24	.43
C	7.47	5.05	.28	5.23	.11
BC	.53	.18	.48	.24	.43
D	.53	.18	.48	.24	.43
BD(=AC)	3.20	1.93	.74	1.71	.96
CD	.53	.18	.48	.24	.43

Based on the AA statistics, the original analysis declared the main effects for factors A and C and the BD interaction significant (actually BD is aliased with AC and is referred to as AC in the following). However, using Z_1^2 or SS(**l**), factor C main effect is the only significant location effect, although the BD interaction effect appears to be relatively important. As was mentioned for the previous experiment, the results in Section 4 show that AA detects spurious interactions when the corresponding main effects are significant. The difference in the conclusion for the factor A main effect can be explained by the fact that for this experiment, factor A main effect is confounded with the AC×C interaction effect, where C and AC are relatively important. Thus for this experiment, AA spuriously detected a factor A main effect. Moreover, while Z_2^2 or SS($\mathbf{d}$) seem to identify main effect A as an important dispersion effect, related results in Section 5 suggest that this apparent important effect for this experiment is also spurious.

We can use the $\mathbf{s_1}$ scores from Z_1^2 or the $\mathbf{l}$ scores from SS($\mathbf{l}$) to choose the "optimum" level of C. By adding up the product of the category frequencies with their corresponding scores for each level, choose the level whose quantity is smallest (since the "none to very slight" category is the most desired). For this data, the high level of factor C is the optimum level. Because AA also detected a factor A main effect and a BD interaction effect, the original recommendations (Lear and Stanton, 1985) chose the high level of factor A and the low levels of factor B and D. Based on our correct statistical analysis, these additional recommendations are suspect. Although these recommendations do not affect the contact stain characteristic, they may have significant financial implications. The difference in our recommended settings from that of the original authors cannot be judged solely by statistical analysis. Engineering judgement and compromise settings when several quality characteristics are involved also need to be considered.

A cautionary note about both experiments is warranted: only one observation was taken per run and the responses were classified into three ordered categories. The small number of replications and categories may not provide enough information for detecting smaller yet important factors.

4. Pitfalls of AA and Related Methods

In this and the next section we explain why AA and the dispersion tests spuriously detected effects in the real experiments. We review some previous results in Section 4.1 and then present new results for AA in Section 4.2 and 4.3. Section 5 concentrates on the location and dispersion tests in the multifactor setting.

4.1. Some Previous Results

Nair (1986) and Hamada and Wu (1986) exposed some inherent problems with analyzing ordinal data as well as problems with AA and the dispersion tests. The salient findings are summarized in the following.

(1) Since the scores $\mathbf{s_2}$ and $\mathbf{d}$ in $Z_2{}^2$ and SS($\mathbf{d}$) are functions of only the marginal frequencies, neither test can simultaneously account for the different locations of the distributions at the I levels of the factor. *They can only be interpreted as tests for dispersion in the single factor setting when there is no location effect.* A simple example to illustrate this problem was given in Hamada and Wu (1986). A simulation study found in Hamada and Wu (1988) further demonstrates this problem under a more general situation.

(2) Unlike continuous data, there is an inherent problem with detecting dispersion effects with ordered categorical data. Nair (1986) pointed out that it is *hopeless to make any reasonable inference about dispersion effects in the presence of strong location effects.* He observed that with strong location effects some levels of the factor appear to have less dispersion than others since most of the observations are pushed into the extreme category. Thus, any reasonable method would detect a spurious dispersion effect in this situation.

(3) Since the AA statistic for a factor in the multifactor setting is computed by collapsing the design onto the factor, by using the interpretation of AA following (2.2) Hamada and Wu (1986) noted that AA is no longer comparing I distributions but I *mixtures* of distributions. This suggests the possibility of the I *mixtures* of distributions being different even when there is no factor effect. Thus, AA has the *potential for detecting spurious effects.* A simple example to demonstrate this was given in Hamada and Wu. To further understand this problem, we study the performance of AA and the two dispersion tests under more realistic situations via simulation studies for underlying location and dispersion models. Note that in industry latent models are often natural ones.

Example 1: Consider an experiment with a 2^{3-1} fractional factorial design (I=123). For the underlying location model, suppose the ordinal response is generated by an underlying continuous random variable Y, where $Y = A + B + C + \epsilon$ and $(-\infty, -1, 0, 1, +\infty)$ defines four ordered categories. Let A=0, B=±1, C=±.5, and ϵ be $N(0, \sigma^2)$. For the underlying dispersion model, $Y = (\exp(A + B + C))\epsilon$, where A=0, B=±.75, C=±.5, ϵ is $N(0, \sigma^2)$, and $(-\infty, -1.25, 0, 1.25, +\infty)$ defines the categories. Ten observations are taken at each design setting. Table 4.1 summarizes the size of the .05 and .10 tests for factor A based on 10,000 simulations. A $d\chi^2_v$ and χ^2_1 approximation (Nair 1986) was used for AA and the dispersion tests, respectively. Note that factor A has neither a location nor a dispersion effect in both models.

Table 4.1: Size of .05 and .10 Tests for
Factor A Effect Using AA, Z_2^2, and SS(**d**)

σ^2	Nominal Level	AA		Z_2^2 & SS(**d**)	
		Location Model	Dispersion Model	Location Model	Dispersion Model
1/4	.05	.85	.13	1.00	.45(.49*)
	.10	.94	.33	1.00	.64(.66*)
1/2	.05	.39	.12	.92	.34
	.10	.60	.26	.96	.49
1	.05	.08	.06	.55	.12
	.10	.19	.13	.67	.22

* Size of SS(**d**)

For these two scenarios, AA detects the difference between the dispersions of the two mixtures of distributions. The second component Z_2^2 detects this difference as demonstrated by Table 4.1 The substantial dispersion difference between the mixtures of distributions explains the dramatic results for the location model. For the dispersion model, AA appears to be performing quite well for large σ^2. However, there is still a (smaller) dispersion difference between the mixtures of distributions. AA does not detect this difference because the ten observations per run and four ordered categories do not provide enough information to do so.

The results for the two models for the 2^{3-1} design above also suggest what could happen in the full factorial setting. Since factor A has no effect, these examples can be viewed as a 2^2 factorial experiment in factors B and C. The test for factor A would correspond to a test for BC interaction. The examples above show that *AA can falsely detect an interaction effect*. This can seriously mislead the experimenter to consider a more complicated situation than is necessary.

4.2. Order Reversal of Factor Importance

For the purpose of product or process improvement, the interest is often in identifying factorial effects that are more important in a relative sense. Formal testing of significance of these effects is less relevant here to achieve this and may not be adopted in practice. We demonstrate a problem with AA and related methods when they are used to assess the *relative importance* of the factors. The factor associated with the largest AA statistic is identified as the most important and so on. The next two examples serve as warnings that this procedure can yield misleading results. The first example demonstrates that not only can a spurious factor be detected but it can be erroneously identified as more important than two real factors. The second example demonstrates that the least important real factor can be erroneously identified as the second most important factor.

Example 2: For an experiment with a 2^{7-4} fractional factorial design (D=ABC, E=AB, F=AC, G=BC), suppose that the ordered categorical random variable is generated by an underlying continuous random variable Y, where $Y = A + B + C + D + E + F + G + \epsilon$. Let A=$\pm$1, B=$\pm$.75, C=$\pm$.5, D=$\pm$.45, E=F=G=0, and ϵ be N(0, 1). Four ordered categories are defined by ($-\infty$, –1, 0, 1, $+\infty$) and five observations are taken at each run. Table 4.2 presents the probability of declaring factors A through G significant based on 5000 simulations. Clearly, AA declares factor E to be more important than both real factors C and D. Because of the confounding relation E=AB=CD, the dispersion difference in the mixtures of distributions for factor E which AA detects is attributed to the relatively large A and B effects as well as the somewhat smaller C and D effects.

Table 4.2: Probability of Detecting Factors
for Example 2 Using .10 Tests

	Factor						
	A	B	C	D	E	F	G
AA	1.00	.82	.25	.16	.30	.11	.10
Z_1^2	1.00	.85	.28	.18	.03	.02	.02
Z_2^2	.03	.06	.08	.08	.80	.48	.46
SS(**l**)	1.00	.84	.28	.17	.01	.02	.02
SS(**d**)	.03	.06	.07	.07	.81	.49	.46

Example 3: Consider an experiment with the same setup as in Example 2 but with A=±1, B=±.5, C=±.55, D=±.45, E=±.40, F=G=0. Table 4.3 presents the probability of declaring factors A through G significant based on 5000 simulations. AA identifies the least important real factor E as the second most important factor. This can be explained by the confounding pattern E=AB=CD as was discussed in the previous example.

Table 4.3: Probability of Detecting Factors
for Example 3 Using .10 Tests

	Factor						
	A	B	C	D	E	F	G
AA	1.00	.40	.45	.21	.53	.15	.11
Z_1^2	1.00	.37	.48	.23	.40	.03	.03
Z_2^2	.03	.24	.09	.08	.52	.55	.44
SS(**l**)	1.00	.34	.48	.24	.36	.03	.03
SS(**d**)	.03	.26	.08	.07	.56	.55	.43

From Tables 4.2 and 4.3, observe that it is primarily the dispersion component which accounts for AA's reversing the order of factor importance although from Table 4.3 the location component contributes as well. This order reversal property is worst for the dispersion tests, not as serious for AA, and least serious for the location tests. AA's performance reflects the fact that it is a combination of the two tests.

4.3. AA and Its Components Depend on All Factors

The simulation study in Section 4.1 showed that the distribution of the AA statistic for a *factor can depend on the other factors.* In fact, this dependence on the other factors can be given a theoretical justification.

We use the 2^{3-1} fractional factorial design to show this dependence. The four design settings for (A, B, C) are given in Table 4.4.

Table 4.4: 2^{3-1} Fractional Factorial Design

Run	Factor		
	A	B	C
1	+	+	+
2	+	-	-
3	-	+	-
4	-	-	+

Let $\mathbf{n_i} = (n_{i1}, \ldots, n_{iK})^T$, where $\sum_{j=1}^{K} n_{ij} = n$. Denote the category probabilities at the ith run by $\mathbf{p_i} = (p_{i1}, \ldots, p_{iK})^T$ and the average of the $\mathbf{p_i}$ by $\mathbf{p_\bullet}$. For I = 2, the AA statistic T can be expressed as

$$T = \mathbf{Z}^T \mathbf{\Lambda} \mathbf{Z} = \sum_{j=1}^{K-1} \lambda_j Z_j^2 ,$$

where $\mathbf{Z} = \mathbf{Q}^T (1/\sqrt{n})(\mathbf{y_1} - \mathbf{y_\bullet})$, $\mathbf{\Lambda} = \text{diag}(\lambda_j)$, $(\mathbf{y_1} - \mathbf{y_\bullet}) = \sum_{i=1}^{4} \pm \mathbf{n_i}/2$ and the sign of $\mathbf{n_i}$ for each factor is given in Table 4.4. For factors A and B, T_A and T_B can be expressed as $T_A = v_A{}^T v_A$ and $T_B = v_B{}^T v_B$, where

$$v_A = (1/2)\boldsymbol{\Lambda}^{1/2}\mathbf{Q}^T(1/\sqrt{n})(\mathbf{n_1} + \mathbf{n_2} - \mathbf{n_3} - \mathbf{n_4})$$

and

$$v_B = (1/2)\boldsymbol{\Lambda}^{1/2}\mathbf{Q}^T(1/\sqrt{n})(\mathbf{n_1} - \mathbf{n_2} + \mathbf{n_3} - \mathbf{n_4}) \ .$$

Then the covariance of v_A and v_B is easily calculated as

$$\mathrm{cov}(v_A \, , v_B) = (1/4)\boldsymbol{\Lambda}^{1/2}\mathbf{Q}^T(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_3 + \boldsymbol{\Sigma}_4)\mathbf{Q}\ \boldsymbol{\Lambda}^{1/2} \ .$$

Thus v_A and v_B are dependent except when there are no factor effects (all $\mathbf{p}_i$ equal so that all $\boldsymbol{\Sigma}_i$ equal) or when only one factor (either A or B) has an effect. Generally, T_A and T_B are dependent. By using the same argument, one can show an analogous dependence property of Z_1^2, Z_2^2, SS($\mathbf{l}$), and SS($\mathbf{d}$).

In general, if none of the factors has an effect, i.e., the $\mathbf{p}_i$ are equal, then as $n \rightarrow \infty$, Z_j^2 are independent χ^2_1 and the AA statistic T is distributed as a linear combination of independent χ^2_1 random variables. Under the more general assumption that $\mathbf{p}_i$ approaches a common value at the rate $n^{-1/2}$, i.e.,

$$\mathbf{p}_i = \mathbf{p}_\bullet + n^{-1/2}\boldsymbol{\delta} \tag{4.1}$$

where $\boldsymbol{\delta}$ is a constant independent of n, as $n \rightarrow \infty$ Z_j^2 are independent noncentral χ^2_1 and T is a weighted sum of these independent random variables. The noncentrality parameter of Z_j^2 for a factor does not depend on the other factors. Both results can be found in Nair (1987). Note however that (4.1) is an unrealistic assumption for most factorial experiments.

5. Location and Dispersion Tests in the Multifactor Setting

The results in Section 4 provide a compelling reason to consider alternatives to AA. Nair (1986) proposed using the first two components of the AA statistic separately as well as their simpler alternatives to increase the power of detecting location and dispersion effects. In Section 4.1, we noted that there already was a problem with the dispersion tests in the single factor setting as well an inherent problem of analyzing dispersion effects from ordinal data when strong location effects are present. In Section 5.1, we show that the dispersion tests are rarely applicable in the multifactor setting which suggests that they would have detected spurious dispersion effects for the real experiments in Section 3. In Section 5.2, we investigate the location tests and show that they are generally useful. Here we provide an argument why these tests are especially useful for typical industrial experiments with ordinal data.

5.1. Dispersion Tests Are Rarely Applicable

In point (3) of Section 4.1, we noted that AA was comparing mixtures of distributions in the multifactor setting. A similar result follows for the dispersion tests Z_2^2 and SS(**d**). In the multifactor setting they are rarely applicable as dispersion tests as the following restrictive conditions show:

(1) the factor has no location effect

(2) at most one of the other factors has a location effect, and

(3) at most one of the other factors has a dispersion effect.

The conditions must hold so that a difference in the mixtures of distributions implies a real dispersion effect. The examples in Section 4 demonstrate the serious consequence of detecting spurious dispersion effects when these stringent conditions are violated. For these examples AA detected spurious effects when a dispersion difference between the mixtures of distributions was picked up by the large Z_2^2 component. Table 4.1 demonstrates that spurious dispersion effects can be detected quite often when the dispersion tests are used. Further evidence is provided in Tables 4.2 and 4.3 in which three and four spurious dispersion effects are often detected, respectively.

Spurious detection of effects for the dispersion model from Table 4.1 provides additional insight into the performance of the dispersion tests. Recall that in the presence of strong location effects, any reasonable method would detect spurious dispersion effects; here, spurious detection is an intrinsic problem with the data, not something that can be attributed to any particular method. However, this is not true for data from the dispersion model where there are no location effects. Whereas the dispersion tests which are based on analyzing marginal tables, detect spurious effects, any method not based on marginal tables, such as modeling frequency counts (McCullagh 1980) or analyzing an analogue of standard error based on scoring the ordered categories, would not.

5.2. Location Tests Are Generally Useful

Next we consider the merits of the location tests Z_1^2 and SS(l) by simulation studies. While there are some problems for the dispersion model, the location tests are generally useful for the location model. The next two examples demonstrate that the location tests can detect spurious effects. First we consider their performance for an underlying dispersion model.

Example 4: For the same 2^{3-1} fractional factorial design from Example 1 of Section 4.1, let $Y = (\exp(A + B + C))\epsilon$, where ϵ has a standard logistic distribution. The ordered categories are defined by $(-\infty, -.08, .28, .85, 2.09, +\infty)$. Let A=0, (B, C) = $(\pm.5, \pm.25)$, and $(\pm1, \pm.5)$. Note that factor B has no location effect. Table 5.1 presents the size of .05 and .10 tests for a factor B location effect using Z_1^2 and SS(l) based on 10,000 simulations. The results show that the tests can detect spurious location effects quite often.

Table 5.1: Size of .05 and .10 Tests for Factor B Location Effect Using Z_1^2 and SS(l) Under the Dispersion Model of Example 4

(B, C)	Nominal Level	Statistic	
		Z_1^2	SS(l)
(±1, ±.5)	.05	.35	.16
	.10	.50	.25
(±.5, ±.25)	.05	.14	.10
	.10	.23	.17

These results can be explained by the choice of category boundaries. Although the underlying distributions have the same location, the asymmetric boundaries (with respect to the location) cause the observed distributions to appear to have different locations. We call this the *categorization* effect. It is this apparent difference that Z_1^2 and SS(l) are detecting. Since the first category contains approximately 50% of the distribution, the observed location tends to shift to the right as the variance is increased. The location tests detect more spurious effects for (B, C) = (±1, ±.5) since the variances vary more, causing more disparate observed locations. Observe that SS(l) is less sensitive than Z_1^2 to this categorization effect. Based on other studies not reported here, the location tests did not detect spurious effects if the boundaries are symmetric.

Next we consider an example that demonstrates the categorization effect can cause the detection of spurious effects for the underlying location model.

Example 5: Table 5.2 presents the simulation results for the .05 and .10 tests using Z_1^2 and SS(l) under the location model from Example 1 of Section 4.1 except that $(-\infty, 0, 1, 2, +\infty)$ defines the category boundaries. We see the categorization effect here: the asymmetric boundaries cause the locations of the mixtures of distributions of factor A to appear different when they are not. It is this observed difference in locations that Z_1^2 detects so that a factor A effect can be spuriously detected often. However, SS(**l**) performs differently. The size of the factor A test is somewhat conservative. This suggests that SS(**l**) is less sensitive to the categorization effect as was also seen in Table 5.1 for the dispersion model.

Table 5.2: Size of Tests for Factor A and Power of Tests for Factors B and C Using Z_1^2 and SS(**l**) Under the Location Model of Example 5

Category Boundaries = $(-\infty, 0, 1, 2, +\infty)$

σ^2	Level	Factor				
		A		B	C	
		Z_1^2	SS(**l**)	Z_1^2 & SS(**l**)	Z_1^2	SS(**l**)
1/4	.05	.24	.01	1.00	.65	.31
	.10	.54	.06	1.00	.85	.58
1/2	.05	.15	.01	1.00	.58	.42
	.10	.34	.05	1.00	.77	.64
1	.05	.08	.01	1.00	.45	.41
	.10	.18	.05	1.00	.62	.58

It was pointed out in Section 4 that Z_1^2 and SS(**l**) for a factor depend on all the other factors. This dependence on the other factors causes small location effects to be missed in the presence of factors with larger effects. This can be seen by examining the noncentrality parameters of the approximate χ^2 distribution of these statistics. For two-level factors, the noncentrality parameter depends on $\mathbf{c}^T\boldsymbol{\eta}$, where $\mathbf{c}$ represents the scores $\mathbf{s}_1$ or $\mathbf{l}$ and $\boldsymbol{\eta} = (1/n)E(\mathbf{y}_1 - \mathbf{y}_\bullet)$. Note that $\pm\boldsymbol{\eta}$ compares how different the mixtures of distributions at the levels of the factor are from a reference distribution. Since $(1/n)E(\mathbf{y}_\bullet)$ is independent of which factor is being considered, $(1/n)E(\mathbf{y}_1)$ is more different from $(1/n)E(\mathbf{y}_\bullet)$ for a factor with a large effect than it is for a factor with a small effect. Therefore, the noncentrality parameter for a factor with a large effect is larger than that for a factor with a small effect. In fact, suppose that one factor has a substantially larger effect than another. Then the mixtures of distributions of the factor with the smaller effect is very similar. That is, $\boldsymbol{\eta} \approx \mathbf{0}$. Therefore, factors with small effects in the presence of factors with large effects can be missed.

In Section 4.2, we exposed the problem that AA can reverse the order of factor importance. For many of the situations studied here, the location tests maintain the order of factor importance. (See Tables 4.1 and 4.2.) Note however that from Table 4.3, the location tests can reverse the order. In Example 3, the least important factor E was identified as the third most important factor. Here and in general the location tests still perform better than AA. Note that SS(l) reverses the order less often than Z_1^2, thus suggesting a preference for SS(l) over Z_1^2.

Finally, we give an explanation why the location tests are especially useful for typical industrial experiments with ordinal data. Because the desired category is usually one of the extreme categories, it is sufficient to look for strong location effects which result in most of the observations being pushed into the desired category. Moreover, this is the very situation where it is hopeless to look for dispersion effects.

6. The Arc Welding Experiment Revisited

In this section, we consider some alternative methods of analysis for the arc welding data of Section 3.1. We consider (1) fixed scoring, (2) a mean response model using weighted least squares (WLS) (Grizzle, Starmer, Koch 1969), (3) logistic regression, and (4) a proportional odds model (McCullagh 1980). While one example does not provide a comprehensive assessment of these methods, nevertheless it does suggest some advantages and disadvantages of these methods.

6.1. Fixed Scoring

An alternative yet simple way to analyze this data for location effects is to assign scores to the categories based on judgement, calculate effects, and use a normal or half-normal plot of the effects to identify important effects. Using different scoring schemes can lead to different conclusions. Figures 6.1 and 6.2 display the half-normal plots using the scores (0, 1, 2) and (0, 1, 5), respectively. (The two error effects are identified in the plots by z.) The latter scores might be chosen because the "Difficult" category is more costly. Using (0, 1 2), we identify the same effects as we did when using SS(l) and Z_1^2. However, using (0, 1, 5) results in the identification of an additional location G effect. If these latter scores are appropriate, then the G effect is a valid one. This demonstrates that the results of the analysis depend critically on the scores used.

The location tests can also viewed as data-based scoring methods. For this data, the scores for Z_1^2 and SS(l) are $(0, 1.36, 2.57)$ and $(0, 1.16, 2.54)$, respectively, which explains their almost identical values and agreement with the half-normal plot based on the scores $(0, 1, 2)$. From this perspective, the location tests use data-based scores which generally have no optimal properties. This suggests that if these methods are used, a necessary part of the analysis should be an inspection of the scores for reasonableness. Graubaud and Korn (1987) give an example where the data-based scores are not reasonable.

6.2. A Mean Response Model Using WLS

Agresti (1986) considered analyzing ordinal data by a mean response model (Grizzle, Starmer, and Koch 1969) which uses WLS to account for the categorical nature of the response. We used both sets of scores as above, (0, 1, 2) and (0, 1, 5). Table 6.1 displays the chi-squared statistics obtained from PROC CATMOD in SAS. Note that adding 0.01 to empty cells was necessary to obtain a solution. Inspection of Table 6.1 reveals that the mean response model gives the same conclusions as was obtained from the half-normal plots; D and F are important when using (0, 1, 2) and in addition G becomes important when using (0,1,5).

Table 6.1: Chi-Square Statistics from the Mean Response Model for the Welding Experiment Using Scores (0, 1, 2) and (0, 1, 5)

Factor	(0,1,2)	(0,1,5)
A	1.5	1.6
B	1.5	3.5
C	15.4	7.8
D	40.3	45.5
E	5.1	6.6
F	78.7	59.8
G	1.5	23.8
H	1.5	3.1
I	22.9	12.7
AG	0.1	2.1
AH	1.5	5.1
AC	1.5	0.8
GH	1.5	0.6

6.3. Logistic Regression

Another alternative is to combine the ordered categories into two categories and analyze this "binary" data by logistic regression. For this example, two separate analyses can be done by first combining the first two categories and then combining the last two categories. Hamada and Tse (1989) find that estimability problems can occur often because of the highly fractionated nature of the designs used and the paucity of data. A check of the existence conditions given in Hamada and Tse for both "binary" data sets revealed that even when only a main effects model is entertained (not even the full model with an additional four interactions), the estimates do not exist. Thus, logistic regression which might provide reasonable results if there were more data, is not applicable for this data set.

6.4. Proportional Odds Model

Rather than create "binary" data from the ordinal data, we might try to fit the data directly using a proportional odds model for ordinal data (McCullagh 1980). Generally the results obtained can also be sensitive to the model assumptions; for example, this method implicitly scores the categories since the underlying category boundaries are estimated. For the welding data, we also encountered estimability problems when fitting even the main effects model. As we found with logistic regression, there is not enough data to apply this method.

7. Summary and Recommendations

Unlike continuous data, there are several inherent problems with analyzing ordered categorical data. The data can suggest spurious dispersion effects in the presence of strong location effects. Furthermore, the dispersion tests have two deficiencies which cause spurious effects to be detected. They do not account for the different locations of the distributions being compared. Moreover, they analyze marginal tables so that mixtures of distributions are being compared in the multifactor setting. Since a difference between the mixtures of distributions need not imply a factor effect, they seldom detect legitimate dispersion effects. Therefore, we do not recommend using these dispersion tests. Because of the inherent problems with detecting dispersion effects, we doubt that any method will work. One final problem with ordered categorical data is that when the response has few categories, little information is available to detect real effects. While one would like to refine the categories as much as possible, there is a tradeoff between more information and increased difficulty in correctly classifying the response.

Taguchi (1987) promotes accumulation analysis (AA) as a central method of his analysis strategy. He proposes AA not only for handling ordered categorical data but also for off-scale data (see Phadke et al. (1983) for an application). While AA's ANOVA-like appearance perhaps makes it appealing, it shares none of ANOVA's desirable properties. A fundamental problem with AA in the multifactor setting is that it analyzes marginal tables. In fact, decomposing the AA statistic reveals that one of its components is a dispersion test based on marginal tables. Two serious consequences of using AA are detecting spurious effects and reversing the order of factor importance. It is interesting in this regard that in a discussion of how to reduce interaction effects, Taguchi indicates the need for proper analysis methods and singles out AA for handling ordered categorical data (Taguchi 1987, p. 171). Reanalysis of data from real experiments demonstrated that AA does indeed detects spurious effects in practice. Therefore, we do not recommend AA.

The problems encountered with the location tests revealed another inherent problem with analyzing these data rather than exposing a deficiency with these methods. Referred to as a categorization effect is the phenomenon that the mixtures of observed distributions can appear to have different locations, although the mixtures of the underlying distributions might have the same location. It is this apparent difference that the location tests detect as would any other method. The categorization effect is usually small so that the location tests are generally useful. Interestingly, the location test which is equivalent to the Kruskal-Wallis test is less sensitive to this phenomenon. Coupled with its simplicity, the Kruskal-Wallis test is recommended for detecting location effects. Moreover, since the desired category is usually one of the extreme categories in the industrial context, we are especially interested in finding strong location effects. By choosing the appropriate factor levels, most of the observations can be pushed into the desired extreme category.

Besides the location tests, what else can be recommended? The simple alternative of scoring the categories converts the categorical data into numerical data so that usual methods like half-normal plots of the effects can be employed. Although the results can be sensitive to the scores used, this method is generally useful if the scores are reliable. Recall that the location tests use data-based scores (see Section 6) so that if these scores are not reasonable measures for the categories, the results can be misleading. This highlights the importance of knowing the relative weights of the categories in a practical situation.

We considered a mean response model which gave similar results as the methods above. Perhaps this is not surprising as commented by a referee that a variety of location tests usually give similar results.

Another alternative is to model the frequencies. Since these methods are based on maximum likelihood estimation, there can be a problem with the existence of the estimates. This is especially so in the industrial context because of the highly fractionated nature of the designs and the paucity of data. Moreover, nonexistence of these estimates are more likely to occur when there are strong location effects, a situation of great interest for quality improvement. Another problem with this approach is computation. Unlike ANOVA for continuous data, many model fittings may be required for identifying suitable models. The results can also be sensitive to the model assumptions. For example, a proportional odds model (McCullagh 1980) implicitly scores the categories since the underlying category boundaries are estimated.

Two possible remedies for the nonexistence of estimates are to take a likelihood approach (Lawless and Singhal 1978) or Bayesian approach. For the likelihood approach, model selection is based on percent contribution of each factor rather than point estimation. Current research is focused on the feasibility of this approach. Regarding the Bayesian approach, the issues of computation and sensitivity to priors should be addressed.

Acknowledgements

We thank the editor and the referees for helpful comments on an earlier version. Our work benefitted from many insightful remarks and suggestions from Vijay Nair whom we especially thank. This research was supported by National Science Foundation Grant DMS-8420968, the Natural Sciences and Engineering Research Council of Canada, and a research grant from the Manufacturing Research Corporation of Ontario.

References

Agresti, A. (1984), *Analysis of Ordinal Categorical Data,* New York, NY: John Wiley.

Agresti, A. (1986), Discussion of "Testing in Industrial Experiments with Ordered Categorical Data," *Technometrics,* 28, 292-294.

Box, G. and Jones, S. (1986), Discussion of "Testing in Industrial Experiments with Ordered Categorical Data," *Technometrics,* 28, 295-301.

Graubard, B. I. and Korn, E. L. (1987), "Choice of Column Scores for Testing Independence in Ordered 2 x K Contingency Tables," *Biometrics,* 43, 471-476.

Grizzle, J.E, Starmer, C. F., and Koch, G. G. (1969), "Analysis of Categorical Data by Linear Model," *Biometrics,* 25, 489-504.

Hamada, M. and Tse, S. K. (1989), "When Do Estimates Exist from Censored Data in Industrial Experiments?," University of Waterloo Institute for Improvement in Quality and Productivity Research Report.

Hamada, M. and Wu, C. F. J. (1986), "Should Accumulation Analysis and Related Methods Be Used for Industrial Experiments?," Discussion of "Testing in Industrial Experiments with Ordered Categorical Data", *Technometrics,* 28, 302-306.

Lawless, J. F. and Singhal, K. (1978), "Efficient Screening of Nonnormal Regression Models," *Biometrics,* 34, 318-327.

Lear, C. and Stanton, J. (1985), "Contact Stain Requirement," in *Third Symposium on Taguchi Methods,* Romulus, MI: American Supplier Institute, Inc., 117-134.

McCullagh, P. (1980), "Regression Models for Ordinal Data," *Journal of the Royal Statistical Society,* Ser. B, 109-142.

Nair, V. N. (1986), "Testing in Industrial Experiments with Ordered Categorical Data," *Technometrics,* 28, 283-291.

Nair, V. N. (1987), "Chi-Squared-Type Tests for Ordered Alternatives in Contingency Tables," *Journal of the American Statistical Association,* 82, 283-291.

Phadke, M. S., Kackar, R. N., Speeney, D. V. and Grieco, M. H. (1983), "Off-Line Quality Control in Integrated Circuit Fabrication Using Experimental Design Optimization," *The Bell System Technical Journal,* 1, 1273-1309.

Taguchi, G. (1974), "A New Statistical Analysis for Clinical Data, the Accumulating

Analysis, in Contrast with the Chi-Square Test," *Saishin Igaku,* 29, 806-813.

Taguchi, G. (1987), *System of Experimental Design Engineering Methods to Optimize Quality and* Minimize Costs, White Plains, NY: UNIPUB/Kraus International Publications.

Taguchi, G. and Wu, Y. (1980), *Introduction to Off-Line Quality Control,* Nagoya, Japan: Central Japan Quality Control Association.

Appendix 1: Design and Data for Arc Welding Experiment

Factor									Workability*	X-ray Response		
A	B	C	D	E	F	G	H	I	Response	Good	Normal	Bad
1	1	1	1	1	1	1	1	1	N	3	1	0
1	1	2	2	2	2	1	1	2	N	3	1	0
1	2	2	1	1	1	1	2	1	N	4	0	0
1	2	1	2	2	2	1	2	2	E	2	2	0
1	2	2	1	1	2	2	1	2	E	2	2	0
1	2	1	2	2	1	2	1	1	D	4	0	0
1	1	1	1	1	2	2	2	2	E	2	2	0
1	1	2	2	2	1	2	2	1	D	3	1	0
2	2	2	1	2	1	1	1	2	N	2	1	1
2	2	1	2	1	2	1	1	1	N	3	1	0
2	1	1	1	2	1	1	2	2	N	3	1	0
2	1	2	2	1	2	1	2	1	N	4	0	0
2	1	1	1	2	2	2	1	1	E	0	3	1
2	1	2	2	1	1	2	1	2	D	3	0	1
2	2	2	1	2	2	2	2	1	N	1	3	0
2	2	1	2	1	1	2	2	2	N	4	0	0

* E=easy, N=normal, D=difficult

Appendix 2: Design and Data for Contact Stain Experiment

Factor				Stain*
A	B	C	D	Response
1	1	1	1	2
1	1	2	2	2
1	2	1	2	3
1	2	2	1	2
2	1	1	2	3
2	1	2	1	1
2	2	1	1	3
2	2	2	2	1

* 1=none to very slight, 2=slight to moderate, 3=moderately severe to severe

Figure 6.1: Half-Normal Plot for Workability Data
Using Scores (0,1,2)

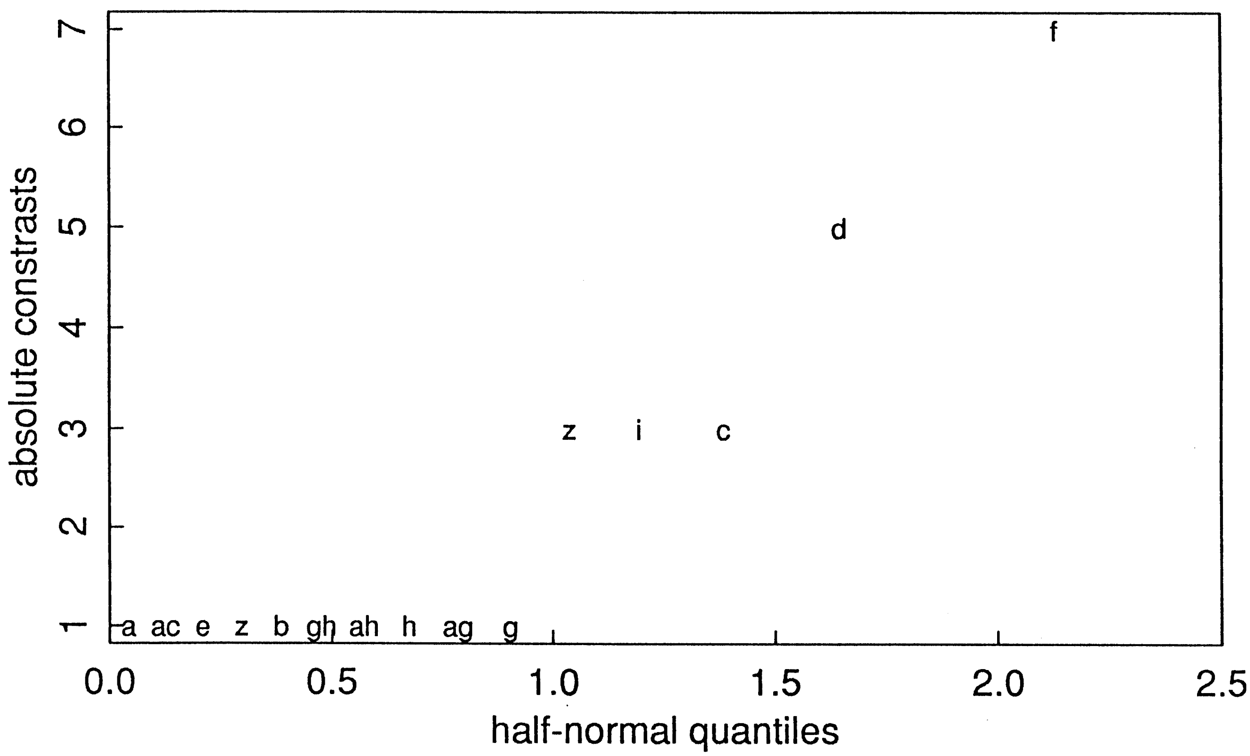

Figure 6.2: Half-Normal Plot for Workability Data
Using Scores (0,1,5)