Augmented Half Normal Effects Plots in the Presence of a Few Error Degrees of Freedom

Ulrike Grömping

Beuth University of Applied Sciences Berlin, Germany

This is the peer reviewed version of the following article:

Grömping, U. (2015). Augmented Half Normal Effects Plots in the Presence of a Few Error Degrees of Freedom. *Quality and Reliability Engineering International* **31**(7), 1185-1196,

which has been published in final form at http://onlinelibrary.wiley.com/doi/10.1002/qre.1842/abstract]. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Abstract

Fractional factorial 2-level experiments are often conducted without any error degrees of freedom. In such cases, a half-normal effects plot – also called Daniel plot according to its inventor Cuthbert Daniel[1,2] – can be used for assessing effect significance. Half-normal effects plots are often accompanied by a numeric method for assessing effect significance, most prominently Lenth's method[3]. There are, however, also situations for which a few error degrees of freedom are available, for example from a replicated center point run. For such cases, besides the obvious possibilities of either ignoring the few replicates (i.e. using half-normal effects plot and Lenth's method, as if they were not there) or using analysis of variance with the replicates for estimating the error variance, several further proposals for assessing effect significance exist (Larntz and Whitcomb[7], Edwards and Mee[8], JMP software[4]). This paper compares the published methods for significance testing, proposes an additional one (that might be very close to what JMP does) and advocates the use of an augmented half-normal effects plot that shows error points and a null reference line along with the effects. It is argued that such a plot can be useful in a fully-replicated experiment for assessing whether the replication process was of adequate quality. The method is available in the R package DoE.base[9].

Key words: Half-normal effects plot, Lenth's method, unreplicated fractional factorial 2-level design, center points, pure error

1 Introduction

In replicated experiments, significance is usually assessed with a linear model analysis, called Anova in the following: Assessment of significance relies on relating effect mean squares to the mean square error (*MSE*) as an estimate for the error variance; the term Anova, as used here, encompasses also the use of t-tests for 2-level factors. In unreplicated fractional factorial 2-level experiments, the *MSE* from

Anova is often undefined, because there are no degrees of freedom (df) for error. The most widely used methods for assessing effect significance for such experiments are visual inspection of (half-) normal effects plots as proposed by Daniel[1,2] and numeric assessment with Lenth's method[3]. Statistical software also sometimes offers Pareto charts of effects (e.g. JMP software[4], Minitab[5] or StatEase[6]); these do not provide a graphical assessment of significance for classical experiments, but can help in assessing relative effect sizes. If the experiment is not fully replicated but has only little replication information, Anova is often used but lacks power. Suggestions have been made to use modified versions of Lenth's method instead (Larntz and Whitcom[7]; Edwards and Mee[8]), and to enhance half-normal effects plots by error points (Larntz and Whitcomb[7]). This paper will make another proposal for that situation, based on half-normal effects plots in conjunction with the application of Lenth's method to an augmented set of estimates.

Mee[10] mentioned the need to take care of the additional df's in the analysis. According to him, most software switches to Anova for assessing effect significance, whenever there is at least one error df, in spite of the aforementioned lack of power. Occasionally, the analysis is accompanied by a half-normal effects plot, e.g. in Minitab®[5], where the plot shows t-values instead of coefficients or effects whenever error df are available, and adds a null reference line. There are two notable exceptions, where commercial software treats replications differently: Both Design-Expert® Version 8 (by Stat-Ease[6]) and JMP® (by SAS Institute CRT[4]) show error points on half-normal effects plots and adapt the significance assessment to account for the replications without switching to Anova. Design-Expert® uses the method published by Larntz and Whitcomb[7], while the JMP® screening platform, according to its documentation and the account by Mee[10], bases the analysis on partitioning the residual space into single df contrasts and obtains p-values via simulation, presumably applying a method which is quite similar to the one proposed here. The method proposed here has been implemented in the open-source software environment R[11] in the package DoE.base[9].

Section 2 provides notation and presents the existing methods, using an example by Montgomery[12]. Section 3 presents the proposed method and details two specific scenarios, a replicated center point and a fully replicated design; the latter is demonstrated using data published by Ding et al.[13] and unpublished data of an experiment from the author's industrial practice. Based on a simulation study, some detail for which is reported in the supplementary material, Section 4 discusses the properties of the proposed method in comparison to its competitors. The final section discusses the implications of the findings for statistical practice and statistical software.

2 Notation and Basics

2.1 Example data

Montgomery[12] reported a seven factor experiment on an injection molding application in 20 runs (16 cube runs plus four center point runs) with the four control factors A="temperature", B="screw

speed", C="holding time" and D="gate size" and the three noise factors a="cycle time", b="moisture content" and c="holding pressure" (notation from Wu and Hamada[15], exercise 22 of Chapter 11). The response variable (denoted as y) was the shrinkage in % of the molded product. The experimental plan was constructed by assigning the noise factors to three-factor interaction columns in a full factorial design in the control factors: a=ABC, b=BCD, c=ACD. The resulting design has resolution IV, and its coded data and 2-factor interaction columns are given in Table 1. The table also includes the response values, as well as two further columns (ABD and non-linearity) that saturate the model.

TABLE 1 ABOUT HERE

A model with main effects and two-factor interactions has seven df for the main effects and seven df for two-factor interactions (21 such interactions, grouped in seven confounded columns). Thus, together with the one df for the intercept, there are p=15 model matrix columns, N_{distinct}=17 different experimental setups and N=20 runs. Hence, there are $\text{df}_{\text{error}} = 20–15 = 5$ error degrees of freedom; two of these (df_{lof}=17–15=2) are lack-of-fit degrees of freedom that can be modeled away by including the three-factor interaction ABD and the nonlinearity check contrast (the last two columns in Table 1; see e.g. Wu and Hamada[15], Ch. 10.3 for construction of the non-linearity check contrast; in addition, it has been normalized to the same variance as the other coefficients), while there are three (df_{pe}=20–17) pure error degrees of freedom. This example will be used for explaining all methods.

2.2 The linear model, coefficients, effects, and the *MSE*

For formulae, we will consider the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where the Nxp model matrix $\mathbf{X}$ contains orthogonal columns for all effects under consideration, the fixed unknown px1 coefficient vector $\boldsymbol{\beta}$ contains the coefficients for the columns of $\mathbf{X}$ in the true underlying model, the Nx1 vector $\boldsymbol{\varepsilon}$ contains random error, independent and normally distributed with expectation 0 and variance σ^2, and $\mathbf{Y}$ is the Nx1 random vector of responses, an instance $\mathbf{y}$ of which has been observed in the experiment. It will be assumed that the first column of $\mathbf{X}$ is a constant column of "+1" values, i.e. that the linear model contains an intercept. Coefficient numbering starts with "0", i.e. $\boldsymbol{\beta} = \left(\beta_0, \beta_1, ..., \beta_{p-1}\right)^{\mathrm{T}}$, where β_0 denotes the intercept, which is usually not of interest in the statistical analysis. The estimates for the unknown parameters are denoted as b_i, i=0,…,p–1. These are also called the coefficients of the model; some people prefer to report the effects instead, which are twice the coefficients. Here, the presentation is in terms of coefficients. The two approaches are equivalent.

The error df are $\text{df}_{\text{error}} = N - p$. If $N > p$, i.e. if $\text{df}_{\text{error}} > 0$, it is standard to use the *MSE* from the linear model (1) for estimating the error variance σ^2:

$$MSE = \frac{1}{N-p}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\,, \qquad (2)$$

with $\hat{y}_i = b_0 + b_1 x_{i1} + ... + b_{p-1} x_{i,p-1}$ the estimated response from the model. Anova can further decompose the sum of squares in (2) into a pure error and a lack-of-fit portion of df_{pe} and df_{lof} degrees of freedom, respectively: If N_{distinct} denotes the number of distinct runs (example: $N_{\text{distinct}} = 17$), $\text{df}_{\text{error}} = N - p = \text{df}_{\text{lof}} + \text{df}_{\text{pe}}$, with $\text{df}_{\text{lof}} = N_{\text{distinct}} - p$ and $\text{df}_{\text{pe}} = N - N_{\text{distinct}}$. If the residuals properly represent experimental error, the *MSE* is an unbiased estimate of the overall error variance σ^2. In most part of this paper, it is assumed that there is no lack of fit portion, since the model is saturated as much as possible. If all columns of **X** have equal length (a pre-requisite for use of the half-normal effects plot, see next section), the coefficient variance is a known constant multiple $K \bullet MSE$ of the *MSE*, i.e. the standard error can be estimated based on the *MSE*. The assessment of effect significance relies on the size of the absolute coefficient relative to that estimated standard error.

Table 2 shows the linear model analysis for the injection molding example. Control factors A, B and their interaction, noise factor c and the interaction column for AD (confounded with Cc and ab) is declared statistically significant at the 5% level, no other effect would become significant by increasing the level to 10%. Note that the effect "center[T.TRUE]" is scaled different from what Table 1 suggests, which does not matter for its significance, but for its estimated standard error, which is different from the others in Table 2, and would be the same as the others, when scaled like in Table 1.

TABLE 2 ABOUT HERE

For the simple linear model and its related Anova-based tests to yield valid results, several assumptions are crucial: The experimental runs need to be conducted independently, a normal distribution of the errors should not be completely off, omitted effects (in this case none) are truly inactive, and experimental replication reproduces the full range of the experimental variability – i.e. "proper replication" has been conducted.

Especially the proper replication assumption is sometimes critical in experiments – between repeating measurements (clearly not proper replication) and replicating experimental runs with all aspects that could cause variability (clearly proper replication), there are cases for which some potential sources of variability have been overlooked, intentionally kept constant (split-plot situation) or erroneously declared negligible. Early in the author's career as a statistician in industry, an experiment involved a case of the latter situation: the engineers involved were very concerned with a particular source of variability; they considered another source of variability as completely negligible in comparison. and therefore did not reset this latter one for each replication. It turned out that they were wrong, because

the allegedly negligible source of variability was very strong indeed under the experimental setting (which was far from reaching the usual manufacturing standard); consequently, the experiment had to be analyzed as a repeated measurement experiment and did not turn out any useful results; this example will be revisited in Section 3.2.2.

2.3 Half-normal effects plots and Lenth's method

Half-normal effects plots (Daniel[1,2]) have first been proposed for unreplicated (fractional) factorial 2-level designs as a method for assessing effect significance in the absence of direct estimates for experimental error: this is the case whenever $p=N$, i.e. the model is saturated and there are no pure error df. The idea further requires a design matrix $\mathbf{X}$ such that $\mathbf{X}^T\mathbf{X}$ is a diagonal matrix with constant diagonal elements, implying that all coefficient estimates are uncorrelated and have the same variance. For unreplicated regular fractional factorial 2-level designs, this can be achieved by -1/+1 coding of the variables, combined with product coding of interaction effects and – of course – inclusion of a rich enough model so that $\mathbf{X}$ is a square matrix. Considering the first 16 rows of the design in Table 1 as an unreplicated fractional factorial 2-level design, this means that the ABD interaction column needs to be included. A half-normal effects plot plots all absolute coefficients from the saturated model against an appropriate set of half normal quantiles. Under an assumption of effect sparsity, i.e. if only few of the true β_i are non-zero (called "active"), many b_i have expectation 0. Due to their common variance and uncorrelatedness, their absolute values are a sample from a half-normal distribution. If all effects are inactive, the points line up on a straight line rooted in the origin with the slope depending on their standard error. Daniel proposed to consider as active those effects that stick out from the line formed by the majority of inactive effects.

Figure 1 shows the half-normal effects plot produced from the first 16 rows of Table 1, i.e. completely ignoring the center points, from a model with main effects and two-factor interactions (i.e., without the ABD interaction). The labeled effects have been declared significant at the 5% level by Lenth's method (see below); they are also clearly identified by visual inspection of the half-normal plot as sticking out from the line of inactive effects. In this example, the half-normal effects plot identifies the same effects as the linear model for all 20 runs (see Table 2).

FIGURE 1 ABOUT HERE

Due to the subjectivity of assessing a half-normal effects plot and due to the difficulty in automating the assessment, it is customary to combine half-normal effects plotting with a numeric assessment of effect significance, and many proposals have been made for robust ways to obtain such an assessment in the absence of error df, see e.g. Haaland and O'Connell[14], Hamada and Balakrishnan[17] or Chen and Kunert[18] for comparative overviews.

One of the most widely used methods – consistently coming up with reasonable overall performance in comparison studies – is the method proposed by Lenth[3], which relies on the following three definitions:

$$s_0 = 1.5 \text{ median } (\{| b_i |: i=1,\ldots,p-1\}) \quad (3)$$

$$PSE = 1.5 \text{ median } (\{| b_i |: i=1,\ldots,p-1, | b_i | < 2.5 s_0\}) \quad (4)$$

$$ME = PSE \bullet t_{d,1-\alpha/2}, \text{ with } d=(p-1)/3 \quad (5)$$

s_0 is the initial estimate of the coefficient standard error, *PSE* is the pseudo-standard error as proposed by Lenth, and *ME* is the "margin of error". Lenth proposed to declare all effects that exceed *ME* as worthy of investigation. Lenth also proposed a bound SME for multiple testing adjustment. In agreement with Wu and Hamada[15], the SME is not considered here, as it is considered much more critical to miss an opportunity than to find too many of them in the screening task in the usual engineering application. Under an assumption of effect sparsity, i.e. if most β_i are in fact 0, the *PSE* is a reasonable estimate for coefficient standard errors.

The *t*-quantile in (5) has been repeatedly criticized for yielding an inappropriate approximation (too conservative), and various authors have proposed use of calibrated tables or simulation of p-values instead (e.g. Haaland and O'Connell[16], Loughin[19], Ye and Hamada[20], Schoen and Kaul[21], Edwards and Mee[8]). This paper uses simulated critical values from a million simulated experiments, as provided with the R package DoE.base[9]. The critical values relevant for this paper are listed in Table 3.

TABLE 3 ABOUT HERE

For the injection molding example ignoring the center points, we have s_0 = 0.103125, *PSE*=0.046875, with an *ME* = 0.10110104 (for α=5%, 15 coefficients, simulated value from Table 3 instead of $t_{d,1-\alpha/2}$). This compares to a larger estimated standard error of 0.05543 and a much larger margin of error of 0.1764 = 0.05543•$t_{3,0.975}$ from the Anova of Table 2 (which includes the center points and the non-linearity check contrast). If we include the appropriately scaled non-linearity check contrast into Lenth's method (see Figure 2(a) below), we obtain a changed s_0 = 0.08719983, but an unchanged *PSE*, and an only slightly reduced *ME* = 0.10023099 (for α=5%, 16 coefficients).

2.4 Modifications by Larntz and Whitcomb[7] and Edwards and Mee[8]

PSE relies on effect sparsity, while an *MSE*-based standard error estimate relies on proper replication. Larntz and Whitcomb[7] proposed to replace (4) with

$$CPSE = \sqrt{\frac{d \cdot PSE^2 + df_{\text{pe}} \cdot K \cdot MSE}{d + df_{\text{pe}}}}, \quad (6)$$

where the common variance of all coefficient estimators is $K \bullet \sigma^2$ (see Section 2.2). *CPSE* is thus based on a weighted average of PSE^2 and the variance estimate from a linear model, with weights based on

the df contributions (remember that $d=(p-1)/3$ was Lenth's proposal for the df to use in the t-quantile). Larntz and Whitcomb continued by incorporating this proposal into the margin of error as

$$CME = CPSE \bullet t_{d+\mathrm{df}_{\mathrm{pe}},1-\alpha/2}. \quad (7)$$

Edwards and Mee[8] seconded this proposal and additionally suggested to incorporate the *MSE* into (3) of Lenth's method by replacing s_0 with

$$\tilde{s}_0 = \sqrt{\frac{d \cdot s_0^2 + m \cdot df_{\mathrm{pe}} \cdot K \cdot MSE}{d + m \cdot df_{\mathrm{pe}}}} \text{ for a chosen constant } m. \quad (8)$$

Based on a simulation study, they recommended to choose $m = 5$, aiming at a good power under not so sparse models, whereas they preferred $m = 1$ for sparse models. We call the sequence (3), (4), (6) and (7) LM98, the sequence (3), (8), (4), (6) and (7) EM08; EM08_1 and EM08_5 denote the variants with m=1 and m=5, respectively. While Larntz and Whitcomb appear to have used the t-quantile in (7), Edwards and Mee suggested to replace it with a simulated critical value. Here, simulated critical values have been used for the two combined Lenth methods (implemented with R package **DoE.base**[9]).

For the injection molding example, both LW98 or EM08 *(m*=1 and *m*=5) yield the significant effects from the linear model at the 5% level and would add the Aa interaction (confounded with BC and Db) at the 10% level, but not the AC interaction that would also be added by the original Lenth's method; the LW98 *CPSE* (0.05012484) is slightly larger than that of the original Lenth's method, the EM08 *CPSE* coincides with it for both *m*=1 and *m*=5 (in spite of decreasing $\tilde{s}_0$).

3 Incorporation of error contrasts into half-normal effects plots and Lenth's method

The proposal presented in this section attempts another compromise between Lenth's method and Anova, i.e. between relying solely on either *PSE* or *MSE*. Before going into any technical detail, the general idea is explained: the N x p model matrix **X** from the linear model (1) is augmented with an orthogonal complement matrix **R**, so that the augmented model matrix (**X R**) is a matrix of full column rank. Often (**X R**) will be a square matrix of rank N; in some situations, it may make sense to omit some lack-of-fit related columns from the analysis (e.g. for external reference points; not detailed here). Section 3.1 provides a general algorithm for obtaining **R**, Section 3.2 details the matrix **R** for two practically relevant scenarii. Based on the augmented model matrix (**X R**), the proposed method is very simple:

a) calculate all coefficients $b_0,b_1,\ldots,b_k$ (with b_0 the overall mean, k+1 the number of columns in (**X R**)),

b) plot $b_1,\ldots,b_k$ on the half-normal effects plot, indicating the type of plot point by the plot symbol (i.e. the first p–1 points are marked as experimental effects, and the remaining points as error effects, with a distinction between lack-of-fit and pure error).

If $df_{pe} > 0$, a null reference line with slope 1/standard error (as estimated from pure error effects) can also be included.

c) apply Lenth's method to $b_1,\ldots,b_k$ for a numeric assessment of effect significance.

Instead of c), using a different method for significance assessment may be appropriate, depending on the situation.

3.1 Orthogonalization of the residual space

The idea presented below is straightforward and was rediscovered by the author, before also finding most of it implemented in JMP[4] as well as suggested in Langsrud[22] or Mee[10]. The residual space can be partitioned into df_{pe} pure error contrasts and df_{lof} lack-of-fit contrasts by two steps that can be formalized, based on model (1): the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ orthogonally projects the response vector onto the p-dimensional column space $\mathcal{R}(\mathbf{X})$ of matrix $\mathbf{X}$, whereas the matrix $\mathbf{M} = \mathbf{I}_N - \mathbf{H}$ is the orthogonal projector onto the $N–p$-dimensional orthogonal complement $\mathcal{R}(\mathbf{X})^{\perp}$, i.e. it spans the $N–p$ dimensional residual space, which may consist of lack-of-fit and/or pure error portions, whenever $\mathbf{X}$ has $p < N$ columns. Applying $\mathbf{M}$ is also called "projecting $\mathbf{X}$ out of" the recipient of the application. Now, denote by $\mathbf{S}$ a model matrix for a saturated model with $N_{distinct}$ orthogonal columns, $p \le N_{distinct} \le N$; note that $\mathbf{S}$ can always be chosen as a matrix of dummy variables for all distinct rows of $\mathbf{X}$, even though that may not be the best choice. Then, $\mathbf{M_S} = \mathbf{I}_N - \mathbf{H_S}$ with $\mathbf{H_S} = \mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T$ is an orthogonal projector onto the df_{pe} dimensional pure error contrast space, i.e. $\mathbf{M_S}$ has rank df_{pe}. $\mathcal{R}(\mathbf{M_S}) = \mathcal{R}(\mathbf{S})^{\perp}$ is a subspace of $\mathcal{R}(\mathbf{X})^{\perp}$, while $\mathcal{R}(\mathbf{X})$ is a subspace of $\mathcal{R}(\mathbf{S})$. With adequately chosen $\mathbf{S}$, an augmentation matrix $\mathbf{R}$ with $(\mathbf{X}\ \mathbf{R})^T(\mathbf{X}\ \mathbf{R}) = \text{const}\bullet\mathbf{I}_N$ is obtained as follows:

Augmentation algorithm

a) For the pure error df, add an appropriately scaled principal component of $\mathbf{M_S}$ for each non-zero eigen value; if $N_{distinct}=N$, i.e. $df_{pe}=0$, all eigen values are zero, so that no pure error columns are added to $\mathbf{R}$.

b) For the lack-of-fit df, add an appropriately scaled principal component of $\mathbf{MH_S}$ for each non-zero eigen value. Again, if $df_{lof}=0$, all eigen values are zero, so that no lack of fit columns are added to $\mathbf{R}$.

   For situations in which some lack-of-fit contrasts are not of interest – e.g. an external replication point for which a block factor is needed but is not of interest for the experiment – $\mathbf{S}$ should be chosen such that these are captured by dedicated columns that do not change under application of $\mathbf{M}$; then, these columns can simply be omitted from $\mathbf{S}$ for the calculation of $\mathbf{H_S}$ *in this step*.

For the pure error columns, the matrix $\mathbf{M_S}$ will usually have several eigen values of the same size, which makes the pure error portion of the matrix $\mathbf{R}$ non-unique (unique up to rotation). Non-uniqueness in the lack-of-fit portion is also possible in case of more than one df for lack of fit; if certain lack-of-fit aspects are of explicit interest, the lack-of-fit columns of $\mathbf{R}$ should be created

manually (like in Table 1) rather than by applying the above algorithm. Section 3.2 will spell out some detail for two specific and practically relevant situations: (i) a single center point with n_{ce} replications augments a saturated fractional factorial 2-level design in n_{cu} runs (ii) a fully replicated design for which there are doubts that the replications captured all relevant experimental variability.

3.2 Specific scenarii

3.2.1 A replicated center point

Consider a fractional factorial 2-level design with n_{ce} center points in addition to n_{cu} cube runs, like in Table 1. Let $\mathbf{X}_{cu}$ denote the non-intercept columns of the model matrix for a saturated model of the cube data (e.g. Table 1, rows 1 to 16, non-intercept columns augmented with the ABD interaction column). Then the matrix $\mathbf{X}$ is given by the portion to the left of the dashed line in Equation (9). As was discussed in Section 2.1, the n_{ce} center points add n_{ce} df, one of which is a lack-of-fit df, which can be captured by an appropriately scaled nonlinearity check contrast (e.g. last column of Table 1). The matrix $\mathbf{R}$ thus consists of one nonlinearity and $n_{ce}-1$ pure error contrast columns. The nonlinearity check contrast is uniquely determined up to direction, the $n_{ce}-1$ df pure error contrasts are uniquely determined up to rotation of the $(n_{ce}-1)$-dimensional sub space; a suitable set of such contrasts can be obtained from an appropriately rescaled set of orthonormal contrasts for a factor in n_{ce} levels, e.g. orthogonal polynomials or Helmert contrasts. Equation (9) shows the resulting augmented matrix:

$$\left(\mathbf{X} \,\vdots\, \mathbf{R}\right) = \left(\begin{array}{cc:cc} \mathbf{1}_{n_{cu}} & \mathbf{X}_{cu} & -a\mathbf{1}_{n_{cu}} & \mathbf{0}_{n_{cu}\times(n_{ce}-1)} \\ \mathbf{1}_{n_{ce}} & \mathbf{0}_{n_{ce}\times(n_{cu}-1)} & \dfrac{an_{cu}}{n_{ce}}\mathbf{1}_{n_{ce}} & \sqrt{n_{cu}}\mathbf{C} \\ \text{Intercept} & \text{experiment} & \text{nonlinearity} & \text{pure error} \end{array}\right), \tag{9}$$

with $a = \sqrt{n_{ce}/\left(n_{cu}+n_{ce}\right)}$ and $\mathbf{C}$ an arbitrary n_{ce}x$(n_{ce}-1)$ orthonormal contrast matrix. Note that Equation (9) preserves the usual effect scaling by normalizing all contrasts to squared length n_{cu} (except for the first column, which is not relevant because the intercept is usually excluded from plotting and testing); another natural choice would be the use of squared length $N = n_{cu}+n_{ce}$, which would sacrifice the usual effect scaling to a common length for all columns including the first. Furthermore, note that Equation (9) is valid for a saturated cube model only. If the matrix $\mathbf{X}$ has fewer than n_{cu} columns, there is more than one lack-of-fit df, and automatic orthogonalization of the lack-of-fit space may mix the nonlinearity check column with other lack-of-fit dimensions. For example, when using the columns of Table 1 without augmenting them with the ABD interaction, the lack-of-fit columns may arise from any rotation of the two-dimensional lack-of-fit space spanned by the ABD and the nonlinearity column.

This method is now applied to the injection molding example used before. Figure 2 (a) shows a plot with the additional nonlinearity contrast but without the pure error points (these are included via the null reference line only), while parts (b) and (c) of Figure 2 additionally include the pure error points in two different codings. In all plots, significance assessment is based on applying Lenth's method to

the plotted points ((a): $s_0 = 0.08719983$, $PSE = 0.046875$; (b): $s_0 = 0.07127467$, $PSE = 0.0375$; (c): $s_0 = 0.07127467$, $PSE = 0.049954$; the *ME* values are obtained by multiplying *PSE* with the appropriate entry from Table 3 (16 or 31 coefficients, respectively).

FIGURE 2 ABOUT HERE

Figure 2 illustrates that half-normal effects plots and test results from added pure error points are indeed coding dependent: the pure error points are in different positions in Figures 2 (b) and (c), and the numeric part of the method renders more effects significant when applied to the effects as shown in Figure 2 (b) than to those as shown in Figure 2 (c). Note that this is an example, for which Anova is the least liberal method, whereas the original Lenth's method is most liberal (seen best at the 10% significance level, for which Lenth's method adds two effects, Anova adds none, and the hybrid methods LW98 and EM08 add one each). The proposed method, when applied at the 10% level, would add the D main effect for the coding of Figure 2 (b), but would not add any further effect for the coding of Figure 2 (c). Thus, it is helpful to consider not only numeric methods but also the (augmented) half-normal effects plot: In this example, in both plots, the picture is quite clear in terms of which effects lie on a line of inactive effects. Thus, the combination of the numeric method with the plot adds the right degree of care to the perhaps somewhat liberal significance result obtained here.

3.2.2 A replicated design

Suppose each of n experimental runs has been conducted r times, yielding a total of $N = nr$ runs. If the r runs are proper replicates, it would usually be considered best to analyze the data by Anova. However, there might be some doubts whether the proper replication assumption holds. In that case, considering a half-normal effects plot analysis with pure error points included is a viable supplement to the linear model analysis; the visual information from the plot helps to better understand the strength of the evidence for the active effects, and an analysis with Lenth's method as proposed in this article pools the pure error information with the information from inactive effects. Denoting as $\mathbf{X}_n$ the $\mathbf{X}$ matrix of an individual replicate (arranged in the same order for all replicates, although this should not be the way for conducting the experiment), and assuming that $\mathbf{X}_n$ is a square full rank matrix saturated with effects including the intercept, one valid way of augmenting the overall experimental matrix $\mathbf{X}_N$ is

$$\left(\mathbf{X}_N \;\vdots\; \mathbf{R}\right) = \left(\mathbf{1}_r \otimes \mathbf{X}_n \;\vdots\; \sqrt{N}\mathbf{C} \otimes \mathbf{I}_n\right), \tag{10}$$

with $\otimes$ the Kronecker product, $\mathbf{C}$ any r x (r–1) matrix of orthonormal contrasts for an r-level factor. For example, if r=2, $\mathbf{C}$ would be the 2 x 1 vector $\left(-\sqrt{1/2}, +\sqrt{1/2}\right)^{\mathrm{T}}$. It is straightforward to see that the n pure error contrasts are then the differences between the two replicates for the same run, divided by

$n/2$ for normalization; this seems quite intuitive. Again, Formula (10) shows just one of infinitely many possible rotations of the residual space.

Ding et al.[13] reported an experiment on five drugs for treating herpes simplex infections: the experiment was run on a batch of cells; each drug constituted an experimental factor (levels none or minimum effective dose), and each run was replicated twice (r=2, replication described in detail in the original paper). The original design was a combination of a resolution V 16 run design with an 18 run orthogonal array, which is also analysed in Xu et al.[14]; here, we only look at the 16 run portion. For each run, a well of infected cells was treated with a particular level combination of the five drugs, and the response is the percentage of infected cells after treatment (as evidenced by a color marker). Following Ding et al. and Xu et al., the response is analysed on the square root scale. For these data, an Anova analysis would very clearly be the standard choice; however, its correctness of course depends on the proper replication assumption. Anova declares effects D, E, DE, B, A, and BC significant at the 5% level, effects AC and AD would be added at the 10% level.

FIGURE 3 ABOUT HERE

Assuming there might be doubt regarding the proper replication assumption, all the Lenth-based tests were also applied, including the original version without any consideration of pure error and the proposed method: The classical Lenth method shows D, E and DE only at the 5% level and 10% level (Figure 3 (a)). All other Lenth-based methods discussed in Section 3 show D, E and DE at the 5% level and add B, A and BC at the 10% level, i.e. their 10% significances coincide with Anova 5% significances. Figures 3 (b) and (c) show half-normal effects plots with pure error effects in two different codings and significance assessed by the proposed application of Lenth's method to the augmented set of effects; in this example, the different pure error codings are visible (different positions of the star points), but do not interfere with active effects. The visual impression from the half-normal effects plots in Figure 3, even when including the pure error points, suggests that the effects A, B, and BC are very close to the null reference line and are on a line of inactive – or weak – effects. Most pure error points are clustered at the bottom left end of the half-normal effects plot (particularly for plot (c)), i.e. do not mix well with inactive – or weak – experimental effects. This can be indicative of mainly two things: proper replication might be violated so that replication variance is smaller than experimental variance, and/or most experimental effects are active, some a lot smaller than others, i.e. effect sparsity is violated. In this particular case, the latter seems more likely than the former, because all five drugs are known to have at least some effect. While drugs D and E are obviously most powerful, it is likely that the other drugs are also effective, and it is not unlikely that many interactions are also at least slightly active. Furthermore, the variability between well plates (each run was randomized to a well on a well plate) and wells could be large enough to be in the range of the smaller experimental effects.

Figure 4 shows two half-normal effects plots for the example data from the author's industrial practice mentioned in the introduction, for which a severe violation of proper replication was identified after looking at a Minitab effects normal plot and discussing with the experimenters. While things were very clear-cut in this example – a strong source of variability was not changed between replicates because it was erroneously believed to be negligible beforehand – in general a picture similar to the one in Figure 4 is also conceivable if almost all effects are strong relative to experimental error.

FIGURE 4 ABOUT HERE

The ideal half-normal effects plot for a replicated design – some active effects, some inactive effects, proper replication – shows pure error and experimental effects mixed reasonably well. Any deviation from such a pattern makes one think about violation of proper replication and/or effect sparsity. It would make sense to explore in more detail the behavior of half-normal effects plots for replicated designs under different simulated scenarios – this would be helpful for using these plots as diagnostic tools. Generally, a plot alone does not help, but needs to be combined with knowledge about the experiment and its background.

4 Properties of the proposed method

A simulation study investigated all numeric methods regarding

- the rejection probabilities for inactive effects, if the assumptions effect sparsity and proper replication are fulfilled or violated
- and their power for detecting active effects.

The simulation study has been conducted on scenarii analogous to those of Section 3.2, varying effect sizes, number of active effects (i.e. degree of effect sparsity), and degree of violation of proper replication. For all Lenth-related methods, simulated critical values have been used, for Lenth's method and the proposed method from a million simulation runs, for the combined PSE-MSE methods from 10 000 simulation runs (see the supplementary material for more detail on the simulated scenarios and simulation results).

Note that any numeric method should always be consulted in conjunction with an augmented half-normal effects plot only, which has of course not been part of the simulation process. The combination of the numeric part of the proposed method or another of the numeric combination methods with the augmented half-normal plot offers the chance to separate active from inactive effects in more situations and with greater conviction than possible with Lenth's method applied to experimental effects only, numeric modifications of Lenth's method without plotting pure error points, or Anova. The in part anti-conservative behavior of the methods in case proper replication is violated can be cautioned against by the augmented half-normal plot, as was for example observed for the Ding et al. example.

4.1 Type I error

First of all, all methods have to hold the individual effect type I error rate

- under a null model (no active effects) with proper replication satisfied,
- for inactive effects (in the presence of active effects) with proper replication satisfied.

Under the null model, all methods keep the type I error well, and all methods except for Anova were conservative for the inactive effects in the presence of active effects, particularly if these were weak or many.

Second, the methods should not be too sensitive to moderate violation of proper replication. The simulations show that all methods except Lenth's original method get anticonservative in models without any active effects, with a large difference between Anova and the other methods for a situation with four center point runs and a much worse robustness performance for all combination methods for fully replicated designs. Among the combination methods, the proposed method is most robust for weak to moderate violation of proper replication for the center point situation. It is therefore a good choice in terms of robustness against violations of proper replication, as it is most important to safeguard against moderate cases: serious violation of proper replication is usually more easy to spot because unexpected similarity of the replicates and resulting unexpected significance results arouse the analyst's suspicion, especially for fully-replicated designs. In case of serious violation of proper replication, the original Lenth's method is the most adequate approach.

4.2 Power

The methods should be powerful for as many situations as possible, but particularly for the target situation of a few error df with moderate violation of proper replication and/or effect sparsity:

- Clearly, the power for 9 active effects in a 16 run experiment with four center point runs is strong for Anova, closely followed by the proposed method if the effects are strong enough. Lenth's method and the other combination methods break down, because effect sparsity is too severely violated. EM08_5 is the only combination method besides the proposed method that at least achieves a reasonable power for large values of α. Thus, the proposed method has succeeded in moving out the boundary for effect sparsity.
- For four center point runs, the proposed method is competitive: while it has lower power than the other combination methods for small effect sizes, it outperforms them with increasing number of effects and increasing effect sizes; this is partly due to decreasing power of the other combination methods, if effect sizes increase too much. There is no clear ranking among the combination methods; preference depends on the situation.
- For a fully replicated experiment with proper replication in place, Anova is of course the most appropriate analysis method, and the other methods except for Lenth's method approach its power for increasing effect sizes; Lenth's original method is of course weakest, because it ignores the replicate information; the proposed method is also rather weak, which may be the

price to pay for its relatively good robustness against violation of proper replication. For the replication situation, the most powerful combined method clearly is EM08_5, and both EM08 methods are more powerful than LW98.

The proposed method has succeeded in enhancing performance for not so sparse situations for a few error df; for small effect sizes, its performance in terms of power is not too impressive, especially in the replicated scenario. Nevertheless, in the light of its better robustness against violation of proper replication, it remains a reasonable method, especially for only few error df.

4.3 Choice of rotation

There is one issue that is unique to the proposed method and also affects the augmented half-normal effects plot: as was mentioned before, the plot points are not unique in case of multiple eigen vectors with the same eigen value (see Section 3.1), which will often happen for pure error points (see e.g. Sections 3.2.1 and 3.2.2) and can also happen for lack-of-fit error points (also possible for the example of Section 3.2.1, but not shown there; for lack-of-fit, it is a good idea to manually determine the chosen contrasts wherever possible). This phenomenon was investigated in the simulation study by exploiting the fact that the rotations are different in the standard implementations of the R software[11] on MacOS and Windows systems. This has been exploited to get an idea about the order of magnitude of the dependence on the rotation: The percentage of simulation runs with equal decisions on all experimental effects is between 52.4% (minimum over all scenarios) and 94.4% (maximum over all scenarios); the percentage of equal experimental effect decisions is between 93.8% and 99.5% (again minimum and maximum over all scenarios). These percentages clearly indicate that the actual numeric significance decision is not arbitrary but nevertheless subject to an influence outside of the experimental outcomes alone. While this non-uniqueness is not a desirable behavior, it is tolerable, as long as one is prepared to only consider significance results together with half-normal effects plots, to tie significant results for lack-of-fit or pure error contrasts to the actual contrast they refer to, and to be pragmatic rather than dogmatic regarding the assessment of effect significance.

Note that Larntz and Whitcomb[7] also recommended visualization of the residual variation with plotted points, however by a quite different method which does not use the individual pure error data points directly but uses the *MSE* only and iteratively calculates points that lie on the null line (slope as expected for pure-error based effect standard errors) in a half-normal effects plot for the data; the author considers it more appropriate to visualize the null line by an actual line, and to provide pure error points that have a direct connection to the individual data points from the experiment, in the way shown above, and presumably also used in SAS JMP®.

5 Discussion

This article has proposed a method to incorporate pure error information into half-normal effects plots as well as into automatic numeric assessment of effect significance. The proposed method appears to be close to what is implemented in the screening platform of JMP® statistical software[4]. Similar to

methods by Larntz and Whitcomb[7] and Edwards and Mee[8], the proposed assessment of effect significance offers a compromise between Lenth's method's dependence on effect sparsity and Anova's dependence on proper replication for the estimation of error variation.

The augmented half-normal effects plot is the most important tool provided here. It can serve as a diagnostic tool for assessing the proper replication assumption, while the proposed application of Lenth's method to the augmented set of effects is its natural companion for numeric assessment of effect significance. Usage of other numeric methods like EM08 or LW98 can be preferable at least when effect sparsity holds or effects are relatively weak (see also below). The half-normal effects plot augmented by error points is important for assessing the validity of the results, since methods that rely on error estimates from a small number of pure error contrasts for assessing effect significance may be liberal under violations of the proper replication assumption. The liberal behavior in conjunction with the plot is acceptable, as for factor screening (a) it is more important to detect active effects than to avoid detection of inactive effects and (b) it is perfectly acceptable to play with significance levels until the picture makes sense – this pragmatic approach can of course not be built into a simulation study.

The proposed numeric method is a very natural extension of Lenth's method; experimenters do not need to learn a new technique, but just have to accept the addition of error points to the half-normal effects plot for grasping this method. Furthermore, it is the most powerful method apart from Anova, if effect sparsity is an issue, and, contrary to Anova, it allows to assess proper replication in connection with the augmented half-normal effects plot. On the down side, the numeric assessment depends on the arbitrary choice of rotation; this is not a strong drawback, since dependence on the rotation usually causes relatively few changes in significance assessment only, and since significance testing in the screening situation should generally not be dogmatic about a particular significance level. A further drawback is a weaker power than that of EM08_5 for situations with weak effects. This may be a reason for using EM08_5 instead of the proposed numeric method together with the augmented half-normal effects plot. Note that the appearance of the half-normal effects plot also depends on the arbitrary choice of rotation, i.e. the placement of pure error and/or lack-of-fit points must be considered with some reservation.

Generally, if the variability of center point runs or other replicated runs is artificially low, application of any significance assessment method – except for the original Lenth's method – may render random experimental effects active with unacceptably high probability. One should be skeptical, if the error points are clustered together at the bottom left of the augmented half-normal effects plot, like in Figure 3 (b) or (c) or – even more drastically – in Figure 4 (b); however, the assessment whether the reason for such behavior is violation of proper replication or rather a large number of active effects requires inclusion of knowledge about the experimental situation. With only few pure error points, like in Figure 2, very small absolute pure error coefficients also create suspicion, but are much less conclusive of violations of any assumption. Likewise, moderate violations of proper replication will

not always be evident from a plot by small pure error coefficients. It would be worthwhile to collect experience on this matter, both from the real world and from simulations.

In spite of all issues that have been mentioned in relation to LW98, EM08 or the method proposed here, using alternatives to Anova is clearly recommended in case of only a few error df. At the very least, significance testing from Anova should be accompanied by an augmented half-normal effects plot; often a different approach to significance assessment may also be more reasonable, because variance estimates from only few error df are highly variable. It is time that the text books highlight the issue of making appropriate use of a few replications. It is also useful to make the augmented half-normal effects plot available as a tool for routine use in assessing the plausibility of the proper replication assumption even for fully replicated designs.

Acknowledgements

Hongquan Xu drew the author's attention to the data for Section 3.2.2 and helped with information regarding the experimental procedures and background. Discussions with Pat Whitcomb clarified the rationale for the Larntz and Whitcomb plotting method.

Part of this work was supported by Deutsche Forschungsgemeinschaft Grant GR 3843/1-1.

References

1. Daniel, C. Use of Half-normal effects plots in Interpreting Two Level Experiments. *Technometrics* 1959; **1**:311–340.
2. Daniel, C. *Application of Statistics to Industrial Experimentation*. Wiley, New York, 1976.
3. Lenth, R.V. Quick and easy analysis of unreplicated factorials. *Technometrics* 1989; **31**:469–473.
4. SAS Institute CRT *JMP 10 Modeling and Multivariate Methods*. SAS Institute, Cary, NC, 2012. Specifically: URL: http://www.jmp.com/support/help/Technical_Details_3.shtml (Accessed March 26 2015).
5. Minitab Inc. *Meet Minitab 16*. Minitab Inc., State College, PA, 2010.
6. Stat-Ease Inc. Design-Expert, Version 8. Software, 2012.
7. Larntz, K. and Whitcomb, P. Use of replication in almost unreplicated factorials. Manuscript of a presentation given at the 42nd ASQ Fall Technical conference in Corning, New York, 1998. URL: http://www.statease.com/pubs/use-of-rep.pdf [26 April 2013].
8. Edwards, D. J. and Mee, R. W. Empirically Determined p-Values for Lenth t-Statistics. *Journal of Quality Technology* 2008; **40**:368–380.
9. Grömping, U. DoE.base: Full factorials, orthogonal arrays and base utilities for DoE packages. R package version 0.26-3, 2014; In R Core Team[11].
10. Mee, R. *A Comprehensive Guide to Factorial Two-Level Experimentation*. Springer, New York, 2009 (Chapter 14).

11. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2014.
12. Montgomery, D.C. Using fractional factorial designs for robust process development. *Quality Engineering* 1990; **3**:193-205.
13. Ding, X., Xu, H., Hopper, C., Yang, J. and Ho, C.M. Use of Fractional Factorial Designs in Antiviral Drug Studies. *Quality and Reliability Engineering International* 2013; **29**:299-304.
14. Xu, H., Jaynes, J., and Ding, X. Combining Two-Level and Three-Level Orthogonal Arrays for Factor Screening and Response Surface Exploration. *Statistica Sinica* 2014; **24**:269-289.
15. Wu, C.F.J. and Hamada, M. *Experiments: Planning, Analysis, and Parameter Design Optimization*. Wiley, New York, 2009.
16. Haaland, P.D. and O'Connell, M.A. Inference for Effect-Saturated Fractional Factorials. *Technometrics* 1995; **37**:82-93.
17. Hamada, M. and Balakrishnan, N. Analyzing unreplicated factorial experiments: a review with some new proposals. *Statistica Sinica* 1998; **8**:1-41.
18. Chen, Y. and Kunert, J. A New Quantitative Method for Analysing Unreplicated Factorial Designs. *Biometrical Journal* 2004; **46**:125-140.
19. Loughin, T.M. Calibration of the Lenth test for unreplicated factorial designs. *Journal of Quality Technology* 1998; **30**:171-175.
20. Ye, K., and Hamada, M.S. Critical Values of the Lenth Method for Unreplicated Factorial Designs. *Journal of Quality Technology* 2000; **32**:57-66.
21. Schoen, E.D. and Kaul, E.A.E. Three robust scale estimators to judge unreplicated experiments. *Journal of Quality Technology* 2000; **32**:276-283.
22. Langsrud, Ø. Identifying Significant Effects in Fractional Factorial Multiresponse Experiments. *Technometrics* 2001; **43**:415-424.

Table 1: The model matrix for the Montgomery[12] data

(y for ID 5, 9 and 14 corrected vs. Table 11.14 in Wu and Hamada[15])

		Model matrix															Response	Lack of fit columns	
			Factor settings							Two-factor interactions									
ID	Run order	I	A	B	C	D	a	b	c	AB Ca bc	AC Ba Dc	AD Cc ab	Aa BC Db	Ab Bc Da	Ac Bb CD	BD Cb ac	y	ABD	Non-linearity
1	8	1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	0,6	-1	$-1/\sqrt{5}$
2	16	1	1	-1	-1	-1	1	-1	1	-1	-1	-1	1	-1	1	1	1,0	1	$-1/\sqrt{5}$
3	18	1	-1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	-1	3,2	1	$-1/\sqrt{5}$
4	17	1	1	1	-1	-1	-1	1	1	1	-1	-1	-1	1	1	-1	6,0	-1	$-1/\sqrt{5}$
5	3	1	-1	-1	1	-1	1	1	1	1	-1	1	-1	-1	-1	1	0,4	-1	$-1/\sqrt{5}$
6	5	1	1	-1	1	-1	-1	1	-1	-1	1	-1	-1	1	-1	1	1,5	1	$-1/\sqrt{5}$
7	10	1	-1	1	1	-1	-1	-1	1	-1	-1	1	1	1	-1	-1	2,6	1	$-1/\sqrt{5}$
8	2	1	1	1	1	-1	1	-1	-1	1	1	-1	1	-1	-1	-1	6,0	-1	$-1/\sqrt{5}$
9	9	1	-1	-1	-1	1	-1	1	-1	1	1	-1	1	-1	-1	-1	0,8	1	$-1/\sqrt{5}$
10	15	1	1	-1	-1	1	1	1	1	-1	-1	1	1	1	-1	-1	1,2	-1	$-1/\sqrt{5}$
11	12	1	-1	1	-1	1	1	-1	-1	-1	1	-1	-1	1	-1	1	3,4	-1	$-1/\sqrt{5}$
12	6	1	1	1	-1	1	-1	-1	1	1	-1	1	-1	-1	-1	1	6,0	1	$-1/\sqrt{5}$
13	13	1	-1	-1	1	1	1	-1	1	1	-1	-1	-1	1	1	-1	1,6	1	$-1/\sqrt{5}$
14	19	1	1	-1	1	1	-1	-1	-1	-1	1	1	-1	-1	1	-1	0,5	-1	$-1/\sqrt{5}$
15	11	1	-1	1	1	1	-1	1	1	-1	-1	-1	1	-1	1	1	3,7	-1	$-1/\sqrt{5}$
16	1	1	1	1	1	1	1	1	-1	1	1	1	1	1	1	1	5,2	1	$-1/\sqrt{5}$
17	20	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2,5	0	$4/\sqrt{5}$
18	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2,9	0	$4/\sqrt{5}$
19	14	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2,4	0	$4/\sqrt{5}$
20	7	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2,7	0	$4/\sqrt{5}$

Table 2: Linear model analysis for the example data

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.73125    0.05543  49.270 1.84e-05 ***
A            0.69375    0.05543  12.515  0.00110 **
B            1.78125    0.05543  32.133 6.62e-05 ***
C           -0.04375    0.05543  -0.789  0.48760
D            0.06875    0.05543   1.240  0.30306
a            0.01875    0.05543   0.338  0.75747
b            0.01875    0.05543   0.338  0.75747
c           -0.24375    0.05543  -4.397  0.02180 *
centerTRUE  -0.10625    0.12395  -0.857  0.45437
A:B          0.59375    0.05543  10.711  0.00174 **
A:C         -0.08125    0.05543  -1.466  0.23898
A:D         -0.26875    0.05543  -4.848  0.01675 *
A:a         -0.09375    0.05543  -1.691  0.18938
A:b          0.03125    0.05543   0.564  0.61233
A:c         -0.00625    0.05543  -0.113  0.91735
B:D         -0.00625    0.05543  -0.113  0.91735
A:B:D        0.00625    0.05543   0.113  0.91735
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2217 on 3 degrees of freedom
Multiple R-squared:  0.9978,    Adjusted R-squared:  0.986
F-statistic:  84.7 on 16 and 3 DF,  p-value: 0.001831
```

Table 3: Simulated Lenth critical values (Source: R package DoE.base[9])

No. of coefficients α	Simulated critical values		t-values with d df	
	0.05	0.1	0.05	0.1
15	2.156822	1.701684	2.570582	2.015048
16	2.138261	1.694899	2.523002	1.987559
19	2.122981	1.695491	2.416031	1.925017
31	2.065203	1.680734	2.218435	1.806546

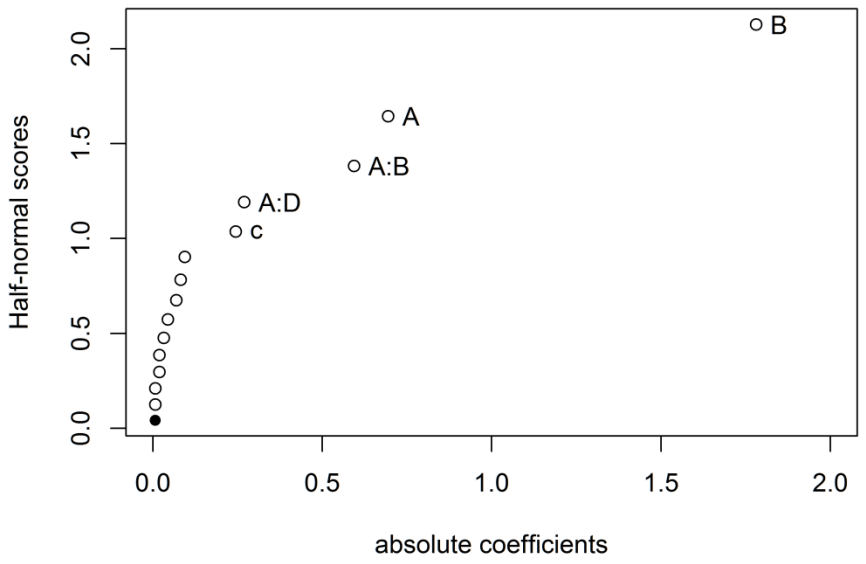

Figure 1: Half-normal effects plot for the injection molding example from Montgomery[12].
The filled dot is the 3-factor interaction that was added for saturating the model matrix.
Effects significant at the 5% level are labeled (Lenth's method, see below).
(Aa and AC would be added for α=10%.)

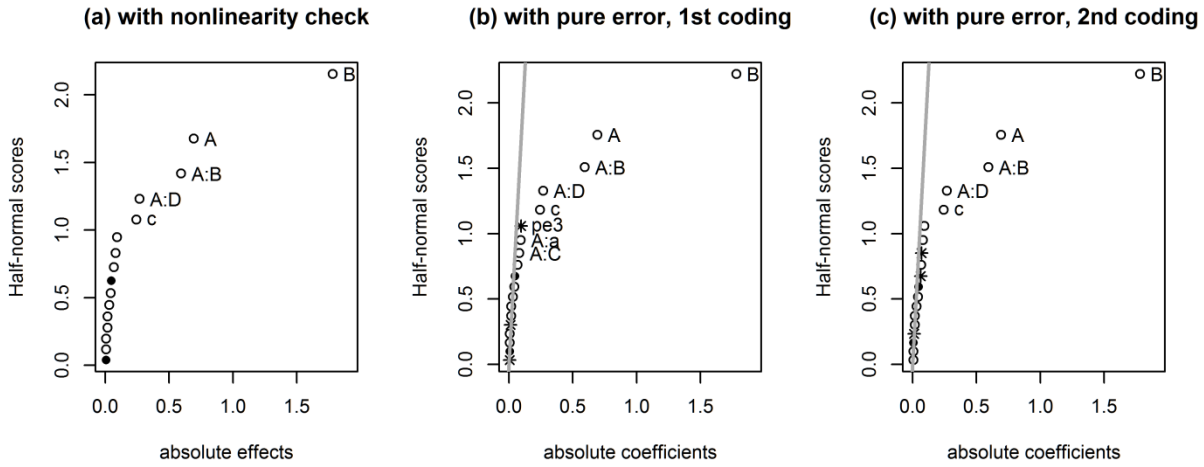

Figure 2: Half-normal effects plots for Montgomery[12] data, α=0.05:
(a) without pure error and (b) / (c) including pure error (*) in two different codings;
the two filled dots are the lack-of-fit points (ABD and nonlinearity contrasts)
The line indicates an approximate no effect line under the estimated pure error.
Note: the three smallest effects are tied, i.e. their order is the result of small rounding differences, which leads to the different placement of the ABD lack-of-fit point in the three plots.

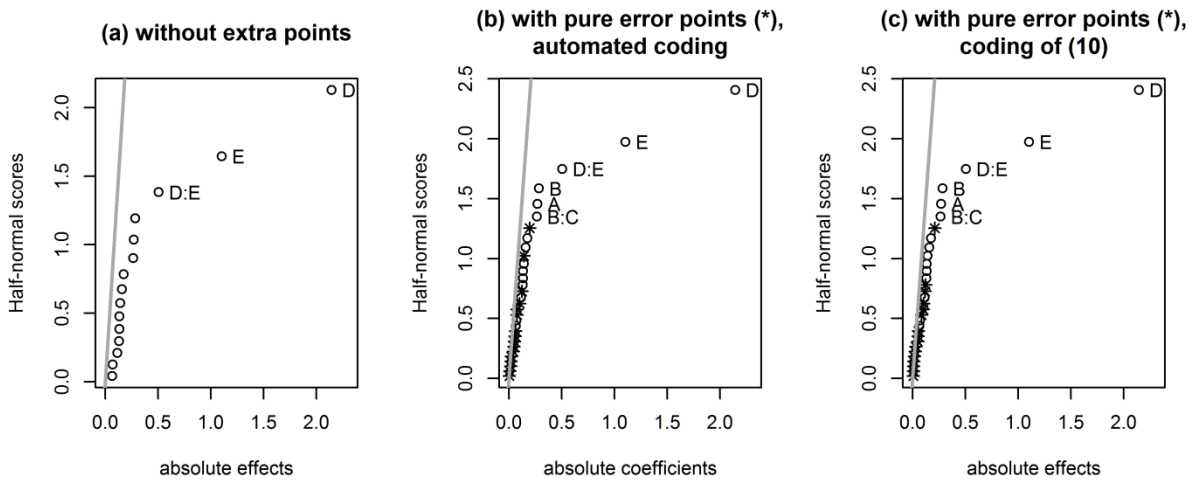

Figure 3: Half-normal effects plots for the Ding et al.[13] data, α=10%
(at α=5%, all plots show D, E, and D:E, plot (b) also B)
The line indicates a null reference line under the estimated pure error.

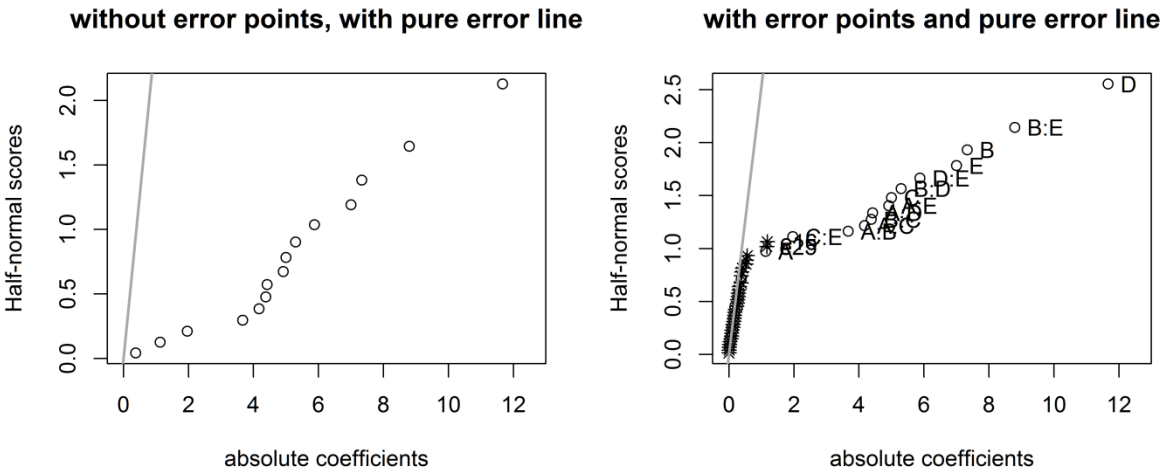

Figure 4: Half-normal effects plots for an experiment from the author's industrial practice with three replicates, but serious violation of proper replication, α=5% (Lenth)
The line indicates a null reference line under the estimated pure error.
(Anova would declare all but the smallest absolute coefficient significant)