Submitted to „The Annals of Statistics”

Generalized resolution for orthogonal arrays

By Ulrike Grömping[1] and Hongquan Xu[2]

Beuth University of Applied Sciences Berlin and University of California Los Angeles

The generalized word length pattern of an orthogonal array allows a ranking of orthogonal arrays in terms of the generalized minimum aberration criterion (Xu and Wu 2001). We provide a statistical interpretation for the number of shortest words of an orthogonal array in terms of sums of R^2 values (based on orthogonal coding) or sums of squared canonical correlations (based on arbitrary coding). Directly related to these results, we derive two versions of generalized resolution for qualitative factors, both of which are generalizations of the generalized resolution by Deng and Tang (1999) and Tang and Deng (1999). We provide a sufficient condition for one of these to attain its upper bound, and we provide explicit upper bounds for two classes of symmetric designs. Factor wise generalized resolution values provide useful additional detail.

1. Introduction

Orthogonal arrays (OAs) are widely used for designing experiments. One of the most important criteria for assessing the usefulness of an array is the generalized word length pattern (GWLP) as proposed by Xu and Wu (2001): A_3, A_4, … are the numbers of (generalized) words of lengths 3, 4, …, and the design has resolution R, if $A_i = 0$ for all $i < R$ and $A_R > 0$. Analogously to the well-known minimum aberration criterion for regular fractional factorial designs (Fries and Hunter 1980), the quality criterion based on the GWLP is generalized minimum aberration (GMA; Xu and Wu 2001): a design D_1 has better generalized aberration than a design D_2, if its resolution is higher or – if both designs have resolution R – if its number A_R of shortest words is smaller; in case of ties in A_R, frequencies of successively longer words are compared, until a difference is encountered.

The definition of the A_i in Xu and Wu is very technical (see Section 2). One of the key results of this paper is to provide a statistical meaning for the number of shortest words, A_R: we will show that A_R is the sum of R^2 values from linear models with main effects model matrix columns in orthogonal coding as dependent variables and full models in $R{-}1$ other factors on the explanatory side. For

[1] Supported in part by Deutsche Forschungsgemeinschaft Grant GR 3843/1-1.

[2] Supported by National Science Foundation Grant DMS-1106854.

AMS 2000 subject classifications: Primary 62K15, secondary 05B15, 62K05, 62J99, 62H20

Key words and phrases: orthogonal arrays, generalized word length pattern, generalized minimum aberration, generalized resolution, qualitative factors, complete confounding, canonical correlation, weak strength t

arbitrary factor coding, the "sum of R^2" interpretation cannot be upheld, but it can be shown that A_R is the sum of squared canonical correlations (Hotelling 1936) between a factor's main effects model matrix columns in arbitrary coding and the full model matrix from R–1 other factors. These results will be derived in Section 2.

For regular fractional factorial 2-level designs, the GWLP coincides with the well-known word length pattern (WLP). An important difference between regular and non-regular designs is that factorial effects in regular fractional factorial designs are either completely aliased or not aliased at all, while non-regular designs can have partial aliasing, which can lead to non-integer entries in the GWLP. In fact, the absence of complete aliasing has been considered an advantage of non-regular designs (e.g. those by Plackett and Burman 1946) for screening applications. Deng and Tang (1999) and Tang and Deng (1999) defined "generalized resolution" (*GR*) for non-regular designs with 2-level factors, in order to capture their advantage over complete confounding in a number. For example, the 12 run Plackett-Burman design has GR=3.67, which indicates that it is resolution III, but does not have any triples of factors with complete aliasing. Evangelaras et al. (2005) have made a useful proposal for generalizing *GR* (called *GRes* by them) for designs in quantitative factors at 3 levels; in conjunction with Cheng and Ye (2004), their proposal can easily be generalized to cover designs with quantitative factors in general. However, there is so far no convincing proposal for designs with qualitative factors. The second goal of this paper is to close this gap, i.e. to generalize Deng and Tang's / Tang and Deng's *GR* to OAs for qualitative factors. Any reasonable generalization of *GR* has to fulfill the following requirements: (i) it must be coding-invariant, i.e. must not depend on the coding chosen for the experimental factors (this is a key difference vs. designs for quantitative factors), (ii) it must be applicable for symmetric and asymmetric designs (i.e. designs with a fixed number of levels and designs with mixed numbers of levels), (iii) like in the 2-level case, $R+1 > GR \geq R$ must hold, and $GR = R$ must be equivalent to the presence of complete aliasing somewhere in the design, implying that $R+1 > GR > R$ indicates a resolution R design with no complete aliasing among projections of R factors. We offer two proposals that fulfill all these requirements and provide a rationale behind each of them, based on the relation of the GWLP to regression relations and canonical correlations among the columns of the model matrix.

The paper is organized as follows: Section 2 formally introduces the GWLP and provides a statistical meaning to its number of shortest words, as discussed above. Section 3 briefly introduces generalized resolution by Deng and Tang (1999) and Tang and Deng (1999) and generalizes it in two meaningful ways. Section 4 shows weak strength R (in a version modified from Xu 2003 to imply strength R–1) to be sufficient for maximizing one of the generalized resolutions in a resolution R design. Furthermore, it derives an explicit upper bound for the proposed generalized resolutions for two classes of symmetric designs. Section 5 derives factor wise versions of both types of generalized resolution and demonstrates that these provide useful additional detail to the overall values. The paper closes with a discussion and an outlook on future work.

Throughout the paper, we will use the following notation: An orthogonal array of resolution R = strength R–1 in N runs with n factors will be denoted as OA(N, $s_1 \ldots s_n$, R–1), with $s_1, \ldots, s_n$ the numbers of levels of the n factors (possibly but not necessarily distinct), or as OA(N, $s_1^{n_1} \ldots s_k^{n_k}$, R–1) with n_1 factors at s_1 levels, …, n_k factors at s_k levels ($s_1, \ldots, s_k$ possibly but not necessarily distinct), whichever is more suitable for the purpose at hand. A subset of k indices that identifies a k-factor projection is denoted by $\{u_1,\ldots,u_k\}$ ($\subseteq \{1,\ldots,n\}$). The unsquared letter R always refers to the resolution of a design, while R^2 denotes the coefficient of determination.

2. Projection frequencies and linear models

Consider an OA(N, s_1 … s_n, R–1). The resolution R implies that main effects can be confounded with interactions among R–1 factors, where the extent of confounding of degree R can be investigated on a global scale or in more detail: Following Xu and Wu (2001), the factors are coded in orthogonal contrasts with squared column length normalized to N. We will use the expression "normalized orthogonal coding" to refer to this coding; on the contrary, the expressions "orthogonal coding" or "orthogonal contrast coding" refer to main effects model matrix columns that have mean zero and are pairwise orthogonal, but need not be normalized. For later reference, note that for orthogonal coding (whether normalized or not) the main effects model matrix columns for an OA (of strength at least 2) are always uncorrelated.

We write the model matrix for the full model in normalized orthogonal coding as

$$\mathbf{M} = (\mathbf{M}_0, \mathbf{M}_1, \ldots, \mathbf{M}_n), \tag{1}$$

where $\mathbf{M}_0$ is a column of "+1"s, $\mathbf{M}_1$ contains all main effects model matrices, and $\mathbf{M}_k$ is the matrix of all $\binom{n}{k}$ k-factor interaction model matrices, k = 2,…,n. The portion $\mathbf{X}_{u_1\ldots u_k}$ of $\mathbf{M}_k = (\mathbf{X}_{1\ldots k},\ldots,\mathbf{X}_{n-k+1\ldots n})$ denotes the model matrix for the particular k-factor interaction indexed by $\{u_1,\ldots,u_k\}$ and is obtained by all products from one main effects contrast column each from the k factors in the interaction. Note that the normalized orthogonal coding of the main effects implies that all columns of $\mathbf{M}_k$ have squared length N for $k \le R$–1. Now, on the global scale, the overall number of words of length k can be obtained as the sum of squared column averages of $\mathbf{M}_k$, i.e., $A_k = \mathbf{1}_N^{\mathrm{T}}\, \mathbf{M}_k\, \mathbf{M}_k^{\mathrm{T}}\, \mathbf{1}_N / N^2$. Obviously, this sum can be split into contributions from individual k-factor projections for more detailed considerations, i.e.,

$$A_k = \sum_{\substack{\{u_1,\ldots,u_k\}\\ \subseteq\{1,\ldots,n\}}} \mathbf{1}_N^{\mathrm{T}} \mathbf{X}_{u_1\ldots u_k} \mathbf{X}_{u_1\ldots u_k}^{\mathrm{T}} \mathbf{1}_N \big/ N^2 =: \sum_{\substack{\{u_1,\ldots,u_k\}\\ \subseteq\{1,\ldots,n\}}} a_k\left(u_1,\ldots,u_k\right), \tag{2}$$

where $a_k(u_1,\ldots,u_k)$ is simply the A_k value of the k-factor projection $\{u_1,\ldots,u_k\}$. The summands $a_k(u_1,\ldots,u_k)$ are called "projection frequencies".

<u>Example 1</u>. For 3-level factors, normalized polynomial coding has the linear contrast coefficients $-\sqrt{3/2}$, 0, $\sqrt{3/2}$ and the quadratic contrast coefficients $\sqrt{1/2}$, $-\sqrt{2}$, $\sqrt{1/2}$. For the regular design OA(9, 3^3, 2) with the defining relation C=A+B (mod 3), the model matrix $\mathbf{M}$ has dimensions 9x27, including one column for $\mathbf{M}_0$, six for $\mathbf{M}_1$, twelve for $\mathbf{M}_2$ and eight for $\mathbf{M}_3$. Like always, the column sum of $\mathbf{M}_0$ is N (here: 9), and like for any orthogonal array, the column sums of $\mathbf{M}_1$ and $\mathbf{M}_2$ are 0, which implies A_0=1, A_1=A_2=0. We now take a closer look at $\mathbf{M}_3$, arranging factor A as (0 0 0 1 1 1 2 2 2), factor B as (0 1 2 0 1 2 0 1 2) and factor C as their sum (mod 3), denoting linear contrast columns by the subscript l and quadratic contrast columns by the subscript q. Then,

$$
\mathbf{M}_3 = \left(\begin{array}{l|cccccccc}
\text{contrast} & A_lB_lC_l & A_qB_lC_l & A_lB_qC_l & A_qB_qC_l & A_lB_lC_q & A_qB_lC_q & A_lB_qC_q & A_qB_qC_q \\
\hline
 & -\sqrt{\frac{27}{8}} & \sqrt{\frac{9}{8}} & \sqrt{\frac{9}{8}} & -\sqrt{\frac{3}{8}} & \sqrt{\frac{9}{8}} & -\sqrt{\frac{3}{8}} & -\sqrt{\frac{3}{8}} & \sqrt{\frac{1}{8}} \\
 & 0 & 0 & 0 & 0 & 0 & 0 & -\sqrt{6} & \sqrt{2} \\
 & -\sqrt{\frac{27}{8}} & \sqrt{\frac{9}{8}} & -\sqrt{\frac{9}{8}} & \sqrt{\frac{3}{8}} & -\sqrt{\frac{9}{8}} & \sqrt{\frac{3}{8}} & -\sqrt{\frac{3}{8}} & \sqrt{\frac{1}{8}} \\
 & 0 & 0 & 0 & 0 & 0 & -\sqrt{6} & 0 & \sqrt{2} \\
 & 0 & 0 & 0 & \sqrt{6} & 0 & 0 & 0 & \sqrt{2} \\
 & 0 & \sqrt{\frac{9}{2}} & 0 & \sqrt{\frac{3}{2}} & 0 & -\sqrt{\frac{3}{2}} & 0 & -\sqrt{\frac{1}{2}} \\
 & -\sqrt{\frac{27}{8}} & -\sqrt{\frac{9}{8}} & \sqrt{\frac{9}{8}} & \sqrt{\frac{3}{8}} & -\sqrt{\frac{9}{8}} & -\sqrt{\frac{3}{8}} & \sqrt{\frac{3}{8}} & \sqrt{\frac{1}{8}} \\
 & 0 & 0 & \sqrt{\frac{9}{2}} & \sqrt{\frac{3}{2}} & 0 & 0 & -\sqrt{\frac{3}{2}} & -\sqrt{\frac{1}{2}} \\
 & 0 & 0 & 0 & 0 & -\sqrt{\frac{9}{2}} & -\sqrt{\frac{3}{2}} & -\sqrt{\frac{3}{2}} & -\sqrt{\frac{1}{2}} \\
\hline
\text{column sum} & -\sqrt{\frac{243}{8}} & \sqrt{\frac{81}{8}} & \sqrt{\frac{81}{8}} & \sqrt{\frac{243}{8}} & -\sqrt{\frac{81}{8}} & -\sqrt{\frac{243}{8}} & -\sqrt{\frac{243}{8}} & \sqrt{\frac{81}{8}}
\end{array}\right)
$$

Half of the squared column sums of $\mathbf{M}_3$ are 243/8 and 81/8, respectively. This implies that the sum of the squared column sums is $A_3 = a_3(1,2,3) = (4*243/8+4*81/8)/81 = 2$.

Table 1. A partially confounded OA(18, $2^1 3^2$, 2) (transposed)

A	0 1 1 0 0 1 0 1 0 1 1 0 1 0 0 1 0 1
B	0 0 0 0 0 0 1 1 1 1 1 1 2 2 2 2 2 2
C	0 1 2 1 2 0 0 2 0 1 1 2 2 1 2 0 1 0

Example 2. Table 1 displays the only OA(18, $2^1 3^2$, 2) that cannot be obtained as a projection from the L18 design that was popularized by Taguchi (of course, this triple is not interesting as a stand-alone design, but as a projection from a design in more factors only; the Taguchi L18 is for convenience displayed in Table 4 below). The 3-level factors are coded like in Example 1, for the 2-level factor, normalized orthogonal coding is the customary -1/+1 coding. Now, the model matrix $\mathbf{M}$ has dimensions 18x18, including one column for $\mathbf{M}_0$, five for $\mathbf{M}_1$, eight for $\mathbf{M}_2$ and four for $\mathbf{M}_3$. Again, $A_0=1$, $A_1=A_2=0$. The squared column sums of $\mathbf{M}_3$ are 9 (1x), 27 (2x) and 81 (1x), respectively. Thus, $A_3 = a_3(1,2,3) = (9+2*27+81)/324 = 4/9$.

The projection frequencies $a_k(u_1,\ldots,u_k)$ from Equation (2) are the building blocks for the overall A_k. The $a_R(u_1,\ldots,u_R)$ will be instrumental in defining one version of generalized resolution. Theorem 1 provides them with an intuitive interpretation. The proof is given in Appendix A.

Theorem 1. In an OA(N, $s_1 \ldots s_n$, R–1), denote by $\mathbf{X}_c$ the model matrix for the main effects of a particular factor $c \in \{u_1,\ldots,u_R\} \subseteq \{1,\ldots,n\}$ in normalized orthogonal coding, and let $\mathrm{C} = \{u_1,\ldots,u_R\} \setminus \{c\}$. Then, $a_R(u_1,\ldots,u_R)$ is the sum of the R^2-values from the s_c–1 regression models that explain the columns of $\mathbf{X}_c$ by a full model in the factors from C.

Remark 1. (i) Theorem 1 holds regardless which factor is singled out for the left-hand side of the model. (ii) The proof simplifies by restriction to normalized orthogonal coding, but the result holds whenever the factor c is coded by any set of orthogonal contrasts, whether normalized or not. (iii) Individual R^2 values are coding dependent, but the sum is not. (iv) In case of normalized orthogonal coding for all factors, the full model in the factors from C can be reduced to the R–1 factor interaction only, since the matrix $\mathbf{X}_c$ is orthogonal to the model matrices for all lower degree effects in the other R–1 factors.

Example 1 continued. The overall $a_3(1,2,3) = 2$ is the sum of two R^2 values which are 1, regardless which factor is singled out as the main effects factor for the left-hand sides of regression. This reflects that the level of each factor is uniquely determined by the level combination of the other two factors.

Example 2 continued. The R^2 from regressing the single model matrix column of the 2-level factor on the four model matrix columns for the interaction among the two 3-level factors is 4/9. Alternatively, the R^2-values for the regression of the two main effects columns for factor B on the AC interaction columns are 1/9 and 3/9 respectively, which also yields the sum 4/9 obtained above for $a_3(1,2,3)$. For factor B in dummy coding with reference level 0 instead of normalized polynomical coding, the two main effects model matrix columns for factor B have correlation 0.5; the sum of the R^2 values from full models in A and C for explaining these two columns is $1/3 + 1/3 = 2/3 \neq a_3(1,2,3) = 4/9$. This demonstrates that Theorem 1 is not applicable if orthogonal coding (see Remark 1 (ii)) is violated.

Corollary 1. In an OA(N, $s_1 \ldots s_n$, R–1), let $\{u_1,\ldots,u_R\} \subseteq \{1,\ldots,n\}$, with $s_{\min} = \min_{i=1,\ldots,R}(s_{u_i})$.

(i) A factor $c \in \{u_1,\ldots,u_R\}$ in s_c levels is completely confounded by the factors in $C = \{u_1,\ldots,u_R\} \setminus \{ c\}$, if and only if $a_R(u_1,\ldots,u_R) = s_c - 1$.

(ii) $a_R(u_1,\ldots,u_R) \leq s_{\min} - 1$.

(iii) If several factors in $\{u_1,\ldots,u_R\}$ have $s_{\min}$ levels, either all of them are or none of them is completely confounded by the respective other R–1 factors in $\{u_1,\ldots,u_R\}$.

(iv) A factor with more than $s_{\min}$ levels cannot be completely confounded by the other factors in $\{u_1,\ldots,u_R\}$.

Part (i) of Corollary 1 follows easily from Theorem 1, as $a_R(u_1,\ldots,u_R) = s_c - 1$ if and only if all R^2 values for columns of the factor c main effects model matrix are 100%, i.e. the factor c main effects model matrix is completely explained by the factors in C. Part (ii) follows, because the sum of R^2 values is of course bounded by the minimum number of regressions conducted for any single factor c, which is $s_{\min} - 1$. Parts (iii) and (iv) follow directly from parts (i) and (ii). For symmetric s-level designs, part (ii) of the Corollary has already been proven by Xu, Cheng and Wu (2004).

Table 2. An OA(8, $4^1 2^2$, 2) (transposed)

A	0	0	0	0	1	1	1	1
B	0	0	1	1	0	0	1	1
C	0	2	1	3	3	1	2	0

Example 3. For the design of Table 2, $s_{\min} = 2$, and $a_3(1,2,3) = 1$, i.e. both 2-level factors are completely confounded, while the 4-level factor is only partially confounded. The individual R^2 values for the separate degrees of freedom of the 4-level factor main effect model matrix depend on the coding (e.g. 0.2, 0 and 0.8 for the linear, quadratic and cubic contrasts in normalized orthogonal polynomial coding), while their sum is 1, regardless of the chosen orthogonal coding.

Theorem 2. In an OA(N, $s_1 \ldots s_n$, R–1), let $\{u_1,\ldots,u_R\} \subseteq \{1,\ldots,n\}$ with $s_{\min} = \min_{i=1,\ldots,R}(s_{u_i})$. Let $c \in \{u_1,\ldots,u_R\}$ with $s_c = s_{\min}$, $C = \{u_1,\ldots,u_R\} \setminus \{c\}$. Under normalized orthogonal coding denote by $\mathbf{X}_c$ the main effects model matrix for factor c and by $\mathbf{X}_C$ the R–1 factor interaction model matrix for the factors in C.

If $a_R(u_1,\ldots,u_R) = s_{\min} - 1$, $\mathbf{X}_C$ can be orthogonally transformed (rotation and or switching) such that $s_{\min} - 1$ of its columns are collinear to the columns of $\mathbf{X}_c$.

<u>Proof.</u> $a_R(u_1,\ldots,u_R) = s_{\min} - 1$ implies all $s_{\min} - 1$ regressions of the columns of $\mathbf{X}_c$ on the columns of $\mathbf{X}_C$ have R^2=1. Then, each of the $s_{\min} - 1$ $\mathbf{X}_c$ columns can be perfectly matched by a linear combination $\mathbf{X}_C\mathbf{b}$ of the $\mathbf{X}_C$ columns; since all columns have the same length, this linear transformation involves rotation and/or switching only. If necessary, these $s_{\min} - 1$ orthogonal linear combinations can be supplemented by further length-preserving orthogonal linear combinations so that the dimension of $\mathbf{X}_C$ remains intact. ///

Theorems 1 and 2 are related to canonical correlation analysis, and the redundancy index discussed in that context (Stewart and Love 1968). In order to make the following comments digestible, a brief definition of canonical correlation analysis is included without going into any technical detail about the method; details can e.g. be found in Härdle and Simar (2003, Ch.14). It will be helpful to think of the columns of the main effects model matrix of factor *c* as the *Y* variables and the columns of the full model matrix in the *R*–1 other factors from the set C (excluding the constant column of ones for the intercept) as the *X* variables of the following definition and explanation. As it would be unnatural to consider the model matrices from experimental designs as random variables, we directly define canonical correlation analysis in terms of data matrices $\mathbf{X}$ and $\mathbf{Y}$ (*N* rows each) and empirical covariance matrices $\mathbf{S}_{xx} = \mathbf{X}^{*T}\mathbf{X}^*/(N-1)$, $\mathbf{S}_{yy} = \mathbf{Y}^{*T}\mathbf{Y}^*/(N-1)$, $\mathbf{S}_{xy} = \mathbf{X}^{*T}\mathbf{Y}^*/(N-1)$ and $\mathbf{S}_{yx} = \mathbf{Y}^{*T}\mathbf{X}^*/(N-1)$, where the superscript * denotes columnwise centering of a matrix. We do not attempt a minimal definition, but prioritize suitability for our purpose. Note that our $\mathbf{S}_{xx}$ and $\mathbf{S}_{yy}$ are nonsingular matrices, since the designs we consider have strength *R*–1; the covariance matrix $(\mathbf{X}^*\ \mathbf{Y}^*)^T(\mathbf{X}^*\ \mathbf{Y}^*)/(N-1)$ of the combined set of variables may, however, be singular, which does not pose a problem to canonical correlation analysis, even though some accounts request this matrix to be nonsingular.

<u>Definition 1</u>. Consider a set of *p* *X*-variables and *q* *Y*-variables. Let the Nxp matrix $\mathbf{X}$ and the Nxq matrix $\mathbf{Y}$ denote the data matrices of *N* observations, and $\mathbf{S}_{xx}$, $\mathbf{S}_{yy}$, $\mathbf{S}_{xy}$ and $\mathbf{S}_{yx}$ the empirical covariance matrices obtained from them, with positive definite $\mathbf{S}_{xx}$ and $\mathbf{S}_{yy}$.

(i) Canonical correlation analysis creates $k = \min(p, q)$ pairs of linear combination vectors $\mathbf{u}_i = \mathbf{X}\mathbf{a}_i$ and $\mathbf{v}_i = \mathbf{Y}\mathbf{b}_i$ with px1 coefficient vectors $\mathbf{a}_i$ and qx1 coefficient vectors $\mathbf{b}_i$, $i = 1,\ldots,k$, such that
   a) the $\mathbf{u}_1,\ldots,\mathbf{u}_k$ are uncorrelated to each other
   b) the $\mathbf{v}_1,\ldots,\mathbf{v}_k$ are uncorrelated to each other
   c) the pair $(\mathbf{u}_1, \mathbf{v}_1)$ has the maximum possible correlation for any pair of linear combinations of the $\mathbf{X}$ and $\mathbf{Y}$ columns, respectively
   d) the pairs $(\mathbf{u}_i, \mathbf{v}_i)$, i=2,…,k successively maximize the remaining correlation, given the constraints of a) and b).

(ii) The correlations $r_i = \mathrm{cor}(\mathbf{u}_i, \mathbf{v}_i)$ are called “canonical correlations”, and the $\mathbf{u}_i$ and $\mathbf{v}_i$ are called “canonical variates”.

<u>Remark 2</u>. (i) If the matrices $\mathbf{X}$ and $\mathbf{Y}$ are centered, i.e. $\mathbf{X} = \mathbf{X}^*$ and $\mathbf{Y} = \mathbf{Y}^*$, the $\mathbf{u}$ and $\mathbf{v}$ vectors also have zero means, and the uncorrelatedness in a) and b) is equivalent to orthogonality of the vectors. (ii) It is well-known that the canonical correlations are the eigenvalues of the matrices $\mathbf{Q}_1 = \mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}$ and $\mathbf{Q}_2 = \mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$ (the first $\min(p, q)$ eigenvalues of both matrices are

the same; the larger matrix has the appropriate number of additional zeroes), and the $\mathbf{a}_i$ are the corresponding eigenvectors of $\mathbf{Q}_1$, the $\mathbf{b}_i$ the corresponding eigenvectors of $\mathbf{Q}_2$.

According to the definition, the canonical correlations are non-negative. It can also be shown that $\mathbf{u}_i$ and $\mathbf{v}_j$, $i \neq j$, are uncorrelated, and orthogonal in case of centered data matrices; thus, the pairs ($\mathbf{u}_i$, $\mathbf{v}_i$) decompose the relation between $\mathbf{X}$ and $\mathbf{Y}$ into uncorrelated components, much like the principal components decompose the total variance into uncorrelated components. In data analysis, canonical correlation analysis is often used for dimension reduction. Here, we retain the full dimensionality. For uncorrelated *Y* variables like the model matrix columns of $\mathbf{X}_c$ in Theorem 1, it is straightforward to see that the sum of the R^2 values from regressing each of the *Y* variables on all the *X* variables coincides with the sum of the squared canonical correlations. It is well-known that the canonical correlations are invariant to arbitrary nonsingular affine transformations applied to the *X*- and *Y*-variables, which translate into nonsingular linear transformations applied to the centered **X**- and **Y**-matrices (cf. e.g. Härdle and Simar 2003, Theorem 14.3). For our application, this implies invariance of the canonical correlations to factor coding. Unfortunately, this invariance property does not hold for the R^2 values or their sum: according to Lazraq and Cléroux (2001, Section 2) the afore-mentioned redundancy index – which is the average R^2 value calculated as $a_R(u_1,\ldots,u_R)/(s_c-1)$ in the situation of Theorem 1 – is invariant to linear transformations of the centered **X** matrix, but only to *orthonormal* transformations of the centered **Y** matrix or scalar multiples thereof. For correlated *Y*-variables, the redundancy index contains some overlap between variables, as was already seen for example 2, where the sum of the R^2 values from dummy coding exceeded $a_3(1,2,3)$; in that case, only the average or sum of the squared canonical correlations yields an adequate measure of the overall explanatory power of the *X*-variables on the *Y*-variables. Hence, for the case of arbitrary coding, Theorem 1 has to be restated in terms of squared canonical correlations:

<u>Theorem 3</u>. In an OA(N, $s_1 \ldots s_n$, $R-1$), denote by $\mathbf{X}_c$ the model matrix for the main effects of a particular factor $c \in \{u_1,\ldots,u_R\}$ in arbitrary coding, and let $\mathrm{C} = \{u_1,\ldots,u_R\} \setminus \{c\}$. Then, $a_R(u_1,\ldots,u_R)$ is the sum of the squared canonical correlations from a canonical correlation analysis of the columns of $\mathbf{X}_c$ and the columns of the full model matrix $\mathbf{F}_\mathrm{C}$ in the factors from C.

<u>Example 1, continued</u>. $s_{\min}=3$, $a_3(1,2,3) = 2$, i.e. the assumptions of Theorems 2 and 3 are fulfilled. Both canonical correlations must be 1, because the sum must be 2. The transformation of $\mathbf{X}_\mathrm{C}$ from Theorem 2 can be obtained from the canonical correlation analysis: For all factors in the role of *Y*, $\mathbf{v}_i \propto \mathbf{y}_i$ (with $\mathbf{y}_i$ denoting the *i*-th column of the main effects model matrix of the *Y*-variables factor) can be used. For the first or second factor in the role of *Y*, the corresponding canonical vectors on the *X* side fulfill

$\mathbf{u}_1 \propto \mathrm{B}_q\mathrm{C}_l - \mathrm{B}_l\mathrm{C}_q - \sqrt{3}\,\mathrm{B}_l\mathrm{C}_l - \sqrt{3}\,\mathrm{B}_q\mathrm{C}_q$,

$\mathbf{u}_2 \propto \sqrt{3}\,\mathrm{B}_l\mathrm{C}_q - \sqrt{3}\,\mathrm{B}_q\mathrm{C}_l - \mathrm{B}_l\mathrm{C}_l - \mathrm{B}_q\mathrm{C}_q$ (or B replaced by A for the second factor in the role of *Y*), with the indices *l* and *q* denoting the *normalized* linear and quadratic coding introduced above.

For the third factor in the role of *Y*,

$\mathbf{u}_1 \propto -\sqrt{3}\,\mathrm{A}_l\mathrm{B}_l + \mathrm{A}_q\mathrm{B}_l + \mathrm{A}_l\mathrm{B}_q + \sqrt{3}\,\mathrm{A}_q\mathrm{B}_q$,

$\mathbf{u}_2 \propto -\mathrm{A}_l\mathrm{B}_l - \sqrt{3}\,\mathrm{A}_l\mathrm{B}_q - \sqrt{3}\,\mathrm{A}_q\mathrm{B}_l + \mathrm{A}_q\mathrm{B}_q$.

<u>Example 1, now with dummy coding</u>. When using the design of Example 1 for an experiment with qualitative factors, dummy coding is much more usual than orthogonal contrast coding. This example shows how Theorem 3 can be applied for arbitrary non-orthogonal coding: A_1 is 1 for A=1 and 0

otherwise, A_2 is 1 for A=2 and 0 otherwise, B and C are coded analogously; interaction matrix columns are obtained as products of the respective main effects columns. The main effect and two-factor interaction model matrix columns in this coding do not have column means zero and have to be centered first by subtracting 1/3 or 1/9, respectively. As canonical correlations are invariant to affine transformations, dummy coding leads to the same canonical correlations as the previous normalized orthogonal polynomial coding. We consider the first factor in the role of *Y*; the centered model matrix columns $\mathbf{y}_1 = A_1-1/3$ and $\mathbf{y}_2 = A_2-1/3$ are correlated, so that we must not choose both canonical variates for the *Y* side proportional to the original variates. One instance of the canonical variates for the *Y* side is $\mathbf{v}_1 = -\mathbf{y}_1/\sqrt{2}$, $\mathbf{v}_2 = (\mathbf{y}_1 + 2\mathbf{y}_2)/\sqrt{6}$; these canonical vectors are unique up to rotation only, because the two canonical correlations have the same size. The corresponding canonical vectors on the *X* side are obtained from the centered full model matrix

$$\mathbf{F}_C = \left(\left(B_1 - \tfrac{1}{3}\right), \left(B_2 - \tfrac{1}{3}\right), \left(C_1 - \tfrac{1}{3}\right), \left(C_2 - \tfrac{1}{3}\right), \left(B_1C_1 - \tfrac{1}{9}\right), \left(B_2C_1 - \tfrac{1}{9}\right), \left(B_1C_2 - \tfrac{1}{9}\right), \left(B_2C_2 - \tfrac{1}{9}\right)\right)$$

as $\mathbf{u}_1 = (-\mathbf{f}_2 - \mathbf{f}_3 + \mathbf{f}_5 + 2\mathbf{f}_6 - \mathbf{f}_7 + \mathbf{f}_8)/\sqrt{2}$ and $\mathbf{u}_2 = (2\mathbf{f}_1 + \mathbf{f}_2 + \mathbf{f}_3 + 2\mathbf{f}_4 - 3\mathbf{f}_5 - 3\mathbf{f}_7 - 3\mathbf{f}_8)/\sqrt{6}$, with $\mathbf{f}_j$ denoting the *j*-th column of $\mathbf{F}_C$.
Note that the canonical vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ now contain contributions not only from the interaction part of the model matrix but also from the main effects part, i.e. we do indeed need the full model matrix as stated in Theorem 3.

Example 2, continued. $s_{\min} = 2$, $a_3(1,2,3) = 4/9$, i.e. the assumption of Theorem 2 is not fulfilled, the assumption of Theorem 3 is. The canonical correlation using the one column main effects model matrix of the 2-level factor A in the role of *Y* is 2/3, the canonical correlations using the main effects model matrix for the 3-level factor B in the role of *Y* are 2/3 and 0; in both cases, the sum of the squared canonical correlations is $a_3(1,2,3) = 4/9$. For any other coding, for example the dummy coding for factor B considered earlier, the canonical correlations remain unchanged (2/3 and 0, respectively), since they are coding invariant; thus, the sum of the squared canonical correlations remains 4/9, even though the sum of the R^2 values was found to be different. Of course, the linear combination coefficients for obtaining the canonical variates depend on the coding (see e.g. Härdle and Simar 2003 Theorem 14.3).

Canonical correlation analysis can also be used to verify that a result analogous to Theorem 2 cannot be generalized to sets of *R* factors for which $a_R(u_1,\ldots,u_R) < s_{\min} - 1$. For this, note that the number of non-zero canonical correlations indicates the dimension of the relationship between the *X*- and the *Y*-variables.

Table 3 displays the R^2 values from two different orthogonal codings and the squared canonical correlations from the main effects matrix of the first factor (*Y*-variables) vs. the full model matrix of the other two factors (*X*-variables) for the ten non-isomorphic GMA OA(32,4^3,2) obtained from Eendebak and Schoen (2013). These designs have one generalized word of length 3, i.e. they are non-regular. There are cases with one, two and three non-zero canonical correlations, i.e. neither is it generally possible to collapse the linear dependence into a one-dimensional structure nor does the linear dependence generally involve more than one dimension.

Table 3. Main effects matrix of factor A regressed on full model in factors B and C for the 10 non-isomorphic GMA OA(32, 4^3, 2)

R^2 values from polynomial coding			R^2 values from Helmert coding			Squared canonical correlations				Designs
L	Q	C	1	2	3	1	2	3	A_3	
0.8	0	0.2	0	2/3	1/3	1	0	0	1	1
0.65	0	0.35	1/8	13/24	1/3	0.75	0.25	0	1	2
0.5	0	0.5	1/4	5/12	1/3	0.5	0.5	0	1	3,6,8,10
0.45	0.25	0.3	1/4	5/12	1/3	0.5	0.25	0.25	1	4,5,7
0.375	0.25	0.375	5/16	17/48	1/3	0.375	0.375	0.25	1	9

3. Generalized resolution

Before presenting the new proposals for generalized resolution, we briefly review generalized resolution for symmetric 2-level designs by Deng and Tang (1999) and Tang and Deng (1999). For 2-level factors, each effect has a single degree of freedom (df) only, i.e. all the **X**'s in any $\mathbf{M}_k$ (cf. Equation (1)) are one-column matrices. Deng and Tang (1999) looked at the absolute sums of the columns of **M**, which were termed *J*-characteristics by Tang and Deng (1999). Specifically, for a resolution *R* design, these authors introduced *GR* as

$$GR = R + 1 - \frac{\max J_R}{N}, \tag{3}$$

where $J_R = |\mathbf{1}_N^{\mathrm{T}}\mathbf{M}_R|$ is the row vector of the *J*-characteristics $|\mathbf{1}_N^{\mathrm{T}}\mathbf{X}_{u_1 \ldots u_R}|$ obtained from the $\binom{n}{R}$ *R*-factor interaction model columns $\mathbf{X}_{u_1 \ldots u_R}$. For 2-level designs, it is straightforward to verify the following identities:

$$GR = R + 1 - \sqrt{\max_{(u_1,\ldots,u_R)} a_R(u_1,\ldots,u_R)} = R + 1 - \max_{(u_1,\ldots,u_R)} \left|\rho(X_{u_1}, X_{u_2 \ldots u_R})\right|, \tag{4}$$

where ρ denotes the correlation; note that the correlation in (4) does not depend on which of the u_i takes the role of u_1. Deng and Tang (1999, prop. 2) proved a very convincing projection interpretation of their *GR*. Unfortunately, Prop. 4.4 of Diestelkamp and Beder (2002), in which a particular OA(18, 3^3, 2) is proven to be indecomposable into two OA(9, 3^3, 2), implies that Deng and Tang's result cannot be generalized to more than two levels.

The quantitative approach by Evangelaras et al. (2005, their eq. (4)) generalized the correlation version of (4) by applying it to single df contrasts for the quantitative factors. For the qualitative factors considered here, any approach based on direct usage of single df contrasts is not acceptable because it is coding dependent. The approach for qualitative factors taken by Evangelaras et al. is unreasonable, as will be demonstrated in Example 5. Pang and Liu (2010) also proposed a generalized resolution based on complex contrasts. For designs with more than 3 levels, permuting levels for one or more factors will lead to different generalized resolutions according to their definition, which is unacceptable for qualitative factors. For 2-level designs, their approach boils down to omitting the square root from $\sqrt{\max_{(u_1,\ldots,u_R)} a_R(u_1,\ldots,u_R)}$ in (4), which implies that their proposal does not simplify to the well-grounded generalized resolution of Deng and Tang (1999) / Tang and Deng (1999) for 2-level designs. This in itself makes their approach unconvincing. Example 5 will compare their approach to ours for 3-level designs. The results from the previous section can be used to create two adequate generalizations of *GR* for qualitative factors. These are introduced in the following two definitions.

For the first definition, an *R* factor projection is considered as completely aliased, whenever all the levels of at least one of the factors are completely determined by the level combination of the other *R*–1 factors. Thus, generalized resolution should be equal to *R*, if and only if there is at least one *R* factor projection with $a_R(u_1,\dots,u_R) = s_{\min}-1$. The *GR* defined in Definition 2 guarantees this behavior and fulfills all requirements stated in the introduction:

<u>Definition 2</u>. For an OA(N, $s_1 \dots s_n$, R–1),

$$GR = R+1-\sqrt{\max_{\substack{\{u_1,\dots,u_R\}\\ \subseteq\{1,\dots,n\}}} \frac{a_R(u_1,\dots,u_R)}{\min_{i=1,\dots,R} s_{u_i}-1}}.$$

In words, *GR* increases the resolution by one minus the square root of the worst case average R^2 obtained from any *R* factor projection, when regressing the main effects columns in orthogonal coding from a factor with the minimum number of levels on the other factors in the projection. It is straightforward to see that (4) is a special case of the definition, since the denominator is 1 for 2-level designs. Regarding the requirements stated in the introduction, (i) *GR* from Def. 2 is coding invariant because the $a_R()$ are coding invariant according to Xu and Wu (2001). (ii) The technique is obviously applicable for symmetric and asymmetric designs alike, and (iii) $GR < R + 1$ follows from the resolution, $GR \geq R$ follows from part (ii) of Corollary 1, $GR = R$ is equivalent to complete confounding in at least one *R*-factor projection according to part (i) of Corollary 1.

<u>Example 4</u>. The *GR* values for the designs from Examples 1 and 3 are 3 (*GR*=*R*), the *GR* value for the design from Example 2 is $3+1-\sqrt{4/9}=3.33$, and the *GR* values for all designs from Table 3 are $3+1-\sqrt{1/3}=3.42$.

Now, complete aliasing is considered regarding individual degrees of freedom (df). A coding invariant individual df approach considers a factor's main effect as completely aliased in an *R* factor projection, whenever there is at least one pair of canonical variates with correlation one. A projection is considered completely aliased, if at least one factor's main effect is completely aliased in this individual df sense. Note that it is now possible that factors with the same number of levels can show different extents of individual df aliasing within the same projection, as will be seen in Example 5 below.

<u>Definition 3</u>. For an OA(N, $s_1 \dots s_n$, R–1) and tuples (c, C) with C = $\{u_1,\dots,u_R\} \setminus \{c\}$,

$$GR_{\text{ind}} = R+1-\max_{\{u_1,\dots,u_R\}\subseteq\{1,\dots,n\}} \max_{c\in\{u_1,\dots,u_R\}} r_1\left(\mathbf{X}_c;\mathbf{F}_{\text{C}}\right)$$

with $r_1(\mathbf{X}_c;\ \mathbf{F}_{\text{C}})$ the largest canonical correlation between the main effects model matrix for factor c and the full model matrix of the factors in C.

In words, GR_{ind} is the worst case confounding for an individual main effects df in the design that can be obtained by the worst case coding (which corresponds to the $\mathbf{v}_1$ vector associated with the worst canonical correlation). Obviously, GR_{ind} is thus a stricter criterion than *GR*. Formally, Theorem 3 implies that *GR* from Def. 2 can be written as

$$GR = R + 1 - \sqrt{\max_{\substack{(u_1,\ldots,u_R):\\ \{u_1,\ldots,u_R\}\subseteq\{1,\ldots,n\}}} \frac{\sum_{j=1}^{s_{u_1}-1} r_j\left(\mathbf{X}_{u_1};\mathbf{F}_{\{u_2,\ldots,u_R\}}\right)^2}{\min_i s_{u_i} - 1}}\ . \tag{5}$$

Note that maximization in (5) is over tuples, so that it is ensured that the factor with the minimum number of levels does also get into the first position. Comparing (5) with Def. 3, $GR_{\text{ind}} \leq GR$ is obvious, because r_1^2 cannot be smaller than the average over all r_i^2 (but can be equal, if all canonical correlations have the same size). This is stated in a Theorem:

<u>Theorem 4</u>. For GR from Def. 2 and GR_{ind} from Def. 3, $GR_{\text{ind}} \leq GR$.

<u>Remark 3</u>. (i) Under normalized orthogonal coding, the full model matrix $\mathbf{F}_{\text{C}}$ in Definition 3 can again be replaced by the R–1 factor interaction matrix $\mathbf{X}_{\text{C}}$. (ii) Definition 3 involves calculation of $R\binom{n}{R}$ canonical correlations (R correlations for each R factor projection). In any projection with at least one 2-level factor, it is sufficient to calculate one single canonical correlation obtained with an arbitrary 2-level factor in the role of Y, because this is necessarily the worst case. Nevertheless, calculation of GR_{ind} carries some computational burden for designs with many factors.

Obviously, (4) is a special case of GR_{ind}, since the average R^2 coincides with the only squared canonical correlation for projections of R 2-level factors. GR_{ind} also fulfills all requirements stated in the introduction: (i) GR_{ind} is coding invariant because the canonical correlations are invariant to affine transformations of the X and Y variables, as was discussed in Section 2. (ii) The technique is obviously applicable for symmetric and asymmetric designs alike, and (iii) $GR_{\text{ind}} < R+1$ again follows from the resolution, $GR_{\text{ind}} \geq R$ follows from the properties of correlations, and $GR_{\text{ind}} = R$ is obviously equivalent to complete confounding of at least one main effects contrast in at least one R factor projection, in the individual df sense discussed above.

Table 4. The Taguchi L18 (transposed)

Row / Col.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
2	0	0	0	1	1	1	2	2	2	0	0	0	1	1	1	2	2	2
3	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
4	0	1	2	0	1	2	1	2	0	2	0	1	1	2	0	2	0	1
5	0	1	2	1	2	0	0	1	2	2	0	1	2	0	1	1	2	0
6	0	1	2	1	2	0	2	0	1	1	2	0	0	1	2	2	0	1
7	0	1	2	2	0	1	1	2	0	1	2	0	2	0	1	0	1	2
8	0	1	2	2	0	1	2	0	1	0	1	2	1	2	0	1	2	0

<u>Example 5</u>. We consider the three non-isomorphic OA(18, 3^3, 2) that can be obtained as projections from the well-known Taguchi L18 (see Table 4) by using columns 3, 4 and 5 (D_1), columns 2, 3 and 6 (D_2) or columns 2, 4 and 5 (D_3). We have $A_3(D_1) = 0.5$, $A_3(D_2) = 1$ and $A_3(D_3) = 2$, and consequently $GR(D_1) = 3.5$, $GR(D_2) = 3.29$ and $GR(D_3) = 3$. For calculating GR_{ind}, the largest canonical correlations of all factors in the role of Y are needed. These are all 0.5 for D_1 and all 1 for D_3, such that $GR_{\text{ind}} = GR$ for these two designs. For D_2, the largest canonical correlation is 1 with the first factor (from column 2 of the L18) in the role of Y, while it is $\sqrt{0.5}$ with either of the other two factors in the role of Y;

thus, $GR_{ind} = 3 < GR = 3.29$. The completely aliased 1 df contrast of the first factor is the contrast of the third level vs. the other two levels, which is apparent from Table 5: the contrast A = 2 vs A in (0,1) is fully aliased with the contrast of one level of B vs. the other two, given a particular level of C. Regardless of factor coding, this direct aliasing is reflected by a canonical correlation "one" for the first canonical variate of the main effects contrast matrix of factor A.

Table 5. Frequency table of columns 2(=A), 3(=B) and 6(=C) of the Taguchi L18

```
, , C = 0          , , C = 1          , , C = 2
   B                  B                  B
A    0 1 2        A    0 1 2        A    0 1 2
  0  1 0 1          0  1 1 0          0  0 1 1
  1  1 0 1          1  1 1 0          1  0 1 1
  2  0 2 0          2  0 0 2          2  2 0 0
```

Using this example, we now compare the *GR* introduced here to proposals by Evangelaras et al. (2005) and Pang and Liu (2010): The *GRes* values reported by Evangelaras et al. (2005) for designs D_1, D_2 and D_3 in the qualitative case are 3.75, 3.6464, 3.5, respectively; especially the 3.5 for the completely aliased design D_3 does not make sense. Pang and Liu reported values 3.75, 3.75 and 3, respectively; here, at least the completely aliased design D_3 is assigned the value "3". Introducing the square root, as was discussed in connection with equation (4), their generalized resolutions become 3.5, 3.5 and 3, respectively, i.e. they coincide with our *GR* results for designs D_1 and D_3. For design D_2, their value 3.5 is still different from our 3.29 for the following reason: our approach considers $A_3 = a_3(1,2,3)$ as a sum of two R^2-values and subtracts the square root of their average or maximum (*GR* or GR_{ind}, respectively), while Pang and Liu's approach considers it as a sum of 2^3=8 summands, reflecting the potentially different linear combinations of the three factors in the Galois field sense, the (square root of the) maximum of which they subtract from R+1.

4. Properties of *GR*

Let G be the set of all runs of an $s_1 \times \ldots \times s_n$ full factorial design, with $|\mathrm{G}| = \prod_{i=1}^{n} s_i$ the cardinality of G. For any design D in N runs for n factors at $s_1, \ldots, s_n$ levels, let $N_{\mathbf{x}}$ be the number of times that a point $\mathbf{x} \in \mathrm{G}$ appears in D. $\bar{N} = N / |\mathrm{G}|$ denotes the average frequency for each point of G in the design D. We can measure the goodness of a fractional factorial design D by the uniformity of the design points of D in the set of all points in G, that is, the uniformity of the frequency distribution $N_{\mathbf{x}}$. One measure, suggested by Tang (2001) and Ai and Zhang (2004), is the variance

$$\mathrm{V}(D) = \frac{1}{|\mathrm{G}|} \sum_{\mathbf{x} \in \mathrm{G}} \left(N_{\mathbf{x}} - \bar{N}\right)^2 = \frac{1}{|\mathrm{G}|} \sum_{\mathbf{x} \in \mathrm{G}} N_{\mathbf{x}}^{\,2} - \bar{N}^2 .$$

Let $N = q|\mathrm{G}| + r$ with nonnegative integer q and r and $0 \le r < |\mathrm{G}|$ (often q=0), i.e. $r = N \bmod |\mathrm{G}|$ is the remainder of N when divided by $|\mathrm{G}|$. Note that $\sum_{\mathbf{x} \in \mathrm{G}} N_{\mathbf{x}} = N$, so V($D$) is minimized if and only if each $N_{\mathbf{x}}$ takes values on q or $q + 1$ for any $\mathbf{x} \in$ G. When r points in G appear $q + 1$ times and the remaining $|\mathrm{G}| - r$ points appear q times, V(D) reaches the minimal value $r(|\mathrm{G}| - r) / |\mathrm{G}|^2$. Ai and Zhang (2004) showed that V(D) is a function of GWLP. In particular, if D has strength n–1, their result implies that $V(D) = \bar{N}^2 A_n(d)$. Combining these results, and using the following definition, we obtain an upper bound for GR for some classes of designs and provide a necessary and sufficient condition under which this bound is achieved.

Definition 4 (modified from Xu 2003).

(i) A design D has maximum t-balance, if and only if the possible level combinations for all projections onto t columns occur as equally often as possible, i.e. either q or q+1 times, where q is an integer such that $N = q|G_{proj}| + r$ with G_{proj} the set of all runs for the full factorial design of each respective t-factor-projection and $0 \le r < |G_{proj}|$.

(ii) An OA(N, $s_1 \dots s_n$, t–1) with $n \ge t$ has weak strength t if and only if it has maximum t-balance. We denote weak strength t as OA(N, $s_1 \dots s_n$, t^-).

Remark 4. Xu (2003) did not require strength t–1 in the definition of weak strength t, i.e. the Xu (2003) definition of weak strength t corresponds to our definition of maximum t-balance. For the frequent case, for which all t-factor projections have q=0 or q=1 and r=0 in Def. 4 (i), maximum t-balance is equivalent to the absence of repeated runs in any projection onto t factors. In that case, maximum t-balance implies maximum k-balance for $k > t$, and weak strength t is equivalent to strength t–1 with absence of repeated runs in any projection onto t or more factors.

Theorem 5. Let D be an OA(N, $s_1 \dots s_R$, R–1). Then $A_R(D) \ge \frac{r\left(\prod_{i=1}^{R} s_i - r\right)}{N^2}$, where r is the remainder when N is divided by $\prod_{i=1}^{R} s_i$. The equality holds if and only if D has weak strength R.

As all R factor projections of any OA(N, $s_1 \dots s_n$, R^-) fulfill the necessary and sufficient condition of Theorem 5, we have the following corollary:

Corollary 2. Suppose that an OA(N, $s_1 \dots s_n$, R) does not exist. Then any OA(N, $s_1 \dots s_n$, R^-) has maximum GR among all OA(N, $s_1 \dots s_n$, R–1).

Corollary 3. Suppose that an OA(N, s^n, R) does not exist. Let D be an OA(N, s^n, R–1). Then $GR(D) \le R + 1 - \sqrt{\frac{r\left(s^R - r\right)}{N^2\left(s-1\right)}}$, where r is the remainder when N is divided by s^R. The equality holds if and only if D has weak strength R.

Example 6. (1) Any projection onto three 3-level columns from an OA(18, $6^1 3^6$, 2) has 18 distinct runs (q=0, r=N=18) and is an OA of weak strength 3, so it has A_3=1/2 and $GR = 4 - \sqrt{18 \bullet 9 / \left(18^2 \bullet 2\right)}$ =3.5. (2) Any projection onto three or more s-level columns from an OA(s^2, s^{s+1}, 2) has GR = 3, since $N = r = s^2$, so that the upper limit from the corollary becomes $GR = R = 3$.

Using the following lemma according to Mukerjee and Wu 1995, Corollary 3 can be applied to a further class of designs.

Lemma 1 (Mukerjee and Wu 1995). For a saturated OA(N, $s_1^{n_1} s_2^{n_2}$, 2) with $n_1(s_1-1) + n_2(s_2-1) = N - 1$, let $\delta_i(a, b)$ be the number of coincidences of two distinct rows a and b in the n_i columns of s_i levels, for i = 1, 2. Then

$$s_1\, \delta_1(a, b) + s_2\, \delta_2(a, b) = n_1 + n_2 - 1.$$

Consider a saturated OA($2s^2$, $(2s)^1 s^{2s}$, 2), where $r = N = 2s^2$, $s_1 = 2s$, $s_2 = s$, $n_1 = 1$, $n_2 = 2s$. From Lemma 1, we have $2\delta_1(a, b) + \delta_2(a, b) = 2$. So any projection onto three or more s-level columns has no repeated runs, and thus it achieves the upper limit $GR = 4 - \sqrt{(s-2)/(2s-2)}$ according to Corollary 3.

Corollary 4. For a saturated OA($2s^2$, $(2s)^1 s^{2s}$, 2), any projection onto three or more s-level columns has $GR = 4 - \sqrt{(s-2)/(2s-2)}$, which is optimum among all possible OAs in $2s^2$ runs.

Example 7. Design 1 of Table 3 is isomorphic to a projection from a saturated OA(32, $8^1 4^8$, 2). A_3 attains the lower bound from Theorem 5 (32•(64–32)/32^2 = 1), and thus GR attains the upper bound $4-(1/3)^{½}$ = 3.42 from the corollary.

Because of Theorem 4, any upper bound for GR is of course also an upper bound for GR_{ind}, i.e. Corollaries 3 and 4 also provide upper bounds for GR_{ind}. However, for GR_{ind} the bounds are not tight in general; for example, GR_{ind} = 3 for the design of Example 7 (see also Example 9 in the following section).

Butler (2005) previously showed that all projections onto s-level columns of OA(s^2, s^{s+1}, 2) or OA($2s^2$, $(2s)^1 s^{2s}$, 2) have GMA among all possible designs.

5. Factor wise GR values

In Section 3, two versions of overall generalized resolution were defined: GR and GR_{ind}. These take a worst case perspective: even if a single projection in a large design is completely confounded – in the case of mixed level designs or GR_{ind} affecting perhaps only one factor within that projection – the overall metric takes the worst case value R. It can therefore be useful to accompany GR and GR_{ind} by factor specific summaries. For the factor specific individual df perspective, one simply has to omit the maximization over the factors in each projection and has to use the factor of interest in the role of Y only. For a factor specific complete confounding perspective, one has to divide each projection's $a_R()$ value by the factor's df rather than the minimum df, in order to obtain the average R^2 value for this particular factor. This leads to

Definition 5. For an OA(N, $s_1 \ldots s_n$, R–1), define

(i) $$GR_{\text{tot}(i)} = R + 1 - \sqrt{\max_{\substack{\{u_2,\ldots,u_R\} \\ \subseteq \{1,\ldots,n\} \setminus \{i\}}} \frac{a_R(i, u_2, \ldots, u_R)}{s_i - 1}}$$

(ii) $$GR_{\text{ind}(i)} = R + 1 - \max_{\{i,u_2,\ldots,u_R\} \subseteq \{1,\ldots,n\}} r_1\left(\mathbf{X}_i; \mathbf{X}_{u_2 \ldots u_R}\right),$$ with $\mathbf{X}_i$ the model matrix of factor i and $\mathbf{X}_{u_2 \ldots u_R}$ the R–1 factor interaction model matrix of the factors in $\{u_2,\ldots,u_R\}$ in normalized orthogonal coding, and $r_1(\mathbf{Y};\mathbf{X})$ the first canonical correlation between matrices $\mathbf{X}$ and $\mathbf{Y}$.

It is straightforward to verify that GR and GR_{ind} can be calculated as the respective minima of the factor specific GR values from Definition 5:

Theorem 6. For the quantities from Definitions 2, 3 and 5, we have

(i) $GR = \min_i GR_{\text{tot}(i)}$

(ii) $GR_{\text{ind}} = \min_i GR_{\text{ind}(i)}$

Example 8. The Taguchi L18 has $GR = GR_{\text{ind}}$ = 3, and the following $GR_{\text{ind}(i)}$ and $GR_{\text{tot}(i)}$ values ($GR_{\text{ind}(i)} = GR_{\text{tot}(i)}$ for all i): 3.18, 3, 3.29, 3, 3, 3.29, 3.29, 3.29. When omitting the second column, the remaining seven columns have $GR = GR_{\text{ind}}$ = 3.18, again with $GR_{\text{ind}(i)} = GR_{\text{tot}(i)}$ and the value for all 3-level factors at 3.42. When omitting the fourth column instead, the then remaining seven columns have GR = 3.18, GR_{ind} = 3, $GR_{\text{tot}(i)}$ values 3.18, 3.29, 3.29, 3.42, 3.29, 3.29, 3.29 and $GR_{\text{ind}(i)}$ values the same, except for the second column, which has $GR_{\text{ind}(2)}$ = 3.

GR from Def. 2 and GR_{ind} from Def. 3 are not the only possible generalizations of (4). It is also possible to define a GR_{tot}, by declaring only those R factor projections as completely confounded for which all factors are completely confounded. For this, the factor wise average R^2 values for each projection – also used in $GR_{\text{tot}(i)}$ – need to be considered. A projection is completely confounded, if these are all one, which can be formalized by requesting their minimum or their average to be one. The average appears more informative, leading to

$$GR_{\text{tot}} = R + 1 - \sqrt{\max_{\substack{\{u_1,\dots,u_R\} \\ \subseteq\{1,\dots,n\}}} \frac{1}{R} \sum_{i=1}^{R} \frac{a_R(u_1,\dots,u_R)}{s_{u_i} - 1}} . \tag{6}$$

It is straightforward to see that $GR_{\text{tot}} \geq GR$, and that $GR_{\text{tot}} = GR$ for symmetric designs. The asymmetric design of Table 2 (Example 3) has GR = 3 and $GR_{tot} = 3 + 1 - \sqrt{(1+1+1/3)/3} = 3.12$ > 3, in spite of the fact that two of its factors are completely confounded. Of course, mixed level projections can never be completely confounded according to (6), which is the main reason why we have not pursued this approach.

The final example uses the designs of Table 3 to show that GR_{ind} and the $GR_{\text{ind}(i)}$ can introduce meaningful differentiation between GMA designs.

Example 9. All designs of Table 3 had A_3=1 and GR=3.42. The information provided in Table 3 is insufficient for determining GR_{ind}. Table 6 provides the necessary information: the largest canonical correlations are the same regardless which variable is chosen as the Y variable for seven designs, while they vary with the choice of the Y variable for three designs. There are five different GR_{ind} values for these 10 designs that were not further differentiated by A_3 or GR, and in combination with the $GR_{\text{ind}(i)}$, seven different structures can be distinguished.

Table 6. Largest canonical correlations, $GR_{\text{ind}(i)}$ and GR_{ind} values for the GMA OA(32, 4^3, 2)

	$r_1(1;23)$	$r_1(2;13)$	$r_1(3;12)$	$GR_{\text{ind}(1)}$	$GR_{\text{ind}(2)}$	$GR_{\text{ind}(3)}$	GR_{ind}
1	1.000	1.000	1.000	3.000	3.000	3.000	3.000
2	0.866	0.866	0.866	3.134	3.134	3.134	3.134
3	0.707	0.707	1.000	3.293	3.293	3.000	3.000
4	0.707	0.707	0.866	3.293	3.293	3.134	3.134
5	0.707	0.707	0.791	3.293	3.293	3.209	3.209
6	0.707	0.707	0.707	3.293	3.293	3.293	3.293
7	0.707	0.707	0.707	3.293	3.293	3.293	3.293
8	0.707	0.707	0.707	3.293	3.293	3.293	3.293
9	0.612	0.612	0.612	3.388	3.388	3.388	3.388
10	0.707	0.707	0.707	3.293	3.293	3.293	3.293

The differentiation achieved by GR_{ind} is meaningful, as can be seen by comparing frequency tables of the first, third and ninth design (see Table 7). The first and third design have GR_{ind}=3, which is due to a very regular confounding pattern: in the first design, dichotomizing each factor into a 0/1 vs. 2/3 design yields a regular resolution III 2-level design (four different runs only), i.e. each main effect contrast 0/1 vs. 2/3 is completely confounded by the two-factor interaction of the other two 0/1 vs. 2/3 contrasts; the third design shows this severe confounding for factor C only, whose 0/1 vs. 2/3 contrast

Table 7. Frequency tables of designs 1, 3 and 9 from Table 6

Design 1

```
, , C = 0          , , C = 1          , , C = 2          , , C = 3

   B                  B                  B                  B
A   0 1 2 3        A   0 1 2 3        A   0 1 2 3        A   0 1 2 3
  0 1 1 0 0          0 1 1 0 0          0 0 0 1 1          0 0 0 1 1
  1 1 1 0 0          1 1 1 0 0          1 0 0 1 1          1 0 0 1 1
  2 0 0 1 1          2 0 0 1 1          2 1 1 0 0          2 1 1 0 0
  3 0 0 1 1          3 0 0 1 1          3 1 1 0 0          3 1 1 0 0
```

Design 3

```
, , C = 0          , , C = 1          , , C = 2          , , C = 3

   B                  B                  B                  B
A   0 1 2 3        A   0 1 2 3        A   0 1 2 3        A   0 1 2 3
  0 1 1 0 0          0 1 1 0 0          0 0 0 1 1          0 0 0 1 1
  1 1 0 1 0          1 1 0 1 0          1 0 1 0 1          1 0 1 0 1
  2 0 1 0 1          2 0 1 0 1          2 1 0 1 0          2 1 0 1 0
  3 0 0 1 1          3 0 0 1 1          3 1 1 0 0          3 1 1 0 0
```

Design 9

```
, , C = 0          , , C = 1          , , C = 2          , , C = 3

   B                  B                  B                  B
A   0 1 2 3        A   0 1 2 3        A   0 1 2 3        A   0 1 2 3
  0 1 1 0 0          0 1 0 1 0          0 0 1 0 1          0 0 0 1 1
  1 1 0 1 0          1 0 1 0 1          1 1 0 0 1          1 0 1 1 0
  2 0 0 1 1          2 1 1 0 0          2 0 1 1 0          2 1 0 0 1
  3 0 1 0 1          3 0 0 1 1          3 1 0 1 0          3 1 1 0 0
```

is likewise completely confounded by the interaction between factors A and B. Design 9 is the best of all GMA designs in terms of GR_{ind}. It does not display such a strong regularity in behavior. GR_{ind} treats Designs 1 and 3 alike, although Design 1 is clearly more severely affected than Design 3, which can be seen from the individual $GR_{\mathrm{ind}(i)}$. However, as generalized resolution has always taken a “worst case” perspective, this way of handling things is appropriate in this context.

6. Discussion

We have provided a statistically meaningful interpretation for the building blocks of GWLP and have generalized generalized resolution by Deng and Tang (1999) and Tang and Deng (1999) in two meaningful ways for qualitatitve factors. The complete confounding perspective of *GR* of Definition 2 appears to be more sensible than the individual df perspective of GR_{ind} as a primary criterion. However, GR_{ind} provides an interesting new aspect that may provide additional understanding of the structure of OAs and may help in ranking tied designs. The factor wise values of Section 5 add useful detail. It will be interesting to pursue concepts derived from the building blocks of $GR_{\text{tot}(i)}$ and $GR_{\text{ind}(i)}$ for the ranking of mixed level designs. As was demonstrated in Section 5, *GR* from Def. 2 and GR_{ind} from Def. 3 are not the only possible generalizations of (4) for qualitative factors. The alternative given in Equation (6) appears too lenient and has therefore not been pursued. The concept of weak strength deserves further attention: For symmetric designs with weak strength *t* according to Def. 4, Xu (2003, Theorem 3) showed that these have minimum moment aberration (MMA) and consequently GMA (as MMA is equivalent to GMA for symmetric designs) if they also have maximum *k*-balance for $k = t+1,\ldots,n$. In particular, this implies that an OA(N, s^n, t^-) with $N \le s^t$ has GMA, because of Remark 4. Here, we showed that designs of the highest possible resolution *R* maximize *GR* if they have weak strength *R*. It is likely that there are further beneficial consequences from the concept of weak strength.

Appendix A: Proof of Theorem 1

Proof. Let $\mathbf{M}_{\text{C}} = (\mathbf{1}_N\ \mathbf{M}_{1;\text{C}}\ \ldots\ \mathbf{M}_{R-1;\text{C}})$, with $\mathbf{M}_{k;\text{C}}$ the model matrix for all *k*-factor interactions, $k=1,\ldots,R-1$. The assumption that the resolution of the array is *R* and the chosen orthogonal contrasts imply $\mathbf{X}_c^{\text{T}}\mathbf{M}_{k;\text{C}} = \mathbf{0}$ for $k < R-1$, with $\mathbf{X}_c$ as defined in the theorem. Denoting the R–1-factor interaction matrix $\mathbf{M}_{R-1;\text{C}}$ as $\mathbf{X}_{\text{C}}$, the predictions for the columns of $\mathbf{X}_c$ can be written as

$$\hat{\mathbf{X}}_c = \mathbf{X}_{\text{C}}\left(\mathbf{X}_{\text{C}}^{\text{T}}\mathbf{X}_{\text{C}}\right)^{-1}\mathbf{X}_{\text{C}}^{\text{T}}\mathbf{X}_c = \frac{1}{N}\mathbf{X}_{\text{C}}\mathbf{X}_{\text{C}}^{\text{T}}\mathbf{X}_c,$$

since $\mathbf{X}_{\text{C}}^{\text{T}}\mathbf{X}_{\text{C}} = N\,\mathbf{I}_{\text{df(C)}}$. As the column averages of $\hat{\mathbf{X}}_c$ are 0 because of the coding, the nominators for the R^2 values are the diagonal elements of the matrix

$$\hat{\mathbf{X}}_c^{\text{T}}\hat{\mathbf{X}}_c = \frac{1}{N^2}\mathbf{X}_c^{\text{T}}\mathbf{X}_{\text{C}}\mathbf{X}_{\text{C}}^{\text{T}}\mathbf{X}_{\text{C}}\mathbf{X}_{\text{C}}^{\text{T}}\mathbf{X}_c \underset{\mathbf{X}_{\text{C}}^{\text{T}}\mathbf{X}_{\text{C}}=N\mathbf{I}_{\text{df(C)}}}{=} \frac{1}{N}\mathbf{X}_c^{\text{T}}\mathbf{X}_{\text{C}}\mathbf{X}_{\text{C}}^{\text{T}}\mathbf{X}_c.$$

Analogously, the corresponding denominators are the diagonal elements of

$$\mathbf{X}_c^{\text{T}}\mathbf{X}_c = N\mathbf{I}_{\text{df}(c)},$$

which are all identical to *N*. Thus, the sum of the R^2 values is the trace of $\frac{1}{N^2}\mathbf{X}_c^{\text{T}}\mathbf{X}_{\text{C}}\mathbf{X}_{\text{C}}^{\text{T}}\mathbf{X}_c$, which can be written as

$$\text{tr}\left(\frac{1}{N^2}\mathbf{X}_c^{\text{T}}\mathbf{X}_{\text{C}}\mathbf{X}_{\text{C}}^{\text{T}}\mathbf{X}_c\right) = \frac{1}{N^2}\text{vec}\left(\mathbf{X}_{\text{C}}^{\text{T}}\mathbf{X}_c\right)^{\text{T}}\text{vec}\left(\mathbf{X}_{\text{C}}^{\text{T}}\mathbf{X}_c\right), \tag{7}$$

where the vec operator stacks the columns of a matrix on top of each other, i.e. generates a column vector from all elements of a matrix (see e.g. Bernstein 2009 for the rule connecting trace to vec). Now, realize that

$$\mathrm{vec}\left(\mathbf{X}_{\mathrm{C}}{}^{\mathrm{T}}\mathbf{X}_{c}\right)^{\mathrm{T}} = \mathrm{vec}\left(\left(\sum_{i=1}^{N}\mathbf{X}_{\mathrm{C}(i,f)}\mathbf{X}_{c(i,g)}\right)_{(f,g)}\right)^{\mathrm{T}} = \mathbf{1}_{1\times N}\,\mathbf{X}_{u_1\ldots,u_R},$$

where an index pair (i, j) stand for the i-th row and j-th column, respectively, and the columns in $\mathbf{X}_{u_1\ldots,u_R}$ are assumed to appear in the order that corresponds to that in $\mathrm{vec}\left(\mathbf{X}_{\mathrm{C}}{}^{\mathrm{T}}\mathbf{X}_{c}\right)^{\mathrm{T}}$ (w.l.o.g.). Then, (7) becomes

$$\frac{1}{N^2}\mathbf{1}_{1\times N}\,\mathbf{X}_{u_1\ldots,u_R}\mathbf{X}_{u_1\ldots,u_R}{}^{\mathrm{T}}\mathbf{1}_{1\times N}{}^{\mathrm{T}} = a_R(u_1\ldots,u_R),$$

which proves the assertion. ///

References

Ai, M.-Y. and Zhang, R.C. (2004). Projection justification of generalized minimum aberration for asymmetrical fractional factorial designs. *Metrika* **60**, 279-285.

Bernstein, D.S. (2009). *Matrix Mathematics: Theory, Facts, and Formulas with Application to Linear Systems Theory*. 2nd ed. Princeton University Press, Princeton.

Butler,N.A. (2005). Generalised minimum aberration construction results for symmetrical orthogonal arrays. *Biometrika* **92**, 485-491.

Cheng, S.W. and Ye, K.Q. (2004). Geometric Isomorphism and minimum aberration for factorial designs with quantitative factors. *Ann. Statist.* **32**, 2168–2185.

Deng, L.Y. and Tang, B. (1999). Generalized resolution and minimum aberration criteria for Plackett-Burman and other nonregular factorial designs. *Statist. Sinica* **9**, 1071-1082.

Diestelkamp, W.S. and Beder, J.H. (2002). On the decomposition of orthogonal arrays. *Util. Math.* **61**, 65-86.

Eendebak, P. and Schoen, E. (2013). Complete series of non-isomorphic orthogonal arrays. URL: http://pietereendebak.nl/oapage/ Accessed May 24, 2013.

Evangelaras, H., Koukouvinos, C., Dean, A.M. and Dingus, C.A. (2005). Projection properties of certain three level orthogonal arrays. *Metrika* **62**, 241-257.

Fries, A.; Hunter, W.G. (1980). Minimum aberration 2**(k-p) designs. *Technometrics* **22**, 601-608.

Härdle, W. and Simar, L. (2003). *Applied Multivariate Statistical Analysis*. Springer, New York.

Hotelling, H. (1936). Relations Between Two Sets of Variates. *Biometrika* **28**, 321–377.

Lazraq, A. and Cléroux, R. (2001). Statistical Inference Concerning Several Redundancy Indices. *J. Multivariate Anal.* **79**, 71-88.

Mukerjee, R. and Wu, C.F.J. (1995). On the existence of saturated and nearly saturated asymmetrical orthogonal arrays. *Ann. Statist.* **23**, 2102-2115.

Pang, F. and Liu, M. Q. (2010). Indicator function based on complex contrasts and its application in general factorial designs. *J. Statist. Plann. Inference* **140**, 189-197.

Plackett, R.L.; Burman, J.P. (1946). The design of optimum multifactorial experiments. *Biometrika* **33**, 305-325.

Stewart, D. and Love, W. (1968). A general canonical correlation index. *Psychological Bulletin* **70**, 160-163.

Tang, B. (2001), Theory of *J*-characteristics for fractional factorial designs and projection justification of minimum G_2 aberration. *Biometrika* **88**, 401-407.

Tang, B. and Deng, L.Y. (1999). Minimum G_2-aberration for nonregular fractional factorial designs. *Ann. Statist.* **27**, 1914-1926.

Xu, H. (2003). Minimum moment aberration for non-regular designs and supersaturated designs. *Statist. Sinica* **13**, 691-708.

Xu, H., Cheng, S.-W. and Wu, C.F.J. (2004). Optimal Projective Three-Level Designs for Factor Screening and Interaction Detection. *Technometrics* **46**, 280-292.

Xu, H. and Wu, C.F.J. (2001). Generalized Minimum Aberration for Asymmetrical Fractional Factorial Designs. *Ann. Statist.* **29**, 1066-1077.

Ulrike Grömping,
Department II – Mathematics, Physics, Chemistry,
Beuth University of Applied Sciences Berlin,
Luxemburger Str. 10,
13353 Berlin,
Germany.
E-mail: groemping@bht-berlin.de.
URL: http://prof.beuth-hochschule.de/groemping

Hongquan Xu,
UCLA, Department of Statistics
8125 Math Sciences Bldg
Box 951554
Los Angeles
CA 90095-1554
USA.
E-mail: hqxu@stat.ucla.edu.
URL: http://www.stat.ucla.edu/~hqxu/.