Journal of Chromatography B, 877 (2009) 2252–2261

Contents lists available at ScienceDirect

Journal of Chromatography B

journal homepage: www.elsevier.com/locate/chromb

Identification of significant effects from an experimental screening design in the absence of effect sparsity☆

B. Dejaegher, A. Durand, Y. Vander Heyden*

Analytical Chemistry and Pharmaceutical Technology (FABI), Pharmaceutical Institute, Vrije Universiteit Brussel – VUB, Laarbeeklaan 103, 1090 Brussels, Belgium

ARTICLE INFO

Article history:
Received 7 August 2008
Accepted 3 October 2008
Available online 18 October 2008

Keywords:
Screening design
Effects significance
Algorithm of Dong
Effect sparsity

ABSTRACT

This paper describes an attempt to derive a new methodology to determine the significance of effects estimated from a screening design, starting from the algorithm of Dong but overcoming its drawbacks. Especially in situations where effect sparsity does not occur and the number of significant effects approaches 50%, the currently often applied algorithm of Dong leads to many important effects incorrectly considered non-significant, i.e. to false negative results. For these situations, a new methodology is recommended. Based on the algorithm of Dong, several alternative approaches were explored and compared. From all approaches, the one using the 75% lowest absolute factor effects to calculate the initial error estimation s_0, i.e. $s_0 = 1.5 \times \text{median}|E_{75\%}|$, resulted in the highest number of correct decisions on effects significance. After its definition, the new methodology was tested on a bioanalysis application data set. This study confirmed the earlier conclusions on literature and semisimulated data. The new methodology is especially interesting to be applied in minimal screening designs, for which other error estimates (e.g. based on interaction or dummy effects) cannot be applied.

1. Introduction

Extensive method validation is often required, especially in pharmaceutical industries, in order to meet the strict regulations set by the regulatory bodies. Robustness testing is part of such validation. The ICH (International Conference on Harmonisation of Technical Requirements for the Registration of Pharmaceuticals for Human Use) guidelines define robustness as: "*The robustness of an analytical procedure is a measure of its capacity to remain unaffected by small, but deliberate variations in method parameters and provides an indication of its reliability during normal usage*."[1].

In the evaluation of method robustness, two-level screening designs, such as fractional factorial (FF) or Plackett-Burman (PB) designs, are usually applied [2,3]. These designs are also often used to identify the most important factors during the first phase of method optimization [4]. They allow screening a relatively high number of factors f in a rather small number of experiments ($N \geq f + 1$) [2,3]. FF designs perform only a fraction of a full factorial design. The number of experiments is a power of two, and depending on the design, two-factor interaction effects can or cannot be estimated unconfounded from the main effects [2,3]. PB designs are factorial designs that examine up to N – 1 factors in N (multiple of four) experiments [2,3,5]. When less than N – 1 factors are to be examined, the remaining PB design columns are defined as dummy factor columns [3].

In robustness testing, two-factor interaction or dummy factor effects estimated from FF and PB designs, respectively, are assumed to be negligible, and can therefore be used in the statistical interpretation of the estimated factor effects (see further) [3]. In the case of screening during optimization, the negligibility cannot be assumed anymore and these effects should not be used to draw decisions on the factor effects significance.

To analyze screening design results, first the factor effects are estimated [2,3]. Afterwards usually an approach is selected to identify the important or significant effects. Both graphical, such as normal [2,3,6] or half-normal [2,3,7] probability plots, and statistical methods can be used. Most statistical methods use the t-test, where the t statistic requires an estimation of error. The error estimation can be made in several ways, e.g. using the variance of replicated experiments (e.g. at center point level, or duplicated design experiments) [2,3], from a priori declared negligible effects (interaction or dummy effects) [2,3], or from effects defined a posteriori negligible by the algorithms of Lenth [8] or Dong [3,9].

Two types of erroneous decisions can occur. A false positive result is obtained when a non-significant effect is considered significant, and a false negative result when a significant effect is

☆ This paper is part of a special issue entitled "Method Validation, Comparison and Transfer", guest edited by Serge Rudaz and Philippe Hubert.

* Corresponding author. Tel.: +32 2 477 47 34; fax: +32 2 477 47 35.
E-mail address: yvanvdh@vub.ac.be (Y. Vander Heyden).

1570-0232/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jchromb.2008.10.019

indicated as being non-significant. The latter is considered worst, seen from a practical point of view.

Needless to say, it is frequently desired to reduce time and costs when performing a screening or a robustness test. For this purpose, often minimal designs are applied, in which the factors are examined in the smallest design possible. These designs do not contain enough interactions or dummy factors unconfounded with the main effects, to estimate the error properly. The use of the algorithm of Dong is then recommended [3]. However, the algorithms of Lenth and Dong require effect sparsity, i.e. ≪50% significant effects. In situations where about 50% of the effects are significant, the algorithm of Dong becomes incapable to indicate the significant effects correctly [10]. The algorithm of Dong then is overestimating the critical effect, resulting in a situation where significant effects incorrectly are considered non-significant. In these situations, it is advised to perform a design that allows estimating the effects of enough a priori declared negligible terms to estimate the error or the critical effect [10]. However, a screening design that complies to this condition might still require rather many experiments, especially when a high number of factors needs to be evaluated.

Therefore, in this study, it was tried to adapt the algorithm of Dong, in order to obtain a more accurate estimation of the critical effect in all situations, especially in those with about 50% significant effects. Nine data sets or designs were taken from literature [7,11–17], with different numbers of examined factors, responses, and design experiments (N = 8, 12, 16, 24). The designs were either FF or PB designs. Eleven cases represented situations where about 50% of the considered effects are significant. These cases were either taken from the literature, or partially simulated starting from the literature data. Different approaches to adapt the algorithm of Dong were considered. A new approach was proposed to obtain a more accurate estimation of the critical effect. This new approach was then applied to a bioanalysis data set, taken from literature [18], containing the robustness study results of a high-performance liquid chromatographic (HPLC) method to analyze three anti-epileptic drugs, pregabalin, gabapentin, and vigabatrin, in human serum.

2. Theory

Factor effects in two-level screening designs are calculated as follows [2,3]

$$E_X = \frac{\sum Y(+) - \sum Y(-)}{N/2} \quad (1)$$

where E_X is the effect of factor X, $\sum Y(+)$ and $\sum Y(-)$ are the sums of the responses where factor X is at (+) or (−) level, respectively, and N is the number of design experiments.

In the statistical interpretation, a t-test statistic [2,3] is calculated to evaluate whether a given E_X is significantly different from zero, i.e. factor X has a significant effect.

$$t = \frac{|E_X|}{(SE)_e} \Leftrightarrow t_{tab} \quad (2)$$

where $(SE)_e$ is the standard error of the effect.

The calculated test statistic (Eq. (2)) is compared with a tabulated t-value, t_{tab}. The number of degrees of freedom (df) for $(SE)_e$ and the applied significance level α will determine t_{tab}. An effect with a $t \geq t_{tab}$ is considered significant, while $t < t_{tab}$ suggests a non-significant effect.

The algorithm of Dong [3,9] is an approach to estimate the error $(SE)_e$. This algorithm calculates from an initial estimate of error based on all effects, s_0 (Eq. (3)), a final estimation of the standard error, s_1 (Eq. (4)), based on the effects that are considered not important. The estimated s_1 allows determining a critical effect for a response, E_{crit}, called the Margin of Error (ME) (Eq. (5)).

$$s_0 = 1.5 \times \text{median}|E_i| \quad (3)$$

$$s_1 = \sqrt{m^{-1} \sum E_j^2} \quad (4)$$

where E_i is the effect of factor i, E_j an effect that in absolute value is smaller than or equal to $2.5^* s_0$, and m the number of such effects. The elimination of effects exceeding $2.5^* s_0$ is derived from $P(|E_X| > 2.5^* s_0) \approx 0.01$.

$$E_{crit} = ME = t_{(1-(\alpha/2),\, df)} \times s_1 \Leftrightarrow |E_X| \quad (5)$$

In Eq. (5), df = m and usually α = 0.05. Values of $|E_X|$ that are larger than or equal to ME are considered significant. For the above-mentioned minimal designs, the algorithm of Dong is appropriate to determine significant effects, except in situations where about 50% of the considered effects is significant [3,10]. The algorithm of Dong is further in the text indicated as approach D.

Because of the above drawback, in this study, several approaches were considered to adapt the algorithm of Dong. The *first approach* does not take into account s_0 to determine the number of effects m to include in the estimation of s_1. The approach simply consists of calculating the Margins of Error for all different numbers of effects, from 1 to i, included in the estimation of s_1. The consecutively included effects are those obtained after sorting and starting with the absolute smallest. Then, the MEs are plotted as a function of the number of included effects. In this approach, the lowest ME at a given significance level α is then considered the critical effect. This approach is indicated as L further on.

The *second approach* also does not calculate s_0 to determine the number of effects m to include in the estimation of s_1. Here a half-normal probability or Birnbaun plot is drawn and visually evaluated to decide on the number of effects, m_B, to include in the estimation of s_1. More information concerning the construction and interpretation of these plots can be found in [2,3,7]. Non-significant effects tend to fall on a straight line through zero, whereas significant effects deviate from this straight line. In this approach, m in Eq. (4) is replaced by m_B, and E_j by E_B, where E_B represent the m_B lowest absolute effects, which were considered non-significant from the plot, thus $s_1 = \sqrt{m_B^{-1} \sum E_B^2}$. The corresponding ME is then calculated with Eq. (5), where df = m_B. This approach is indicated as B further on.

The *third approach*, indicated as P, replaces E_j in Eq. (4) by E_P, and m by m_P, where E_P is an effect that in absolute value is smaller than or equal to $2.0^* s_0$, and m_P the number of such effects, thus $s_1 = \sqrt{m_P^{-1} \sum E_P^2}$. In this approach, the elimination of effects exceeding $2.0^* s_0$ is derived from $P(|E_X| > 2.0^* s_0) \approx 0.05$. The corresponding ME is then calculated with Eq. (5), where df = m_P.

Whereas the first three approaches focus on adapting the final estimation of error s_1, a *fourth approach* reconsiders the initial estimate of error s_0. To estimate s_0, the algorithm of Dong uses the median of all absolute effects. In this fourth approach, the median is arbitrarily calculated from the 95%, 90%, 85%, 80%, or 75% lowest absolute effects. In Eq. (3), E_i is then replaced by $E_{\%}$, which represents the considered percentage of lowest absolute effects, thus $s_0 = 1.5 \times \text{median}|E_{\%}|$. This approach is indicated as 95, 90, 85, 80 or 75%, depending on the percentage included effects. In Table 1, the numbers of included and excluded effects for s_0 are given for different designs and for different percentages of lowest absolute effects. The corresponding s_1 estimation is then made with Eq. (4), and the corresponding ME with Eq. (5).

The results of these four approaches were evaluated critically for a number of case studies, and are discussed further.

Table 1
Numbers of included (# included) and of excluded (# excluded) effects for s_0, considering different designs and percentages of lowest absolute effects (95%-90%-85%-80%-75%). N = number of design experiments.

N	%	# included	# excluded
8	95-90-85-80	6	1
	75	5	2
12	95-90	10	1
	85-80	9	2
	75	8	3
16	95-90	14	1
	85	13	2
	80	12	3
	75	11	4
24	95	22	1
	90	21	2
	85	20	3
	80	18	5
	75	17	6

3. Experimental

3.1. Data sets to test the new approaches

The nine data sets applied in [10] were also used here in order to study the above approaches to determine the effects significance from a screening design. These nine data sets were taken from literature [7,11–17] and resulted from the use of screening designs during robustness testing (data sets 1 till 7) or method optimization (data sets 8 and 9). The numbers of design experiments were 8 (data sets 1 till 4), 12 (data sets 5 and 6), 16 (data sets 7 and 8) or 24 (data set 9). Both fractional factorial (data sets 1, 2, 4, and 8) and Plackett-Burman (data sets 3, 5, 6, 7, and 9) designs were considered. Table 2 gives an overview of the nine data sets. The numbers of design experiments (N), of examined factors (f), of interactions (I) (estimated unconfounded with the main effects) or dummy (d) effects, and of responses (r) are given for each data set.

Amongst these $\sum r = 34$ responses, eleven represent situations where about 50% of the effects are significant (Table 3). These cases were either taken from the literature or partially simulated. For data sets 1 and 3 (both N = 8), one response with $3/7 \times 100 = 43\%$ significant effects, i.e. responses 1 and 7, respectively, was retained and evaluated. For data set 6 (N = 12), three responses (21–23) were retained, each with $5/11 \times 100 = 45\%$ significant effects. For data sets 4 (N = 8), 7 (N = 16), and 9 (N = 24), such responses were partially simulated [10]. For data set 4 (N = 8), two partially simulated responses (12 and 13) were created in which $3/7 \times 100 = 43\%$ and $4/7 \times 100 = 57\%$ significant effects, respectively, occur. For data set 7 (N = 16), two partially simulated responses (28 and 29) were created with $6/15 \times 100 = 40\%$ and $8/15 \times 100 = 53\%$ significant effects, respectively. For data set 9 (N = 24), two partially simulated situations (33 and 34) were created with $7/23 \times 100 = 30\%$ and $12/23 \times 100 = 52\%$ significant factor effects, respectively.

The above numbers of significant effects, that were used as reference further in the text, were obtained from a graphical (data sets 1 till 9) and, where possible, an alternative statistical (data sets 4 till 9) interpretation of the data. Graphically, a half-normal probability plot was drawn, whereas statistically critical effects were calculated based on a priori considered negligible effects, i.e. interactions or dummies for FF or PB designs, respectively. More explanation and the results of both reference approaches (R) can be found in [10] and Table 3. In case these two methods lead to a different result, the result from the negligible effects was used.

For the exact design set-ups of the data sets, we refer to [10], where for each data set, the factors, the experimental design, the responses, the effects on each response, and the critical effects, estimated from the algorithm of Dong, are presented. For original information regarding the nine data sets, we refer to [7,11–17].

The ratio $E_{crit}/E_{crit,D}$ in Table 3 allows situating the critical effect of a given approach relative to that obtained by the algorithm of Dong. When $E_{crit}/E_{crit,D} < 1$ the alternative approach uses a critical effect that is smaller than that used by Dong's approach, i.e. some larger effects were excluded from the error estimation.

3.2. Application data set

The best new approach to determine the effects significance from a screening design set-up was subsequently applied to analyze the results of an application data set [18].

The robustness of a high-performance liquid chromatographic method to analyze three anti-epileptic drugs, pregabalin, gabapentin, and vigabatrin, in human serum, was tested in [18]. First the robustness of the chromatographic analysis was evaluated, and secondly that of the sample preparation procedure. For our study, the first part of the robustness study was retained. The robustness of the chromatographic analysis was performed using a 12-experiments Plackett-Burman design. The effects of six factors on several responses were evaluated. The retained responses for our study were the % relative standard deviations (%RSD) of the vigabatrin ($\%RSD_v$) and gabapentin ($\%RSD_g$) concentrations, the resolution between the vigabatrin peak and its nearest neighbour peak (Rs_v), the asymmetry factor of the gabapentin peak (Asf_g), and the plate number of gabapentin ($(N/mm)_g$).

Table 2
Overview of the data sets: the origin with the applied experimental design, the number of experiments (N), the number of examined factors (f), the number of interactions (I) (estimated unconfounded with the main effects) or dummy effects (d), and the number of responses (r).

Data set	Origin of case study	Design	N	f	I	d	r
1	Robustness test of chemical filtration process in an industrial plant [7]	2^{7-4} fractional factorial	8	7	0	–	1
2	Robustness test of HPLC assay for triadimenol [11]	2^{6-3} fractional factorial	8	6	1	–	3
3	Robustness test of HPLC assay for ketoconazole in antidandruff shampoo [12]	7-factors 8-experiments PB	8	6	–	1	3
4	Robustness test of HPLC assay for tetracycline and its impurities [13]	2^{4-1} fractional factorial	8	4	3	-	6
5	Robustness test of CE assay to separate enantiomers of praziquantel and warfarin [14]	11-factors 12-experiments PB	12	8	-	3	3
6	Robustness test of HPLC assay for a drug substance [15]	11-factors 12-experiments PB	12	6	–	5	7
7	Robustness test of HPLC assay for formaldehyde in antidandruff shampoo [12]	15-factors 16-experiments PB	16	11	–	4	6
8	Optimization of FIA assay for compounds with a secondary amine or an amide function [16]	2^{6-2} fractional factorial	16	6	9	–	2
9	Optimization of FIA assay for fluticasone propionate [17]	23-factors 24-experiments PB	24	8	–	15	3

Table 3
Evaluation of different approaches to determine effects significance; (#) number of factors included in the estimation of s_1, ($E_{crit}/E_{crit,D}$) the critical effect relative to that of Dong, both at $\alpha = 0.05$, and ($E_{\alpha=0.05}$) the effects considered significant at $\alpha = 0.05$. For the meaning of R, D, L, B, P, 95, 90, 85, 80, and 75%, see text. The results are presented for each data set (*a*–*i*) and each response. * responses with about 50% significant effects and (. . .) their number.

(a) Data set 1

Response	*Approach*	*#*	$E_{crit}/E_{crit,D}$	$E_{\alpha=0.05}$
1(3)*	R	/	/	F/A/C
	D	5	1	F/A
	L	4	0.54	F/A/C
	B	4	0.54	F/A/C
	P	5	1	F/A
	95-90-85-80%	5	1	F/A
	75%	4	0.54	F/A/C

(b) Data set 2

Response	*Approach*	*#*	$E_{crit}/E_{crit,D}$	$E_{\alpha=0.05}$
2	R	/	/	F
	D	6	1	F
	L	2	0.43	F/I_1/C
	B	6	1	F
	P	6	1	F
	95-90-85-80%	6	1	F
	75%	6	1	F
3	R	/	/	F/B
	D	5	1	F/B
	L	3	0.29	F/B/C
	B	5	1	F/B
	P	4	0.42	F/B/C
	95-90-85-80%	4	0.42	F/B/C
	75%	4	0.42	F/B/C
4	R	/	/	E
	D	6	1	E
	L	3	0.68	E/A
	B	6	1	E
	P	6	1	E
	95-90-85-80%	6	1	E
	75%	6	1	E

(c) Data set 3

Response	*Approach*	*#*	$E_{crit}/E_{crit,D}$	$E_{\alpha=0.05}$
5	R	/	/	Nothing
	D	7	1	Nothing
	L	4	0.87	Nothing
	B	7	1	Nothing
	P	7	1	Nothing
	95-90-85-80%	7	1	Nothing
	75%	7	1	Nothing
6	R	/	/	C
	D	6	1	C
	L	2	0.53	C/E/F
	B	6	1	C
	P	6	1	C
	95-90-85-80%	6	1	C
	75%	6	1	C
7(3)*	R	/	/	B/C/A
	D	5	1	B
	L	3	0.34	B/C/A
	B	4	0.40	B/C/A
	P	4	0.40	B/C/A
	95-90-85-80%	4	0.40	B/C/A
	75%	4	0.40	B/C/A

(d) Data set 4

Response	*Approach*	*#*	$E_{crit}/E_{crit,D}$	$E_{\alpha=0.05}$
8	R	/	1.52	B/C
	D	5	1	B/C

Table 3 (*Continued*)

(d) Data set 4

Response	*Approach*	*#*	$E_{crit}/E_{crit,D}$	$E_{\alpha=0.05}$
	L	3	0.65	B/C
	B	5	1	B/C
	P	5	1	B/C
	95-90-85-80%	5	1	B/C
	75%	5	1	B/C
9	R	/	1.21	B/C
	D	5	1	B/C
	L	3	0.94	B/C
	B	5	1	B/C
	P	5	1	B/C
	95-90-85-80%	5	1	B/C
	75%	5	1	B/C
10	R	/	1.73	Nothing
	D	7	1	Nothing
	L	4	0.76	Nothing
	B	7	1	Nothing
	P	7	1	Nothing
	95-90-85-80%	7	1	Nothing
	75%	7	1	Nothing
11	R	/	1.09	C
	D	6	1	C
	L	2	0.39	C/B/I_3/A
	B	6	1	C
	P	6	1	C
	95-90-85-80%	6	1	C
	75%	6	1	C
12(3)*	R	/	1.30	B/A/C
	D	4	1	B/A/C
	L	2	0.74	B/A/C
	B	4	1	B/A/C
	P	4	1	B/A/C
	95-90-85-80%	4	1	B/A/C
	75%	4	1	B/A/C
13(4)*	R	/	0.30	B/A/C/D
	D	7	1	Nothing
	L	3	0.30	B/A/C/D
	B	3	0.30	B/A/C/D
	P	7	1	Nothing
	95-90-85-80%	7	1	Nothing
	75%	3	0.30	B/A/C/D

(e) Data set 5

Response	*Approach*	*#*	$E_{crit}/E_{crit,D}$	$E_{\alpha=0.05}$
14	R	/	0.71	H/D
	D	11	1	Nothing
	L	3	0.36	H/D/F/d_1/A
	B	9	0.63	H/D
	P	10	0.85	H/D
	95-90%	11	1	Nothing
	85-80%	11	1	Nothing
	75%	11	1	Nothing
15	R	/	1.21	B
	D	10	1	B
	L	3	0.36	B/A/C/D/d_1/E/d_2
	B	10	1	B
	P	10	1	B
	95-90%	10	1	B
	85-80%	10	1	B
	75%	10	1	B
16	R	/	0.95	D
	D	11	1	D
	L	2	0.29	D/E/C/A/F/d_3/d_2/d_1/B
	B	10	0.75	D
	P	10	0.75	D
	95-90%	11	1	D
	85-80%	11	1	D
	75%	11	1	D

Table 3 (*Continued*)

(f) Data set 6

Response	*Approach*	*#*	$E_{crit}/E_{crit,D}$	$E_{\alpha=0.05}$
17	R	/	0.81	F/D
	D	9	1	F/D
	L	2	0.11	F/D/B/A/d_3/E/d_2/d_5
	B	10	1.36	F/D
	P	9	1	F/D
	95-90%	9	1	F/D
	85-80%	9	1	F/D
	75%	9	1	F/D
18	R	/	0.82	F
	D	10	1	F
	L	4	0.15	F/B/D/C/d_4/d_5
	B	10	1	F
	P	10	1	F
	95-90%	10	1	F
	85-80%	6	0.37	F/B/D/C/d4
	75%	6	0.37	F/B/D/C/d4
19	R	/	0.88	F/C
	D	10	1	F
	L	3	0.51	F/C/D
	B	10	1	F
	P	10	1	F
	95-90%	10	1	F
	85-80%	10	1	F
	75%	10	1	F
20	R	/	1.10	F
	D	10	1	F
	L	2	0.21	F/A/d_3/E/d_2/d_4/C
	B	10	1	F
	P	9	0.81	F/A
	95-90%	10	1	F
	85-80%	9	0.81	F/A
	75%	8	0.58	F/A/d_3
21(5)*	R	/	0.95	A/F/D/C/B
	D	6	1	A/F/D/C/B
	L	1	0.03	A/F/D/C/B/d_4/E/d_3/d_1/d_2
	B	6	1	A/F/D/C/B
	P	6	1	A/F/D/C/B
	95-90%	6	1	A/F/D/C/B
	85-80%	6	1	A/F/D/C/B
	75%	6	1	A/F/D/C/B
22(5)*	R	/	0.35	A/C/D/B/F
	D	9	1	A
	L	3	0.09	A/C/D/B/F/d_4/d_5/d_1
	B	6	0.30	A/C/D/B/F
	P	7	0.47	A/C/D/B
	95-90%	7	0.47	A/C/D/B
	85-80%	7	0.47	A/C/D/B
	75%	7	0.47	A/C/D/B
23(5)*	R	/	0.40	E/D/F/B/C
	D	10	1	E
	L	3	0.14	E/D/F/B/C/d_4
	B (1)	10	1	E
	B (2)	5	0.19	E/D/F/B/C/d_4
	P	10	1	E
	95-90%	9	0.83	E/D
	85-80%	7	0.51	E/D/F/B
	75%	6	0.35	E/D/F/B/C

(g) Data set 7

Response	*Approach*	*#*	$E_{crit}/E_{crit,D}$	$E_{\alpha=0.05}$
24	R	/	1.13	Nothing
	D	15	1	Nothing
	L	2	0.14	C/E/d_4/I/A/F/D/B/H/d_3/d_1
	B	15	1	Nothing
	P	15	1	Nothing
	95-90%	15	1	Nothing
	85%	15	1	Nothing
	80%	15	1	Nothing
	75%	15	1	Nothing

Table 3 (*Continued*)

(g) Data set 7

Response	*Approach*	*#*	$E_{crit}/E_{crit,D}$	$E_{\alpha=0.05}$
25	R	/	0.99	C/E/B
	D	12	1	C/E/B
	L	4	0.46	C/E/B/G/J/F/H
	B	12	1	C/E/B
	P	12	1	C/E/B
	95-90%	12	1	C/E/B
	85%	12	1	C/E/B
	80%	12	1	C/E/B
	75%	12	1	C/E/B
26	R	/	1.10	C/E
	D	14	1	C/E
	L	2	0.32	C/E/d_2/B/H/d_3/A/J
	B	13	0.75	C/E
	P	13	0.75	C/E
	95-90%	14	1	C/E
	85%	13	0.75	C/E
	80%	13	0.75	C/E
	75%	13	0.75	C/E
27	R	/	1.43	C
	D	14	1	C
	L	3	0.57	C/E/d_2
	B	14	1	C
	P	14	1	C
	95-90%	14	1	C
	85%	14	1	C
	80%	14	1	C
	75%	14	1	C
28(6)*	R	/	0.63	A/D/C/H/E/J
	D	13	1	A/D
	L	2	0.18	A/D/C/H/E/J/d_2/B/d_3
	B	9	0.43	A/D/C/H/E/J
	P	11	0.75	A/D/C/H
	95-90%	13	1	A/D
	85%	12	0.88	A/D/C
	80%	11	0.75	A/D/C/H
	75%	9	0.43	A/D/C/H/E/J
29(8)*	R	/	0.40	F/C/A/H/E/J/D/G
	D	15	1	Nothing
	L	2	0.24	F/C/A/H/E/J/D/G
	B	7	0.29	F/C/A/H/E/J/D/G
	P	15	1	Nothing
	95-90%	15	1	Nothing
	85%	10	0.75	Nothing
	80%	9	0.66	F/C
	75%	8	0.52	F/C/A/H/E/J/D/G

(h) Data set 8

Response	*Approach*	*#*	$E_{crit}/E_{crit,D}$	$E_{\alpha=0.05}$
30	R	/	0.86	A
	D	15	1	Nothing
	L	5	0.19	A/E/I_5/B/I_9/I_1/D/I_6
	B (1)	15	1	Nothing
	B (2)	9	0.38	A/E/I_5/B/I_9/I_1
	P	14	0.89	A
	95-90%	14	0.89	A
	85%	12	0.71	A/E/I_5
	80%	12	0.71	A/E/I_5
	75%	12	0.71	A/E/I_5
31	R	/	0.87	D/E
	D	14	1	D
	L	2	0.18	D/E/C/I_6/B/I_1/I_9/I_4/I_3/I_7/I_8/A
	B	14	1	D
	P	14	1	D
	95-90%	14	1	D
	85%	14	1	D
	80%	14	1	D
	75%	14	1	D

Table 3 (*Continued*)

(i) Data set 9

Response	*Approach*	*#*	$E_{crit}/E_{crit,D}$	$E_{\alpha=0.05}$
32	R	/	0.86	C/G/A/d_2
	D	21	1	C/G/A
	L	2	0.29	C/G/A/d_2/B/d_7/H/d_{12}/E/d_6/d_4/d_{13}
	B	21	1	C/G/A
	P	20	0.91	C/G/A/d_2
	95%	21	1	C/G/A
	90%	21	1	C/G/A
	85%	21	1	C/G/A
	80%	19	0.82	C/G/A/d_2/B
	75%	18	0.74	C/G/A/d_2/B/d_7
33(7)*	R	/	0.63	C/A/B/G/H/J/I
	D	18	1	C/A/B/G/H/J/I
	L	2	0.31	C/A/B/G/H/J/I/d_{12}/E/d_6/d_4/d_{13}
	B	16	0.62	C/A/B/G/H/J/I
	P	17	0.79	C/A/B/G/H/J/I
	95%	17	0.79	C/A/B/G/H/J/I
	90%	17	0.79	C/A/B/G/H/J/I
	85%	17	0.79	C/A/B/G/H/J/I
	80%	16	0.62	C/A/B/G/H/J/I
	75%	16	0.62	C/A/B/G/H/J/I
34(12)*	R	/	0.20	C/A/B/F/K/E/D/G/H/J/L/I
	D	23	1	Nothing
	L	4	0.13	C/A/B/F/K/E/D/G/H/J/L/I/d_6
	B	11	0.21	C/A/B/F/K/E/D/G/H/J/L/I
	P	23	1	Nothing
	95%	22	0.93	C
	90%	15	0.52	C/A/B/F/K/E/D/G
	85%	15	0.52	C/A/B/F/K/E/D/G
	80%	14	0.46	C/A/B/F/K/E/D/G/H/J
	75%	14	0.46	C/A/B/F/K/E/D/G/H/J

Since no responses occurred with about 50% significant effects, two such situations (R_6 and R_7) were created, by introducing 5 and 6 significant effects on responses (Asf_g) and ($\%RSD_g$), respectively. These responses R_6 and R_7 with $5/11 \times 100 = 45\%$ and $6/11 \times 100 = 55\%$ significant effects, respectively, were then also considered. To create these responses, some dummy effects were replaced by significant effects. It also should be noticed that the resulting responses have no physical meaning anymore.

In Table 4, the experimental design, the considered responses, and the effects are given for this data set.

4. Results and discussion

4.1. Evaluation of the different approaches

In Table 3, for each data set (1 till 9) and each response, the results are presented for the reference criteria (R), the algorithm of Dong (D), and the alternative approaches (L, B, P, 95-90-85-80-75%). For each approach, the number of factors included in the estimation of s_1, the critical effect relative to that of Dong ($E_{crit}/E_{crit,D}$) both at $\alpha = 0.05$, and the effects considered significant at $\alpha = 0.05$ ($E_{\alpha=0.05}$), are given.

The *algorithm of Dong (D)* usually works fine in cases where the effect sparsity principle is fulfilled. For responses 2 till 6, 8 till 12, 15 till 18, 20 till 21, 24 till 27, and 33, the same conclusions on effects significance are made as when using the reference approaches (half-normal probability plot and/or the negligible effects) (Table 5). Compared to the reference criteria, false negative results are obtained for responses 14 ($2/11 \times 100 = 18\%$ significant effects), 19 ($2/11 \times 100 = 18\%$ significant effects), 30 ($1/15 \times 100 = 6.7\%$ significant effects), 31 ($2/15 \times 100 = 13\%$ significant effects), and 32 ($4/23 \times 100 = 17\%$ significant effects). In these cases, the algorithm of Dong already seems to slightly overestimate the experimental error, leading to a somewhat higher $E_{crit,D}$, compared to $E_{crit,R}$ (Tables 3 and 5).

In 8 out of the 11 cases where the effect sparsity principle was violated and where thus about 50% of the effects are significant, the algorithm of Dong fails in determining the significant effects correctly. This is the case for responses 1, 7, 13, 22, 23, 28, 29, and 34. In these situations, many false negative results are obtained relative to the reference criteria (Table 5).

For all 34 responses, 45 effects less than with the reference(s) are indicated as significant. This is due to the fact that Dong overestimates the experimental error, leading to a too high $E_{crit,D}$, because a number of significant effects are included in the error estimate (Tables 3 and 5, $E_{crit,D} > E_{crit,R}$).

For the *first alternative approach (L)*, the plots of ME as a function of the number of included effects (m) in the estimation of s_1 usually have a profile similar to Fig. 1, i.e. a profile that shows a minimum. Only in one situation, i.e. response 21, the plot does not show a minimum, but ME continuously increases with the number of included effects. This is probably caused by the fact that the smallest effect ($|E_X| = 0.00008$) is much smaller than all other (range $|E_X| = 0.004–0.411$).

For designs with N = 8, this approach was found appropriate in those situations where the number of significant effects approaches 50%. For example, all significant effects on responses 1, 7, 12, and 13 are correctly found, while the algorithm of Dong only indicated them correctly for response 12 (Table 3). However, for situations where the number of significant effects is far below 50%, this approach leads to a relatively high number of false positive results. This is, for example, the case, amongst others, for response 11, where besides one significant effect, three non-important effects are incorrectly indicated as significant.

For larger designs (N = 12, 16, 24), this approach also leads to many false positive results for situations where the number of significant effects is low. This is, for example, the case for responses 15, 24, and 32, amongst others. In situations where the effect sparsity principle is violated, this approach indicates the significant effects correctly. For example, all significant effects on responses 21, 22, 23, 28, 29, 33 and 34 were found important, while Dong only indicates them correctly for responses 21 and 33. However, additional false positive results are found in these situations. For example, for responses 21, 22, 23, 28, 33, and 34, some non-important effects are also considered significant (Table 3).

The large number of false positive results when using this approach can be explained by the fact that often the lowest ME, which is used here, seriously underestimated the error on an effect (see Table 3), leading to many non-significant factors that are incorrectly considered significant. It can be concluded that considering

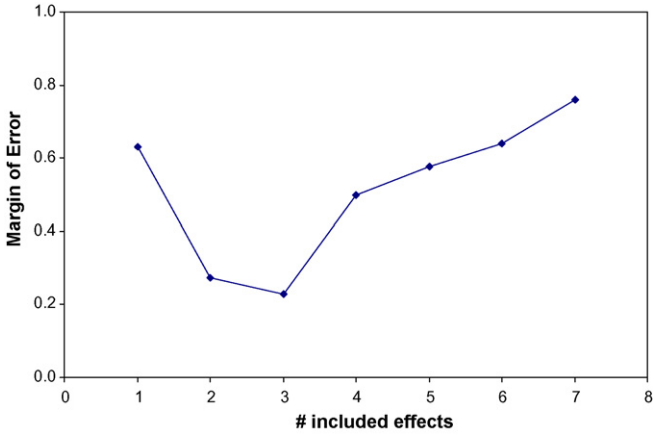

Fig. 1. Margin of Error ($ME_{\alpha=0.05}$) plotted as a function of the number (#) of included effects in the estimation of s_1 for response 13.

Table 4
Application data set: 12-experiments Plackett-Burman design to evaluate the effects of six real factors (A till F) and five dummies (d_1 till d_5), on seven responses, i.e. the % relative standard deviations (%RSD) of vigabatrin ($\%RSD_v$) and gabapentin ($\%RSD_g$) concentrations, the resolution between the vigabatrin peak and its nearest neighbour peak (Rs_v), the asymmetry factor of the gabapentin peak (Asf_g), the plate number of gabapentin ($(N/mm)_g$), and the simulated responses R_6 and R_7.

Exp	*Factors*											*Responses*						
	A	*B*	*C*	*D*	*E*	*F*	$G=d_1$	$H=d_2$	$I=d_3$	$J=d_4$	$K=d_5$	$\%RSD_v$	Rs_v	$\%RSD_g$	Asf_g	$(N/mm)_g$	R_6	R_7
1	1	1	−1	1	1	1	−1	−1	−1	1	−1	7.0	1.23	0.8	1.54	59.6	1.50	−0.7
2	1	−1	1	1	−1	1	1	1	−1	−1	−1	0.8	1.14	0.3	1.52	68.8	1.56	−1.2
3	−1	−1	1	−1	1	1	−1	1	1	1	−1	1.0	1.33	1.0	1.41	71.8	1.37	1.5
4	1	−1	−1	−1	1	−1	1	1	−1	1	1	0.6	1.51	0.9	1.42	63.4	1.42	−1.6
5	1	1	1	−1	−1	−1	1	−1	1	1	−1	1.8	1.22	1.3	1.56	63.7	1.60	0.8
6	1	−1	1	1	1	−1	−1	−1	1	−1	1	1.4	1.12	1.6	1.38	75.7	1.34	2.1
7	−1	1	1	−1	1	1	1	−1	−1	−1	1	1.5	1.92	0.8	1.41	68.0	1.41	−0.7
8	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	1.5	1.99	0.5	1.41	77.0	1.41	−1.0
9	−1	−1	−1	1	−1	1	1	−1	1	1	1	0.6	1.26	0.8	1.42	83.1	1.46	2.3
10	−1	1	1	1	−1	−1	−1	1	−1	1	1	1.3	1.38	4.0	1.44	73.9	1.44	6.5
11	−1	1	−1	1	1	−1	1	1	1	−1	−1	5.0	1.02	6.9	1.48	64.7	1.48	11.4
12	1	1	−1	−1	−1	1	−1	1	1	−1	1	3.7	2.06	2.5	1.48	64.6	1.48	2.0

Responses	*Effects*										
$\%RSD_v$	0.73	2.40	−1.77	1.00	1.13	0.50	−0.93	−0.23	0.13	−0.27	−1.33
Rs_v	−0.10	0.08	−0.16	−0.48	−0.15	0.12	−0.17	−0.05	−0.19	−0.22	0.22
$\%RSD_g$	−1.10	1.87	−0.57	1.23	0.43	−1.50	0.10	1.63	1.13	−0.63	−0.03
Asf_g	0.055	0.058	−0.005	0.015	−0.032	0.015	0.025	0.005	−0.002	0.018	−0.062
$(N/mm)_g$	−7.12	−7.55	1.58	2.88	−4.65	−0.42	−1.82	−3.32	2.15	−0.55	3.85
R_6	0.055	0.058	−0.005	0.015	−0.072	0.015	0.065	0.005	−0.002	0.018	−0.062
R_7	−3.10	2.87	−0.57	3.23	0.43	−2.50	0.10	2.63	3.13	−0.63	−0.03

Table 5
Number of significant effects (# Sign) from the half-normal probability plot (HNPP), and from the critical effects based on negligible effects (NE), i.e. interaction (I) or dummy (d) effects, from Dong (D) and from the 75% approach (75%), for the different responses. For the D and 75% approaches, the number of false positive (# FP) and false negative (# FN) results, relative to HNPP and/or NE, are also indicated. * see Table 3.

Response	*HNPP*	*NE*		*D*			*75%*		
	# Sign	*I or d*	*# Sign*	*# Sign*	*# FP*	*# FN*	*# Sign*	*# FP*	*# FN*
*1**	3	/	/	2	0	1	3	0	0
2	1	/	/	1	0	0	1	0	0
3	2	/	/	2	0	0	3	1	0
4	1	/	/	1	0	0	1	0	0
5	0	/	/	0	0	0	0	0	0
6	1	/	/	1	0	0	1	0	0
*7**	3	/	/	1	0	2	3	0	0
8	2	I	2	2	0	0	2	0	0
9	2	I	2	2	0	0	2	0	0
10	0	I	0	0	0	0	0	0	0
11	1	I	1	1	0	0	1	0	0
*12**	3	I	3	3	0	0	3	0	0
*13**	4	I	4	0	0	4	4	0	0
14	2	d	2	0	0	2	0	0	2
15	1	d	1	1	0	0	1	0	0
16	1	d	1	1	0	0	1	0	0
17	1	d	2	2	0	0	2	0	0
18	1	d	1	1	0	0	5	4	0
19	1	d	2	1	0	1	1	0	1
20	1	d	1	1	0	0	3	2	0
*21**	5	d	5	5	0	0	5	0	0
*22**	5	d	5	1	0	4	4	0	1
*23**	5	d	5	1	0	4	5	0	0
24	0	d	0	0	0	0	0	0	0
25	3	d	3	3	0	0	3	0	0
26	2	d	2	2	0	0	2	0	0
27	1	d	1	1	0	0	1	0	0
*28**	6	d	6	2	0	4	6	0	0
*29**	8	d	8	0	0	8	8	0	0
30	0	I	1	0	0	1	3	2	0
31	1	I	2	1	0	1	1	0	1
32	2	d	4	3	0	1	6	2	0
*33**	7	d	7	7	0	0	7	0	0
*34**	12	d	12	0	0	12	10	0	2
					$\sum=0$	$\sum=45$		$\sum=11$	$\sum=7$

the lowest ME as critical effect was not found appropriate to determine effects significance.

The *second approach (B)*, i.e. evaluating visually the half-normal probability plot, was more appropriate to determine the number of effects to include in the calculation of s_1. When all effects deviating from the non-significance straight line through zero are discarded, more accurate estimations of the critical effects are obtained, also for situations with about 50% significant effects and for larger designs. In all 11 responses with about 50% significant effects, the important effects are correctly indicated. In contrast with the algorithm of Dong (D), no false negative results are obtained using this approach (B), and compared to the first approach (L), this method (B) usually also leads to less false positive results. The observed is not unexpected as the graphical interpretation was part of the reference approaches.

However, this approach also has some drawbacks. The interpretation of the half-normal probability plot is not always straightforward, i.e. it is not always evident to properly draw the straight line through the non-significant effects. Occasionally, different lines could be chosen, resulting in considerably different conclusions. In Table 3, it is indicated as B (1) and B (2) when the decision was based on two different lines. This was, for example, the case for response 23. When using B (1), four of the five significant effects are considered non-significant. On the other hand, when using B (2), all five effects are considered significant. The number of effects included in the calculation of s_1 thus depends on a graphical, i.e. a subjective, approach, which hardly can be generalized as a statistical or mathematical one can.

Therefore, the search for a better methodology was continued and other approaches were explored. From these possibilities (third and fourth approaches), the best result, i.e. that having the critical effect corresponding best to the reference approaches, was usually obtained when calculating the median based on the 75% lowest absolute effects, i.e. when using $s_0 = 1.5 \times \text{median}|E_{75\%}|$.

In situations with about 50% significant effects, the *third approach (P)* does not always correctly indicate the significant effects, and false negative results are obtained. This is the case in 7 out of the 11 situations, i.e. for responses 1, 13, 22, 23, 28, 29, and 34. The reason for the observed is that in the above situations, the approach usually still leads to an overestimated error, and thus a high critical effect (Table 3).

When using the *fourth approach*, different numbers of effects included for s_0 were evaluated. When using the *95*, *90*, *85*, *or 80%* smallest effects in situations with about 50% significant effects, often effects still were incorrectly considered non-significant, and thus false negative results are obtained (Table 3). This is seen for 7 out of the 11 cases, i.e. for responses 1, 13, 22, 23, 28, 29, and 34.

The use of $s_0 = 1.5 \times \text{median}|E_{75\%}|$ (*75% approach*) leads to more correct decisions on the effects significance than the algorithm of Dong and the other alternative approaches. Especially in situations where the number of significant effects approaches the 50% limit, this adapted algorithm is better capable of correctly identifying the significant effects. Less effects are incorrectly considered non-significant and thus less false negative results are obtained. With this approach, the important effects are correctly indicated without the occurrence of any false positive results in 9 out of the 11 responses with about 50% significant effects, i.e. for responses 1, 7, 12, 13, 21, 23, 28, 29, and 33. For the two other cases, i.e. responses 22 and 34, this approach allows determining all significant effects, except one (of five) and two (of twelve), respectively. However, although in situations where the effect sparsity principle is violated, only some rare false negative results are obtained, much more effects are correctly indicated as significant, compared to the algorithm of Dong (see Tables 3 and 5).

On the other hand, also in situations where the number of significant effects is far below 50%, the adapted algorithm is capable of identifying the significant effects, and usually without leading to many false negative or false positive results. Compared to the reference criteria, false negative results are obtained only for responses 14, 19, and 31. However, the algorithm of Dong does not indicate these effects as significant either.

Thus, in general, much less false negative results are obtained using this approach than when using the algorithm of Dong, i.e. 7 versus 45 (Tables 3 and 5).

False positive results are seen for responses 3, 18, 20, 30, and 32. Compared to the algorithm of Dong, this approach thus leads to a somewhat higher occurrence of false positive results, i.e. 11 versus 0 (Tables 3 and 5). However, from a practical point of view, this situation is much less problematic than when ignoring significant effects. Here one will react when in fact it is not necessary, while in the case of false negative results, one should react but does not.

In summary, the algorithm of Dong leads to many false negative results, especially in cases where the effect sparsity principle is violated. The first approach (L or using the lowest ME) indicates all significant effects correctly, but usually way too much false positive results are obtained. Although the second approach (B or visual inspection of the half-normal probability plot) correctly indicates all significant effects, with much less occurrence of false positive results, this approach remains based on a graphical, thus a subjective, interpretation, and is therefore less recommended. When using the third approach (P) or some of the fourth approaches (95-90-85-80%), still too many false negative results are again obtained. The fourth approach based on the 75% lowest effects leads to the best decisions on effects significance. Compared to the algorithm of Dong, much less false negative results are obtained. A drawback is that somewhat more false positive results are found, but as mentioned above, this is less problematic than the false negative results.

Therefore, when a Dong-like approach is used to identify significant effects, e.g. in minimal screening designs, we suggest, based on the results of the case studies, to use the adapted algorithm, where the 75% lowest absolute factor effects are used to calculate the initial error estimate s_0, i.e. $s_0 = 1.5 \times \text{median}|E_{75\%}|$.

4.2. Application

In [18], the robustness of a high-performance liquid chromatographic method to analyze three anti-epileptic drugs, pregabalin, gabapentin, and vigabatrin, in human serum was examined using the 12-experiments Plackett-Burman design, presented in Table 4. The results of the robustness test were treated with the above proposed methodology to determine the effects significance. In Table 6, the results are presented for the algorithm of Dong and the approach using $s_0 = 1.5 \times \text{median}|E_{75\%}|$, for each response of the application data set. As reference criterion, also the critical effects based on the dummy effects are presented.

For the responses Rs_v, $\%RSD_g$, and $(N/mm)_g$, the same results were obtained for both approaches (D and 75%). Compared to the approach based on the dummy effects ($d_{1\text{-}2\text{-}3\text{-}4\text{-}5}$), the same conclusions were drawn, except for $(N/mm)_g$, where two effects ($E_B = -7.55$ and $E_A = -7.12$) are considered borderline significant ($E_{crit} = 6.71$). For Dong ($s_0 = 4.33$, $ME_{\alpha=0.05} = 8.76$) and the 75% approach ($s_0 = 2.98$, $ME_{\alpha=0.05} = 7.63$), these two effects are considered non-significant, though in the latter case it is borderline. Thus, the effects of B and A are to be considered borderline cases.

For the other responses, i.e. $\%RSD_v$, Asf_g, R_6 and R_7, different results were obtained for both approaches (D and 75%). They are further considered in more detail. For $\%RSD_v$, the 75% approach indicates factor B ($E_B = 2.40$) as borderline significant ($s_0 = 0.93$, $ME_{\alpha=0.05} = 2.11$), while Dong considers it as borderline

Table 6
Evaluation of the proposed methodology (75% approach) to determine effects significance, and comparison with the critical effects (E_{crit}) based on dummy effects and on the algorithm of Dong. For symbols used, see Tables 3–5. The results are presented for the responses of the application data set.

Response	*Approach*	#	E_{crit}	$E_{\alpha=0.05}$
$\%RSD_v$	NE ($d_{1\text{-}2\text{-}3\text{-}4\text{-}5}$)	/	1.92	B
	D	11	2.55	Nothing
	75%	10	2.11	B
Rs_v	NE ($d_{1\text{-}2\text{-}3\text{-}4\text{-}5}$)	/	0.47	D
	D	11	0.46	D
	75%	11	0.46	D
$\%RSD_g$	NE ($d_{1\text{-}2\text{-}3\text{-}4\text{-}5}$)	/	2.40	Nothing
	D	11	2.42	Nothing
	75%	11	2.42	Nothing
Asf_g	NE ($d_{1\text{-}2\text{-}3\text{-}4\text{-}5}$)	/	0.080	Nothing
	NE ($d_{1\text{-}2\text{-}3\text{-}4}$)	/	0.044	d_5/B/A
	D	11	0.075	Nothing
	75%	9	0.056	d_5/B
$(N/mm)_g$	NE ($d_{1\text{-}2\text{-}3\text{-}4\text{-}5}$)	/	6.71	B/A
	D	11	8.76	Nothing
	75%	10	7.63	Nothing
R_6	NE ($d_{2\text{-}3\text{-}4}$)	/	0.035	E/G/K/B/A
	D	11	0.095	Nothing
	75%	7	0.056	E/G/K/B
R_7	NE ($d_{1\text{-}4\text{-}5}$)	/	1.18	D/I/A/B/H/F
	D	11	4.79	Nothing
	75%	5	1.10	D/I/A/B/H/F

non-significant ($s_0 = 1.40$, $ME_{\alpha=0.05} = 2.55$). From the half-normal probability plot (not shown), B does not clearly deviate from the straight line of the non-significant effects. When using the dummies ($d_{1\text{-}2\text{-}3\text{-}4\text{-}5}$), B is also considered borderline significant ($E_{crit} = 1.92$). Thus, the effect of B is a borderline case.

For Asf_g, the algorithm of Dong does not indicate any effect as significant ($s_0 = 0.028$, $ME_{\alpha=0.05} = 0.075$), while the 75% approach considers d_5 ($E_{d_5} = -0.062$) and B ($E_B = 0.058$) as borderline significant ($s_0 = 0.023$, $ME_{\alpha=0.05} = 0.056$) and A ($E_A = 0.055$) as borderline non-significant. When evaluating the half-normal probability plot (Fig. 2a), factors d_5, B, and A deviate from the straight line. From the critical effect based on the dummies ($d_{1\text{-}2\text{-}3\text{-}4\text{-}5}$) ($E_{crit} = 0.080$), no effects are considered significant. However, since the effect of the fifth dummy (d_5) is the largest, it should be excluded. The critical effect based on four dummies ($d_{1\text{-}2\text{-}3\text{-}4}$) was $E_{crit} = 0.044$. Then factors d_5, B, and A are considered significant. It might be concluded that in this situation, the algorithm of Dong includes too many effects to estimate s_1, thus overestimating $ME_{\alpha=0.05}$ (see Table 6) and indicating too few effects as significant. The alternative approach leads to more correct decisions on the significance, because here the two highest effects are excluded from the s_1 error estimation, resulting in a lower $ME_{\alpha=0.05}$. These observations confirm the conclusions from the earlier case studies.

For the simulated responses, R_6 and R_7, which are in the context of this paper most interesting, the number of significant effects approaches 50%. The algorithm of Dong incorrectly indicates no effects as being significant, since $s_0 = 0.028$, $ME_{\alpha=0.05} = 0.095$ and $s_0 = 3.75$, $ME_{\alpha=0.05} = 4.79$ for R_6 and R_7, respectively (Table 6). The half-normal probability plots for these two responses (Fig. 2b and c) clearly indicate the presence of important effects, which are obviously deviating from the non-significant effects. For R_6, factors E, G, K, B, and A are important, while for R_7, factors D, I, A, B, H, and F are. Their effects are also considered significant when evaluating the critical effects based on the dummies ($E_{crit} = 0.035$ and $E_{crit} = 1.18$ for R_6 and R_7, respectively) (Table 6). When using the 75% approach, four of the five significant effects are correctly indicated for response R_6 ($s_0 = 0.023$, $ME_{\alpha=0.05} = 0.056$) and all six for R_7 ($s_0 = 0.90$, $ME_{\alpha=0.05} = 1.10$). Thus, in these situations, the new approach again leads to much less false negative results than the algorithm of Dong. The latter included all 11 effects for the s_1 error estimation, while the alternative methodology only 7 and 5 for R_6 and R_7, respectively, leading to lower and more accurate estimations of the error and the $ME_{\alpha=0.05}$ (Table 6). These results again confirm the earlier case studies conclusions.

In general, it can be concluded that the new methodology to determine the effects significance from a screening design provides good decisions on the importance of the estimated effects. In cases where the number of significant effects approaches 50%, the approach based on $s_0 = 1.5 \times \text{median}|E_{75\%}|$ leads to more accurate results on the effects significance than the original algorithm of Dong, i.e. less false negative results are obtained. On the other

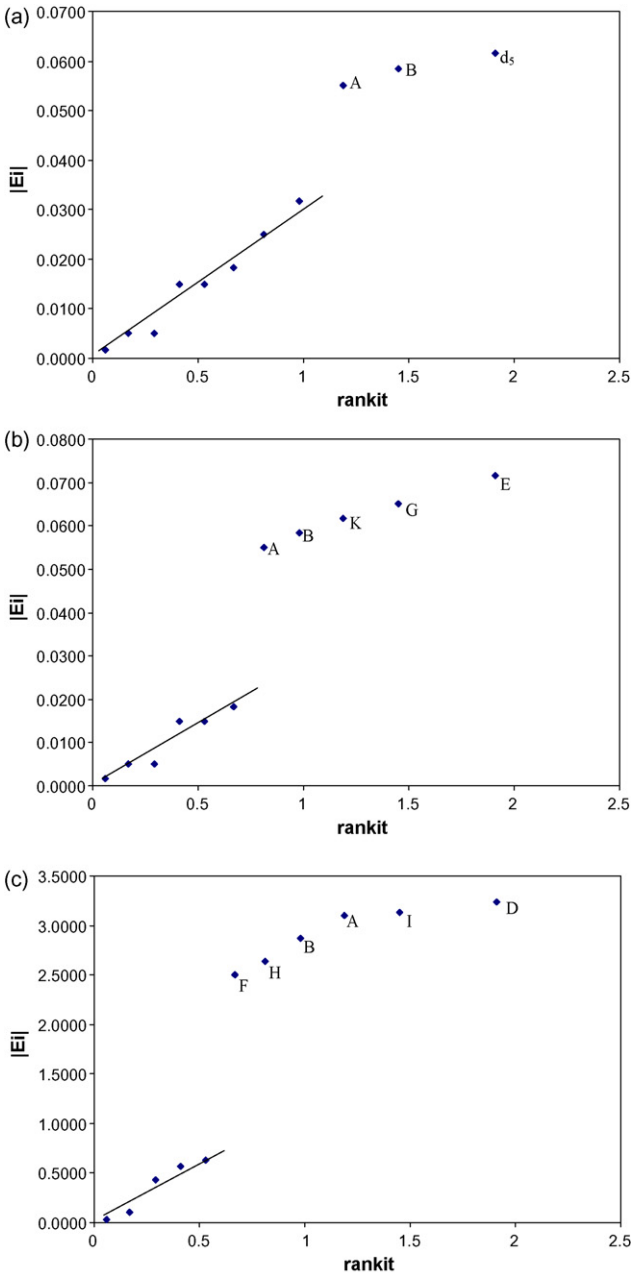

Fig. 2. Half-normal probability plot of the 11 effects on the responses (a) Asf_g, (b) R_6, and (c) R_7 (application data set, Table 4). The straight line of the non-significant effects is drawn.

hand, in case of effect sparsity, the 75% approach usually leads to the same results as the algorithm of Dong.

5. Conclusions

In this paper, a new methodology to determine significance of effects, estimated from a screening design, was developed starting from the algorithm of Dong. It is suggested to apply the 75% lowest absolute factor effects to calculate the initial error estimation s_0, i.e. $s_0 = 1.5 \times \text{median}|E_{75\%}|$.

This new methodology was compared with the algorithm of Dong. Especially in situations with about 50% significant effects, Dong includes too many effects in the s_1 error estimation, leading to many important effects incorrectly considered non-significant, i.e. false negative results. In these situations, the new methodology includes a more appropriate number of effects, leading to more accurate estimations of the error. In situations where effect sparsity occurs, the new methodology usually leads to the same results as the original algorithm of Dong, although a somewhat higher occurrence of false positive results is observed.

This new methodology was then successfully applied on a bioanalysis application data set. The results confirmed those found with the earlier case studies.

The new methodology is especially interesting to be applied in minimal screening designs, i.e. situations where no or too few dummies or negligible interaction effects can be estimated to be used in the statistical interpretation of the effects.

Acknowledgement

Bieke Dejaegher is a post-doctoral fellow of the Fund for Scientific Research (FWO) – Vlaanderen, Belgium.

References

[1] Guidelines prepared within the International Conference on Harmonisation of Technical Requirements for the Registration of Pharmaceuticals for Human Use (ICH), Validation of Analytical Procedures: Text and Methodology, Q2(R1), 2005, 1–13, http://www.ich.org/ (accessed on July 23th 2008).
[2] Y. Vander Heyden, D.L. Massart, in: M.W.B. Hendriks, J.H. de Boer, A.K. Smilde (Eds.), Robustness of Analytical Chemical Methods and Pharmaceutical Technological Products, Elsevier, Amsterdam, 1996, pp. 79–147.
[3] Y. Vander Heyden, A. Nijhuis, J. Smeyers-Verbeke, B.G.M. Vandeginste, D.L. Massart, J. Pharm. Biomed. Anal. 24 (2001) 723.
[4] Y. Vander Heyden, C. Perrin, D.L. Massart, in: K. Valkó (Ed.), Handbook of Analytical Separations, Vol. 1, Separation Methods in Drug Synthesis and Purification, Elsevier, Amsterdam, 2000, pp. 163–212.
[5] R.L. Plackett, J.P. Burman, Biometrika 33 (1946) 303.
[6] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics Part A, Elsevier, Amsterdam, 1997, pp. 367–369, 659–699.
[7] A. Nijhuis, H.C.M. van der Knaap, S. de Jong, B.G.M. Vandeginste, Anal. Chim. Acta 391 (1999) 187.
[8] R.V. Lenth, Technometrics 31 (1989) 469.
[9] F. Dong, Stat. Sin. 3 (1993) 209.
[10] B. Dejaegher, X. Capron, J. Smeyers-Verbeke, Y. Vander Heyden, Anal. Chim. Acta 564 (2006) 184.
[11] E. Hund, Y. Vander Heyden, M. Haustein, D.L. Massart, J. Smeyers-Verbeke, J. Chromatogr. A 874 (2000) 167.
[12] A. Nguyen Minh Nguyet, L. Tallieu, J. Plaizier-Vercammen, D.L. Massart, Y. Vander Heyden, J. Pharm. Biomed. Anal. 32 (2003) 1.
[13] Y. Vander Heyden, A. Bourgeois, D.L. Massart, Anal. Chim. Acta 347 (1997) 369.
[14] C. Perrin, H. Fabre, M. Maftouh, D.L. Massart, Y. Vander Heyden, J. Chromatogr. A 1007 (2003) 165.
[15] Y. Vander Heyden, C. Hartmann, D.L. Massart, L. Michel, P. Kiechle, F. Erni, Anal. Chim. Acta 316 (1995) 15.
[16] C. Vannecke, M.S. Bloomfield, Y. Vander Heyden, D.L. Massart, Anal. Chim. Acta 455 (2002) 117.
[17] C. Vannecke, A. Nguyen Minh Nguyet, M.S. Bloomfield, A.J. Staple, Y. Vander Heyden, D.L. Massart, J. Pharm. Biomed. Anal. 23 (2000) 291.
[18] T.A.C. Vermeij, P.M. Edelbroek, J. Chromatogr. B 810 (2004) 297.