ELSEVIER

Journal of Pharmaceutical and Biomedical Analysis
17 (1998) 153–168

JOURNAL OF
PHARMACEUTICAL
AND BIOMEDICAL
ANALYSIS

A ruggedness test strategy for procedure related factors: experimental set-up and interpretation

Y. Vander Heyden, F. Questier, D.L. Massart *

Vrije Universiteit Brussel, ChemoAC, Pharmaceutical Institute, Laarbeeklaan 103, B-1090 Brussels, Belgium

Received 14 April 1997; received in revised form 27 July 1997

Abstract

A strategy to perform ruggedness tests for mainly procedure related factors is described. The different steps in the set-up of the experiments and in the interpretation of the results are given. The described strategy is based on a number of case studies and allows a statistical interpretation of the significance of the effects. It was implemented in a software tool. This original strategy was completed with a number of minimal screening designs which reduce the number of experiments to perform, but in consequence only allow a limited or no statistical interpretation of the effects. Some of the minimal designs are expandable to designs with characteristics similar to those of the original strategy. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Ruggedness test strategy; Method validation; Experimental design; Factorial design; Plackett–Burman design

1. Introduction

Ruggedness of an analytical method can be described as the ability to reproduce the method under different circumstances (different days, different technicians, different laboratories) without the occurrence of unexpected differences in the results obtained. A ruggedness test is a part of method validation which is becoming increasingly important, particularly in the pharmaceutical industry. However, we observed that several laboratories have problems with creating the experimental set-up and with the interpretation of the results obtained. Performing a ruggedness test in general implies the execution of an experimental design and some analytical chemists are reserved in applying it. Moreover the fact that significant effects can be determined in different ways is not encouraging either. Therefore, after examining the literature [1] and performing a number of case studies ([2–6], personal communications) a strategy was selected to determine the ruggedness of an analytical method. It should encourage the analytical chemists that are reserved in applying experimental design to use it anyway. The strategy is meant as a guide to the less experienced users of experimental design, in the choices that can be made in the set-up and interpretation of a ruggedness test.

* Corresponding author. Tel.: + 32 2 4774737; fax: + 32 2 4774735; e-mail: fabi@vub.vub.ac.be

PII S0731-7085(97)00196-9

Two definitions for ruggedness or for a ruggedness test can be distinguished. The first one is the definition most frequently used in the chemical literature: "A ruggedness test is an intralaboratory experimental study in which the influence of small changes in the operating or environmental conditions, called factors, on measured or calculated responses are evaluated. The changes introduced reflect the changes that can occur when a method is transferred between different laboratories, different experimentators, different devices, etc." [7,8]. The factors one can examine according to this definition are mainly procedure related factors and they are most often examined with screening designs [9–11]. These factors are in general defined in the operating procedure of the method.

A definition for ruggedness less frequently used is the following one "The ruggedness of an analytical method is the degree of reproducibility of test results obtained by the analysis of the same sample under a variety of normal test conditions, such as different laboratories, different analysts, different instruments, different lots of reagents, different elapsed assay times, different assay temperatures, different days, etc. Ruggedness is a measure of reproducibility of test results under normal, expected operational conditions from laboratory to laboratory and from analyst to analyst [12]". The factors here are mainly non-procedure related factors. If several of these factors need to be examined, a nested design [13,14] could be applied.

The main difference between both definitions is that in the second one the procedure is executed as described in its operating procedure (most procedure related factors are kept constant) but on different days, by different analysts, on different instruments, etc. and according to the first definition the procedure-related factors are varied while the non-procedure related ones are mainly kept constant. This latter approach in general applies a screening design to examine the influence of the factors. The use of such a design to examine ruggedness according to the second definition is often not possible because impossible factor combinations are required [1] which destroy the well-balanced properties of the screening designs.

In this paper a strategy is described for performing a ruggedness test for (mainly) procedure-related factors. One can identify the following steps in a ruggedness test: (a) identification of the factors to be tested; (b) definition of the different levels for the factors; (c) selection of the experimental design and of the complete experimental set-up; (d) carrying out the experiments described in the experimental set-up and determining the responses of the method; (e) calculation of effects; (f) statistical and/or graphical analysis of the effects; and (g) drawing chemically relevant conclusions from the statistical analysis and if necessary give advice for improving the performance of the method. This text describes the choices that were made for the different steps mentioned above and that were implemented into a software tool. However, it is not the aim of this paper to focus on the description of the software tool and its possibilities, but on the choices made for the different steps in ruggedness testing. This should allow the user to make the same choices using his own (commercial) software and should also make him aware of occasional differences implemented in his software.

2. Experimental

A number of case studies were performed to evaluate the selected strategy. The experimental set-up of these case studies is described in Refs. [2–6] (personal communications). It concerns experiments in different application fields as high performance liquid chromatography (HPLC), size exclusion chromatography (SEC) and galenics. The selected strategy was implemented in a software program. The software environment was Microsoft Windows and Matlab 4.0 (MathWorks, Natick, MA). The hardware used was a computer with an Intel Pentium 75 MHz CPU processor and with 8 Mb RAM.

3. Discussion

The strategy described was selected after examining both the literature [1] and evaluating the

results of a number of case studies ([2–6], personal communications). The above mentioned software program contains the compilation of the selections made and which are described below. Commercially available software packages such as for instance Statgraphics plus 6.1 (Manugistics, Rockville, MD), MODDE for Windows (Umetri AB, Umeå, Sweden), SAS (SAS Institute, Cary, NC), The Unscrambler (Camo A/S, Trondheim, Norway), Design-Ease v.5 and Design-Expert v.5 (Stat-Ease, Minneapolis, MN) and Minitab for Window 10.5 (Minitab, State College, PA) also allow to generate designs and to analyse the results. The above list is anyway not limitative and probably several other software tools exist which allow the same. However, these commercial packages were designed for much broader application areas, they contain lots of possibilities and they therefore often require a more thorough theoretical background from the user to set-up and interpret for instance a ruggedness test.

Also some expert systems specifically designed for ruggedness testing are also available such as for instance RES (commercialised under the name 'Shaiker') created by Van Leeuwen et al. [7,15,16] and the Ruggedness Method Manager (Merck, Clevenot, France). In RES statistically significant effects are identified by comparing with predefined critical values. These are however dependent on the method and analytical technique applied. The Ruggedness Method Manager focuses on the generation of a fractional factorial design with a well-defined confounding pattern which again requires from the user a thorough knowledge about the method and about screening designs.

3.1. Definition of factors to be tested

The factors to be examined are mainly selected from the procedure as it is documented in for instance a standard operating procedure. A maximum of twenty factors is foreseen in our strategy. However, when the number of factors exceeds eight, we recommend to consider to split the set of factors into two, e.g. one set containing factors of the sample preparation and one containing factors of the determination. A larger number of factors requires a larger experimental design, which can be time-consuming and rather complicated to perform in practice. Examples of factors examined in a ruggedness test can be found in [1–7,17]. The factors can be quantitative or qualitative. Quantitative factors in the experimental design context are factors that vary on a continuous scale, while qualitative factors only can take discrete levels such as the origin or the manufacturer of the stationary phase in chromatography (Table 1). The factors one can select depends on the analytical technique used and on the complexity of the method. In Table 1 for example some potential factors from an HPLC method having a sample preparation step and a post-column derivatisation, are given.

3.2. Selection of factor levels

Factors are examined at two or three levels. If the factors are examined at only two levels (the extreme levels), they are normally situated around those specified in the operating procedure (nominal level). When examined at three levels, the nominal and two extremes are selected. The interval chosen between the levels represents the (often somewhat exaggerated) boundaries between which one expects the factors to vary. Examples of levels used in different case studies can be found in [1–7,17] and in Table 1.

The selection of the nominal level in the middle of the interval is what is usually done but what is not necessary and sometimes also not the most logical choice. Let us consider for instance in Table 1 the factor reaction time. If 30 min is the required reaction time, then normally the analyst will not use a much shorter time. However a longer reaction time is much more probable. The analyst is for instance busy with something else when the 30 min reaction time is passed; supposes or decides that a somewhat longer reaction time will not influence the result and stops the reaction only after finishing what he was busy with.

Therefore, in this case, a non-symmetrical interval around the nominal level is, seen from a practical point of view, more logical.

Concerning the definition (formulation) of factors to be tested and their levels it has to be

Table 1
Some potential factors and their possible levels extracted from an HPLC method with a sample preparation procedure and a post-column derivatisation

Factors	Nominal level	Extreme levels	
Sample pretreatment in plasma			
Concentration of pretreatment reagent	0.050 M	0.045 M	0.055 M
Concentration of pH adjusting solution	0.10 M	0.09 M	0.11 M
Reaction time	30 min	25 min	45 min
Reaction temperature	37°C	25°C	45°C
Concentration deproteinating solution	1.7 M	1.5 M	1.9 M
HPLC part of the method			
Amount organic modifier in mobile phase	10%	8%	12%
pH of the aqueous part in the mobile phase	3.0	2.8	3.2
Ionic strength of aqueous part of mobile phase	0.010	0.008	0.012
Flow rate	1.0 ml min^{-1}	0.9 ml min^{-1}	1.1 ml min^{-1}
Column temperature	30°C	25°C	35°C
Origin/manufacturer mobile phase (RP-18)	LiChrosorb	LiChrospher	Superspher
Post-column derivatisation (reaction of substance to determine with 2 reagents)			
pH of derivatisation solution	11.0	10.8	11.2
Concentration of borate buffer	0.050 M	0.045 M	0.055 M
Concentration of derivatisation reagent 1	1.0×10^{-3} M	0.9×10^{-3} M	1.1×10^{-3} M
Concentration of derivatisation reagent 2	2.0×10^{-3} M	1.9×10^{-3} M	2.1×10^{-3} M
Flow of derivatisation solution	0.50 ml min^{-1}	0.45 ml min^{-1}	0.55 ml min^{-1}

remarked that an appropriate knowledge of the physical properties of the factors and a thorough understanding of the used experimental design can increase considerably the information gained from the experiments. The factor levels have to be achievable in combination with each other and they have to be capable of being set and reset reproducibly between the different design experiments. Besides this, the way of defining the factors can be important to gain additional information. Suppose a phosphate buffer is defined in an operating procedure by the concentrations of Na_2HPO_4 and NaH_2PO_4. If both concentrations are defined as two factors there is no practical problem since they are achievable with each other and they can be reproducibly set and reset. However, one is not capable to give an immediate physical meaning to the calculated effects for these factors since, for instance, the pH and the ionic strength are related to both these factors. When both factors would have been combined a new factor describing the pH or the ionic strength (μ) could be created and the calculated effect then represents the effect of pH or of μ on the considered response. Therefore, though the first approach is not incorrect, the second one leads to better interpretable results. A more detailed discussion about this problem is given in a paper which is in preparation. Other factors are also discussed there as well as the selection of factors levels.

3.3. Selection of the experimental design

The influences of the factors are examined in an experimental design, which is selected as a function of the number of factors to investigate. The designs applied are fractional factorial [9,10] or Plackett–Burman designs [11]. Those originally implemented in the program are shown in Table 2. All designs applied are so-called two-level screening designs which allow to screen a relatively large number of factors in a relatively small number of experiments.

The designs used take into account requirements for statistical interpretation. We started from the idea that a statistical interpretation of the effects was necessary. Based on experimental

Table 2
Screening designs applied in the strategy described

No. of factors	Selected design	Generators	No. of dummies	No. of experiments (N)
2	Full factorial for three factors: 2^3	—	1	8
3	Full factorial for three factors: 2^3	—	0	8
4	Half-fraction factorial for four factors: 2^{4-1}	D = ABC	0	8
5	Half-fraction factorial for five factors: 2^{5-1}	E = ABCD	0	16
6	Quarter-fraction factorial: 2^{6-2}	E = ABC, F = BCD	0	16
7	Eighth-fraction factorial: 2^{7-3}	E = ABC, F = BCD, G = ABD	0	16
8	Sixteenth-fraction factorial: 2^{8-4}	E = ABC F = BCD, G = ABD, H = ACD	0	16
5–8	Plackett–Burman design for 11 factors	—	6-3	12
9–12	Plackett–Burman design for 15 factors	—	6-3	16
13–16	Plackett–Burman design for 19 factors	—	6-3	20
17–20	Plackett–Burman design for 23 factors	—	6-3	24

results [2] it was decided to estimate the experimental error on the effects from the design itself (see Section 3.6). This involves that the design performed is not always the one with the smallest number of experiments possible for a given number of factors, because a minimal number of degrees of freedom to estimate the experimental error was taken into account. The Plackett–Burman designs were chosen so that at least three dummy factors [2,3] can be included. In the fractional factorial designs used, (i) the two-factor interactions [9,10] are not confounded with the main effects, i.e. the design resolution is at least IV; and (ii) at least three two-factor interactions can be calculated. The designs given in Table 2 are those with the lowest number of experiments possible while still fulfilling these requirements. For more detailed background information about the generation of the different designs used we would like to refer to [1,9,10].

For examining five to eight factors the choice between a fractional factorial design with 16 experiments or a Plackett–Burman design with 12 experiments was included. The fractional factorial designs allow to estimate main effects without being confounded with two-factor interactions. This is not the case with the Plackett–Burman designs for which however a lower number of experiments needs to be performed. The above mentioned choice was mainly added because some analysts do not like using Plackett–Burman designs because of the confounding with the two-factor interactions. However, we observed in practice that in ruggedness testing both types of designs gave similar results since the two-factor interactions can be considered negligible.

It has to be stressed that in a ruggedness test one is mainly concerned about the main effects of factors. Interaction effects are much less of interest at least as long as they do not disturb the estimation of the main effects. In Plackett–Burman designs, two-factor interactions are confounded with the main effects. The two-factor interactions occurring in a ruggedness test can be considered negligible since the main effects estimated from Plackett–Burman designs on the one hand and those from fractional factorial designs with resolution IV on the other were found to be similar notwithstanding the differences in confounding. For that reason, (i) the Plackett-Burman designs were included in our ruggedness test strategy; and (ii) the two-factor interactions effects which can be estimated in the applied fractional factorial designs, were used to estimate the experimental error on the effects in these latter designs (see Section 3.6).

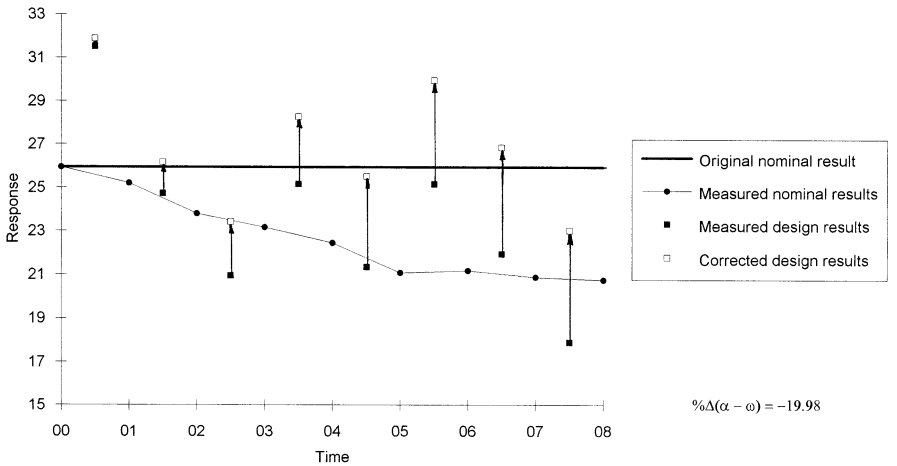

Fig. 1. Check and correction for drift based on replicate nominal experiments.

If the factors are examined at three levels, the designs of Table 2 are reflected [7], i.e. the designs are performed twice, namely once with an extreme level and the nominal one and once with the other extreme level and the nominal one. Three-level design were not used by us. A full three-level factorial design requires $3f$ experiments, with f being the number of factors, which is too high to be practically executed. Three-level (Plackett–Burman) factorial designs are described [18]. The smallest one examines four factors in nine experiments and the larger one, up to 13 factors in 27 experiments. Their use would have been a possibility but they were not incorporated in our strategy. We chose to use the reflected designs in analogy with [7]. Though these two-level designs do not cover the whole experimental space formed by the three levels of the factors, the reflected designs give acceptable estimates for the effects since the interactions between factors are in general negligible in a ruggedness test [2,6]. The use of the reflected designs also allows us to limit the number of designs incorporated in the strategy since they were the same as those used for examining the factors at two levels.

3.3.1. Selection of the complete experimental set-up

Beside the design experiments, a number of experiments at nominal levels can be included. These replicated nominal experiments are then performed at regular times between the ruggedness test experiments (Fig. 1). They are used to check if the nominal response is drifting as a function of time and to correct the design results for such an occasionally occurring drift [6]. How this correction is performed is schematised in Fig. 1. In the program only a limited number of possibilities is provided to add these nominal experiments (Table 3). However, if one is setting-up ones own ruggedness test, other possibilities can be chosen.

The check and correction for drift was added to the strategy to avoid that in certain cases wrong factor effects would be estimated. From the design results the influence (effect) of the examined factors is estimated. This involves that no other factor than those examined, is expected to affect the change in response between design experiments. However, if the nominal response (i.e. the one measured with all factors constant at method conditions) is drifting as a function of time, due to whatever cause, the consequence will be that the estimation of one or more factor effects will be influenced and that wrong estimates can be made [6].

The design experiments are normally performed in a random sequence. For practical reasons experiments are often blocked (sorted) by one or more factors. This means that for the blocked

Table 3
The possibilities for the different designs (N), provided in the program, to perform nominal experiments every n designs experiments

Design	No. of design experiments to perform	Nominal experiment possibilities every n design experiments						
Two levels examined								
$N=8$	$N_{exp}=8$	1	2	4				
$N=12$	$N_{exp}=12$	1	2	3	4	6		
$N=16$	$N_{exp}=16$	1	2	4	8			
$N=20$	$N_{exp}=20$	1	2	4	5	10		
$N=24$	$N_{exp}=24$	1	2	3	4	6	8	12
Three levels examined								
$N=8$	$N_{exp}=14$	1	2	7				
$N=12$	$N_{exp}=22$	1	2	11				
$N=16$	$N_{exp}=30$	1	2	3	5	6	10	15
$N=20$	$N_{exp}=38$	1	2	19				
$N=24$	$N_{exp}=46$	1	2	23				

factor first all experiments where it is at one level are performed followed afterwards by those at the other. The possibility is offered to block a maximum of three factors where the second is blocked within the first and the third within the second. Within the block(s) the experiments are randomised. Even though this blocking facility is offered and is regularly used one has to be aware that this way of working can contain some pitfalls. If drift occurs the estimated effect(s) of the blocked factor(s) will namely be affected most by the drift as can be observed in [1,6,19]. If blocking is performed, at least a minimal check for drift is recommended.

The blocking which occurs here is a blocking by one or more of the examined factors. This has nothing to do with a blocking by external factors not tested in the design such as for instance time (different days). When for example a design cannot be performed within one day, it can be executed in blocks on different days. This kind of blocking can also cause a blocking effect which is confounded with one or more effects estimated for the design factors. Which effects are confounded in that case depend on the sequence the design experiments are performed, as can be observed in [6]. For this latter kind of confounding (from an external blocking factor) no check nor correction was included.

If no regular check of the drift is performed the possibility to add two nominal experiments, one before and one after the designs experiments, is still provided. These experiments allow (i) to check if the method performs well at the beginning and at the end of the experiments; (ii) to obtain a first guess for drift; and (iii) to normalise the effects (see further).

The above described steps (Sections 3.1, 3.2 and 3.3) allow to define the experimental set-up, i.e. the sequence of the experiments to be performed. The creation of this experimental set-up forms a first part of the program as can be seen in the flow chart of Fig. 2.

3.4. Determining responses

From the experiments performed, a number of responses are determined. One will look in the first instance at responses describing a quantity (called quantitative responses further) such as for instance the contents of main substance and by-products, and to a smaller extent peak areas or peak heights in chromatographic methods. Secondly, one can also consider some responses describing the quality of the separation or of the analysis, such as for example in chromatographic methods the resolution, relative retention, capacity factors and asymmetry factors (called qualitative responses further).

To avoid confusion we would like to point out the difference between qualitative and quantitative factors, defined in Section 3.1 on the one

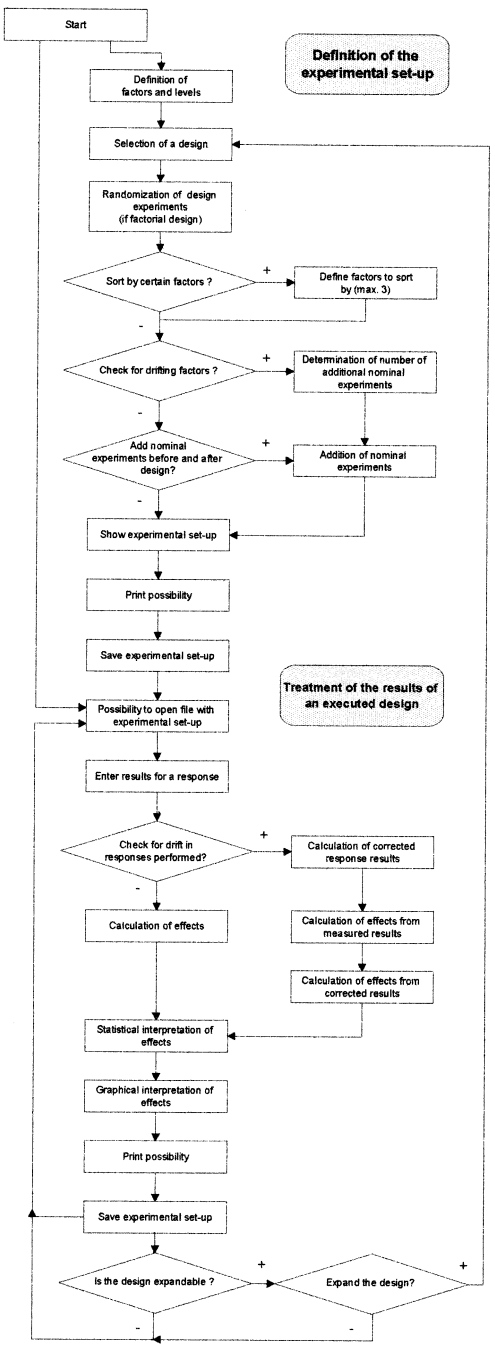

Fig. 2. Flow chart of the software program containing our strategy.

hand, and what we have called above qualitative and quantitative responses on the other.

3.4.1. Calculation of corrected responses

If one checked for drift, corrected response results can be calculated from the measured ones as was seen in Fig. 1. For this correction it is assumed that the different experiments are performed at equal time intervals. The corrected design results are calculated as

$$y_{i,\text{corrected}} = y_{i,\text{measured}} + y_{\text{nom,begin}} - \left(\frac{(n+1-i)*y_{\text{nom,before}} + i*y_{\text{nom,after}}}{n+1}\right) \quad (1)$$

where $i = 1, 2, \ldots, n$ and n is the number of design experiments between two consecutive nominal experiments, $y_{i,\text{corrected}}$ is a corrected design result, $y_{i,\text{measured}}$ the corresponding measured design result, $y_{\text{nom,begin}}$ the nominal result at the beginning of the experiments (before design), $y_{\text{nom,before}}$ and $y_{\text{nom,after}}$ the nominal results measured before and after the design result for which one is correcting.

3.5. Calculation of effects

Effects are calculated both from the measured and the corrected response results. When the factors are examined at two levels, for each factor one effect is calculated according to the equation

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

where X can represent (i) the factors A, B, C,…; (ii) the two-factor interactions for fractional factorial designs; and (iii) the dummies for Plackett–Burman designs, E_X is the effect of X on response Y; $\Sigma\, Y(+1)$ and $\Sigma\, Y(-1)$ are the sums of the responses where X is at the extreme levels $(+1)$ and (-1), respectively, and N is the number of experiments of the design.

This also means that when a response was corrected, two sets of effects are obtained, one from the measured results and one from the corrected results (Table 4). The effects were also normalised relative to the nominal result (y_n), in those cases where nominal experiments were performed

Table 4
Example of interpretation tables of the effects both from measured and corrected results for a design in which four factors were examined at two levels

Factors (interactions)	Effect	Normalised effect	Significance
Effects from measured results			
A	−0.705	−2.72	—
B	−4.195	−16.18	*
C	−4.000	−15.43	*
D	−4.710	−18.16	**
AB+CD	0.785	3.03	—
AC+BD	0.590	2.28	—
AD+BC	0.870	3.36	—
Critical effect ($\alpha = 0.05$)	2.410	9.30	
Critical effect ($\alpha = 0.01$)	4.424	17.06	
Effects from corrected results			
A	−0.058	−0.22	—
B	−2.987	−11.52	**
C	−1.100	−4.24	—
D	−4.715	−18.18	**
AB+CD	0.578	2.23	—
AC+BD	0.365	1.41	—
AD+BC	0.185	0.71	—
Critical effect ($\alpha = 0.05$)	1.300	5.02	
Critical effect ($\alpha = 0.01$)	2.387	9.20	

The effects were calculated from the results presented in Fig. 1.

$$\%E_X = \frac{E_X \cdot 100}{y_n} \quad (3)$$

The result y_n was defined as the average result of the two nominal experiments, performed before and after the design, respectively, if no systematic check for drift was done and as the nominal result measured before the design experiments, in case drift was checked.

The effects, as calculated in Eq. (2), were preferred to the use of linear regression coefficients [20]. The regression coefficients can be considered as an estimate of the change in response that occurs when the factor is changed from the nominal level to an extreme level, while the effect of Eq. (2) describes the change that occurs when the factor is changed from one extreme level to the other. The relation between the effect Eq. (2) and the regression coefficient is that the latter is a factor two smaller [20]. While the effect describes simply the observed average difference between the two extreme levels, the regression coefficient describes an interpolation (measurements performed at the extreme levels, but conclusions drawn for a part of this interval).

The conclusions drawn from the regression coefficients are therefore only valid if a number of assumptions is fulfilled: (a) the factor is quantitative and not qualitative; (b) the nominal level is situated in the middle of the interval between the two extreme levels; (c) the response is linear in the interval between the two extreme levels. Since these assumptions are not always fulfilled, as for instance can be observed in Sections 3.1 and 3.2, the use of Eq. (2) was preferred.

If a reflected design is performed two effects for each factor are calculated

$$E_{X(-1)} = \frac{\Sigma Y(0)}{N/2} - \frac{\Sigma Y(-1)}{N/2} \quad (4)$$

$$E_{X(+1)} = \frac{\Sigma Y(+1)}{N/2} - \frac{\Sigma Y(0)}{N/2} \quad (5)$$

where for the estimation of $E_{X(-1)}$ the half of the experiments where the factors were examined at levels (−1) and (0) are used and for $E_{X(+1)}$ the other half with levels (0) and (+1). This implies that when responses are corrected four sets of

effects are calculated, two with factor levels (0) and (-1) and two with factor levels (0) and $(+1)$.

If the effect of a factor is linearly related to its level and if the nominal level is situated centrally in the interval, then $E_{X(-1)}$ will be identical to $E_{X(+1)}$ and they will only differ due to experimental error. However, our strategy does not foresee in a systematic test to verify if $E_{X(-1)}$ is significantly different from $E_{X(+1)}$, nor in the option to build a single model using the whole experimental set-up which would lead to the estimation of one coefficient (or effect) instead of two. The latter is because we do not prefer to use regression models as already mentioned higher. If the effects $E_{X(-1)}$ and $E_{X(+1)}$ are different one could visualise the results by drawing so-called effect plots [20,21] which give an idea about the change of the response as a function of the factor levels. Comparison of these plots for the different factors also allow to visualise the relative importance of the effects. Our strategy does not foresee in the drawing of these effect plots. The consequence of a difference between $E_{X(-1)}$ and $E_{X(+1)}$ will depend on the factor examined and on the method (technique) involved. The conclusions to be drawn in that case are left to the analyst.

3.5.1. Verification for drift

To verify if a response is drifting, the replicate nominal results are plotted as a function of time (Fig. 1). Besides this plot, the percent change of the response can also be calculated

$$\%\Delta(\alpha-\omega)=\frac{y_{\text{nom,end}}-y_{\text{nom,begin}}}{y_{\text{nom,begin}}}*100 \tag{6}$$

where $y_{\text{nom,end}}$ is the nominal result measured after the last design experiment. Based on the plot and on the value for $\%\Delta(\alpha-\omega)$ the analyst decides if the response indeed is drifting and whether the effects estimated from the measured or from the corrected results (Table 4) will be interpreted.

3.6. Statistical interpretation of effects

To identify statistically significant effects a t-test is performed [1–7,17,22–26]

$$t=\frac{|E_X|}{(\text{SE})_e}\Leftrightarrow t_{\text{critical}} \tag{7}$$

with $(\text{SE})_e$ here being estimated by $\sqrt{\Sigma\, E^2_{X_iX_j}/n_{X_iX_j}}$ for fractional factorial designs and by $\sqrt{\Sigma\, E^2_{\text{dummy}_i}/n_{\text{dummy}}}$ for Plackett–Burman designs. The symbol $(\text{SE})_e$ stands for the standard error on an effect and represents the experimental variability within the design. It is estimated from at least three two-factor interaction ($E_{X_iX_j}$) or dummy (E_{dummy}) effects.

The number of degrees of freedom used for t_{critical} is the number of effects used to estimate $(\text{SE})_e$, i.e. $n_{X_iX_j}$ and n_{dummy}.

The calculation of $(\text{SE})_e$, as above, assumes that the interaction or dummy effects are a measure for the experimental error. It was shown in several case studies that in general in ruggedness testing this is the case [2–6]. Calculation of $(\text{SE})_e$ using significant dummy or interaction effects would (i) increase the estimated value for $(\text{SE})_e$; (ii) decrease t; and (iii) make it more difficult to identify significant factors. Occasional significant interaction or dummy effects can be identified from the normal probability plots (see Section 3.6.1) and can, by the analyst, be deleted from the estimation of $(\text{SE})_e$. However, in the latter case one better does not estimate $(\text{SE})_e$ with less than three dummy or interaction effects since then t_{critical} increases strongly which again makes it more difficult to identify significant effects.

The test given in Eq. (7) was rewritten as

$$|E_X|\Leftrightarrow E_{\text{critical}}=t_{\text{critical}}\cdot(SE)_e \tag{8}$$

or

$$|\%E_X|\Leftrightarrow\%E_{\text{critical}}=\frac{E_{\text{critical}}\cdot 100}{y_n} \tag{9}$$

A critical effect (E_{critical}) is calculated at a significance level α of 0.05 and 0.01. An effect is considered significant at a given a level if $|E_X|>E_{\text{critical}}$. Significance of an estimated effect at a $\alpha=0.05$ is indicated with * and at $\alpha=0.01$ with ** (Table 4). The results of the statistical interpretation are also represented graphically as shown in Fig. 3.

3.6.1. Graphical interpretation of effects

A graphical interpretation of effects is done by drawing normal probability plots (Fig. 4) [9]. Non-significant effects are normally distributed around zero so that in a normal probability plot they tend to fall on a straight line through zero, while significant effects deviate from this line.

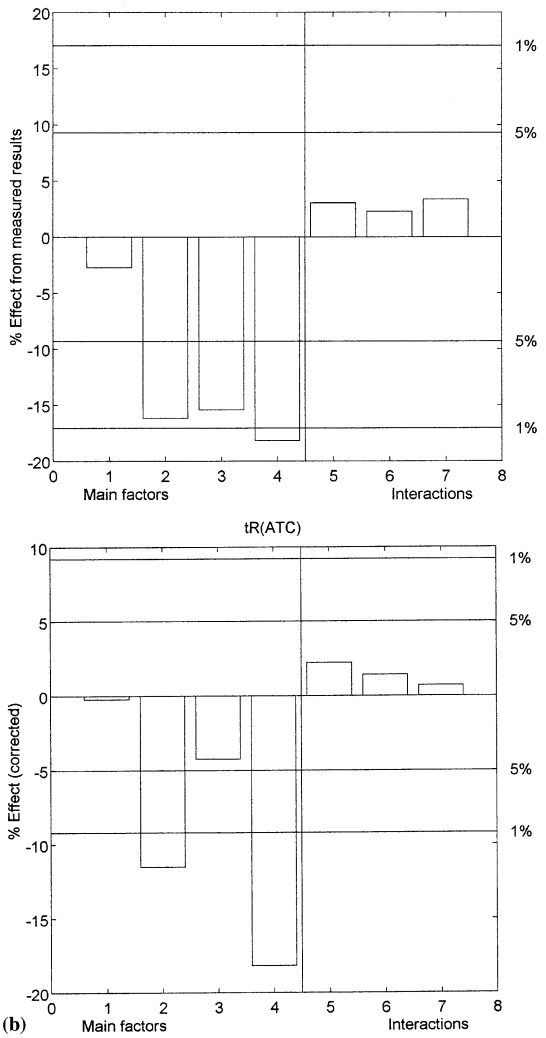

Fig. 3. Graphical representation of the statistical interpretation for (a) effects from measured results, and (b) effects from corrected results. The data from Table 4 was used.

3.7. Drawing chemically relevant conclusions

The results of the ruggedness test should lead to the identification of the factors that could cause problems when a method is transferred to another instrument or laboratory or when an interlaboratory study is performed on the method in order to determine its reproducibility.

To decide on the ruggedness of the method, the first responses of interest are the quantitative ones as already mentioned earlier. If the contents of the substances remain unaffected by the factors a method can be considered rugged.

Besides the quantitative responses a large number of qualitative ones can also be considered for most methods. Normally several significant effects will be observed in these responses. This however does not a priori means that the method is not robust. The knowledge of these significances can be interesting to add to the validation documents of the method and to prove to regulatory bodies (e.g. to the Food and Drug Administration for pharmaceutical methods) ones thorough understanding and documenting of the method.

4. Inclusion of minimal designs in the strategy

The above described strategy proved to work in a number of case studies. However, it can require a relatively large number of experiments which may exceed the number that analysts are willing to perform. A statistical interpretation of the effects is not always wanted neither. It is often sufficient to have an idea about the magnitude of effects. Therefore, the above strategy was completed by considering different possibilities for the selection of a design as a function of the number of factors to examine. A first one was the inclusion of the minimal screening designs for a given number of factors. A second one was the inclusion of fractional factorial designs that allow extension. One can then start from a smaller fraction (less experiments) and go, after a first evaluation of the effects, to a larger one by performing additional experiments. A third possibility is the inclusion of supersaturated designs [27–29]. This latter possibility is not included yet

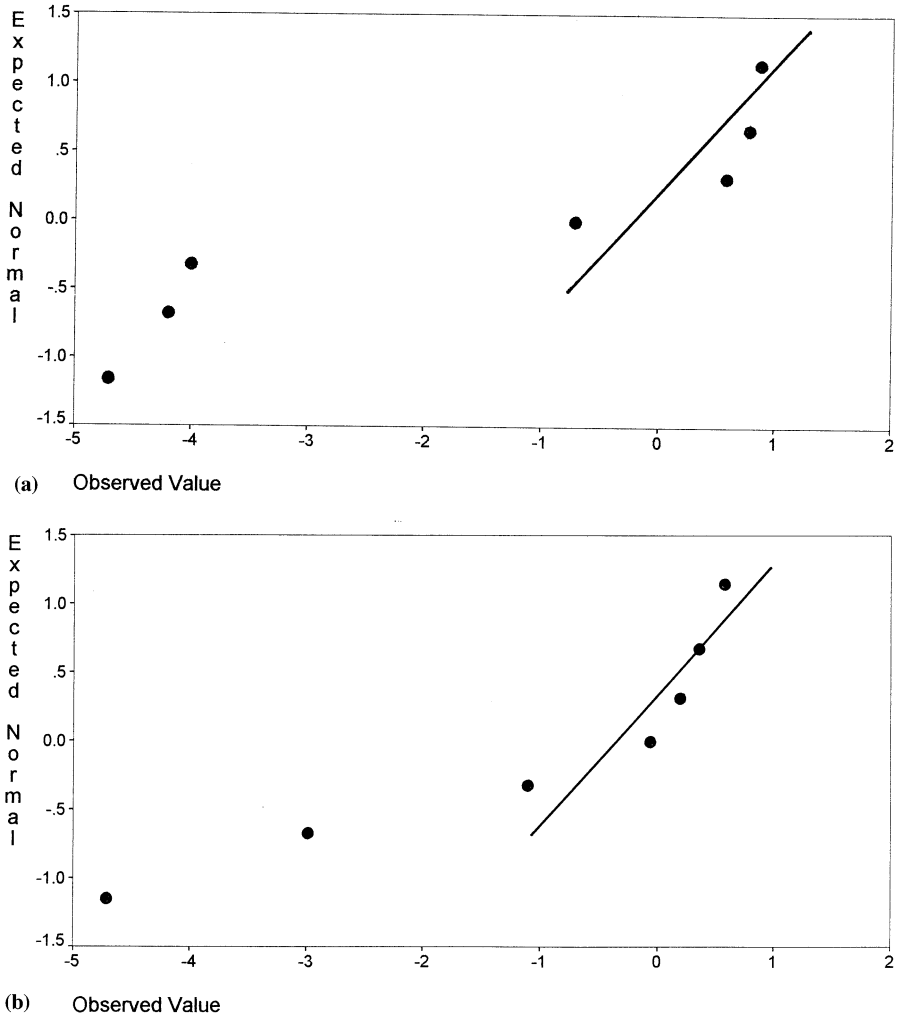

Fig. 4. Graphical interpretation: normal probability for (a) effects from measured results, and (b) effects from corrected results. The data from Table 4 was used. The straight lines of non-significant effects were drawn arbitrarily.

since the applicability of those designs in ruggedness testing still needs to be examined.

To the designs of the original strategy (Table 2) a number of minimal screening designs and expandable fractional factorial designs was added. Those selected are shown in Table 5. The statistical interpretation described above cannot be performed any more or is only based on a limited number of interactions or dummy factor effects as presented in Table 5.

After a first evaluation of the effects some of these designs can be expanded to a design similar to the one described in Table 2 for that number of factors. This is done by performing additional experiments which also form a design (Table 6). For the expandable fractional factorial designs another fraction of the full factorial is performed. Combination of both designs gives a new one with characteristics similar to those of Table 2. Therefore, the fraction that is added needs to fulfil some requirements. Combination of the original and of the new fraction should give a design of

Table 5
Minimal designs added to the strategy

No. of factors	Selected design	Generators	No. of interactions	No. of dummies	Expandable?	No. of experiments (N)
2	Full factorial for two factors: 2^2	—	1	0	Yes	4
3	Half-fraction factorial for three factors: 2^{3-1}	C = AB	0	0	Yes	4
4	Half-fraction factorial for four factors: 2^{4-1}	D = ABC	3	0	No	8
5	Quarter-fraction factorial for five factors: 2^{5-2}	D = AB, E = AC	2	0	Yes	8
6	Eighth-fraction factorial: 2^{6-3}	D = AB, E = AC, F = BC	1	0	Yes	8
7	Sixteenth-fraction factorial: 2^{7-4}	D = AB, E = AC, F = BC, G = ABC	0	0	Yes	8
8–11	Plackett–Burman design for 11 factors	—	—	3–0	No	12
12–15	Plackett–Burman design for 15 factors	—	—	3–0	No	16
16–19	Plackett–Burman design for 19 factors	—	—	3–0	No	20
20–23	Plackett–Burman design for 23 factors	—	—	3–0	No	24

resolution IV as is the case in Table 2. For more detailed theoretical background information about fractional factorial designs, we again would like to refer to [1,9,10].

Let us consider, for example, the 2^{5-2} quarter fraction design with generators D = AB and E = AC. Three other quarter fractions of the full factorial can be considered, namely those with generators, (i) D = −AB and E = −AC; (ii) D = −AB and E = AC; and (iii) D = AB and E = −AC. Only combination of the fractions with generators D = AB; E = AC and D = −AB; E = −AC gives a 2^{5-1} design with resolution IV (generator E = BCD). Combination of D = AB; E = AC with one of the other fractions gives a 2^{5-1} design with only resolution III.

The defining relations for the 2^{5-2} minimal design are I = ABD = ACE = BCDE. Those for the other fractions of the full factorial are (i) I = −ABD = −ACE = BCDE; (ii) I = −ABD = ACE = −BCDE; (iii) I = ABD = −ACE = −BCDE, respectively. When combining the different fractions, the defining relation(s) of the new design can be derived from the defining relations of those combined. Combination of the minimal design with fraction (i) gives as defining relation I = BCDE. This new defining relation can be obtained from the original ones by maintaining those occurring with the same sign (BCDE here) and deleting those with opposite signs (ABD and ACE here). The new design has resolution IV as can be observed from the number of terms in its defining relation and the generator of the design can for instance be chosen as E = BCD. Combination of the defining relations of fractions (ii) and (iii) with the ones of the minimal design gives I = ACE and I = ABD respectively which lead in both cases to designs with resolution III.

Similarly for the fractional factorial designs to examine six or seven factors the selected additional fraction was one that allows to create a design with resolution IV (Table 6). Let us consider the six factor case. The generators of the minimal 2^{6-3} design given in Table 6 have the defining relations I = ABD = ACE = BCF = BCDE = ACDF = ABEF = DEF. To create a de-

Table 6
Expansion of the minimal designs

No. of factors	Original minimal design	Additional experiments (expansion design)	Newly formed design
2	2^2	2^2	2^3 (containing one dummy); $N=8$
3	2^{3-1}; Generator: C = AB	2^{3-1}; Generator: C = −AB	2^3; $N=8$
5	2^{5-2}; Generators: D = AB, E = AC	2^{3-1}; Generators: D = −AB, E = −AC	2^{5-1}; Generator: E = BCD; $N=16$
6	2^{6-3}; Generators: D = AB, E = AC, F = BC	2^{6-3}; Generators: D = −AB, E = −AC, F = −BC	2^{6-2}; Generators: E = BCD, F = ACD; $N=16$
7	2^{7-4}; Generators: D = AB, E = AC, F = BC, G = ABC	2^{7-4}; Generators:, D = −AB, E = −AB, F = −BC, G = ABC	2^{7-3}; Generators: E = BCD, F = ACD, G = ABC; $N=16$

N = number of design experiments.

sign of resolution IV by combination of the minimal design with one of the seven other eighth-fractions, the three-factor terms of the defining relations needs to be removed. This will be the case for the fraction with defining relations, I = −ABD = −ACE = −BCF = BCDE = ACDF = ABEF = −DEF and generators D = −AB, E = −AC and F = −BC. The defining relations of the newly created quarter fractional design (2^{6-2}) will be I = BCDE = ACDF = ABEF and the generators of this design are for instance E = BCD and F = ACD.

For the seven factor design (2^{7-4}) the defining relations are I = ABD = ACE = BCF = ABCG = BCDE = ACDF = CDG = ABEF = BEG = AFG = DEF = ADEG = BDFG = CEFG = ABCDEFG. Combination of this minimal design with the 2^{7-4} design with generators D = −AB, E = −AC, F = −BC and G = ABC gives a 2^{7-3} (IV) design with defining relations I = ABCG = BCDE = ACDF = ABEF = ADEG = BDFG = CEFG and generators E = BCD, F = ACD and G = ABC.

For examining the factors at three levels these minimal designs are again reflected. Also for these reflected designs an expansion is provided.

The interest of adding the minimal designs to our strategy is multiple as already mentioned higher. They allow, compared to the originally included designs, to perform a smaller design for given number of experiments for instance when the execution of an experiment is time consuming or expensive, or when no statistical interpretation of the effects is required. The expansion of the minimal designs can be of interest when an effect estimated from a minimal design is doubtful or when a statistical interpretation of the effects becomes necessary. The expanded design has a higher resolution than the minimal one and gives therefore a better estimate of the effect. From an expanded design (or from one of the originally included designs) it is also easier to report the results in an understandable way to a person that is not fully familiar with an experimental design approach, since such design allows to identify statistically significant effects.

5. Conclusions

As conclusion we would like to summarise some practical things to consider during the ruggedness testing of an analytical method. A first topic is the selection and definition of the factors and of their levels. The way of formulating the factors can lead to effects that have less or more a physical meaning which is immediately interpretable. The chosen factor levels should represent a reasonable interval and should not be exaggerated nor taken too small. The situating of the interval around the nominal level should reflect a realistic situation.

A design is then selected based on (i) the number of factors to examine; (ii) the fact a statistical interpretation of the effects derived from the design results is desired or not; and (iii) the time

needed to perform a given number of experiments. The finally executed experimental set-up can depend on (i) practical constraints that require sorting of experiments on one or more factors; and (ii) the addition of nominal experiments to verify for drift in the nominal response and/or to normalise effects.

If a check for drift is performed the analyst will have to decide if the nominal response is indeed drifting and whether or not the effects calculated from the corrected responses will be interpreted.

Finally we would like to point out that the occurrence of significant effects not always leads to a non-rugged method. In a ruggedness test one should focus first on the qualitative responses. The absence of significant effects on these responses indicate that the method remained unaffected by the variations introduced in the experimental design. The occurrence of significant effects on qualitative responses are then in general of a minor importance to decide on its ruggedness.

Acknowledgements

The authors thank the Fund for Scientific Research (FWO) for financial support. Parts of this work were funded by the Research Contract Nr. NO/03/003 of the Belgian government (The Prime Minister Services—Federal Office for Scientific, Technical and Cultural Affairs, Standardisation Programme) and by the European Community Standards, Measurements and Testing research programme. The authors are grateful to V. Reynders for the technical aid.

References

[1] Y. Vander Heyden, D.L. Massart, Review of the use of robustness and ruggedness in analytical chemistry, in: A. Smilde, J. de Boer, M. Hendriks (Eds.), Robustness of Analytical Methods and Pharmaceutical Technological Products, Elsevier, Amsterdam, 1996, pp. 79–147.

[2] Y. Vander Heyden, K. Luypaert, C. Hartmann, D.L. Massart, J. Hoogmartens, J. De Beer, Ruggedness tests on the HPLC assay of the United States Pharmacopeia XXII for tetracycline hydrochloride. A comparison of experimental designs and statistical interpretations, Anal. Chim. Acta 312 (1995) 245–262.

[3] Y. Vander Heyden, C. Hartmann, D.L. Massart, L. Michel, P. Kiechle, F. Erni, Ruggedness tests on an HPLC assay: comparison of tests at two and three levels by using two-level Plackett–Burman designs, Anal. Chim. Acta 316 (1995) 15–26.

[4] Y. Vander Heyden, D.L. Massart, Y. Zhu, J. Hoogmartens, J. De Beer, Ruggedness tests on the HPLC assay of the United States Pharmacopeia XXII for tetracycline hydrochloride: comparison of different columns in an interlaboratory approach, J. Pharm. Biomed. Anal. 14 (1996) 1313–1326.

[5] Y. Vander Heyden, C. Hartmann, D.L. Massart, P. Nuyten, A.M. Hollands, P. Schoenmakers, Ruggedness testing of a size exclusion chromatographic assay for low molecular mass polymers, J. Chromatogr. A 756 (1996) 89–106.

[6] Y. Vander Heyden, A. Bourgeois, D.L. Massart, Influence of the sequence of experiments in a ruggedness test when drift occurs, Anal. Chim. Acta 347 (1997) 369–384.

[7] J.A. Van Leeuwen, L.M.C. Buydens, B.G.M. Vandeginste, G. Kateman, P.J. Schoenmakers, M. Mulholland, RES, an expert system for the set-up and interpretation of a ruggedness test in HPLC method validation. Part 1: the ruggedness test in HPLC method validation, Chemom. Intell. Lab. Syst. 10 (1991) 337–347.

[8] G.T. Wernimont, in: W. Spendley (Ed.), Use of Statistics to Develop and Evaluate Analytical Methods, Association of Official Analytical Chemists, Arlington, VA, pp. 78–82.

[9] E. Morgan, Chemometrics: Experimental Design; Analytical Chemistry by Open Learning, Wiley, Chichester, 1991, pp. 118–188.

[10] G. Box, W. Hunter, J. Hunter, Statistics for Experimenters, an Introduction to Design, Data analysis and Model Building, Wiley, New York, 1978, pp. 306–418.

[11] R.L. Plackett, J.P. Burman, The design of optimum multifactorial experiments, Biometrika 33 (1946) 305–325.

[12] The United States Pharmacopeia XXII, The National Formulary XVII; United States Pharmacopeial Convention, Rockville, MD, 1990, p. 1712.

[13] International Organisation for Standardisation (ISO); Accuracy (trueness and precision) of measurement methods and results—Part 3: Intermediate measures of the precision of a standard measurement method; International Standard ISO 5725-3:1994(E).

[14] R.R. Sokal, F.J. Rohlf, Biometry, 2nd ed., W.H. Freeman, New York, 1981, pp. 271–320.

[15] J.A. Van Leeuwen, L.M.C. Buydens, B.G.M. Vandeginste, G. Kateman, P.J. Schoenmakers, M. Mulholland, RES, an expert system for the set-up and interpretation of a ruggedness test in HPLC method validation. Part 2: the ruggedness expert system, Chemom. Intell. Lab. Syst. 11 (1991) 37–55.

[16] J.A. Van Leeuwen, L.M.C. Buydens, B.G.M. Vandeginste, G. Kateman, A. Cleland, M. Mulholland, C. Jansen, F.A. Maris, P.H. Hoogkamer, J.H.M. van den Berg, RES, an expert system for the set-up and interpretation of a ruggedness test in HPLC method validation. Part 3: the evaluation, Chemom. Intell. Lab. Syst. 11 (1991) 161–174.
[17] J. Caporal-Gautier, J.M. Nivet, P. Algranti, M. Guilloteau, M. Histe, M. Lallier, J.J. N'Guyen-Huu, R. Russotto, Guide de validation analytique, rapport d'une commission SFSTP, STP Pharma Prat. 2 (1992) 205–239.
[18] Y. Vander Heyden, M.S. Khots, D.L. Massart, Three level screening designs for the optimisation and ruggedness testing of analytical procedures, Anal. Chim. Acta 276 (1993) 189–195.
[19] J.L. Goupy, Methods for Experimental Design, Principles and Applications for Physicists and Chemists, Elsevier, Amsterdam, 1993, pp. 159–177, 421–427.
[20] S.N. Deming, S.L. Morgan, Experimental design: a chemometric approach, in: B.G.M. Vandeginste, S.C. Rutan (Eds.), Data Handling in Science and Technology, vol. 11, Elsevier, Amsterdam, 1993, pp. 317–360.
[21] S. Boonkerd, M.R. Detaevernier, Y. Vander Heyden, J. Vindevogel, Y. Michotte, Determination of the enantiomeric purity of dexfenfluramine by capillary electrophoresis: use of a Plackett–Burman design for the optimization of the separation, J. Chromatogr. A 736 (1996) 281–289.
[22] W.J. Youden, E.H. Steiner, Statistical Manual of the Association of Official Analytical Chemists, The Association of Official Analytical Chemists, Arlington, VA, 1975, pp. 33–36, 70–71, 82–83.
[23] J. Vindevogel, P. Sandra, Resolution optimization in micellar electrokinetic chromatography: use of Plackett–Burman statistical design for the analysis of testosterone esters, Anal. Chem. 63 (1991) 1530–1536.
[24] S.F.Y. Li, Capillary electrophoresis: principles, practice and applications, J. Chromatogr. Libr. 52 (1992) 316–318.
[25] Statgraphics® Plus, Statistical Graphics System by Statistical Graphics Corporation, version 6, Reference Manual, Manugistics, Rockville, MD.
[26] D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte, L. Kaufman, Chemometrics: A Textbook, Elsevier, Amsterdam, 1988, pp. 101–106.
[27] K.H.V. Booth, D.R. Cox, Some systematic supersaturated designs, Technometrics 4 (1962) 489–495.
[28] D.K.J. Lin, Generating systematic supersaturated designs, Technometrics 37 (2) (1995) 213–225.
[29] D.K.J. Lin, A new class of supersaturated designs, Technometrics 35 (1) (1993) 28–31.