Journal of Pharmaceutical and Biomedical Analysis
24 (2001) 723–753

JOURNAL OF PHARMACEUTICAL AND BIOMEDICAL ANALYSIS

www.elsevier.com/locate/jpba

Guidance for robustness/ruggedness tests in method validation

Y. Vander Heyden [a,*], A. Nijhuis [b], J. Smeyers-Verbeke [a], B.G.M. Vandeginste [b], D.L. Massart [a]

[a] *Vrije Universiteit Brussel, ChemoAC, Pharmaceutical Institute, Laarbeeklaan 103, 1090 Brussel, Belgium*
[b] *Unilever Research Vlaardingen, P.O. Box 114, 3130 AC Vlaardingen, The Netherlands*

Received 31 May 2000; accepted 11 September 2000

Abstract

This paper is intended to give guidance in setting-up and interpreting a robustness test. The different steps in a robustness test are discussed and illustrated with examples. The recommendations given for the different steps are based on approaches found in the literature, several case studies performed by the authors and discussions of the authors within a commission of the French SFSTP (*Société Française des Sciences et Techniques Pharmaceutiques*). In the end of the paper a worked-out example is given of a robustness test case study set up and interpreted according to the guidelines. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Robustness test; Ruggedness test; Method validation; Method development; System suitability limits

Contents

* Corresponding author. Tel.: +32-2-4774735; fax: +32-2-4774736.
E-mail address: yvanvdh@fabi.vub.ac.be (Y. Vander Heyden).

PII: S0731-7085(00)00529-X

1. Introduction

1.1. Definitions

The definition for robustness/ruggedness applied is: *The robustness/ruggedness 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].

Robustness can be described as the ability to reproduce the (analytical) method in different laboratories or under different circumstances without the occurrence of unexpected differences in the obtained result(s), and a robustness test as an experimental set-up to evaluate the robustness of a method. The term ruggedness is frequently used as a synonym [2–5]. Several definitions for robustness or ruggedness exist which are, however, all closely related [1,6–10]. The one nowadays most widely applied in the pharmaceutical world is the one given by the International Conference on Harmonisation of Technical Requirements for the Registration of Pharmaceuticals for Human Use (ICH) [1] and which was given above. Only in [9] a distinction between the terms ruggedness and robustness is made and ruggedness is defined there as the degree of reproducibility of the test results obtained 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. The latter definition will not be applied since detailed guidelines exist for the estimation of the reproducibility and the intermediate precision [11,12]. The ICH guidelines [1] also recommend that '*one consequence of the evaluation of robustness should be that a series of system suitability parameters* (*e.g. resolution tests*) *is established to ensure that the validity of the analytical procedure is maintained whenever used*'.

The assessment of the robustness of a method is not required yet by the ICH guidelines, but it can be expected that in the near future it will become obligatory.

Robustness testing is nowadays best known and most widely applied in the pharmaceutical world because of the strict regulations in that domain set by regulatory authorities which require extensively validated methods. Therefore, most definitions and existing methodologies, e.g. those from the ICH, can be found in that field, as one can observe from the above. However, this has no implications for robustness testing of analytical methods in other domains and this guideline is therefore, not restricted to pharmaceutical methods.

1.2. Situating robustness in method development and validation

Robustness tests were originally introduced to avoid problems in interlaboratory studies and to identify the potentially responsible factors [2]. This means that a robustness test was performed at a late stage in the method validation since interlaboratory studies are performed in the final stage. Thus the robustness test was considered a part of method validation related to the precision (reproducibility) determination of the method [3,13–16].

However, performing a robustness test late in the validation procedure involves the risk that when a method is found not to be robust, it should be redeveloped and optimised. At this stage much effort and money have already been spent in the optimisation and validation, and therefore, one wants to avoid this. Therefore, the performance of a robustness test has been shifting to earlier points of time in the life of the method. The Dutch Pharmacists Guidelines [6], the ICH Guidelines [7] as well as some authors working in bio-analysis [17] consider robustness a method validation topic performed during the development and optimisation phase of a method, while others [18] consider it as belonging to the development of the analytical procedure.

Therefore, the robustness test can be viewed as a part of method validation that is performed at the end of method development or at the beginning of the validation procedure. The exact position has relatively little influence on how it is performed.

1.3. Objectives of a robustness evaluation

The robustness test examines the potential sources of variability in one or a number of responses of the method. In the first instance, the quantitative aspects (content determinations, recoveries) of the method are evaluated. However, besides these responses also those for which system suitability test (SST) limits can be defined (e.g. resolution, tailing factors, capacity factors, column efficiency in a chromatographic method) can be evaluated (Section 5).

To examine potential sources of variability, a number of factors are selected from the operating procedure (Section 2.1) and examined in an interval (Section 2.2) that slightly exceeds the variations which can be expected when a method is transferred from one instrument to another or from one laboratory to another. These factors are then examined in an experimental design (Section 3) and the effect of the factors on the response(s) of the method is evaluated (Section 6). In this way the factors that could impair the method performance are discovered. The analyst then knows that such factors must be more strictly controlled during the execution of the method.

Another aim of a ruggedness/robustness test may be to predict reproducibility or intermediate precision estimates [9]. In this guideline this kind of ruggedness testing is not considered.

The information gained from the robustness test can be used to define SST limits (Section 7). This allows determining SST limits based on experimental evidence and not arbitrarily on the experience of the analyst.

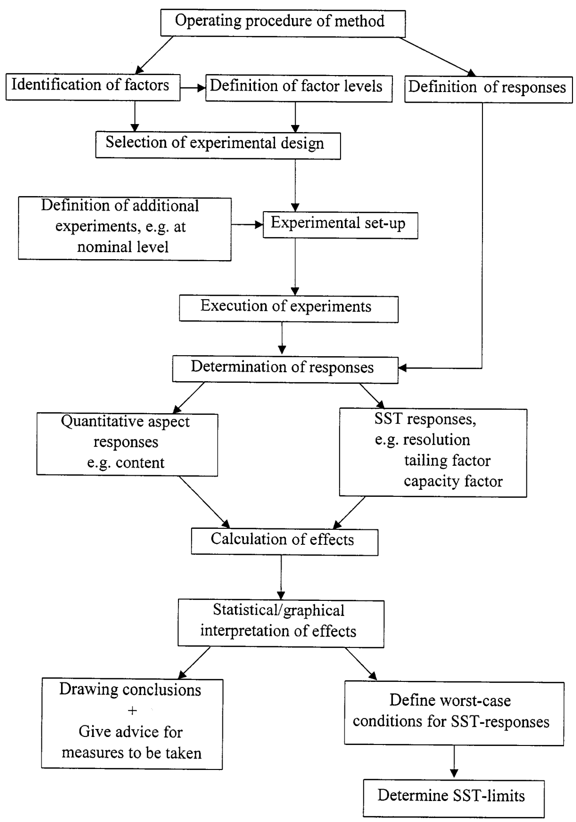

Fig. 1. Schematic representation of the different steps in a robustness test.

1.4. The steps in a robustness test

The following steps can be identified: (a) identification of the factors to be tested; (b) definition of the different levels for the factors; (c) selection of the experimental design; (d) definition of the experimental protocol (complete experimental set-up); (e) definition of the responses to be determined; (f) execution of the experiments and determination of the responses of the method; (g) calculation of effects; (h) statistical and/or graphical analysis of the effects; and (i) drawing chemically relevant conclusions from the statistical analysis and, if necessary, taking measures to improve the performance of the method. These different steps are schematically represented in Fig. 1 and are considered in more detail below. An example of a worked-out robustness test case study is described in Section 8.

2. Selection of factors and levels

2.1. Selection of the factors

The factors to be investigated in a robustness test are related to the analytical procedure (operational factors) and to the environmental conditions (environmental factors). The operational factors are selected from the description of the analytical method (operating procedure), whereas the environmental factors are not necessarily specified explicitly in the analytical method.

The selected factors can be quantitative (continuous), qualitative (discrete) or mixture factors. Table 1 indicates a list of factors that could be considered during robustness testing of chromatographic (liquid, gas or thin-layer chromatography) or electrophoretic methods. The list is not exhaustive, but gives the reader an idea of the factors commonly examined. If a sample preparation procedure (liquid–liquid extraction, solid–liquid extraction, ultrafiltration, dialysis) is required before the chromatographic or electrophoretic analysis, also factors from this procedure should be considered in the robustness testing. The number of factors to be examined further increases when the analytical procedure requires a pre- or a post-column derivatisation step.

Table 1
Potential factors to be examined in the robustness testing of some analytical methods[a]

Method	Factors

HPLC	pH of the mobile phase
	Amount of the organic modifier
	Buffer concentration, salt concentrations or ionic strength
	Concentration of additives (ion pairing agents, competing amine)
	Flow rate
	Column temperature
	For gradient elution
	  initial mobile phase composition
	  final mobile phase composition
	  slope of the gradient
	Column factors
	  batch of stationary phase
	  manufacturer
	  age of the column
	Detector factors
	  wavelength (UV or fluorimetric detection)
	  voltage (electrochemical detection)
	Integration factors
	  sensitivity

Gas Chromatography (GC)	Injection temperature
	Column temperature
	Detection temperature
	For temperature program
	  initial temperature
	  final temperature
	  slope of the temperature gradient
	Flow-rate of the gas
	For flow-program
	  initial flow
	  final flow
	  slope of the flow gradient
	Split-flow
	Type of liner
	Column factors
	  batch of stationary phase
	  manufacturer
	  age of the column

TLC	Eluent composition pH of the mobile phase Temperature Development distance Spot shape Spot size Batch of the plates Volume of sample Drying conditions (temperature, time) Conditions of spot visualisation (spraying of reagent, UV detection, dipping into a reagent)

CE and related techniques	Electrolyte concentration
	Buffer pH
	Concentration of additives (organic solvents, chiral selectors, surfactants)
	Temperature
	Applied voltage
	Sample injection time
	Sample concentration
	Concentration of the liquids to rinse
	Rinse times
	Detector factors
	  wavelength (UV or fluorimetric detection)
	Factors related to the capillary
	  batch
	  manufacturer
	Integration factors

[a] HPLC, high performance liquid chromatography; TLC, thin layer chromatography; and CE, capillary electrophoresis.

Examples of quantitative factors are the pH of a solution, the temperature or the concentration of a solution; of qualitative factors the batch of a reagent or the manufacturer of a chromatographic column, and of a mixture factor the fraction of organic modifier in a mobile phase.

The selected factors should represent those that are most likely to be changed when a method is transferred between laboratories, analysts or instruments and that potentially could influence the response(s) of the method.

2.1.1. Mixture-related factors

Mixtures of solvents are often used in analytical methods [5], e.g. mobile phases in chromatography or buffers in electrophoresis are mixtures. In a mixture of p components only $p-1$ can be changed independently. In HPLC analysis the mobile phase can contain, besides the aqueous phase, one–three organic modifiers, yielding mixtures of two–four components. In robustness testing both mixture and process variables (e.g. flow, temperature, wavelength) need to be combined in the same experimental set-up. The simplest procedure is to select maximally $p-1$ components to be examined as factors in the experimental design. These $p-1$ factors are then mathematically independent, called mixture-related variables [19] and are treated in the design in the same way as the process variables. The pth component is used as adjusting component: its value is determined by that of the $p-1$ mixture related variables. The contributions of the different components in the mixture preferably are expressed as volume fractions. As adjusting component the solvent occurring with the highest fraction in the mixture is selected.

If one component of a mixture is found to be important, this means in practice that the mixture composition as a whole is important. Since it is not possible to control only one of the compo-nents of a mixture, the composition of the mixture as a whole should be more strictly controlled.

Example: For a mobile phase containing methanol/acetonitrile/aqueous buffer with a composition of 10:20:70 (v/v/v), the methanol (MeOH) content and the acetonitrile (ACN) content can be selected as mixture-related variables and entered as factors in the design while the buffer content is used as adjusting component. This latter component is not considered to be a factor. The nominal levels (prescribed method conditions) of MeOH and of ACN are then 0.10 and 0.20, respectively.

2.1.2. Quantitative factors

A set of factors often can be entered in the experimental design in different ways and this can lead to physically more or less meaningful information. Therefore, when setting up a robustness test the analyst should carefully consider how to define or formulate the factors. As an example, consider the compounds of a buffer. The composition of a buffer can be defined by the concentrations of its acidic (C_a) and basic (C_b) compounds (Example (1)). There are several possibilities to examine these two compounds in a design, namely as two different factors or combined to represent the pH and/or the ionic strength (μ). If one wants to maximise the information extracted from a robustness test, it may be preferable to choose the factors in such a way that the effects have a physical meaning. In that case one should use pH and μ.

If the emphasis is only on measuring the robustness of the method then one could use the first approach (Example (2)). This involves that when one of the two factors (C_a or C_b) is found to be important, the second one also needs to be controlled strictly, as was the case with mixtures.

In the second approach, the two concentrations are combined to one factor, C_b/C_a. Depending on the variation introduced in this factor, one will simulate a change in the pH, in the ionic strength or in both. If one keeps the molar ratio constant and changes the concentrations of C_a and C_b then the factor examines a change in ionic strength. When the ratio is changed, then one can introduce, depending on the kind of buffer used, a change in the pH (for instance for a buffer like NaH_2PO_4/H_3PO_4 where only one compound contributes to μ) or both in the pH and the ionic strength (for instance for a buffer like Na_2HPO_4/NaH_2PO_4 where both compounds are contributing to μ) [5] (Example (3)).

Example:

1. A phosphate buffer description extracted from a monograph of [20] is for instance: "Place 6.8 g potassium dihydrogen phosphate, 500 ml water and 1.8 ml phosphoric acid in a 1000 ml volumetric flask. Adjust to volume with water and mix well. Render the contents of the flask homogenous by shaking vigorously until all solids are dissolved".
2. The factors C_a and C_b are then defined as the volume H_3PO_4 per litre buffer and the weight of NaH_2PO_4 per litre buffer, respectively.
3. Factors C_a and C_b are combined to NaH_2PO_4/H_3PO_4 which is then multiplied with a given constant to define the extreme levels for this factor. For the above described buffer two possibilities exist. A first one is $a \times (NaH_2PO_4/H_3PO_4)$ where the ratio between NaH_2PO_4 and H_3PO_4 is kept constant and $a=1$ represents the nominal situation while $a<1$ or >1 are the extreme levels. In this situation the pH is kept constant while the ionic strength changes.

A second possibility is ($a \times NaH_2PO_4$)/ H_3PO_4 with again $a = 1$, the nominal level and $a < 1$ or > 1, the extreme levels. In this situation the ratio is changed which means that the pH is changed.

It can also be remarked that:

1. for a buffer like Na_2HPO_4/NaH_2PO_4 where both compounds are contributing to the ionic strength this latter approach changes both the pH and μ;
2. only one of the two above possibilities can be defined as a factor in an experimental design.

Another regularly used way to describe the preparation of a buffer is to dissolve a given amount of a salt, e.g. NaH_2PO_4 and then to adjust the pH by adding an acid (e.g. H_3PO_4) or a base (e.g. NaOH). In this situation the pH should be chosen as a factor. The concentration of the salt could also be examined as a second factor.

Example: A buffer according to this type of definition is for instance: '10 mM phosphate buffer, pH 3.0'. In this situation the pH and the concentration phosphate (which represents the ionic strength) could be examined as factors.

2.1.3. Qualitative factors

Often qualitative factors are also included. For instance, for chromatographic methods, factors related to the column, such as the 'column manufacturer', the 'batch of the column' or even columns from the same batch are examined. Examining columns from a same batch is done to evaluate if characteristics unique to a single column, e.g. artefacts of the column packing procedure, affect the results. Columns from different batches are used to evaluate the batch-to-batch variations and from different manufacturers to examine the variations between manufacturers.

However, one should be aware of the fact that no observed significant effect for such a qualitative factor does not mean that this factor never has an influence. Examination of a very limited number of representatives (e.g. different columns) does not allow to draw any conclusion about the total population. Only conclusions concerning the robustness of the method with respect to the examined representatives can be made (see also Section 8).

When including several qualitative factors in an experimental design impossible factor combinations should be avoided. An example is the combination of the factors 'manufacturer of column material' and 'batch of material' in one two-level design (Section 3). Selecting two levels for the manufacturer of material would give manufacturers *I* and *J*. Selecting two levels for the batch of material is not possible since one cannot define batches common to both manufacturers *I* and *J*.

2.2. Selection of the factor levels

2.2.1. Quantitative and mixture factors

The factor levels are usually defined symmetrically around the nominal level prescribed in the operating procedure. The interval chosen between the extreme levels represents the (somewhat exaggerated) limits between which the factors are expected to vary when a method is transferred. In most case studies, the analyst defines the levels according to ones personal opinion. However, selection of the levels also can be based on the precision or the uncertainty [21] with which a factor can be set and reset. For instance the uncertainty in the factor 'pH of a solution' will depend on the uncertainty of the pH meter result and on the uncertainty related to the calibration of the pH meter. Suppose one knows, for instance from a systematic determination of the uncertainty [21], that the pH varies with a confidence level of 95% in the interval pH ± 0.02. How to do this is described further in this section. Due to the uncertainty in the pH, one can expect the nominal pH (pH_{nom}) to vary between the levels $pH_{nom} \pm 0.02$. To select the extreme levels in a robustness experiment this interval is enlarged to represent possible variations between instruments or laboratories. This is done by multiplying the uncertainty with a coefficient k which gives as extreme levels pH $\pm k \times 0.02$. The value $k = 5$ is proposed as default value. Other values can be used when the analyst considers larger or smaller intervals for certain factors to be feasible. Since the selected extreme levels are also subjected to uncertainty, $k = 2$ are the strictest conditions for which extreme levels can be evaluated that are clearly different from each other. To be clearly different from the nominal level, $k \geq 3$ is needed to define the extreme levels.

To quantify the uncertainty in analytical measurements detailed Eurachem guidelines are available [21]. They are quite tedious to apply since they try to quantify all sources of variability in a factor level, which is not always obvious to do. Therefore, for the purpose of robustness testing a simpler alternative is proposed [5].

Proposal: For each measured response one so-called absolute uncertainty is defined which quantifies the most obvious source of variation. For instance: (i) for a mass, consider the last number given by the balance or a value specified by the manufacturer, as uncertain, e.g. 0.1 mg for an analytical balance; (ii) for a volume, take the uncertainty in the internal volume of the volumetric recipient, specified by the manufacturer, e.g. 0.08 ml for a 100 ml volumetric flask; (iii) for a pH value, use the last digit of the display or a value specified by the manufacturer of the pH meter. When a response is calculated from a combination of measured components — as for instance is the case with a concentration which is the quotient between a mass and a volume — the following rules are applied: (i) the absolute uncertainty for a sum or a difference is the sum of the absolute uncertainties in the terms; and (ii) the relative uncertainty (i.e. ratio of absolute uncertainty over response value) for a product or a quotient is the sum of the relative uncertainties in the terms.

Table 2
Detailed information about the uncertainties in the mixture factors levels

	MeOH	H_2O	$MeOH+H_2O$
Volume (V)	300 ml	700 ml	1000 ml
Absolute uncertainty (V)	1.88 ml	5 ml	6.88 ml
Relative uncertainty (V)	6.27×10^{-3}	7.14×10^{-3}	6.88×10^{-3}
Fraction (f)	0.3	0.7	
Relative uncertainty (f)	1.32×10^{-2}	1.40×10^{-2}	
Absolute uncertainty (f)	0.004	0.01	

Example: Consider for instance the determination of the uncertainty in the concentration of a solution. Suppose a reagent solution with a nominal concentration of 100 mg l^{-1} is defined in the operating procedure and is prepared in a 100 ml volumetric flask. The concentration C is determined as $C = m/V$ where m is the mass weighed and V the volume. To determine the uncertainty in C the uncertainties in m and V are estimated first. The absolute uncertainty in the mass is defined as 2×0.1 mg (mass obtained from a difference of two measurements) and in the volume as 0.08 ml. This gives relative uncertainties of 0.02 and 0.0008, and for the concentration of 0.0208. The absolute uncertainty in the concentration is then 2 mg l^{-1} and the extreme levels to be examined in a design would be about 90 and 110 mg l^{-1} ($k = 5$).

The introduction of the coefficient k should also compensate for the occasional sources of variability which where not taken into account in the estimation of the absolute uncertainty.

A similar reasoning as for the quantitative factors is valid for mixture factors.

Example: Consider a mobile phase 30:70 v/v $MeOH/H_2O$ prepared using graduated cylinders of 500 ml for MeOH (uncertainty internal volume 1.88 ml) and of 1000 ml for water (uncertainty internal volume 5 ml). The fraction of methanol is calculated as $f_{MeOH} = V_{MeOH}/(V_{MeOH} + V_{H_2O})$ (volumes considered additive for ease of calculation). When applying the alternative estimates (not the Eurachem ones), only the uncertainties in the internal volumes are taken into account. According to these rules the absolute uncertainty in f_{MeOH} is 0.004 and the one in f_{H_2O} 0.01. More detailed information about these uncertainties can be found in Table 2.

2.2.2. Asymmetric intervals for quantitative factors

For quantitative factors, the interval between the extreme levels is usually situated symmetrically around the nominal one. For some factors the selection of an asymmetrical interval can rep-resent more reality. However, the probability or feasibility for the selection of asymmetric intervals needs to be evaluated for each factor separately and this within the context of a given robustness test.

Example: Suppose that a column temperature of 35°C is prescribed. Then it is not unlogical to define as low level a temperature that represents room temperature (e.g. 20 or 25°C) since it is probable that the method in some cases (laboratories) will be executed at this temperature, for instance because one does not have a column oven. To determine the high level, the uncertainty in the column oven temperature multiplied with coefficient k still could be used, giving for instance 40°C.

It could be argued that examining the temperature in an interval between 20 and 40°C is not a small perturbation, as the definitions for robustness require. However, the idea of robustness testing is that the factors are examined for changes that occur in practice when a method is transferred. If it can be excluded that, after transfer of the method from the laboratory where it is developed to those where it will be used, a method prescribed at 35°C will be executed at room temperature then the above given temperature interval could be replaced by a symmetric one, for instance, 30–40°C determined according to the rules of Section 2.2.1.

2.2.3. Qualitative factors

For qualitative factors, the obtained results are not representative for the whole population of the factor to which the selected levels belong but only allow an immediate comparison between the two discrete levels selected.

Example: Inclusion of only two columns in a robustness study does not allow to draw conclusions about the population of columns, i.e. about the robustness of the method on the particular type of columns (e.g. different batches), since it is far from evident that the selected columns represent extreme levels for the whole population. Only conclusions about the robustness of the method on the two examined columns can be drawn and no extrapolation to whatever other column can be made (Section 8).

3. Selection of the experimental designs

The factors are examined in an experimental design, which is selected as a function of the number of factors to investigate. 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 applied are fractional factorial [22–24] or Plackett–Burman designs [24,25]. In a robustness test one is only concerned about the main effects of factors.

In Plackett–Burman designs, two-factor interaction effects, among higher-order interaction effects, are confounded with the main effects [24]. Confounding between effects means that from a given design these effects cannot be estimated separately. The two-factor interactions occurring in a robustness test can however, be considered negligible [26].

Therefore, the Plackett–Burman designs are included in this guideline. For an inexperienced experimental design user, Plackett–Burman designs are easier to construct than fractional factorial designs. The latter are however, also given, for the sake of completeness (see Section 3.1).

For a given number of factors, both within the Plackett–Burman and the fractional factorial designs, two options are presented. The first option consists of using minimal designs, i.e. the designs with the absolute minimal number of experiments for a number of factors, while the second option allows a more extensive statistical interpretation of the effects. The recommended Plackett–Burman designs are described in Table 3. The smallest number of factors to be examined in an experimental design was considered to be three. For statistical reasons concerning effect interpretation,

Table 3
Plackett–Burman designs applied in the guideline

No. of factors	Selected design	No. of dummy factors	No. of experiments (N)
(a) *Minimal designs*			
3–7	Plackett–Burman design for 7 factors	4–0	8
8–11	Plackett–Burman design for 11 factors	3–0	12
12–15	Plackett–Burman design for 15 factors	3–0	16
16–19	Plackett–Burman design for 19 factors	3–0	20
20–23	Plackett–Burman design for 23 factors	3–0	24
(b) *Designs for statistical interpretation of effects from dummy factors*			
3–4	Plackett–Burman design for 7 factors	4–3	8
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

Table 4
First line for the applicable Plackett–Burman designs

Design (N)	First line
8	+ + + − + − −
12	+ + − + + + − − − + −
16	+ + + + − + − + + − − + − − −
20	+ + − − + + + + − + − + − − − − + + −
24	+ + + + + − + − + + − − + + − − + − + − − − −

designs with less than eight experiments are not used, while those with more than 24 experiments are considered unpractical. The designs are constructed as follows. The first line for the designs with $N = 8-24$ as described by Plackett–Burman [25] is in Table 4 with N being the number of experiments and $(+)$ and $(-)$ the levels of the factors. An example of a Plackett–Burman design for $N = 12$ is shown in Table 5. The first row in the design is copied from the list above. The following $N-2$ rows (in our example, 10) are obtained by a cyclical permutation of one position (i.e. shifting the line by one position to the right) compared to the previous row. This means that the sign of the first factor (A) in the second row is equal to that of the last factor (K) in the first row. The signs of the following $N-2$ factors in the second row are equal to those of the first $N-2$ factors of the first row. The third row is derived from the second one in an analogous way. This procedure is repeated $N-2$ times until all but one line is formed. The last (Nth) row consists only of minus signs.

A Plackett–Burman design with N experiments can examine up to $N-1$ factors. After determination of the number of real factors to be examined, the remaining columns in the design are defined as dummy factors. A dummy factor is an imaginary factor for which the change from one level to the other has no physical meaning.

In the minimal designs (Table 3a) the significance of effects is determined based on the distribution of the factor effects themselves (see Section 6.2.2.3), while in the other designs (Table 3b) the standard error on the effects is estimated from dummy factor effects (see Section 6.2.2.2) [26,27]. The latter designs are therefore, not always those 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, i.e. some columns are needed for dummy factors. The Plackett–Burman designs in Table 3b are chosen so that at least three dummy factors are included.

For complex methods, e.g. with an extensive sample pretreatment and/or a post-column derivatisation, it might be necessary to examine a large number of factors and a relatively large design is required which becomes tedious to perform. In such cases it may be more practical to split the factors in two sets and evaluate them in two smaller designs that are easier to execute. For instance, the factors of the derivatisation procedure are examined in one design and those related to the analytical technique in a second. The most commonly used designs consist of 8–16 experiments.

3.1. Selection of a fractional factorial design

For the selection of a fractional factorial design also two possibilities are provided, the minimal designs and those that take into account requirements for statistical interpretation from two-factor interactions. The designs proposed are described in Table 6. For more detailed background information about the generation of the different designs used we refer to [22–24]. Table 6a shows the minimal fractional factorial designs for a given number of factors. Some designs can be expanded to a design with similar characteristics as those described in Table 6b. This expansion possibility allows to execute a smaller fraction, and then, after a first evaluation of the effects, to perform another fraction of the full factorial, for which the combination with the previously performed experiments leads to a design with characteristics analogous to those given in Table 6b (see below) [28]. No expansion design was given for the minimal designs with 16 experiments since a total of 32 experiments is considered not feasible any more. The generators are not proposed when no expansion is foreseen and when it is not possible anymore to create a design with resolution IV [22,24].

Table 5
Plackett–Burman design for 11 factors ($N=12$)

Experiment	Factors											Response
	A	B	C	D	E	F	G	H	I	J	K	
1	+	+	−	+	+	+	−	−	−	+	−	y_1
2	−	+	+	−	+	+	+	−	−	−	+	y_2
3	+	−	+	+	−	+	+	+	−	−	−	y_3
4	−	+	−	+	+	−	+	+	+	−	−	y_4
5	−	−	+	−	+	+	−	+	+	+	−	y_5
6	−	−	−	+	−	+	+	−	+	+	+	y_6
7	+	−	−	−	+	−	+	+	−	+	+	y_7
8	+	+	−	−	−	+	−	+	+	−	+	y_8
9	+	+	+	−	−	−	+	−	+	+	−	y_9
10	−	+	+	+	−	−	−	+	−	+	+	y_{10}
11	+	−	+	+	+	−	−	−	+	−	+	y_{11}
12	−	−	−	−	−	−	−	−	−	−	−	y_{12}

Table 6
Fractional factorial designs applied in the guideline[a]

No. of factors	Selected design	Generators	Expansion generators	No. of experiments (N)
(a) *Minimal designs*				
3	Full factorial: 2^3	–	–	8
4	1/2th fraction factorial: 2^{4-1}	D = ABC	–	8
5	1/4th fraction factorial: 2^{5-2}	D = AB E = AC	D = −AB E = −AC	8
6	1/8th fraction factorial: 2^{6-3}	D = AB E = AC F = BC	D = −AB E = −AC F = −BC	8
7	1/16th fraction factorial: 2^{7-4}	D = AB E = AC F = BC G = ABC	D = −AB E = −AC F = −BC G = ABC	8
8	1/16th fraction factorial: 2^{8-4}	E = ABC F = BCD G = ABD H = ACD	ng	16
9	1/32th fraction factorial: 2^{9-5}	ng	ng	16
10	1/64th fraction factorial: 2^{10-6}	ng	ng	16
11	1/128th fraction factorial: 2^{11-7}	ng	ng	16
12	1/256th fraction factorial: 2^{12-8}	ng	ng	16
13	1/512th fraction factorial: 2^{13-9}	ng	ng	16
14	1/1024th fraction factorial: 2^{14-10}	ng	ng	16
15	1/2048th fraction factorial: 2^{15-11}	ng	ng	16
(b) *Designs for statistical interpretation of effects from two-factor interactions*				
3	Full factorial: 2^3	–		8
4	1/2th fraction factorial: 2^{4-1}	D = ABC		8
5	1/2th fraction factorial: 2^{5-1}	E = ABCD		16
6	1/4th fraction factorial: 2^{6-2}	E = ABC F = BCD		16
7	1/8th fraction factorial: 2^{7-3}	E = ABC F = BCD G = ABD		16
8	1/16th fraction factorial: 2^{8-4}	E = ABC F = BCD G = ABD H = ACD		16

[a] ng, not given.

The designs from Table 6b have the following characteristics: (i) the two-factor interactions [22,23] are not confounded with the main effects, i.e. the design resolution is at least IV; and (ii) at least three two-factor interaction effects can be estimated. The designs described in Table 6b are those with the lowest number of experiments that still fulfil these requirements. Designs with more than eight factors and fulfilling the above requirements are not given because they require at least 32 experiments. The two-factor interaction effects that can be estimated from the fractional factorial designs of Table 6b, are used to estimate the experimental error on the effects in these latter designs (see Section 6.2.2.2).

4. Experimental work

4.1. Execution of trials

Aliquots of the same test sample and standard(s) are examined at the different experimental conditions. In case there is a large range of concentrations to be determined (factor 100 or more) several concentrations could be examined.

The design experiments are preferably performed in a random sequence. For practical reasons experiments may be blocked (sorted) by one or more factors. This means that for the blocked factor first all experiments where it is at one level are performed and afterwards those at the other. Within the blocks the experiments are randomised. Even though blocking is often used this way of working can contain some pitfalls. Indeed, if drift (time effect) occurs the estimated effect(s) of the blocked factor(s) will be affected by the drift [24,29,30]. If blocking is performed, at least a minimal check for drift is recommended. With drift is meant that a response measured at constant conditions (e.g. nominal ones) is changing (increasing or decreasing) as a function of time.

Blocking by external factors not tested in the design such as, for instance, days is also possible. When 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 [30].

4.2. Minimising the influence of uncontrolled factors

A method can be subject to unavoidable drift. For instance, all HPLC columns are ageing and as a consequence some responses drift as a function of time. A robustness test on methods with drifting responses is still useful since it will indicate whether or not other factors affect the response. However, some of the estimated factor effects are corrupted when they are calculated from the measured data without taking some precautions, such as: (i) correcting for the drift using replicated (nominal) experiments; (ii) confounding the time effects (due to the drift of the response) with dummy variables effects; or (iii) confounding the time effects with non-significant interactions in fractional factorial designs.

4.2.1. Using replicated experiments

A number of additional experiments, usually at nominal levels, can be added to the experimental design experiments to complete the experimental set-up. These replicate experiments are performed before, at regular time intervals between, and after the robustness test experiments of the Plackett–Burman design. The simplest possibility is to carry out two replicate experiments, one before and after the design experiments. 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 estimate for drift; (iii) to correct the measured results for possible time effects, such as drift; and occasionally (iv) to normalise the effects (see Section 6.1).

4.2.2. Using dummy variables

Instead of correcting for time effects, one can minimise their influence on the factor effect estimates by executing the design experiments in a well-defined sequence. Such anti-drift sequences have been defined for full factorial designs [29] and for some fractional factorial ones [30]. When the design experiments are executed in this sequence the factor effects are not or minimally influenced by the drift (at least when the drift is linear) because then the time effect is confounded with the interaction effects [30].

In analogy with the fractional factorial designs one could minimise in Plackett–Burman designs the influence of an occasional time effect on the real factor effects by selecting an experimental sequence that confounds the time effect maximally with the dummy factors. In practice, one will construct a Plackett–Burman design and evaluate for that design which columns are most affected by drift. These can be determined by awarding, for each column (factor), to the experiment number its corresponding sign, being (+) or (−), and summing the resulting values. For instance, in Table 5, for factor A, experiment one gets value $+1$; two gets -2; three gets $+3$ and so on (-4, -5, -6, $+7$, $+8$, $+9$, -10, $+11$, -12). The resulting sum for factor A is zero.

The columns with the highest absolute results are affected most by drift and to these columns the dummy factors will be attributed.

Example: For the design shown in Table 5 the values 0, -10, 2, -8, -18, -28, -16, -4, 8, -2 and 10 are obtained for columns A, B, C, D, E, F, G, H, I, J and K, respectively. One will award a first dummy factor to column F, a second to E, a third to G, and so on.

5. Determining responses

5.1. Responses measured in a robustness test

From the experiments performed, a number of responses can be determined. For chromatographic methods, responses describing a quantity such as the content of main substance and by-products, and/or peak areas or peak heights are the more evident. The evaluation of the content can hide certain effects: indeed, when a factor has a similar effect on the peak area/height of the sample and standard(s), this effect will not be seen anymore in the content. Therefore, the evaluation of both peak areas/heights and contents can indicate different factors as important and is to be preferred. An alternative for the study of areas/heights is to calculate the content from the area/height of the sample measured for the different design experiments relative to the result(s) of the standard(s) measured at nominal conditions.

For a separation method one should also consider one or more parameters describing the quality of the separation, such as, for example, the resolution or the relative retention. The evaluation of these separation parameters can also lead to system suitability test (SST) limits as required by the ICH. When determining SST-limits, other responses such as capacity factors or retention times, asymmetry factors and number of theoretical plates can also be studied (see Section 8).

5.2. Corrected response results

If one checked for drift, e.g. by replicate nominal experiments, corrected response results can be calculated from the measured results. The corrected design results are calculated as

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

where $i = 1, 2, \ldots, p$ and p is the number of design experiments between two consecutive nominal experiments, $y_{i,\mathrm{corrected}}$ is a corrected design result, $y_{i,\mathrm{measured}}$ the corresponding measured design result, $y_{\mathrm{nom,begin}}$ the nominal result at the beginning of the experiments (before design), $y_{\mathrm{nom,before}}$ and $y_{\mathrm{nom,after}}$ the nominal results measured before and after the design result for which one is correcting. Eq. (1) is only correct if the hypothesis can be accepted that the experiments were performed equidistant in time.

6. Analysis of the results

6.1. Calculation of effects

Effects can be calculated both from the measured and the corrected response results. The two effects estimated for a factor are similar for factors not affected by drift and different from those that are. For each factor its effect is calculated according to the equation

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

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

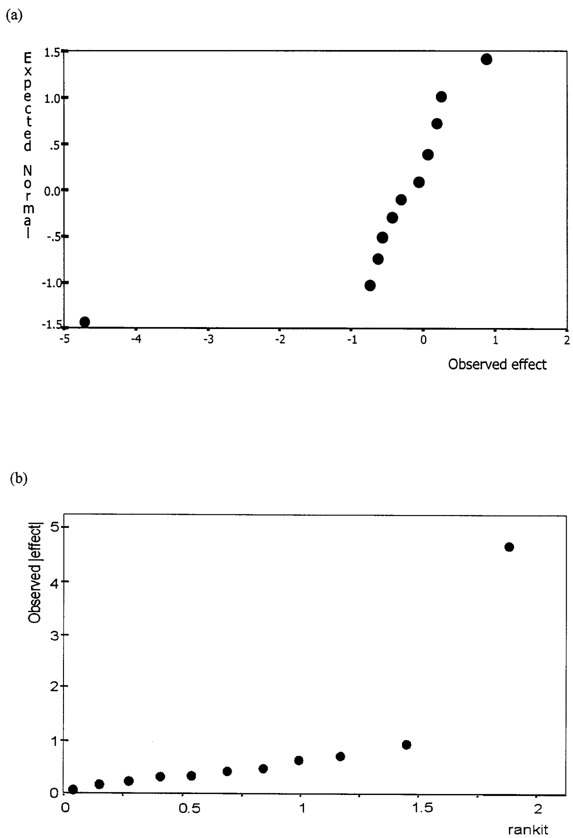

Fig. 2. (a) Normal probability plot; and (b) Half-normal probability plot for the effects estimated from a $N = 12$ Plackett–Burman design.

The effects can also be normalised relative to the average nominal result ($\bar{Y}$), in case the response is not drifting, or to the nominal result measured before the design experiments, when the response is drifting [26]. Usually normalised effects, much more than the regular effect estimates, allow the user of the method to consider the influence of a factor as important, even without statistical interpretation.

$$E_X(\%) = \frac{E_X}{\bar{Y}} \times 100\%. \qquad (3)$$

6.2. Interpretation of effects

If an estimate of the experimental error of the effects is available with a proper number of degrees of freedom, it is recommended to perform a statistical test [13,15,24,26,31]. Graphical approaches can also be used to obtain an acceptable interpretation. However, we recommend using both, if possible.

6.2.1. Graphical interpretation

The graphical identification of important effects is usually applied with a normal probability plot [22–24,28] or a half-normal plot [24,31]. Non-significant effects are normally distributed around zero. Both in a normal probability and a half-normal plot these non-significant effects tend to fall on a straight line through zero, while significant effects deviate from it. An example of a normal probability plot and of a half-normal plot is shown in Fig. 2. The normal probability and the half-normal plots lead to similar conclusions. Only the construction of the half-normal plot is described below.

To create the half-normal plot, the n effects are ranked according to increasing absolute effect size. The rth value of that sequence is plotted against a scale defined by partitioning the right half of the normal distribution in n parts of equal area, and by taking the median of the rth slice. This value is called the rankit. In Table 7 the rankits are given for the most frequently executed Plackett–Burman designs. The effects that are derived from a robustness test design are then plotted against the corresponding rankits. Several

Table 7
Rankits to draw a half-normal plot for the most frequently used screening designs (effect '1' indicates the smallest effect)

Effect	Design size		
	$N=8$	$N=12$	$N=16$
1	0.09	0.06	0.04
2	0.27	0.17	0.12
3	0.46	0.29	0.21
4	0.66	0.41	0.29
5	0.90	0.53	0.38
6	1.21	0.67	0.47
7	1.71	0.81	0.57
8		0.98	0.67
9		1.19	0.78
10		1.45	0.89
11		1.91	1.02
12			1.18
13			1.36
14			1.61
15			2.04

commercial statistical software packages allow the construction of normal probability or of half-normal plots.

6.2.2. Statistical interpretations

The statistical interpretation provides the user a numerical limit value that can be plotted on the graphical representation (e.g. the half-normal plot) and that allows to define, in a less subjective way than the visual one, what is significant and what is not. This limit value to identify statistically significant effects is usually derived from the t-test statistic [2,13,16,24,31,32].

$$t=\frac{|E_X|}{(\mathrm{SE})_\mathrm{e}} \Leftrightarrow t_{\mathrm{critical}}, \tag{4}$$

with $(\mathrm{SE})_\mathrm{e}$, the standard error of an effect, which represents the experimental variability within the design. For robustness experiments, this $(\mathrm{SE})_\mathrm{e}$ can be estimated in different ways (see below). The statistic given in Eq. (4) can be rewritten as

$$|E_X| \Leftrightarrow E_{\mathrm{critical}}=t_{\mathrm{critical}}(\mathrm{SE})_\mathrm{e} \tag{5}$$

or normalised relative to the response value (Y_n)

$$|\%E_X| \Leftrightarrow \%E_{\mathrm{critical}}=\frac{E_{\mathrm{critical}}\cdot 100\%}{Y_n}. \tag{6}$$

The critical effect (E_{critical}) is usually calculated at a significance level $\alpha=0.05$ (occasionally 0.01 or 0.1). An effect is considered significant at a given α level if $|E_X|>E_{\mathrm{critical}}$.

For the statistical interpretation of effects different ways of estimating the error of an effect are described [24,31]. Three approaches are retained in this guideline since they were found to give acceptable and physically relevant results [24,26,27,31,33–35].

1. An intermediate precision estimate of the measurement error is available (Section 6.2.2.1).
2. An error estimate on an effect is obtained from the dummy factors or from two-factor (or multiple-factor) interactions (Section 6.2.2.2).
3. The error of an effect is estimated from the distribution of the effects themselves (Section 6.2.2.3).

6.2.2.1. Estimation of error from intermediate precision estimates.

The standard error on an estimated effect is calculated according to the equation for the standard error on a difference of means:

$$\mathrm{SE}=\sqrt{\frac{s_a^2}{n_a}+\frac{s_b^2}{n_b}}, \tag{7}$$

where s_a^2 and s_b^2 estimate the variances of the two sets of measurements and n_a and n_b are the numbers of measurements of those sets. The standard error of an effect thus becomes

$$(\mathrm{SE})_\mathrm{e}=\sqrt{\frac{s^2}{N/2}+\frac{s^2}{N/2}}=\sqrt{\frac{4s^2}{N}}, \tag{8}$$

since s_a^2 and s_b^2 are estimated by the same variance, s^2, and $n_a=n_b=N/2$.

The variance s^2 can be determined from replicated experiments at nominal levels or from duplicated design experiments. The t_{critical} is a tabulated t-value at R degrees of freedom where R is the number of degrees of freedom with which s^2 is estimated. When using replicated nominal experiments $R = n - 1$ with n being the number of replicated experiments, while for duplicated design experiments $s^2 = \Sigma_{i=1}^{N} d_i^2 \big/ 2N$ and $R = N$, with d_i being the differences between the duplicated experiments.

With this criterion, only estimates of s^2 obtained under intermediate precision conditions lead to relevant conclusions [27]. With intermediate precision conditions is meant here that at least the factor time was varied, i.e. the measurements were performed on different days.

Thus, this criterion can only be used if intermediate precision estimates are available which often is not yet the case when a robustness test is performed. Secondly, it is not always practically feasible to duplicate design experiments under intermediate precision conditions given the increased workload.

6.2.2.2. Estimation of error from dummy effects or from two-factor interaction effects. An estimate of $(SE)_e$ can also be obtained from dummy or interaction effects, i.e. effects that are considered negligible. The following equation is used

$$(SE)_e = \sqrt{\frac{\sum E_{\text{error}}^2}{n_{\text{error}}}}, \tag{9}$$

where $\Sigma E_{\text{error}}^2$ is the sum of squares of the n_{error} dummy or interaction effects. The $(SE)_e$ is then used in Eq. (4) or Eq. (5) to perform the statistical test.

One should be aware of the low power of a statistical test with few degrees of freedom. If dummy factors are used, a design which contains at least three dummy factors should be selected (Table 3). The minimal designs described above have no, or at least not a sufficient, number of degrees of freedom to test the effects on their significance using the dummy effects. As a consequence the power of the t-test to detect any significance is low. In these cases the algorithm of Dong is to be preferred (see Section 6.2.2.3).

One should also take into account the fact that in some situations the dummy effects are potentially affected by the drift (see Section 4.2.2). These should be eliminated from the estimation of $(SE)_e$, since they are not necessarily representing non-significant effects anymore.

6.2.2.3. Estimation of error from the distribution of effects (algorithm of Dong). For small designs the algorithm of Dong [36] is a suitable tool to identify significant effects. This algorithm can be used for all screening designs of Tables 3 and 6, including the minimal designs. Other algorithms were also proposed [37–40] but are not considered here.

In the approach of Dong, an initial estimate of the error on an effect is obtained in the following way.

$$s_a = 1.5 \cdot \underset{i}{\text{median}} \; |E_i|, \tag{10}$$

where E_i is the value of effect i. The value 1.5 in Eq. (10) is appropriate for a random variable that follows a normal distribution $N(0, \sigma^2)$. When the E_i are independent realisations of this distribution the median of the absolute effects $|E|$ is namely about 0.675σ [39].

From s_0, a final estimation of the standard error (s_1) is derived as

$$s_1 = \sqrt{m^{-1} \sum E_i^2} \quad \text{for all } |E_i| \leq 2.5 s_0, \tag{11}$$

where m is the number of absolute effects smaller than $2.5s_0$. By using s_1 instead of s_0, one avoids overestimating the error. The constraint of eliminating effects exceeding the $2.5s_0$ limit follows from the fact that $P(|E| > 2.5\sigma) \approx 0.01$.

Next, the s_1 value is used to calculate a so-called margin of error (ME) which is a critical effect.

$$\text{ME} = t_{(1-\alpha/2,\text{df})} \cdot s_1, \tag{12}$$

where $1 - \alpha/2 = 0.975$ and $\text{df} = m$. The ME is statistically a valid criterion for significance testing when only one effect has to be tested. When multiple effects are tested the chance for non-significant effects that exceed the ME increases. To compensate for these events, statistically, the significance level has to be adjusted and a second limit is defined, the simultaneous margin of error (SME).

$$\mathrm{SME} = t_{(1-\alpha^*/2,\mathrm{df})} \cdot s_1, \qquad (13)$$

where $\alpha^* = 1 - (1-\alpha)^{(1/m)}$, the Sidak adjusted significance level [41].

The ME and SME are critical effects similar to the definition of Eq. (5). According to the literature, an effect that exceeds the ME, but is below the SME, is called to be possibly significant and an effect that is above the SME, is considered to be significant [36]. However, in practice, the ME limit is recommended to be used as the decision criterion for all effects calculated from the robustness test, even though there is an increased chance for false positive decisions. Indeed, using the SME reduces the probability for such false positives (α-error), i.e. the indication of a non-significant effect as significant, but on the other hand it largely increases the number of false negatives (β-error), i.e. the number of significant effects that are considered to be non-significant. This latter means that a method will be considered robust while in fact it is not. Therefore, it can be recommended not to use the SME limit in robustness testing.

If the number of significant effects is large, the algorithm of Dong is iterated, i.e. s_0 is replaced by s_1 and a new s_1^* is calculated. For the robustness strategy as proposed in this guideline, iteration will not be necessary because it is assumed that significant effects are sparse [31]. Therefore, we recommend applying Dong's method in its most simple form for interpretation of the normal probability and half-normal plots.

The assumption of effect sparsity (i.e. $< 50\%$ of the effects are significant) can be a disadvantage. This assumption is usually valid for a response such as the content determination, but it is not necessarily the case for responses describing the separation or the quality of a chromatogram such as for instance resolution, capacity factor, asymmetry factor (see Section 8). If problems with the algorithm of Dong are expected, one should prefer the approach that uses dummy variables for statistical interpretation, although this increases the design size.

Table 8
Effects from a seven-factor Plackett-Burman design (case study extracted from [15])

Factor (definition)	|Effect|	Rankit
Standard concentration (D)	0.075	0.09
Ionic strength buffer (A)	0.795	0.27
pH of buffer (G)	0.860	0.46
Flow mobile phase (B)	0.860	0.66
Mobile phase composition (E)	0.970	0.90
Detection wavelength (F)	1.035	1.21
Age of the standard (C)	4.945	1.71

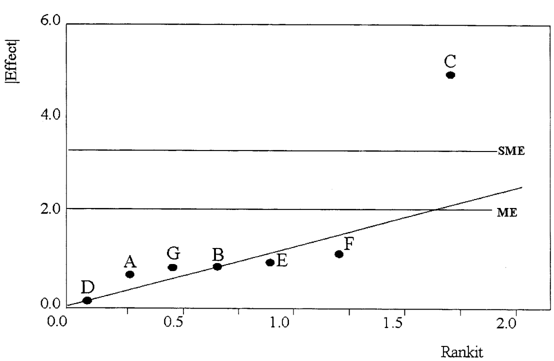

Fig. 3. Half-normal probability plot for the effects of Table 8 with identification of the critical effects ME and SME.

Example of the application of Dong's algorithm: A case study is described to demonstrate the principle of the half-normal plot and the algorithm of Dong. In this study a robustness test is applied to the determination of phenylbutazone with HPLC [15]. In the robustness test a $N = 8$ Plackett–Burman design for seven factors was used. There are no dummy effects available to estimate the $(\mathrm{SE})_e$. A graphical approach in combination with the algorithm of Dong is used to identify significant effects. The effects on the content of phenylbutazone, resulting from the Plackett–Burman design, are reported in Table 8.

The rankits of Table 7 ($N = 8$) are used to build the half-normal plot presented in Fig. 3. The median of the effects for this example is 0.860. From Eqs. (10) and (11) the following results are derived

$s_0 = 1.5 \times 0.860 = 1.29$

and

$$s_1 = \sqrt{(0.075^2 + 0.795^2 + 0.860^2 + 0.860^2 + 0.970^2 + 1.035^2)/6} = 0.830.$$

In the calculation of s_1 the effect of factor C (age of the reference solution) is left out, because this effect is larger than $2.5 \times s_0$. Further, Eqs. (12) and (13) are applied to define the ME and SME limits ($\mathrm{ME} = t_{(0.975,6)} \times 0.830 = 2.45 \times 0.830 = 2.03$). For the SME limit the significance level α is adapted using the adjustment as defined by Sidak [41], which results in a corrected significance level equal to $\alpha = 0.0085$, i.e. $(1-(1-0.05)^{1/6})$. This gives $\mathrm{SME} = t_{(0.996,6)} \times 0.830 = 3.98 \times 0.830 = 3.30$. In Fig. 3 the half-normal plot is presented, with the ME and SME limits included. It indicates that all effects are more or less equal to random error, except for the effect of factor C. This factor has a significant effect when the factor levels are varied in the interval as specified in the robustness test. It is evident, for this example, that factor C deviates from the straight line. However, the interpretation is not always so obvious. The algorithm of Dong or one of the other statistical interpretation methods are then helpful tools.

6.3. Estimating non-significance intervals for significant factors

When a factor has a significant effect on a response, one can wonder in which interval the factor levels should be controlled to eliminate the effect. These levels can be estimated as

$$\left[X_{(0)} - \frac{|X_{(1)} - X_{(-1)}|E_{\mathrm{critical}}}{2|E_X|},\ X_{(0)} + \frac{|X_{(1)} - X_{(-1)}|E_{\mathrm{critical}}}{2|E_X|}\right], \tag{14}$$

where $X_{(0)}$, $X_{(1)}$ and $X_{(-1)}$ are the real values of factor X for the levels $(0), (1)$ and (-1), respectively.

Example: Suppose the factor pH of a buffer was examined in the interval 6.5–7.1 with the nominal pH being 6.8, and its influence on the resolution was evaluated. A significant effect ($E_X = 0.427$) was found with $E_{\mathrm{critical}} = 0.370$. Non-significant factors levels for pH are then estimated as

$$\left[6.8 - \left(\frac{|7.1-6.5|*0.370}{2*|0.427|}\right),\ 6.8 + \left(\frac{|7.1-6.5|*0.370}{2*|0.427|}\right)\right] = [6.54; 7.06].$$

When the pH is controlled within this interval (e.g. within 6.6–7.0) no significant effect of the pH on the resolution will be found anymore.

It is evident that: (i) such levels can be calculated only for quantitative factors; (ii) the extreme levels must be symmetrically situated around the nominal one; and (iii) a linear behaviour of the response as a function of the factor levels is assumed.

7. The derivation of system suitability limits from robustness test results

A system suitability test (SST) is an integral part of many analytical methods [7]. It ascertains the suitability and effectiveness of the operating system [9]. The SST-limits usually are established based on the experimental results obtained during the optimisation and validation of a method and on the experience of the analyst. However, the ICH guidelines state that *one consequence of the evaluation of robustness should be that a series of system suitability parameters (e.g. resolution tests) is established to ensure that the validity of the analytical procedure is maintained whenever used*. Therefore, we propose to use the results of the worst-case situations to define the SST-limits, e.g. for resolution [42].

Notice that we recommend determining the SST-limits only when the method can be considered robust for its quantitative assay. In that case it can be expected that in none of the points of the experimental domain, including those at which certain (system suitability) responses have their worst result, there would be a problem with the quantitative response. Of course, the hypothesis that the worst case conditions do not affect the quantitative results can easily be verified in practice. This is further discussed in Section 8.

Beside the recommendation of the ICH guidelines, there are also practical reasons for defining SST-limits based on the results of a robustness test. From experience it was seen that the SST limits selected independently from the results of a robustness test, frequently are violated when the method is transferred. This is due to the fact that they are chosen too strictly and relatively arbitrarily based on the experience of the analyst in the optimisation laboratory. On the other hand, it is neither considered desirable to choose as SST-limit the most extreme value that still allows a quantitative determination. For instance, when the operational conditions after method development give a resolution of about six, a resolution of two is not considered acceptable, even if quantification still seems possible. It is namely important to maintain the method at all times around the conditions at which it is optimised and validated. Therefore, it is considered preferable to derive the system suitability limits from the robustness test, since there the most extreme variations in the factors that still are probable under acceptable conditions, are examined.

These worst-case conditions are predicted from the calculated effects. The worst-case situation is then the factor (level) combination giving for instance the lowest resolution. For responses like the capacity factor it is the one causing the smallest result, while for the tailing factor it is usually the situation resulting in the highest value. To define the worst-case conditions only the statistically significant factors (at $\alpha = 0.05$) and the ones that come close to it (significant at $\alpha = 0.1$) are considered. The factors not significant at $\alpha = 0.1$ are considered negligible and their effects are considered to originate only from experimental error. As the experimental designs proposed in this guideline are saturated two-level designs, only linear effects for the maintained factors are considered in the prediction of the worst-case situation which is acceptable since in robustness testing only a restricted domain of the response surface is considered. The factor level combination for which the equation

$$Y = b_0 + \frac{E_{F_1}}{2} \cdot F_1 + \frac{E_{F_2}}{2} \cdot F_2 + \cdots + \frac{E_{F_k}}{2} \cdot F_k \qquad (15)$$

predicts the worst result is derived. In Eq. (15) Y represents the response, b_0 the average design result, E_{F_i} the effect of the factor considered for the worst-case experiment and F_i the level of this factor (-1 or $+1$). Non-important factors are kept at nominal value ($F_k = 0$). An example of calculation is given in Section 8.

The SST-limit can experimentally be determined from the result of one or several experiments performed at these conditions, or it can be predicted. When the experiment is replicated the SST-limit can be defined as the upper or lower limit from the one-sided 95% confidence interval [13] around the worst case mean. For resolution and capacity factor, for instance, the lower limit would be chosen, while for the tailing factor it would be the upper one. The confidence interval is defined as $[\bar{Y}_{\text{worst-case}} - t_{\alpha,n-1}(s/\sqrt{n}), \infty]$ when the lower limit has to be considered and as $[0, \bar{Y}_{\text{worst-case}} + t_{\alpha,n-1}(s/\sqrt{n})]$ when it is the upper one.

If no significant effects were occurring for a response then its SST limit can be determined analogously to the above situation but the measurements will be executed at nominal conditions.

A less strict and easier alternative is to define the (average) worst-case result ($\bar{Y}_{\text{worst-case}}$) as the SST-limit.

Finally, one could estimate the SST-limit from the theoretical model of Eq. (15) without performing additional experiments.

8. Example of a robustness test performed according to this guideline

The case study described concerns the robustness testing of the high performance liquid chromatographic (HPLC) method for the identification and assay of an active substance (main compound, MC) and for the detection of two related compounds (RC1 and RC2) in tablets (extracted from [42]).

8.1. Nominal conditions

The analysis method uses an external standard without tablet placebo. The solutions used in this robustness test are: (i) a standard solution containing 125 mg l^{-1} of RC1 and of RC2 in

Table 9
Composition of the mobile phase during the solvent gradient (% volume fractions)[a]

Solvent	Time (min)				
	0	13	15	17	22
A	50	50	50	50	50
B	25	43	43	25	25
C	25	7	7	25	25

[a] A, 0.25% ammonium acetate in water; B, acetonitrile; C, water.

methanol; (ii) a reference solution containing per 100 ml, 25.0 mg of MC reference material and 1.0 ml of 'standard solution (i)' in a mixture methanol/0.25% ammonium acetate in water (9:1, v/v); (iii) a sample solution containing per 100 ml, 25.0 mg of MC reference material, 1.0 ml of the 'standard solution (i)' and 10 placebo formulation tablets in a mixture methanol/0.25% ammonium acetate in water (9:1, v/v); and (iv) a blank solution containing methanol/0.25% ammonium acetate in water (9:1, v/v). The mixtures to prepare solutions (ii) and (iii) are mechanically shaken for 30 min, diluted to volume and filtered through a 0.45 μm chemical resistant Acrodisc-filter. Notice that the sample solution represents a tablet simulation.

Chromatographic conditions: a 10 cm × 4.6 mm I.D. column, packed with Hypersil BDS-C18, 3 μm particle size is prescribed. The substances are eluted in a gradient elution mode at a flow rate of 1.5 ml min^{-1} and at ambient temperature. The solvent gradient used is shown in Table 9. Detection of the eluted substances is done spectrophotometrically at 265 nm. The injection volume is 10 μl.

8.2. Selection of factors and levels

The factors investigated in the robustness evaluation of the HPLC method for identification and assay of MC and its related compounds in tablet simulations are summarised in Table 10.

Both quantitative and qualitative (column manufacturer) factors are examined in the robustness test. For the qualitative factor the nominal column is also used as one of the 'extreme' levels since it is more logical to compare the nominal column with another one, than to compare two columns different from the nominal one with each other. Notice that inclusion of only two columns in the robustness study does not allow to draw any conclusion about the total population of columns, i.e. about the robustness of the method on the particular type of columns to which the two selected ones belong. Only conclusions about the robustness of the method on the two examined columns can be drawn.

The indication, in the operating procedure, of the column temperature as ambient could be insufficient if the temperature is an important factor since ambient temperature can vary largely from

Table 10
Factors and levels investigated in the robustness test[a1]

Factor	Units	Limits	Level (−1)	Level (+1)	Nominal
1. The flow of the mobile phase	ml min^{-1}	±0.1	1.4	1.6	1.5
2. The pH of the buffer	–	±0.3	6.5	7.1	6.8
3. The column temperature	°C	±5	23	33	ambient
4. The column manufacturer			Alltech	Prodigy	Alltech
5. Percentage organic solvent (%B) in the mobile phase at the start of the gradient	%	±1	24	26	25
6.%B in the mobile phase at the end of the gradient	%	±2	41	45	43
7. Concentration of the buffer	% m v^{-1}	±0.025	0.225	0.275	0.25
8. The wavelength of the detector	nm	±5	260	270	265

[a] Alltech, Alltech Hypersyl 3 μm BDS C18; and Prodigy, Phenomenex Prodigy 3 μm ODS (3) 100 A C18.

Table 11
The selected Plackett–Burman design[a1]

Experiment No.	Factors										
	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*
	pH	Column	Dum1	Temp	%*B* begin	%*B* end	Dum2	Flow	Wave-length	Buffer-conc.	Dum3
1	1[c1]	1	1	−1[b1]	1	1	−1	1	−1	−1	−1
2	1	1	−1	1	−1	−1	−1	1	1	1	−1
3	1	−1	1	1	−1	1	−1	−1	−1	1	1
4	1	−1	−1	−1	1	1	1	−1	1	1	−1
5	1	−1	1	−1	−1	−1	1	1	1	−1	1
6	−1	1	1	1	−1	1	1	−1	1	−1	−1
7	−1	1	−1	−1	−1	1	1	1	−1	1	1
8	−1	−1	−1	1	1	1	−1	1	1	−1	1
9	−1	−1	1	1	1	−1	1	1	−1	1	−1
10	−1	1	1	−1	1	−1	−1	−1	1	1	1
11	1	1	−1	1	1	−1	1	−1	−1	−1	1
12	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1

[a] *Abbreviations:* pH, pH of the buffer; Column, column manufacturer; Dum1, dum2, dum3, dummy variables; Temp, column temperature; % *B* begin, percentage B in the mobile phase at the start of the gradient; % *B* end, percentage B in the mobile phase at the end of the gradient; Flow, flow of the mobile phase; Wavelength, wavelength of the detector; Buffer conc., concentration of the buffer.

[b] −1, low factor level.

[c] 1, high factor level.

Table 12
Results of the experiments for the studied responses

Experiment	Responses						
	%MC	%RC1	%RC2	R_s(MC–RC1)	k'(MC)	Asf(MC)	t_R(RC2)
1	101.6	100.9	101.4	5.691	3.800	0.813	11.500
2	101.7	101.2	102.7	7.484	5.083	1.031	13.000
3	101.6	101.7	101.3	5.770	4.000	1.453	9.833
4	101.9	103.0	102.9	5.025	3.167	1.549	9.483
5	101.8	99.3	99.1	5.440	3.800	1.458	10.317
6	101.1	99.9	101.7	5.711	5.817	0.861	12.567
7	101.1	100.8	101.4	5.932	5.250	0.836	12.083
8	101.6	100.2	98.8	4.962	3.200	1.059	8.417
9	98.4	97.1	101.8	5.427	3.367	0.977	9.200
10	99.7	100.5	99.3	6.344	5.350	0.853	13.800
11	99.7	98.6	98.7	6.715	4.783	0.920	13.317
12	102.3	101.1	103.1	5.186	4.933	1.412	11.150
Mean	101.0	100.4	101.0	5.807	4.379	1.102	11.222
RSD	1.15	1.52	1.61				

one laboratory to another and it has to be standardised in that case.

For the selection of the levels the uncertainties in the nominal factor levels multiplied with a constant (k), as specified in Section 2.2.1, are used. When the percentage organic modifier was changed in factors (5) and (6) of Table 10, water (solvent C) was used as adjusting compound.

8.3. Selection of the experimental design

The eight factors that were selected from the operating procedure were examined in a Plackett–Burman design for 11 factors ($N = 12$). Therefore, three dummy factor columns have to be included. This was done randomly. The selected design is shown in Table 11.

Table 5 and Table 11, at first sight, seem to be different. This is however, not the case. Table 5 was constructed, as described in Section 3, starting from the first line given by Plackett–Burman. The design of Table 11 is the one generated by a statistical software package. With this we would like to indicate that when someone is using its own available software to select or create a design, the sequences of rows and columns are not necessarily the ones given in this guideline, though the final designs are equivalent.

8.4. Execution of the trials

No additional nominal experiments were added to the experimental set-up. For each of the 12 experimental design runs, three injections are performed: (i) a blank injection; (ii) an injection of the reference solution; and (iii) an injection of the sample solution. With this set-up it is assumed that in practice the sample and standard, used to determine the sample content are analysed under identical experimental conditions (see Section 5).

8.5. Responses determined

The responses determined in this robustness test are: (i) the percent recoveries of MC, RC1 and RC2; (ii) the resolution (R_s) of the critical peak pair, which is MC and RC1; (iii) the capacity factor (k') of MC; (iv) the tailing or asymmetry factor (Asf) of MC; and (v) the analysis time given as the retention time (t_R) of the last eluting substance RC2. Table 12 shows the experimentally obtained design values for the responses that are studied.

8.6. Calculation of effects

The effects of the different factors on the considered responses are shown in Table 13a. Since in this case study, no additional nominal experiments are performed, no normalised effect values are calculated.

8.7. Graphical interpretation of effects

Normal probability and half-normal plots are drawn with the effects estimated for the different responses. They are shown in Fig. 4. From these plots it can be observed that the interpretation of these plots is not always straightforward and it can be recommended to combine them with a statistical interpretation. This latter interpretation allows to draw the critical effects on the plots, as was for instance done in Fig. 3. In Fig. 4 no critical effects were drawn on the plots since actually they belong to the statistical interpretation of the effects.

In both types of plots, the visual identification of important effects becomes less evident as the total number of plotted effects decreases (i.e. for smaller designs). It is not always obvious to draw the line formed by the non-significant effects. The graphical interpretation becomes more interesting when the number of estimated effects is large and only a limited number is expected to be signifi-cant. The plots also can be used to indicate suspect dummy factor effects which are relatively high and that are possible outliers to the population of non-significant effects (cf. E_{dum3} on %RC2) and therefore occasionally, can be eliminated from the statistical interpretation (cf. further).

Table 13
(a) Effects of the factors on the different responses; and (b) critical effects obtained from the different statistical interpretation methods

Factors	Effects on						
	%MC	%RC1	%RC2	R_s(MC–RC1)	k'(MC)	Asf(MC)	t_R(RC2)
pH	0.683	0.850	0.000	0.427	−0.547	0.204	0.039
Column	−0.450	−0.083	−0.300	1.011	1.269	−0.432	2.978
Dum1	−0.683	−0.917	−0.500	−0.154	−0.047	−0.065	−0.039
Temperature	−0.717	−1.150	−0.367	0.408	−0.008	−0.103	−0.333
%*B* begin	−1.117	−0.617	−1.067	−0.226	−0.869	−0.147	−0.539
%*B* end	0.883	1.450	0.467	−0.584	−0.347	−0.013	−1.150
Dum2	−0.750	−1.150	−0.167	−0.198	−0.030	−0.003	−0.122
Flow	−0.017	−0.883	−0.300	0.031	−0.592	−0.146	−0.939
Wavelength	0.517	0.650	−0.533	0.041	0.047	0.067	0.084
Buffer concentration	−0.617	0.717	1.100	0.380	−0.019	0.029	0.022
Dum3	−0.250	−0.350	−2.500	0.106	0.036	−0.011	0.144

	Critical effects for						
	%MC	%RC1	%RC2	R_s(MC–RC1)	k'(MC)	Asf(MC)	t_R(RC2)
Experimental error estimated from dummy factors							
$\alpha = 0.05$	1.919	2.778	4.694	0.500	0.122	0.121	0.354
Without Dum3			1.604				
$\alpha = 0.1$	1.419	2.054	3.471	0.370	0.090	0.090	0.262
Without Dum3			1.088				
Experimental error estimated from Dong's criterion							
ME ($\alpha = 0.05$)	1.476	1.939	1.307	0.691	0.084	0.228	0.545
ME ($\alpha = 0.1$)	1.205	1.582	1.064	0.562	0.067	0.186	0.440

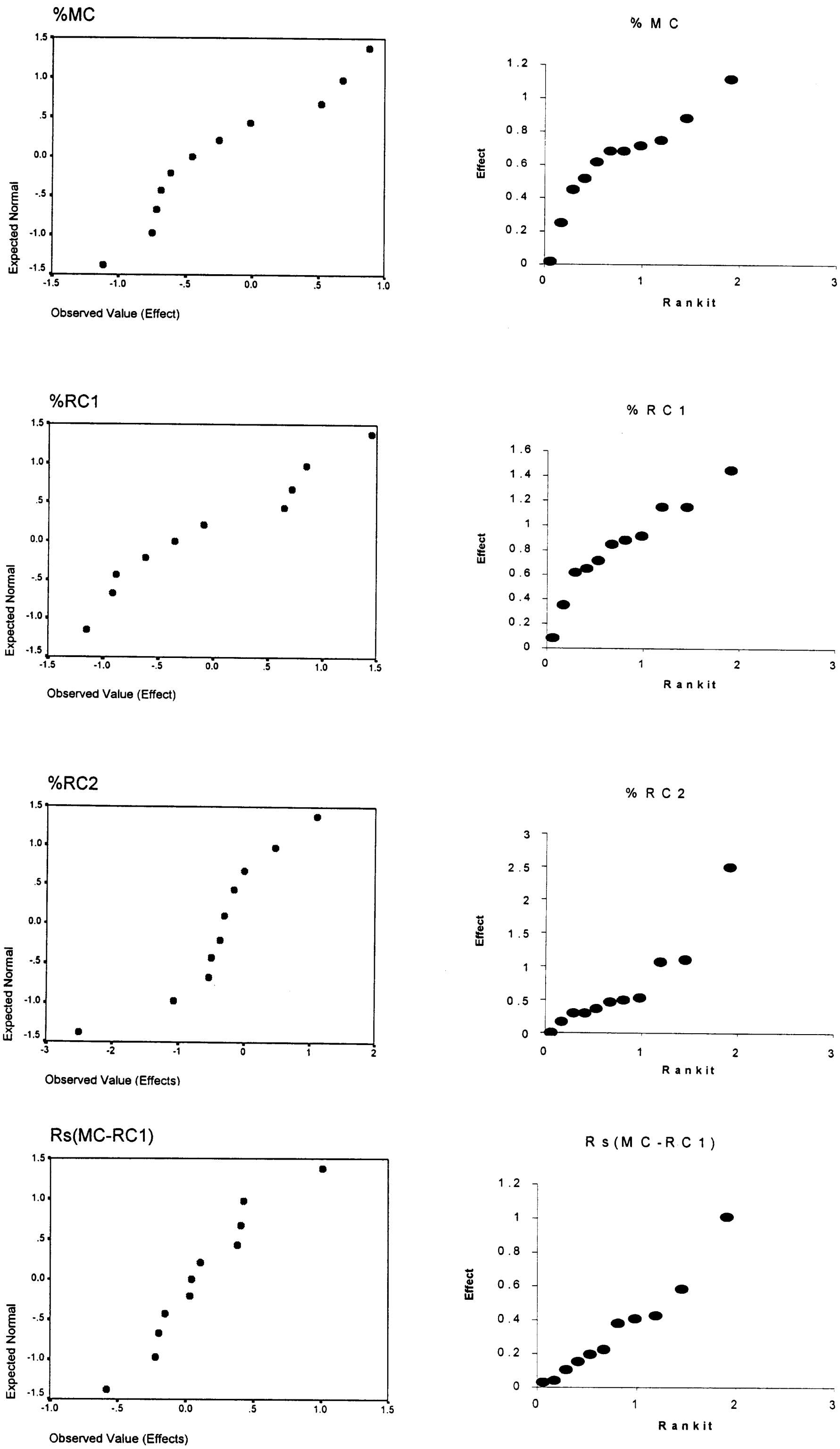

Fig. 4. Normal probability and half-normal plots for the factor effects on the responses of Table 13a.

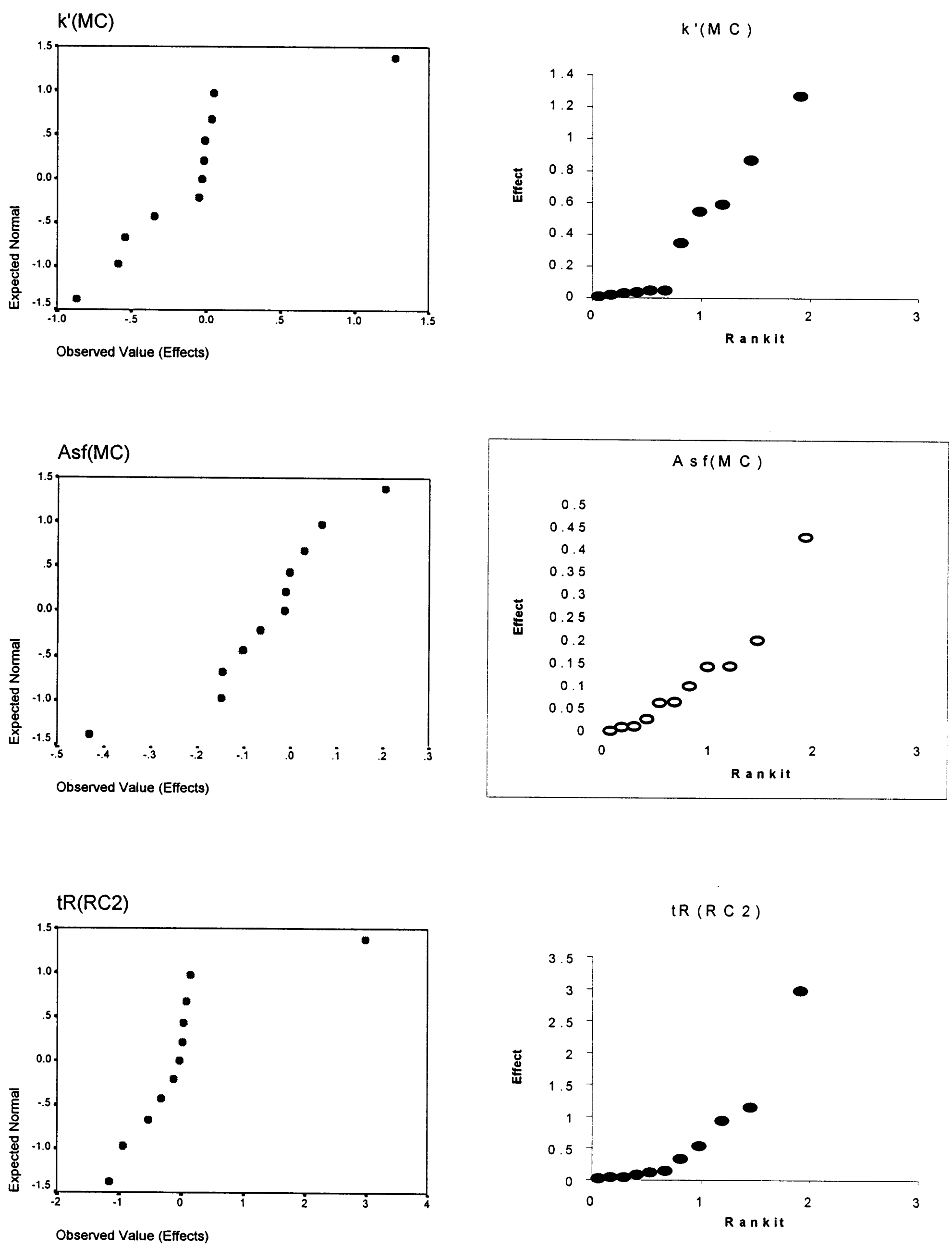

Fig. 4. (*Continued*)

8.8. Statistical interpretation of effects

Two statistical interpretations were performed on this data set: (i) the one in which the experimental error is estimated from the dummy effects (Section 6.2.2.2); and (ii) the one which uses Dong's criterion (Section 6.2.2.3). The criterion based on an intermediate precision estimate (Section 6.2.2.1) is not used since that kind of information was not available at the moment the robustness test was performed. Notice that in general only one statistical interpretation will be applied. Here two are given for comparison. The critical effects obtained with both interpretation methods are shown in Table 13b. The significance of the factor effects according to both interpretation methods is shown in Table 14. It can be observed that the quantitative results of the method, the percent recovery of the substances, are not considered significantly affected by one of the examined factors according to the interpretation using the dummy factor effects to estimate the experimental error.

When Dong's criterion is applied some factors were found to be significant for the%RC2, however, only at $\alpha = 0.1$ level. This difference can be explained by the fact that the effect of one dummy (dum3) is relatively high which affects the critical effect estimated from the dummy effects, while this is not the case for the limit from Dong's criterion. This demonstrates that Dong's criterion is a more robust estimator of the experimental error, when relatively large dummy factor effects occur.

Table 14
Significance of the factor effects on the different responses

Factors	Significance of factor effects on						
	%MC	%RC1	%RC2	R_s(MC–RC1)	k'(MC)	Asf(MC)	t_R(RC2)
(*a*) *When critical effects are estimated using dummy effects*							
pH	–	–	–	a1	b1	b1	–
Column	–	–	–	b1	b1	b1	b1
Dum1	–	–	–	–	–	–	–
Temperature	–	–	–	a1	–	a1	a1
%*B* begin	–	–	–	–	b1	b1	b1
%*B* end	–	–	–	b1	b1	–	b1
Dum2	–	–	–	–	–	–	–
Flow	–	–	–	–	b1	b1	b1
Wavelength	–	–	–	–	–	–	–
Buffer concentration	–	–	– (a1)e1	a1	–	–	–
Dum3	–	–	– (b1)e1	–	–	–	–
(*b*) *When Dong's criterion was used*							
pH	–	–	–	–	d1	c1	–
Column	–	–	–	d1	d1	d1	d1
Dum1	–	–	–	–	–	–	–
Temperature	–	–	–	–	–	–	–
%*B* begin	–	–	c1	–	d1	–	c1
%*B* end	–	–	–	c1	d1	–	d1
Dum2	–	–	–	–	–	–	–
Flow	–	–	–	–	d1	–	d1
Wavelength	–	–	–	–	–	–	–
Buffer concentration	–	–	c1	–	–	–	–
Dum3	–	–	d1	–	–	–	–

[a] Significance at $\alpha = 0.10$ level.
[b] Significance at $\alpha = 0.05$ level.
[c] Significant to ME ($\alpha = 0.10$).
[d] Significant to ME ($\alpha = 0.05$).
[e] Without dum3.

Based on the graphical methods (Fig. 4) one also could have decided to remove dum3 from the statistical interpretation since it seems to be an outlier to the population of non-significant effects. After removal of dum3 from the estimation of $(SE)_e$ the critical effects become comparable to those estimated with Dong's criterion (see Table 13b).

8.9. Evaluation of the robustness of the method

The assay of MC and its related compounds can be considered robust because: (i) none of the factors studied has a significant effect (at $\alpha = 0.05$ level) on the determination of the recovery of the main and related compounds when the dummy effects are used to estimate the experimental error; (ii) using Dong's criterion none of the factor effects is significant neither at $\alpha = 0.05$; (iii) the most extreme results obtained in the design (Table 12), are within the acceptance limits for the recovery (95–105%, which were handled in this case study); and (iv) the percent relative standard deviations of the design results are also considered acceptable for this method (1.2, 1.5 and 1.6% for MC, RC1 and RC2, respectively).

The fact that for the responses such as resolution, capacity factor, retention time or asymmetry factor several significant effects are found does not mean that the method should be considered as non-robust or that the method was not well optimised. When the quantitative aspect of the method is not influenced by the factors examined the method can be considered robust. The standardisation one would have to make to prevent factors from affecting responses such as for instance the capacity factor, would be so strict that execution of the method would not be feasible anymore and moreover, would go beyond the original intention of the robustness test.

8.10. Derivation of system suitability limits from robustness test results

The worst-case factor-level combinations for the responses for which SST limits were desired are shown in Table 15a. The worst-case situation for resolution is the factor combination giving the lowest resolution, for the capacity factor it is the one causing the smallest capacity factor, while for the tailing factor it is the factor combination resulting in the highest value. These worst-case conditions were predicted from the significances observed with the statistical interpretation using the dummy effects (Table 14a).

Example: Consider the response resolution between MC and RC1. Significant effects at $\alpha = 0.10$ are observed for the factors pH, column, temperature %*B* end and buffer concentration (see Table 14a). To define the worst-case conditions the non-significant factors %*B* begin, flow and wavelength are kept at nominal level as described in Section 7. To define the worst-case levels for the significant factors the estimated effects are considered (Table 13a). For pH the effect was estimated to be 0.427. This means that level $(+1)$ gives a higher response than level (-1) as can be derived from the equation for effects (Eq. (2)), and that the worst resolution is obtained at level (-1). For the factors column, temperature, %*B* end and buffer concentration the worst-case levels are defined analogously as being $(-1), (-1), (+1)$ and (-1), respectively. This combination of levels is the one given in Table 15a as the predicted worst-case factor-level combination for R_s(MC–RC1).

The worst-case experiment for a given response was then carried out in three independent replicates. The results and the system suitability limits derived from these experiments are shown in Table 15b.

The results of the two other possibilities to define SST-limits, namely taking the average worst-case result, or estimating them from the theoretical model of the effects, are also shown. It can be observed that the SST-limits calculated from the theoretical model are the least strict ones in the case study.

Remark: As mentioned in Section 7, if one doubts about the hypothesis that the quantitative results of the method are not affected by the

Table 15
(a) Predicted worst-case factor-level combinations for the different responses; and (b) results at these conditions together with the derived SST-limits

Factors	Worst-case factor levels for		
	R_s(MC–RC1)	k'(MC)	Asf(MC)
pH	−1	+1	+1
Column	−1	−1	−1
Temperature	−1	0	−1
%*B* begin	0	+1	−1
%*B* end	+1	+1	0
Flow	0	+1	−1
Wavelength	0	0	0
Buffer concentration	−1	0	0

Run	R_s(MC–RC1)	k'(MC)	Asf(MC)
1	4.870	2.800	1.453
2	4.819	2.800	1.483
3	4.702	2.817	1.429
Mean	4.797	2.806	1.455
Standard deviation (*s*)	0.0861	9.81*E−03	0.0271
N	3	3	3
	SST-limits from worst case results		
	$4.797-2.92(0.0861/\sqrt{3})=4.65$	$2.806-2.92(0.0098/\sqrt{3})$ $=2.79$	$1.455+2.92(0.0271/\sqrt{3})$ $=1.59$
	SST-limits from theoretical model		
	4.40	2.57	1.62

worst-case conditions of Table 15a, quantitative experiments (to determine the recovery of the substances in this example) can be executed at these conditions to confirm.

Acknowledgements

Y. Vander Heyden is a postdoctoral fellow of the Fund for Scientific Research (FWO) — Vlaanderen. B. Boulanger, P. Chiap, Ph. Hubert, G. Caliaro and J.M. Nivet (SFSTP commission); P. Kiechle and C. Hartmann (Novartis, Basel, Switzerland) are thanked for various discussions on the subject.

References

[1] ICH harmonised tripartite guideline prepared within the third international conference on harmonisation of technical requirements for the registration of pharmaceuticals for human use (ICH), Text on Validation of Analytical Procedures, 1994, (http:/www.ifpma.org/ich1.html).
[2] Youden, E.H. Steiner, Statistical manual of the association of official analytical chemists, The Association of Official Analytical Chemists ed., Arlington, 1975, pp. 33–36, 70–71, 82–83.
[3] 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, Chemometrics and Intelligent Laboratory systems 10 (1991) 337–347.

[4] M. Mulholland, Ruggedness testing in analytical chemistry, TRAC 7 (1988) 383–389.

[5] Y. Vander Heyden, F. Questier, D.L. Massart, Ruggedness testing of chromatographic methods: selection of factors and levels, Journal of Pharmaceutical and Biomedical Analysis 18 (1998) 43–56.

[6] F.J. van de Vaart, et al., Validation in pharmaceutical and biopharmaceutical analysis, Het Pharmaceutisch Weekblad 127 (1992) 1229–1235.

[7] ICH harmonised tripartite guideline prepared within the third international conference on harmonisation of technical requirements for the registration of pharmaceuticals for human use (ICH), Validation of Analytical Procedures: Methodology, 1996, 1–8 (http:/www.ifpma.org/ich1.html).

[8] 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 Pratiques 2 (1992) 205–239.

[9] The United States Pharmacopeia, 23rd edition, National Formulary 18, United States Pharmacopeial Convention, Rockville, USA, 1995.

[10] Drugs Directorate Guidelines, Acceptable Methods, Health Protection Branch-Health and Welfare Canada, 1992, 20–22.

[11] International organisation for standardisation (ISO), accuracy (trueness and precision) of measurement methods and results — part 2: basic method for the determination of repeatability and reproducibility of a standard measurement method, International Standard ISO (1994E), 5725–5722, First edition.

[12] 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 1994(E), 5725–3, First edition.

[13] 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.

[14] M. Mulholland, J. Waterhouse, Development and evaluation of an automated procedure for the ruggedness testing of chromatographic conditions in high-performance liquid chromatography, Journal of Chromatography 395 (1987) 539–551.

[15] H. Fabre, V. Meynier de Salinelles, G. Cassanas, B. Mandrou, Validation d'une méthode de dosage par chromatography en phase liquide haute performance, Analusis 13 (1985) 117–123.

[16] J.C. Miller, J.N. Miller, Statistics for Analytical Chemistry, Ellis Horwood, New York, 1993, pp. 53–59.

[17] D. Dadgar, P.E. Burnett, M.G. Choc, K. Gallicano, J.W. Hooper, Application issues in bioanalytical method validation, sample analysis and data reporting, Journal of Pharmaceutical and Biomedical Analysis 13 (1995) 89–97.

[18] H. Fabre, Robustness testing in liquid chromatography and capillary electrophoresis, Journal of Pharmaceutical and Biomedical Analysis 14 (1996) 1125–1132.

[19] D.A. Doornbos, A.K. Smilde, J.H. de Boer, C.A.A. Duineveld, Experimental design, response surface methodology and multicriteria decision making in the development of drug dosage forms, in: E.J. Karjalainen (Ed.), Scientific Computing and Automation (Europe), Elsevier, Amsterdam, 1990, pp. 85–95.

[20] European pharmacopoeia 1997, 3rd edition, European Department for the Quality of Medicines within the Council of Europe, Strasbourg, 1996.

[21] Eurachem, A focus for analytical chemistry in Europe, Quantifying Uncertainty in Analytical Measurement, First Edition 1995.

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

[23] 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.

[24] 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.

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

[26] 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, Analytica Chimica Acta 312 (1995) 245–262.

[27] 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, Analytica Chimica Acta 316 (1995) 15–26.

[28] Y. Vander Heyden, F. Questier, D.L. Massart, A ruggedness test strategy for procedure related factors : experimental set-up and interpretation, Journal of Pharmaceutical and Biomedical Analysis 17 (1998) 153–168.

[29] J.L. Goupy, Methods for Experimental Design, Principles and Applications for Physicists and Chemists, Elsevier, Amsterdam, 1993, pp. 159–177, 421–427.

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

[31] A. Nijhuis, H.C.M. van der Knaap, S. de Jong, B.G.M. Vandeginste, Strategy for ruggedness tests in chromatographic method validation, Analytica Chimica Acta 391 (1999) 187–202.

[32] K. Jones, Optimization of experimental data, International Laboratory 16 (9) (1986) 32–45.

[33] Y. Vander Heyden, D.L. Massart, Y. Zhu, J. Hoogmartens, J. De Beer, Ruggedness tests on the HPLC assay of the United States Pharmacopeia XXIII for tetracycline hydrochloride: comparison of different columns in an interlaboratory approach, Journal of Pharmaceutical and Biomedical Analysis 14 (1996) 1313–1326.

[34] 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, Journal of Chromatography A 756 (1996) 89–106.

[35] Y. Vander Heyden, G.M.R. Vandenbossche, C. De Muynck, K. Strobbe, P. Van Aerde, J.P. Remon, D.L. Massart, Influence of process parameters on the viscosity of Carbopol® 974P dispersions, Ph.D. thesis, personal communication.

[36] F. Dong, On the identification of active contrasts in unreplicated fractional factorials, Statistica Sinica 3 (1993) 209–217.

[37] C. Daniel, Use of half-normal plots in interpreting factorial two-level experiment, Technometrics 1 (1959) 311–341.

[38] D.A. Zahn, An empirical study of the half-normal plot, Technometrics 17 (1975) 201–211.

[39] R.V. Lenth, Quick and easy analysis of unreplicated factorials, Technometrics 31 (1989) 469–473.

[40] P.D. Haaland, M.A. O'Connel, Inference for effect-saturated fractional factorials, Technometrics 37 (1995) 82–93.

[41] Z. Sidak, Rectangular confidence regions for the means of multivariate normal distributions, Journal of the American Statistical Association 62 (1967) 626–633.

[42] Y. Vander Heyden, M. Jimidar, E. Hund, N. Niemeijer, R. Peeters, J. Smeyers-Verbeke, D.L. Massart, J. Hoogmartens, Determination of system suitability limits with a robustness test, Journal of Chromatography A 845 (1999) 145–154.