Organic Process Research & Development

Article

pubs.acs.org/OPRD

Identification of Important Experimental Variables in Organic Synthetic Procedures by Near-Orthogonal Experiments

Rolf Carlson,* Geir Simonsen, and Alexandre Descomps

Department of Chemistry, Faculty of Science, University of Tromsoe, NO-9037 Tromsoe, Norway

Johan E. Carlson

Department of Computer Science, Electrical, and Space Engineering, Luleå University of Technology, SE-971 87 Luleå, Sweden

ABSTRACT: A new strategy is presented for the design of screening experiments in synthetic chemistry when the objective is to identify the important experimental variables from a limited number of experimental runs. The methodology is based on Taylor expansion (response surface) models The experimental design is constructed in such a way that the vector of the variables in the Taylor model in each run are near-orthogonal to each other. This is achieved by laying out a grid of possible experiments in the experimental space, expanding this candidate experimental design matrix to the corresponding model matrix, i.e. the matrix containing columns for all variables in the Taylor expansion. This model matrix is then factorised by singular value decomposition, SVD. The row in the model matrix that is most parallel to the first singular vectors is selected as the first experiment. .The variation displaced by this first experiment is removed from the elements of the model matrix by projections. The resulting matrix is the orthogonal complement to the first selected row. The procedure is repeated until all dimensions of the model space have been spanned by the selected experiments The singular vectors are mutually orthogonal, and selected experiments will be nearly orthogonal and span the dimensions of the model space. The experiments can be run in sequence and thus allow for a systematic search, one experiment at a time. It is shown that subset selections from such designs in combination with PLS modelling can be used to identify the important variables. The principles are illustrated with two examples: (a) a dibromination of an acetyl with four experimental variables and (b) a synthesis of an enamine by condensing a ketone and morpholine in the presence of molecular sieves in which seven experimental variables are involved. In the acetal bromination, it was found that 5 experiments out of 12 were sufficient for identifying the most important variables. In the enamine example, 8 experiments out of 30 were sufficient.

INTRODUCTION

When an experimental procedure is to be developed into a reliable *method*, an early and important step is to identify the critical experimental factors as well as their possible interaction effects. To this end, a variety of different statistical experimental designs are available: Factorial and fractional factorial designs,[1] D-Optimal designs,[2] Plackett−Burman designs.[3] The use of such designs in organic synthesis is thoroughly described in ref 4.

There are, however, situations in which severe time-constraints preclude any attempt to run a screening design with many experimental runs. Two examples are: (1) A new compound turned out to have interesting pharmaceutical properties. For more testing, 200 g of the compound is needed within 4 weeks. The testing is expensive, and no delay can be tolerated. The chemists have to produce the necessary quantity within the time limits. (2) Outsourcing is now very common to produce the active ingredients in drugs. A chemical company is contracted by the customer to run some test experiments of a given procedure and to deliver 200 g of the desired compound. The time limits are strict, and it is not possible to run more than a handful of tests. Common to these problems is that the chemist should run a reaction that is known and has already been used to make small quantities of the desired compound. It can therefore be assumed that a useful experimental domain is known (i.e., the possible ranges of variation of the experimental factors). It can also be assumed that improved results can be obtained in the vicinity of the known experimental conditions.

Under these circumstances it is reasonable to assume that the observed response, y, can be modelled by a truncated Taylor expansion in the scaled experimental variables, x_i, centred around the known experimental conditions and that

$$y = \beta_o + \Sigma\beta_i x_i + \Sigma\Sigma\beta_{ij} x_i x_j + e$$

in which β_0 is the intercept of the response model at the centre point of the experimental domain, and β_i and β_{ij} are the values of the partial derivatives along the variable axes at the centre point. Least squares estimates (b_0, b_i, and b_{ij}) of the Taylor coefficients can be obtained by fitting the polynomial to the experimental results obtained by a proper design.

We have previously shown in this journal[5] that response surface models can be established from designs constructed in such a way that the rows in the model matrix are nearly orthogonal. The construction of such designs is described in ref 5, and we will not repeat these details here. The essence of these designs is that each new experiment selected for the design spans a new dimension of the model space, i.e. the space

Received: February 17, 2012
Published: July 24, 2012

ACS Publications © 2012 American Chemical Society

dx.doi.org/10.1021/op300041p | *Org. Process Res. Dev.* 2012, 16, 1371−1377

spanned by the variables in the Taylor polynomial, and that it is possible to investigate the roles played by the variables and their interactions by a sequential approach in which the experiments are added one by one until the variations in the model space have been mapped.

In a screening, the task is to determine which experimental variables have a real influence on the result. It is often the case that out of many variables initially considered to be potentially important, there are only a few of them that really matter. A discussion of this is given in ref 6.

It came to our minds that a design based on near-orthogonal experiments might be useful in a screening situation. The reasons are the following: The very first experiment in such a design describes the direction showing the largest variation of the variable settings in the model space. The important variables will exert their influence in this experiment. The second experiment is near-orthogonal to the first one, and the important variables will influence this experiment too, *but in a different way*. If there are only a handful of important variables, it might be possible that these can be identified from a handful of experimental runs. In this report, we show two examples along these principles.

EXAMPLES

The first example is the bromination of an acetal fully described in ref 5. The reaction is portrayed in Scheme 1. The variables explored and the design are given in Tables 1 and 2, respectively.

Scheme 1

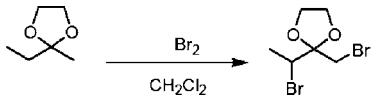

Table 1. Experimental variables in the bromination of the acetal and the levels of their settings[a]

	levels of the settings		
variables	−1	0	+1
x_1: reaction temperature/ °C	0	15	30
x_2: concentration of acetal/M	0.2	0.3	0.4
x_3: stirring rate/rpm	250	325	400
x_4: rate of bromine addition/meq min^{-1}	20	50	70

[a]Reproduced with permission from the American Chemical Society.

The second example is the synthesis of an enamine by a condensation between 4-methyl-2-pentanone and morpholine in the presence of molecular sieves, see Scheme 2.

The variables explored and their settings are shown in Table 3. The experimental design and the yields obtained are shown in Table 4. There are three discrete variables at two levels and four variables at five levels. The candidate experiments were defined by the full $2^3 * 5^4$ full factorial design with a total of 5000 runs. The design was expanded to the candidate model matrix by appending columns of the cross-product terms. The design was generated by an algorithm based on singular value decomposition[7] as described in ref 5.

Data. It was assumed that second-order interaction models would be sufficient for describing the variation in yield, y, as functions of the experimental settings, x_i and the interactions, x_i x_j:

Table 2. Experimental design and yields obtained in the bromination of the acetal[a]

	design				yield
exp #	x_1	x_2	x_3	x_4	y
1	1.0	1.0	1.0	1.0	87.4
2	1.0	−1.0	−1.0	−1.0	95,8
3	−1.0	1.0	−1.0	1.0	79.5
4	−1.0	−1.0	1.0	−1.0	63.7
5	−1.0	−1.0	0.2	1.0	53.9
6	−1.0	1.0	1.0	−1.0	68.7
7	1.0	1.0	1.0	−1.0	58.8
8	1.0	−1.0	1.0	−1.0	93.5
9	1.0	−1.0	−1.0	1.0	94.0
10	−1.0	−1.0	−1.0	−1.0	77.1
11	1.0	−1.0	1.0	1.0	80.9
12	0	0	0	0	88.6

[a]Reproduced with permission from the American Chemical Society.

Scheme 2

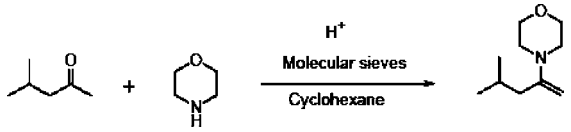

Table 3. Experimental variables and their settings in the enamine synthesis

	settings				
variables	−1	−0.5	0	0.5	1
x_1: type of acid	Nafion				TFA
x_2: temperature/°C	0	10	20	30	40
x_3: type of molecular sieve 5A	powder				pellets
x_4: stirring/rpm	none				300
x_5: ratio morpholine/ketone/ mol/mol	1.0	1.5	2.0	2.5	3.0
x_6: ratio molecular sieves/ ketone/g/mol	200	300	400	500	600
x_7: molar concentration of ketone[a]	2.5	2.9	3.3	4.0	5.0

[a]Actually, the amount of solvent was varied, and the concentrations given are calculated from this.

$$y = \beta_0 + \Sigma\beta_i x_i + \Sigma\Sigma\beta_{ij} x_i x_j + e(i \neq j)$$

where e is a random error term.

In the bromination example there are four variables and 11 unknown parameters in the model. In the enamine synthesis, there are seven variables and 29 unknown parameters in the model. For these reasons, the model spaces will have 11 and 29 dimensions, respectively and the corresponding designs for fitting the full Taylor polynomial by least-squares multiple regression must have at least 11 and 29 experimental runs, respectively.

Data Analysis. We wished to know whether or not a limited number of experimental runs would be sufficient for identifying the important variables. Hence, the number of experiments will be less than the number of coefficients in the model and it will not be possible to estimate the coefficients by least-squares multiple regression. Instead, we have used PLS-modelling to estimate the coefficients. The PLS models can be rotated to give estimates of the Taylor model coefficients. Thorough treatments of PLS modelling are given in refs 8, 9.

dx.doi.org/10.1021/op300041p | Org. Process Res. Dev. 2012, 16, 1371−1377

Table 4. Experimental design and the yields obtained in the enamine synthesis

	variables							yield
exp. no.	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
1	−1	−1	−1	−1	−1	−1	1	0.4
2	1	1	1	1	1	1	1	44.2
3	−1	−1	−1	−1	1	1	1	0.9
4	−1	−1	−1	1	−1	1	1	1.7
5	1	1	1	−1	1	1	−1	51.9
6	1	0.5	1	−1	−1	−1	−1	27.0
7	1	1	1	1	−1	1	−1	30.9
8	1	1	1	−1	−1	1	1	26.3
9	−1	−1	−1	−1	1	−1	−1	0.2
10	−1	−1	−1	1	−1	−1	−1	0.2
11	−1	1	1	1	−1	−1	1	22.7
12	1	−1	−1	1	−1	−1	−1	5.2
13	−1	−1	−1	1	1	−1	1	0.8
14	−1	−1	1	−1	−1	1	−1	25.6
15	1	−1	−1	1	−1	−1	1	4.5
16	1	−1	−1	1	−1	−1	1	4.5
17	−1	1	−1	−1	−1	1	−1	4.8
18	1	−1	−1	−1	−1	1	−1	4.5
19	1	−1	−1	−1	1	−1	1	6.2
20	1	1	−1	1	1	−1	−1	27.4
21	−1	1	−1	1	−1	1	1	11.1
22	1	−1	−1	1	1	1	−1	5.6
23	−1	1	−1	1	1	1	−1	13.3
24	−1	1	1	−1	1	−1	−1	25.2
25	−1	−1	1	1	−1	−1	1	11.5
26	−1	1	1	−1	1	−1	−1	25.2
27	1	1	−1	−1	−1	−1	1	17.0
28	1	−1	1	1	1	1	1	25.4
29	−1	1	1	1	1	1	−1	43.6
30	−1	−1	−1	−1	−1	1	−1	0.4

It is not possible, however, to obtain accurate estimates when the number of experiments is lower than the number of parameters to be estimated; the estimates will be biased. The constant in the model will be the average response in the experimental runs and corresponds to the result obtained in the average point. If the average response is different from the constant in the Taylor polynomial and the difference is $d = \beta_0 - y_{average}$, we have to adjust for this bias prior to fitting the model, and the expectation of the estimated parameter vector, $E[\mathbf{b}]$, will be

$$E[\mathbf{b}] = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{d}$$

in which $\mathbf{d}$ is the bias vector obtained from the bias increment added to each observed response.

The objective of the present investigation is to evaluate if a limited number of experiments can be used to identify the important variables. In this context, even a biased estimate will be indicative.

RESULTS

The results of the PLS modelling are summarised in Table 5.

Bromination of the Acetal. We have analysed the roles played by the variables in two ways: (1) by the cumulative normal probability distribution plots of their coefficients[10] and (2) by the variable influence plots.[8a] Figure 1 shows the cumulative normal probability plots obtained when five, six, and all experiments, respectively, were used to establish the PLS

Table 5. Summary of PLS results[a]

reaction system	experiments	PLS components	R^2	Q^2
acetal	5	2	1.00	0.726
	6	1	0.898	0.382
	12	1	0.808	0.269
enamine	8	2	0.894	0.572
	10	1	0.933	0.449
	16	1	0.841	0.545
	20	1	0.858	0.565
	30	1	0.890	0.560

[a]The fitted model was the second-order interaction model.

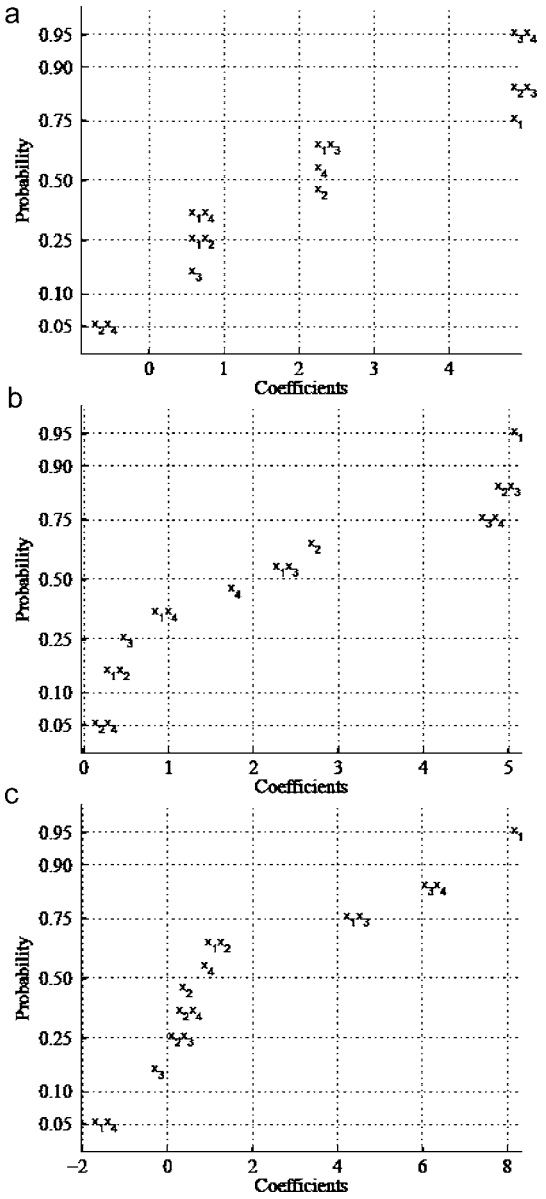

Figure 1. Acetal bromination: cumulative normal probability plots of estimated coefficients from: (a) 5 experiments; (b) 6 experiments; (c) 12 experiments.

model. Figure 2 shows the corresponding variable influence plots. It is clearly seen that one variable, x_1, is visible as an important variable in all plots. This indicates that the reaction temperature is an important variable to control. When 5 and 6 experiments in the design, two interaction effects are indicated

dx.doi.org/10.1021/op300041p | Org. Process Res. Dev. 2012, 16, 1371−1377

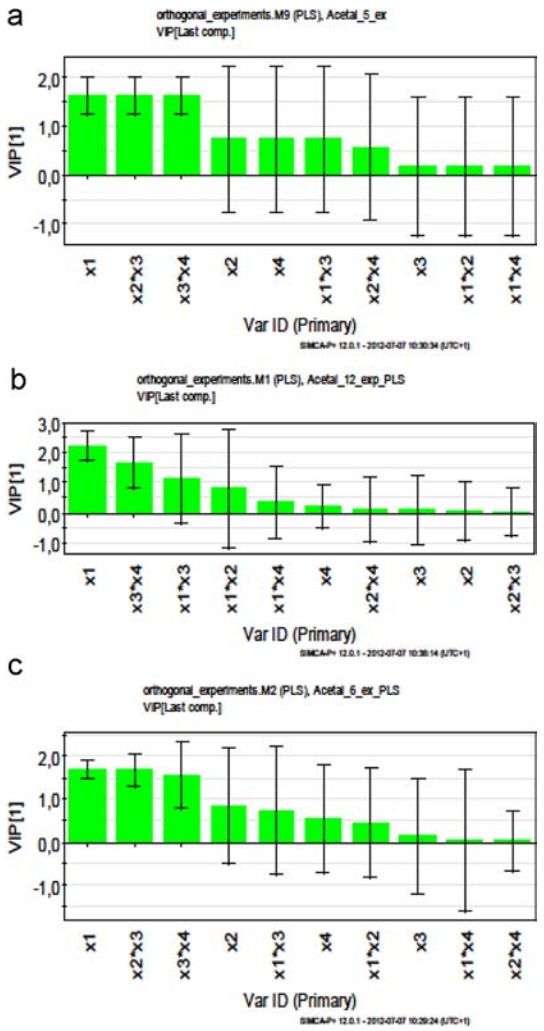

Figure 2. Variable influence plots from (a) 5 experiments; (b) 6 experiments; (c) 12 experiments. A variable influence >1.0 indicates a significant contribution to the model.

as possibly important: between x_2 (the acetal concentration and x_3 (the stirring rate) and between x_3, (the rate of bromine addition) The reaction is exothermic, and this may explain these interaction effects. When all experiments were included in the design, only x_1 and the interaction x_3,x_4 appear to be important. The other variables have only a minor importance. This was also the conclusion reached in ref 5.

Enamine Synthesis. This is a more complicated system. Figure 3 shows the cumulative normal probability distributions of the estimated coefficients obtained when 8, 10, 16, 20, and all experiments, respectively, were used to establish the PLS model. Figure 4 shows the corresponding variable influence plots. The experiments are more easily evaluated from the variable influence plots than from the cumulative normal probability distributions.

With all designs, three variables are indicated as highly important: x_1 (the type of acid catalyst), x_2 (the reaction temperature), and x_3 (the type of molecular sieve). Two more variables: x_6 (the ratio molecular sieves/ketone), and x_7 (the molar concentration of the ketone) are indicated as possibly important. Some interaction effects of minor importance are also seen.

When the first eight experiments were used, the cross-product x_1x_3 is constant, and the corresponding interaction coefficient is confounded with the average, β_0.

A screening experiment is carried out to identify which of many possible variables are likely to have a significant influence on the result. These variables should then be more carefully studied in subsequent experiments. If some variables found to be significant in the first screening should turn out to have only minor influences in the follow-up experiments, no harm is done, and these variables can then be safely removed from further consideration. It is much worse if important variables are overlooked.

The coefficients estimated with PLS are biased, see above. However, even a biased estimate contains information. The objective is not to estimate a response surface model with high precision but to discern which variables are likely to be important, and for this porpose, the PLS estimates will be sufficient.

DISCUSSION

One reviewer checked our models with the MODDE 8.0 software,[11] and the results obtained were different from ours. We have used the SIMCA 12.0 software[12] to establish our models (with cross-validation to avoid overfitting). The models established by the MODDE software included more PLS components than our models, and it was suggested that the difference in results were due to underfitting of our models. We have checked this and refitted our models including as many components as suggested by MODDE. However, the differences remained. We also tried OPLS[13] for fitting the models. By this procedure, the variation in the model matrix, **X**, that is uncorrelated to the response is removed prior to fitting the model. The results were similar to what had been obtained with regular PLS. There is another explanation as to why our results using SIMCA differ from the models obtained by MODDE: the algorithms operate differently. In MODDE the variables are orthogonally scaled (−1 for the low level, +1 for the high level) prior to fitting the PLS model, while *i* SIMCA the variables are autoscaled to unit variance over the set of experiments prior to fitting the PLS model. This may explain the differences in the result pointed out by one of the reviewers.

The cumulative normal probability plot and the variable influence plots were all similar, but not identical, using PLS with cross-validation, PLS with more components, and OPLS, respectively, and yielded the same conclusions. For simplicity, the plots shown in this report were obtained from PLS with cross-validation.

It is evident that the amount of information on the experimental variables is limited when only a few experiments have been run and the estimates of the Taylor polynomial coefficient will be biased by confounding with other effects. It was suggested by one reviewer that, in the first place, a design for fitting a model with only first-order terms should be run. We agree, this is good advice. It is then possible to augment the design for fitting a second-order interaction model and remove the experiments already done. The designs for fitting a first-order linear model are embedded in the designs for fitting the second-order interaction models (see the design matrices given in the Appendix in ref 5. It is therefore possible first to run the experiments for a linear model and then run the remaining experiment to fit an interaction model.

Since the experiments are near-orthogonal, they span the different dimensions in the model space, and running all the

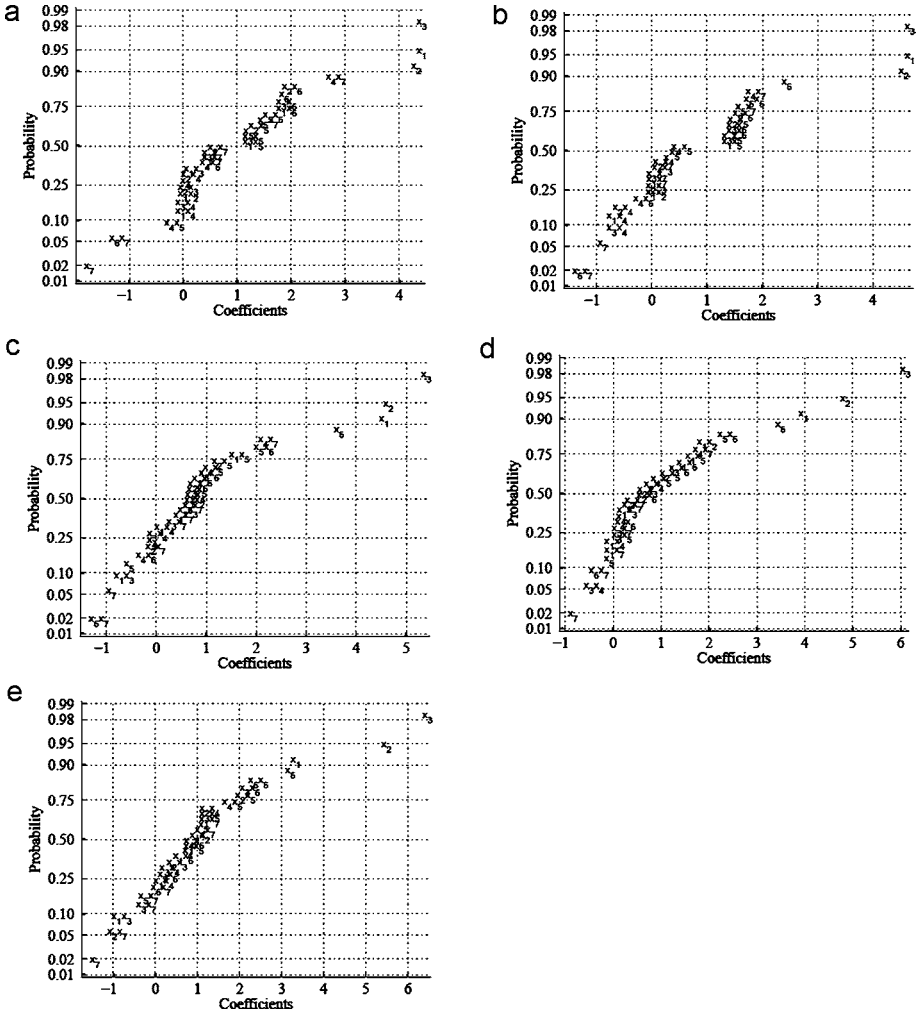

Figure 3. Enamine synthesis: Cumulative normal probability distribution plots of estimated coefficients from: (*a*) 8 experiments; (*b*) 10 experiments; (c) 16 experiments; (d) 20 experiments; (e) 30 experiments.

experiment in the designs will deconfound the estimated Taylor coefficients.

■ CONCLUSIONS

If the objective is to fit a response model with high precision, then of course it is necessary to use a statistical design that permits accurate estimations of the model parameters. However, such designs often contain a fairly large number of individual experimental runs, and sometimes this precludes their use. Under such circumstances, we have suggested that a design based on near-orthogonal experiments can be useful, and in this contribution we have shown that sets of such experiment make it possible to discern the important experimental variables from only a few runs. As the experiments are run sequentially, one by one, it is possible to run the number of experiments necessary to obtain a clear picture. We assume that this will be an appealing technique in the realm of process chemistry.

■ EXPERIMENTAL SECTION

Computations. The experimental designs were generated in MATLAB.[14] The PLS models were obtained with cross validation using the SIMCA P-12 software.[12]

Chemicals. Morpholine (puriss.) was obtained from Fluka. 4-Methyl-2-pentanone (HPLC grade), cyclohexane (99.5%), and phenylcyclohexane (puriss.) internal standard for GC were obtained from Aldrich. They were used as delivered. Molecular sieves, 5Å powder and pellets were obtained from Fluka. They were activated at 300 °C for 24 h prior to use and stored in a desiccator over phosphorus pentoxide. A reference sample of the morpholine enamine from 4-methyl-2-pentanone used for GC-calibration was prepared according to ref 15.

GC Analyses. *Enamine Synthesis.* A Varian 3400 gas chromatograph with a flame ionisation detector coupled to a Varian 4400 integrator was used. The column was SPB-5, 30 m, 0.35 mm i.d., operated with the following temperature program: 70 °C, 5 min; 10 °C min^{-1}; 180 °C. The yields were determined from the integrated peak areas using phenylcyclohexane as internal standard.

General Procedure for the Screening Experiments, Bromoacetal Synthesis. The experimental procedure for the bromination of the acetal is given in ref 5, and it is not reproduced here.

General Procedure for the Screening Experiments, Enamine Synthesis. The settings of the variables are shown in Table 3. The experiments were run in 50 mL test tubes using a heating block reactor system, Bohdane 2080 miniblock from Mettler Toledo. In the experiments, 5 mmol of the ketone, 4-methyl-2-pentanone, was used. The test tube was charged with the ketone, the amount x_6 of the molecular sieves of type x_3, the

dx.doi.org/10.1021/op300041p | *Org. Process Res. Dev.* 2012, 16, 1371−1377

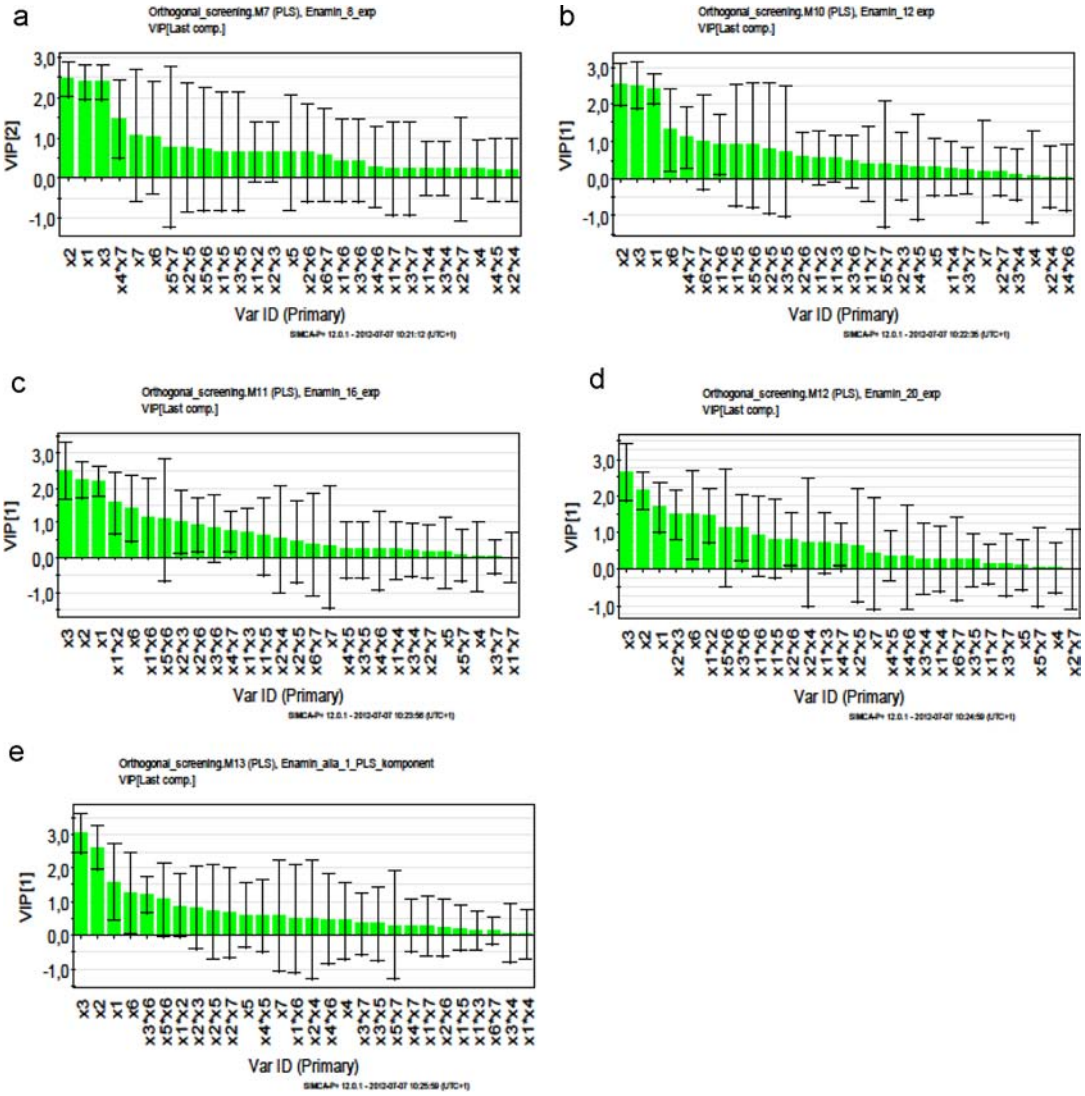

Figure 4. Variable influence plots from (a) 8 experiments; (b) 10 experiments; (c) 16 experiments; (d) 20 experiments; (e) 30 experiments. A variable influence >1.0 indicates a significant contribution to the model.

amount x_5 of morpholine, a carefully weighed amount, ~200 mg, of phenylcyclohexane (internal standard), and 20 mg of the acid x_1. The calculated amount of cyclohexane solvent to give the concentation x_7 was then added to the test tube The temperature, x_2, and the stirring, x_4, were adjusted. The reaction was monitored by gas chromatography. Samples of 0.1 mL were withdrawn, filtered through a plug of cotton, diluted with 2 mL of pentane, and analysed by GC. Integrated peak areas were used for quantification. The yields obtained after 24 h are shown in Table 4.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rolf.carlson@uit.no.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank the American Chemical Society for the kind permission to reproduce details from ref 5. We also thank the Norwegian Research Council for financial support via the KOSK-II program.

REFERENCES

(1) (a) Box, G. E. P.; Hunter, J. S. *Technometrics* **1961**, *3*, 311−351. (b) Box, G. E. P.; Hunter, J. S. *Technometrics* **1961**, *3*, 449−458. (c) Box, G. E. P.; Hunter, J. S. ; Hunter, W. G. *Statistics for Experimenters. Design, Innovation, and Discovery*, 2nd ed.; Wiley: Hoboken, NJ, 2005.

(2) (a) Nalimov, V. V.; Golikova, T. I.; Mikeshina, N. G. *Technometrics* **1970**, *12*, 799−812. (b) Fedorov, V. V. *Theory of Optimal Experiments*; Academic Press: New York, 1972.

(3) Plackett, R. L.; Burman, J. P. *Biometrika* **1946**, *33*, 305−325.

(4) Carlson, R.; Carlson, J. E. *Design and Optimization in Organic Synthesis*, 2nd revised and enlarged ed.; Elsevier: Amsterdam, 2005.

(5) Carlson, R.; Simonsen, G.; Descomps, A.; Carlson, J. E. *Org. Process Res. Dev.* **2009**, *13*, 798−803.

(6) For comments on this, see: (a) Box, G. E. P.; Hunter, J. S. ; Hunter, W. G. *Statistics for Experimenters. Design. Innovation, and Discovery*, 2nd ed.; Wiley: Hoboken, NJ, 2005; pp 243−244; (b) Box, G. E. P.; Draper, N. R. *Response Surfaces, Mixtures, and Ridge Analysis*, 2nd ed.; Wiley: Hoboken, NJ, 2007; pp 148−150.

(7) Strang, G. *Introduction to Linear Algebra*, 3rd ed.; Wellesley-Cambridge Press: Wellesley MA, 1998; Section 6.7; ISBN 0-9614088-5-5.

(8) (a) Wold, S.; Sjöström, M.; Eriksson, L. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130. (b) Carlson, R.; Carlson, J. E. *Design and*

Optimization in Organic Synthesis, 2nd revised and enlarged ed.; Elsevier: Amsterdam, 2005; Chapter 18.
(9) Eriksson, L.; Johansson, E.; Kettaneh-Wold., N.; Wold, S. *Multi- and Megavariate Data Analysis: Principles and Applications*; Umetrics: Umeå, Sweden, 2001.
(10) Daniel, C. *Application of Statistics to Industrial Experimentation*; Wiley: New York, Probability and Statistics, 1976; p 1076.
(11) *MODDE-8*; Umetrics Inc., MKS Instuments: San José, CA, 2006.
(12) *SIMCA-P-12*; Umetrics Inc., MKS Instuments: San José, CA, 2008.
(13) Trygg, J; Wold, S. *J. Chemom.* **2002**, *16*, 119−128.
(14) (a) *MATLAB*, The MathWorks, Inc.: Natick, MA, 2007; (b) For the generation of designs based on near-orthogonal experiments, see NO_Designs:http://jec-solutions.mine.nu.
(15) Carlson, R.; Nilsson, Å.; Strömqvist., M. *Acta Chem. Scand., Ser. B* **1983**, *37*, 7−13.