DEPARTMENT OF ENVIRONMENT,
TECHNOLOGY AND TECHNOLOGY MANAGEMENT

Analyzing Categorical Data from Split-Plot and Other Multi-Stratum Experiments

Peter Goos & Steven G. Gilmour

UNIVERSITY OF ANTWERP
Faculty of Applied Economics

Stadscampus
Prinsstraat 13, B.213
BE-2000 Antwerpen
Tel. +32 (0)3 265 40 32
Fax +32 (0)3 265 47 99
http://www.ua.ac.be/tew

FACULTY OF APPLIED ECONOMICS

DEPARTMENT OF ENVIRONMENT,
TECHNOLOGY AND TECHNOLOGY MANAGEMENT

Analyzing Categorical Data from Split-Plot and Other Multi-Stratum Experiments

Peter Goos & Steven G. Gilmour

RESEARCH PAPER 2010-021
SEPTEMBER 2010

University of Antwerp, City Campus, Prinsstraat 13, B-2000 Antwerp, Belgium
Research Administration – room B.213
phone: (32) 3 265 40 32
fax: (32) 3 265 47 99
e-mail: joeri.nys@ua.ac.be

The papers can be also found at our website:
www.ua.ac.be/tew (research > working papers) &
www.repec.org/ (Research papers in economics - REPEC)

D/2010/1169/021

Analyzing Categorical Data from Split-Plot and Other Multi-Stratum Experiments

Peter Goos

Faculty of Applied Economics & StatUa Center for Statistics, Universiteit Antwerpen,

City Campus - Prinsstraat 13, 2000 Antwerpen, Belgium.

Erasmus School of Economics, Erasmus Universiteit Rotterdam,

Postbus 1738, 3000 DR Rotterdam, The Netherlands.

peter.goos@ua.ac.be

Steven G. Gilmour

Queen Mary University of London, School of Mathematical Sciences,

Mile End Road, London E1 4NS, UK.

s.g.gilmour@qmul.ac.uk

Abstract: Many factorial experiments yield categorical response data. Moreover, the experiments are often run under a restricted randomization for logistical reasons and/or because of time and cost constraints. The combination of categorical data and restricted randomization necessitates the use of generalized linear mixed models. In this paper, we demonstrate the use of Hasse diagrams for laying out the randomization structure of a complex factorial design involving seven two-level factors, four three-level factors and a five-level factor, and three repeated observations for each experimental unit. The Hasse diagrams form the basis of the mixed model analysis of the ordered categorical data produced by the experiment. We also discuss the added value of categorical data over binary data and difficulties with the estimation of variance components and, consequently, with the statistical inference. Finally, we show how to deal with repeats in the presence of categorical data, and describe a general strategy for building a suitable generalized linear mixed model.

Keywords: binary data, cumulative logit regression, generalized linear mixed model, Hasse diagram, ordered categorical data, split-plot analysis.

1 Introduction

Methods for the analysis of categorical data are widely used in medical and biological experiments, in which the treatment structures are often simple, and have been used and recommended for industrial experiments with factorial treatments. At the time of the discussion papers of Nair (1986) and Hamada and Wu (1990), there was still some disagreement about whether the routine use of generalized linear models (GLMs) was straightforward enough to be recommended in place of linear models with scoring methods, i.e. treating the category labels as if they are continuous responses. Scoring-based methods still form the subject of recent quality-oriented research, e.g. Wu and Yeh (2006) and D'Ambra, Köksoy and Simonetti (2009). However, GLMs have become increasingly accepted as the standard analyses, e.g. Myers, Montgomery and Vining (2002), Lee and Nelder (2003). GLMs with categorical responses have been successfully used in practice, for example using the failure amplification method (Joseph and Wu, 2004) to find optimum manufacturing settings for printed circuit boards (Jeng, Joseph and Wu, 2008) and for optimizing the synthesis of cadmium selenide nanostructures (Dasgupta *et al.*, 2008).

Factorial designs continue to make a major contribution to industrial research and it has been increasingly recognized in recent years that many, if not most, industrial experiments have some factors which are hard to set (often called "hard-to-change", although in a completely randomized design, they should be reset between runs even if the factor level does not change). When properly taken into account at the design stage, hard-to-set factors lead naturally to multi-stratum structures, with different factors applied in different strata through restricted randomization, as in split-plot designs. This has been an area of much research in the last decade, e.g. Letsinger, Myers and Lentner (1996), Trinca and Gilmour (2001), Goos (2002), Vining, Kowalski and Montgomery (2005), Gilmour and Trinca (2006), Parker, Kowalski and Vining (2007), Jones and Goos (2007, 2009), Vivacqua and Bisgaard (2009), Gilmour and Goos (2009).

Our approach to analyzing continuous data from experiments with hard-to-set factors is to use the randomization restrictions to motivate an appropriate random effects structure in a mixed model. We model the random effects as being independent and normally distributed. Although these assumptions

cannot be justified by the randomization, they are very commonly made and gross departures from them are easily diagnosed through analysis of residuals. The effects of the treatment factors are then modeled as fixed effects and the choice and selection of an appropriate model proceeds in a similar way as for completely randomized experiments.

In this paper, we describe how to analyze categorical response data obtained from multi-stratum designs. This work was motivated by an experiment on a polypropylene process, which used a somewhat complicated multi-stratum design, and we use this application throughout to illustrate the appropriate analyses. In Section 2 we give some description of the experiment and how it was designed. The different responses (one continuous response as well as binary and ordered categorical responses for five different coatings) are analyzed in Sections 3 to 5, where the details of how to construct appropriate models are described. Further complications arise if we try to analyze all of the coatings together, which is done in Section 6. A general strategy for analysing categorical data from multi-stratum designs is outlined in Section 7. Finally, some conclusions are drawn in Section 8.

2 The Polypropylene Experiment

In 2004 and 2005, four Belgian companies, Domo PolyPropylene Compounds (a producer of thermoplastic materials), Europlasma (a developer of gas plasma systems), Structuplas (a company specializing in the finishing of thermoplastic materials) and Techni-Coat International (a company specializing in applying coatings) ran an experiment to investigate the impact of several additives and a gas plasma surface treatment on the adhesive properties of polypropylene. The experiment was of great interest to car manufacturers who are increasingly using polypropylene because it is inexpensive and light, and because it can be recycled. The experiment was financially supported by Flanders' Drive, a technological platform that stimulates innovation in the automotive industry in Flanders and that itself is supported by the Flemish government.

An undesirable property of polypropylene is that glues and coatings do not adhere well to its surface unless it undergoes a surface treatment, such as a gas

Table 1: Levels of factors studied in the polypropylene experiment

Factor	Units	Levels		
		-1	0	1
EPDM (X_1)	%	0		15
Ethylene (X_2)	%	0		10
Talcum (X_3)	%	0		20
Mica (X_4)	%	0		20
Lubricant (X_5)	%	0		1.5
UV-stabiliser (X_6)	%	0		0.8
EVA (X_7)	%	0		1.5
Power (X_8)	Watts	500	1000	2000
Gas flow rate (X_9)	sccm	1000	1500	2000
Processing time (X_{10})	min	2	8	15
Type of gas (X_{11})		Etching	Activation 1	Activation 2

plasma treatment. The goal of the experiment was to search for economical plasma treatments that lead to a good adhesion for various kinds of coatings. Polypropylene is often compounded with additives to tailor the plastic to a specific end use. Hence the effects of several additives in the polypropylene were studied, as well as several plasma treatment factors.

Seven additives, coded as $X_1 - X_7$, were included in the experiment, each at two levels: ethylene propylene diene monomer (EPDM) rubber, ethylene copolymer, talcum, mica, lubricant, UV-stabilizer and ethylene vinyl acetate (EVA). Four plasma treatment factors, coded as $X_8 - X_{11}$, were included, each at three levels: power, gas flow rate, processing time and type of gas used. The levels and units used for each of these eleven factors are shown in Table 1.

The entire polypropylene experiment involves a complicated randomization but it is based on a D-optimal split-plot design, the construction of which is discussed in Jones and Goos (2007). The complicated randomization is due to the fact that the complete experiment was carried out in several stages:

1. First, 20 different batches of polypropylene plates were produced according to a whole plot design in the seven additives. Each of the batches contains several dozen polypropylene plates with the same set-

tings for the seven additives. Each of the plates was stored individually in identical conditions. For each of the following stages, the appropriate numbers of plates were removed from storage immediately prior to the further processing.

2. Next, three to seven samples were selected from each of these batches and processed according to the sub-plot design. Although no formal randomization took place, the selection was essentially random. The sub-plot design consisted of 100 gas plasma treatments which were applied in 100 independent oven runs, in a random order. After each of the 100 oven runs, the surface tension of the treated sample was measured. This stage of the experiment thus involves a continuous response and a split-plot design with seven whole-plot factors and four sub-plot factors.

3. At a later point in time, three to seven sets of three new samples were randomly selected from each of the 20 batches. The three samples in each set were processed together in one oven run, using one gas plasma treatment from the sub-plot design. Each set of three samples was processed in a separate run of the oven. For logistical reasons, the order in which the gas plasma treatments were applied to the sets of samples was the same as in Stage 2. A fixed number of days after the gas plasma treatment, coating 1 was applied to each of the three samples in a set. A six-level categorical response, related to the success of the coating's adhesion to the plastic, was measured as soon as the coating was dry.

4. Stage 3 was repeated four more times for four different types of coatings.

From this experiment the investigators sought answers to many research questions. For the purpose of this article, the most important question was whether the effects of the different factors were substantially different between coatings. Another research question that we will discuss in this paper was whether the surface tension can be used as a surrogate for the success of a coating. It is obvious that an answer to the first research question requires the combined analysis of the data from the five coatings involved in the experiment. As a result, the type of coating can be regarded as a twelfth experimental factor, having five levels. The polypropylene experiment can

therefore be regarded as a strip-plot type of experiment, with additional complications such as the three repeats, the fact that the order of the oven runs corresponding to the sub-plot design was not randomized separately for every coating, and the categorical nature of the responses.

We now describe our approach to the analysis of the continuous response and the categorical responses for the five coatings. In all the analyses we conducted, we used two contrasts to model the impact of the three-level categorical factor X_{11}. The first contrast, labeled 'Type of gas', was constructed so that it quantifies the difference between the etching gas and the two activation gases. The second contrast, labeled 'Activation gas', measures the extent to which the effect of the first activation gas is different from that of the second activation gas. As we explain in Section 6, we also use specific contrasts for the five-level coating factor when combining the data for all coatings, but we do not need these contrasts for the analysis of the surface tension response or for the analysis of the categorical response for each coating separately.

3 Surface Tension

From each oven run, the initial surface tension was measured on a continuous scale. This can be analysed in what has become a standard way, i.e. by fitting a linear mixed model, with random effects included for each stratum implied by the randomization. In this case there are just two strata, one corresponding to the randomization of batches and one corresponding to the randomization of oven runs. The design used is an unbalanced split-plot design of the type analysed by Letsinger, Myers and Lentner (1996). Hence, the appropriate model has the form

$$Y_{ij} = \beta_0 + \mathbf{x}'_{ij}\boldsymbol{\beta} + \delta_i + \epsilon_{ij}, \tag{1}$$

where Y_{ij} is the surface tension measured for the ith batch and jth oven run, $\mathbf{x}_{ij}$ represents the corresponding levels of the treatment factors, β_0 is a fixed intercept, $\boldsymbol{\beta}$ is a vector of fixed parameters, δ_i is a random batch effect and ϵ_{ij} is a random oven run effect. It is usually assumed that all random effects are independent and that $\delta_i \sim N(0, \sigma^2_\delta)$ and $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$.

It is most common to estimate the random effects using residual maximum likelihood (REML) and the fixed effects using generalized least squares (GLS), although Gilmour and Goos (2009) recently gave a warning about the automatic use of REML-GLS in experiments with small numbers of whole plots. The experiment described here is on the borderline of the size at which the Bayesian analysis described by Gilmour and Goos will be beneficial. This is because the experiment involves 20 batches or whole plots, which should allow for a proper estimation of σ^2_δ, even after fitting the main effects of the additives and some of the interactions involving EPDM. Therefore, we will not pursue a Bayesian approach here.

It is not too difficult to identify the appropriate linear mixed model in this case, since unbalanced split-plot designs have become quite familiar in the industrial experiments literature. However, we find it beneficial to write down an approximate skeleton analysis of variance, ignoring the nonorthogonality of treatment effects with random terms. For the surface tension response the approximate skeleton analysis of variance is given in Table 2. It is only approximate because the nonorthogonality of effects involving $X_8, \ldots, X_{11}$ with the batch effects means that some of the residual degrees of freedom in the batches stratum will be taken up by inter-batch information on these treatment effects. We find it beneficial to sketch such analyses to clarify the relationship between the randomization and the mixed model. In an orthogonal design, this analysis of variance table would give exactly the same analysis as the linear mixed model (1) analysed using REML-GLS.

Another useful tool for understanding the design structure is the Hasse diagram. Its uses for identifying structure in experiments in order to determine appropriate analysis of variance models were described by Taylor and Hilton (1981) and Lohr (1995). However, following Bailey (2007), we prefer to use separate Hasse diagrams to identify the structure in the experimental units, leading to random effects, and to identify structure in the treatments, leading to fixed effects. In the polypropylene experiment, the treatment structure is clear, so we concentrate on the structure in the experimental units.

The Hasse diagram for the surface tension response is very simple and is shown in Figure 1. Each level of randomization creates a node in the graph, while the edges show the nesting. The highest stratum always represents the entire experiment, often referred to as the universe (U). The lowest stratum

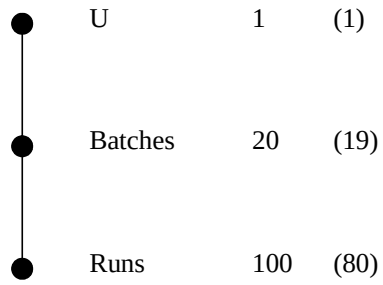
Table 2: Approximate skeleton analysis of variance for the surface tension response

Stratum	Source of Variation	df
Mean	Total	1
Batch	$X_1, \ldots, X_7$	7
	$X_1 * X_2, \ldots, X_1 * X_7$	6
	Residual	6
	Total	19
Run	$X_8, \ldots, X_{11}$	8
	$X_8 * X_9, \ldots, X_{10} * X_{11}$	9
	$X_1 * X_8, \ldots, X_7 * X_{11}$	35
	Residual	28
	Total	80

always represents the observational units. Here, the observational units are identical to the experimental units, since each oven run produced only a single surface tension observation. The first number after each stratum name in the Hasse diagram gives the number of units in that stratum, while the second number (in brackets) gives the degrees of freedom of the corresponding stratum. This is obtained by subtracting the degrees of freedom for higher strata from the number of units in the stratum under consideration.

We find these classical tools useful in understanding the structure of complex

Figure 1: Hasse diagram for the surface tension response



multi-stratum designs. They allow us to go from the specific randomization used, through the Hasse diagram and the approximate skeleton analysis of variance, to the corresponding linear mixed model. This mixed model, obtained as a direct result of the randomization in an orthogonal design, is then the appropriate one for our analysis, as best justified by the randomization.

We analysed the data for the surface tension using the linear mixed model (1). We started by fitting a model including all the effects listed in Table 2. The quadratic effects and most of the two-factor interaction effects were not significant. After a backward stepwise elimination procedure, which respected the marginality of the models, we obtained the model summarized in Table 3. Most of the significant two-factor interaction effects in the final model involve the contrast 'Type of gas' and are quite large. This indicates that the impact of the factors very strongly depends on whether or not the etching gas is used. It turns out that the factor effects are all substantially larger when the etching gas is used rather than an activation gas. We did not find any significant differences between the two types of activation gases in the analysis of the surface tension.

For the final model, σ^2_δ and σ^2_ϵ were estimated to be 2.7148 and 8.9420, respectively, which shows that there is some batch-to-batch variation and confirms that the REML-GLS analysis does not cause any major problems here. Note that, in Table 3, Kenward-Roger degrees of freedom are used, and that the degrees of freedom for the main effects of ethylene, EVA and lubricant are substantially smaller than those for the main effects of power, flow, time and type of gas, and for the two-factor interaction effects. This is in line with the approximate analysis of variance in Table 2, but not identical because the final model in Table 3 involves fewer effects than the model considered in Table 2 and because of the slight nonorthogonality of the D-optimal design used.

4 Success of Coating

The surface tension response was not of direct interest in itself, but was considered a potential surrogate for the success of the coating. In addition to the 100 plates used for measuring surface tension, another several hundred

Table 3: Linear mixed model analysis of the surface tension

Effect	Estimate	SE	DF	t	P-value
Intercept	11.1596				
Ethylene	−2.0930	0.4820	15.4	−4.34	0.0005
EVA	1.2018	0.4838	15.6	2.48	0.0248
Lubricant	−0.3069	0.4843	15.6	−0.63	0.5355
Talcum	−0.4909	0.4968	16.0	−0.99	0.3377
Power	4.9873	0.3554	72.3	14.03	<.0001
Flow	−1.6234	0.3633	75.2	−4.47	<.0001
Time	2.0977	0.3787	72.9	5.54	<.0001
Type of gas	8.9690	0.4609	73.3	19.46	<.0001
Time * Type of gas	3.0333	0.5356	76.9	5.66	<.0001
Power * Type of gas	6.5682	0.5497	78.2	11.95	<.0001
Flow * Type of gas	−2.8431	0.5415	74.6	−5.25	<.0001
Power * Flow	−0.9124	0.3992	74.7	−2.29	0.0251
Time * Talcum	−1.4290	0.3755	73.2	−3.81	0.0003
EVA * Type of gas	1.7659	0.4697	73.3	3.76	0.0003
Lubricant * Type of gas	−1.8145	0.4611	72.7	−3.94	0.0002

were processed using the plasma treatments, with the same design, but on a separate occasion. Several plates were used for each oven run and three of these were randomly selected for coating. Coating was done on the third day after processing and the plates were then dried. The coating was tested using a cross-cut test, which involved carving a regular grid on the plates, applying tape and pulling it off. These were then assessed visually using the American Standard Test Method (ASTM) score, a six point scale (0-5), considered the standard for this type of assessment. The coating was considered acceptable if it resulted in an ASTM score of at least three.

We also find the same steps as used for the linear mixed model useful in identifying the appropriate linear predictor in generalized linear mixed models with categorical response data. Compared with the design used for the surface tension, there is now an additional complication due to the fact that three plates were coated and scored for each of the 100 oven runs. As a result, there are three observational units, often referred to as repeats, for every experimental unit. This leads to the Hasse diagram shown in Figure 2, where we call the experimental units runs and the observational units tests. The corresponding approximate skeleton analysis of variance is given in Table 4. This immediately shows us that, if the response were normally distributed, an appropriate linear mixed model would be

$$Y_{ijk} = \beta_0 + \mathbf{x}'_{ij}\boldsymbol{\beta} + \delta_i + \epsilon_{ij} + \phi_{ijk}, \tag{2}$$

where Y_{ijk} is the response from the kth test on the jth oven run from the ith batch and $\phi_{ijk} \sim N(0, \sigma^2_\phi)$ is the corresponding random effect. Model (2) can be rewritten in the form of a generalized linear mixed model as

$$Y_{ijk}|\delta_i, \epsilon_{ij} \sim N(\mu_{ij}, \sigma^2_\phi), \tag{3}$$

where

$$\mu_{ij} = \beta_0 + \mathbf{x}'_{ij}\boldsymbol{\beta} + \delta_i + \epsilon_{ij},$$

$\delta_i \sim N(0, \sigma^2_\delta)$, $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$ and all random variables are independent.

We deal with the fact that the success of coating is a binary response variable in the usual way, by assuming it has a Bernoulli distribution and using a logistic link to a linear predictor which has the same form as the normal generalized linear mixed model (3). Thus we have a generalized linear mixed

Figure 2: Hasse diagram for the success of coating

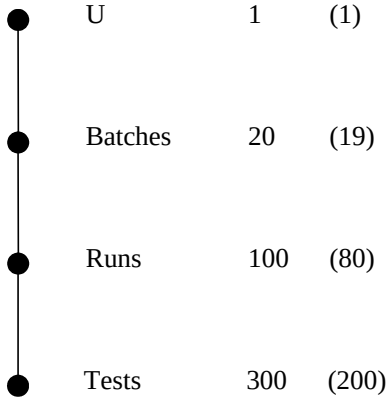


Table 4: Approximate skeleton analysis of variance for the success of coating and the ASTM score

Stratum	Source of Variation	df
Mean	Total	1
Batch	$X_1, \ldots, X_7$	7
	$X_1 * X_2, \ldots, X_1 * X_7$	6
	Residual	6
	Total	19
Run	$X_8, \ldots, X_{11}$	8
	$X_8 * X_9, \ldots, X_{10} * X_{11}$	9
	$X_1 * X_8, \ldots, X_7 * X_{11}$	35
	Residual	28
	Total	80
Test	Total	200

Table 5: Mixed binary logit model analysis of the success of coating 2

Effect	Estimate	SE	DF	t	P-value
Intercept	4.1593				
EPDM	1.1961	0.6016	9.81	1.99	0.0754
Ethylene	1.5689	0.5882	10.25	2.67	0.0231
Talcum	2.1786	0.7443	11.31	2.93	0.0134
Mica	1.3322	0.6840	8.975	1.95	0.0834
Power	−0.8050	0.3737	55.17	−2.15	0.0356
Time	2.6360	0.5009	91.44	5.26	<.0001
Type of gas	2.9872	0.6892	71.99	4.33	<.0001
Activation gas	−1.4446	0.3792	55.10	−3.81	0.0004
Power *Activation gas	1.3652	0.4223	72.20	3.23	0.0018

model for Y_{ijk}, the success (1) or failure (0) of the kth test on the jth observation from the ith batch. The model, which we refer to as a mixed binary logit model, can be written as

$$Y_{ijk}|\delta_i, \epsilon_{ij} \sim \text{Bernoulli}(\pi_{ij}), \tag{4}$$

where

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \mathbf{x}'_{ij}\boldsymbol{\beta} + \delta_i + \epsilon_{ij},$$

π_{ij} is the probability of success for the jth oven run from batch i, $\delta_i \sim N(0, \sigma_\delta^2)$, $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ and all random variables are independent.

For each of the five coatings, we analyzed the success of coating using the mixed binary logit model (4). The results for coatings 1, 2 and 5 are very similar. The simplified model we obtained for coating 2 is displayed in Table 5. The main effects of the factors EPDM, ethylene, talcum, mica, power and time are at least borderline significant, as well as the type of gas (etching gas versus activation gas) and the type of activation gas (activation gas 1 versus 2). We also found a significant interaction effect between power and the type of activation gas.

Comparing the results for the analysis of the surface tension in Table 3 and the ones for the success of coating 2 in Table 5, we observe that the type of activation gas used makes a significant difference for the success of coating,

but not for the surface tension. A striking difference is also that the presence of ethylene has a positive impact on the success of coating, while it has a negative relationship with the surface tension. There are thus substantial differences between the two models, which suggests that surface tension is not a very good surrogate for the success of coating 2.

The two variance components in the binary logit model (4), σ_δ^2 and σ_ϵ^2, are estimated to be 3.6670 and 1.1554, respectively. The positive estimate for σ_δ^2 again indicates batch-to-batch variation, while that for σ_ϵ^2 confirms that the three observational units (named tests in Figure 2) within every experimental unit (named runs in Figure 2) are dependent. The Kenward-Roger degrees of freedom in Table 5 indicate that less information is available about the additives EPDM, ethylene, talcum and mica than about the gas plasma treatment factors power, time and type of gas.

Analyzing the data for the coatings 3 and 4 leads to several differences from the results obtained for coatings 1, 2 and 5. For both coatings 3 and 4, σ_δ^2 is estimated to be zero which suggests that there is no batch-to-batch variation for these coatings. The estimate of the other variance component was similar to that reported for coating 2. Other differences were found too. For example, the type of activation gas used does not have a significant impact on the success of coating 3. However, many of the significant effects for the success of coating 3 are very similar to those for the success of coating 2. An overview of the estimates of significant effects and variance components is given in Table 6. The intercepts in the overview clearly indicate that, all other things being equal, coatings 2 and 3 (which have the highest intercepts) are more likely to be successful than the other coatings. Coating 4 (which has the smallest intercept) is the least likely to be successful.

So far, we have simply used the randomization to define the appropriate random effects and the nature of the responses to define appropriate distributions and link functions. In the skeleton analyses of variance produced, we have assumed that we are interested in estimating all relevant interactions, as well as the main effect of each factor. This is a reasonable aim, but might be rather ambitious for discrete data, given the numbers of parameters involved. Estimating the mixed binary logit model turns out to be a real challenge when interactions are included in the model. The inclusion of several interactions together often leads to convergence problems. Therefore,

the only workable model selection strategy was a forward selection procedure, where a main-effects model is estimated first and interaction effects involving factors with significant main effects are added one by one. Even then, convergence problems are likely to occur when adding certain interactions. For instance, adding the two-factor interaction effect of EPDM and type of gas when analyzing the success of coating 2 causes the convergence to fail. As pointed out by Chipman and Hamada (1996), severe problems with fitting generalized linear models for categorical data are quite common and, in our experience, these problems are exacerbated when fitting generalized linear mixed models.

In the next section, we move from a binary response to an ordered categorical response, with six outcome categories. One might expect that the better quality of the response data would lead to fewer problems with convergence and we investigate whether using that response has a beneficial impact on the analysis.

Table 6: Overview of the separate analyses for the success of coatings 1-5

	Coating 1		Coating 2		Coating 3		Coating 4		Coating 5	
Effect	Est.	P-value	Est.	P-value	Est.	P-value	Est.	P-value	Est.	P-value
Intercept	1.7675		4.1593		4.2681		−2.1076		2.2643	
EPDM	0.9889	0.0112	1.1961	0.0754	1.7732	0.0003			1.2928	0.0037
Ethylene	0.7646	0.0336	1.5689	0.0231	1.6659	0.0009			0.7235	0.0613
Talcum	0.9779	0.0152	2.1786	0.0134	0.9276	0.0548			1.9853	0.0005
Mica			1.3322	0.0834	0.8123	0.1001			1.3500	0.0088
Lubricant					0.8690	0.0376				
Power			−0.8050	0.0356					1.0680	0.0011
Time	1.4733	<.0001	2.6360	<.0001	2.6006	<.0001	0.6896	0.0519	1.5594	0.3069
Type of gas	1.7626	0.0003	2.9872	<.0001	3.3645	0.0023	1.1864	0.0049	1.6576	0.0006
Activation gas			−1.4446	0.0004					−0.0983	0.6696
Power * Type of gas									1.7646	0.0029
Power * Activation gas			1.3652	0.0018					0.4990	0.0755
Time * Type of gas									0.9348	0.0411
Batch (σ^2_δ)	0.5170		3.6670		0.0000		0.0000		1.3487	
Run (σ^2_ϵ)	2.2950		1.1554		3.7815		3.7778		0.5415	

5 ASTM Score

Although the experimenters were able to describe an ASTM score of at least 3 as leading to an acceptable product, they were also interested in how to improve it further. We might also expect to get more information by analyzing the actual ASTM scores. Since these are ordered, a natural model is the extension of model (4) to a cumulative logit model - see, for example, Agresti (2002). Since the design structure implied by the randomization is exactly as above, a suitable model would seem to be

$$Y_{ijk}|\delta_i, \epsilon_{ij} \sim \text{Multinomial}(1, \boldsymbol{\pi}_{ij}), \tag{5}$$

where

$$\log\left(\frac{\sum_{a=1}^{c}\pi_{aij}}{\sum_{b=c+1}^{6}\pi_{bij}}\right) = \beta_{c0} - \mathbf{x}'_{ij}\boldsymbol{\beta} + \delta_i + \epsilon_{ij}$$

and $\boldsymbol{\pi}'_{ij} = [\pi_{1ij} \ \cdots \ \pi_{6ij}]$, with $\pi_{aij} = P(Y_{ijk} = a)$. The parameters β_{c0}, $c = 1, \ldots, 5$, are intercept parameters representing the overall levels falling into each category. Furthermore, $\delta_i \sim N(0, \sigma^2_\delta)$, $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$ and all random variables are independent. Note that the intercepts are the only parameters which depend on which comparison of categories is being made. Hence, we are assuming proportional odds. Although this need not be true, it is a plausible assumption and with such a complex model it will be difficult to detect departures from proportional odds which could not be fixed by changing some other aspect of the model.

We analyzed the ASTM scores for the five coatings using the cumulative binary logit model (5). The simplified model we obtained for coating 2 is displayed in Table 7. The main effects of the factors EPDM, ethylene, talcum and time are clearly highly significant. Also the type of gas (etching gas versus activation gas) and type of activation gas (activation gas 1 versus 2) have significant effects. The factor power does not have a significant main effect, but its interactions with the type of gas and the type of activation gas are highly significant. There is also some indication that the interaction between mica and the type of activation gas has an effect on the ASTM score.

Comparing Table 7 for the mixed cumulative logit analysis with Table 5 for the mixed binary logit analysis shows that more significant effects were found using the ASTM score as the response, suggesting that this analysis

Table 7: Mixed cumulative logit model analysis of the ASTM score of coating 2

Effect	Estimate	SE	DF	t	P-value
Intercept 1	−0.4988				
Intercept 2	1.4035				
Intercept 3	3.1170				
Intercept 4	4.9161				
Intercept 5	6.2085				
EPDM	0.7413	0.3676	12.83	2.02	0.0652
Ethylene	1.3152	0.3646	13.99	3.61	0.0029
Talcum	1.4867	0.4470	14.60	3.33	0.0048
Mica	0.7568	0.4534	12.20	1.67	0.1205
Power	−0.3134	0.2964	55.21	−1.06	0.2950
Time	1.9307	0.3136	67.10	6.16	$< .0001$
Type of gas	2.3826	0.4270	70.15	5.58	$< .0001$
Activation gas	−0.5931	0.3075	48.91	−1.93	0.0596
Power * Type of gas	1.2123	0.4750	63.23	2.55	0.0131
Power * Activation gas	0.8434	0.3333	52.63	2.53	0.0144
Ethylene * Power	−0.5763	0.2886	52.54	−2.00	0.0510
Mica * Activation gas	0.5732	0.3033	48.29	1.89	0.0648

is indeed more powerful. Some of the newly detected effects are small, but the interaction effects involving the factor power are certainly as large as the significant main effects. The estimates of the effects that the binary and the cumulative logit model have in common possess the same signs, indicating that many of the qualitative conclusions from the two models will be similar.

The two variance components in the model, σ_δ^2 and σ_ϵ^2, are estimated to be 1.2507 and 3.7213, respectively, which indicates substantial batch-to-batch variation as well as dependence between the three observational units (the tests) within every experimental unit (i.e. within every oven run). The impact of the randomization can be seen from the degrees of freedom for the significance tests in Table 7, where 12 to 15 degrees of freedom are used for the whole-plot factor effects and between 48 and 68 degrees of freedom for the sub-plot effects. When taking into account the fact that the model in Table 7 is a simplified model, these degrees of freedom are in line with the

approximate skeleton analysis of variance in Table 4.

An overview of the results from the analysis for the ASTM scores for the five coatings is given in Table 8. The table shows that the type of activation gas has a significant impact on the ASTM scores for four of the five coatings, either through a main effect or through one or more interaction effects. Also, the presence of EPDM has a positive impact on four of the ASTM scores. This confirms our earlier results from the binary logit model. Using the cumulative logit model, however, we also find two significant quadratic effects plus several main effects and interaction effects that were not detected using the binary logit model. This demonstrates the added value of using the raw ASTM scores, rather than the binary response, the success of coating. Another difference between the analyses using the cumulative and binary logit models is observed for coating 3. Fitting the cumulative logit model for that coating's ASTM score yields a positive estimate for σ_δ^2, which measures the batch-to-batch variation. In the simplified cumulative logit model, the estimate amounts to 2.4521, whereas it is zero in the simplified binary model. As a result of this, the main effects of talcum, mica and lubricant, which were borderline significant in the binary logit model, are no longer significant in the cumulative logit model.

Remarkably, we encountered almost no convergence problems with the mixed cumulative logit model when using our forward selection procedure, even when including several potentially interesting interaction effects. This is in contrast with the mixed binary logit model where almost no interaction effects were estimable. Clearly, using a six-category ordinal response variable rather than a binary response allows for a more refined analysis. For some of the coatings, σ_δ^2 is still estimated to be zero.

Table 8: Overview of analysis for ASTM score of coatings 1-5

	Coating 1		Coating 2		Coating 3		Coating 4		Coating 5	
Effect	Est.	P-value	Est.	P-value	Est.	P-value	Est.	P-value	Est.	P-value
Intercept 1	0.8248		−0.4988		2.3275		−7.3705		−2.7046	
Intercept 2	1.9447		1.4035		2.7446		−5.8842		1.9182	
Intercept 3	3.1880		3.1170		4.1791		−3.0111		3.0387	
Intercept 4	4.2825		4.9161		5.4898		−1.9059		4.2061	
Intercept 5	4.9474		6.2085		6.4184		−1.2401		5.0899	
EPDM	0.9183	0.0008	0.7413	0.0652	1.6215	0.0076			1.2360	< .0001
Ethylene	0.8533	0.0014	1.3152	0.0029	2.0649	0.0016			0.8856	0.0002
Talcum	1.1160	0.0007	1.4867	0.0048			0.7201	0.0959	1.7214	< .0001
Mica	0.7599	0.0191	0.7568	0.1205					1.1463	0.0001
Lubricant									0.7526	0.0013
UV									−0.9235	0.0001
Power			−0.3134	0.2950	0.3919	0.4391			0.9226	0.0007
Time	1.7875	< .0001	1.9307	< .0001	2.8145	< .0001	1.0154	0.0395	1.6122	< .0001
Type of gas	2.1018	< .0001	2.3826	< .0001	3.5504	0.0001	1.5205	0.0119	1.6407	< .0001
Activation gas	−0.7779	0.0093	−0.5931	0.0596	−0.8873	0.0175			−0.0097	0.9700
Power * Type of gas			1.2123	0.0131	1.8217	0.0461			1.1374	0.0045
Power * Act. gas			0.8434	0.0144					0.7153	0.0191
EPDM * Act. gas	0.5490	0.0614								
EPDM * Ethylene	0.7519	0.0043							−0.5386	0.0193
Ethylene* Power			−0.5763	0.0510						
Mica * Act. gas			0.5732	0.0648						
Power * Power									−1.3002	0.0275
Time * Time	−1.1530	0.0574								
Batch (σ^2_δ)	0.0000		1.2507		2.4521		0.0000		0.0000	
Run (σ^2_ϵ)	3.8584		3.7213		4.2720		11.1528		2.7527	

6 Combined Analysis for Different Coatings

The analyses described in Sections 4 and 5 were carried out for five different coatings on five different occasions, several weeks apart. However, the experimenters were mainly interested in comparing the effects of the factors on different coatings. This necessitates the combined analysis of the data for all five coatings. We start this section by discussing the randomization structure which led to the combined data involving the 11 factors listed in Table 1 and coating as the twelfth experimental factor.

Although there was no formal randomization of coatings to occasions, the sequence in which they were run can be considered as being essentially random (and different random samples of plates were used on each occasion). It would seem natural, and might have been better, to separately randomize the orders of the 100 oven runs on each occasion. This would have lead to the Hasse diagram shown in Figure 3. In this case the randomizations of batches and occasions are crossed, i.e. each batch is used on each occasion, and so the combinations of these define another stratum, as in a strip-plot design. The units in the Batches*Occasions stratum, which would be the lowest stratum in a classical strip-plot design, are split into smaller units (named runs), each of which involves three observational units (the tests).

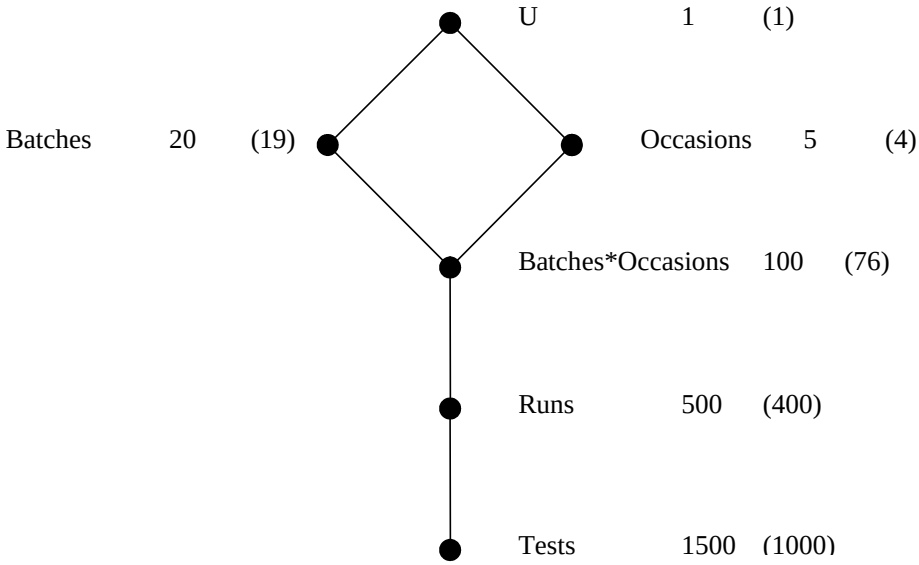
In fact, the 100 oven runs were performed in the same order on each of the five occasions, which means that the orders and occasions are also crossed. This leads to the Hasse diagram shown in Figure 4. In this case the combinations of orders and occasions must be included in the Hasse diagram and these correspond to the 500 oven runs, which are the experimental units in this case. With some familiarity with Hasse diagrams, it becomes very easy to produce them from a clear description of the randomization. They can then be easily translated into the approximate skeleton analysis of variance, that for the combined analysis being shown in Table 9.

The approximate skeleton analysis of variance can then be used to write down the corresponding linear mixed model for an orthogonal design. Expressing this linear mixed model in the form of a generalized linear mixed model gives the appropriate linear predictor which can be combined with a suitable distribution and link function to obtain a suitable model for our categorical response data. The skeleton analysis of variance in Table 9 imme-

Table 9: Approximate skeleton analysis of variance for the combined analysis of the ASTM scores for all coatings

Stratum	Source of Variation	df
Mean	Total	1
Batch	$X_1, \ldots, X_7$	7
	X_1 * X_2, ..., X_1 * X_7	6
	Residual	6
	Total	19
Occasion	Coatings	4
	Residual	0
	Total	4
Batch * Occasion	X_1 * Coating, ..., X_7 * Coating	28
	X_1 * X_2 * Coating, ..., X_6 * X_7 * Coating	24
	Residual	24
	Total	76
Order	$X_8, \ldots, X_{11}$	8
	X_8 * X_9, ..., X_{10} * X_{11}	9
	X_1 * X_8, ..., X_7 * X_{11}	35
	Residual	28
	Total	80
Run	X_8 * Coating, ..., X_{11} * Coating	32
	X_8 * X_9 * Coating, ..., X_{10} * X_{11} * Coating	36
	X_1 * X_8 * Coating, ..., X_7 * X_{11} * Coating	140
	Residual	112
	Total	320
Test	Total	1000

Figure 3: Hasse diagram for combined analysis if orders of oven runs had been randomized



diately shows that we cannot estimate a random occasion effect in a linear mixed model, because all the degrees of freedom in this stratum are used to estimate the effect of the coating. Similarly, it is impossible to estimate the variance component corresponding to the occasions when analyzing our categorical responses, even though the multinomial distribution has a fixed scale parameter, i.e. unlike in the linear mixed model (2), there is no σ^2_ϕ to estimate. Therefore we exclude the random occasion effect from our analysis. Otherwise, we simply follow the structure of this analysis of variance to suggest appropriate random effects for our generalized linear mixed model.

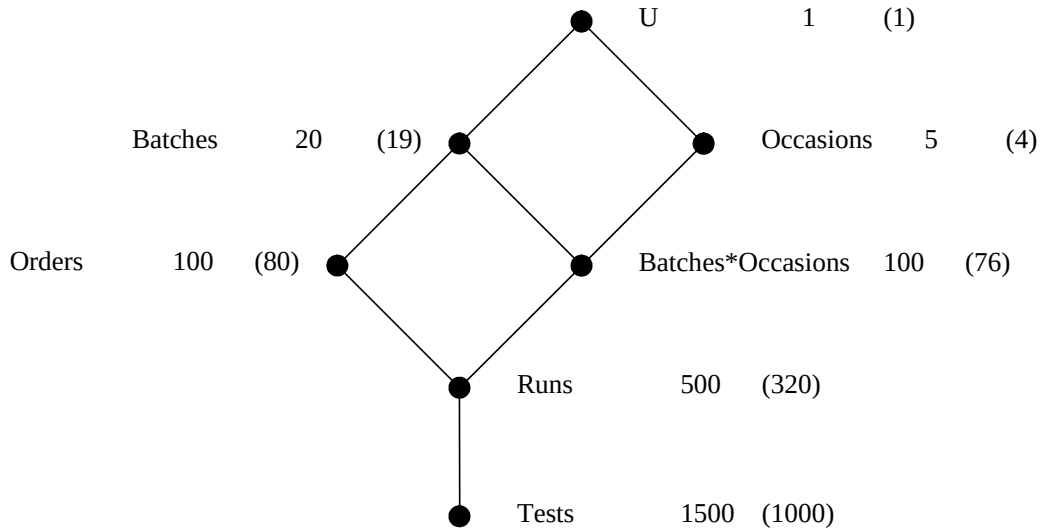
An appropriate model is

$$Y_{ijkl}|\delta_i, \gamma_{ij}, \lambda_{ik}, \epsilon_{ijk} \sim \text{Multinomial}(1, \boldsymbol{\pi}_{ijk}), \tag{6}$$

where

$$\log\left(\frac{\sum_{a=1}^{c}\pi_{aijk}}{\sum_{b=c+1}^{6}\pi_{bijk}}\right) = \beta_{c0} - \mathbf{x}'_{ijk}\boldsymbol{\beta} + \delta_i + \gamma_{ij} + \lambda_{ik} + \epsilon_{ijk},$$

Figure 4: Hasse diagram for combined analysis



$\delta_i \sim N(0, \sigma_\delta^2)$, $i = 1, \ldots, 20$, is a random batch effect, $\gamma_{ij} \sim N(0, \sigma_\gamma^2)$, $j = 1, \ldots, n_i$, is a random effect for the orders within batch i, $\lambda_{ik} \sim N(0, \sigma_\lambda^2)$, $k = 1, \ldots, 5$, is a random effect for the combinations of batches and occasions, $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$ is a random oven run effect, $l = 1, 2, 3$ denotes the test and all random variables are independent.

We analyzed the ASTM scores for the five different coatings simultaneously using the mixed cumulative logit model (6). In our analysis, we used four different contrasts to capture the effects of the categorical factor coating. The first contrast, which we named 'Solvent-based coat 1 vs. 2' (C_1) compares the solvent-based two-layer coating 1 (1-component base coat + 2-component top coat) with coating 2, which is a solvent-based single-layer coating (using a 2-component coat). The second contrast, labeled 'Water- vs. solvent-based' (C_2), compares the water-based two-layer coating 3 (1-component base coat + 2-component top coat) with the two solvent-based coatings 1 and 2. The third contrast, named 'UV coat vs. traditional' (C_3), compares coating 5,

which is a coating that is dried using UV light, with the traditional solvent- and water-based coatings 1, 2 and 3. The final contrast, labeled 'Low-end coat vs. rest' (C_4), compares coating 4, which is a water-based single-layer low-end product, with the other four coatings, each of which are high-quality coatings.

The final model we obtained is summarized in Table 10. This model was obtained using a manual stepwise regression, starting from an initial model involving the main effects of the 11 experimental factors listed in Table 1, the four contrasts for the coatings and the interactions between these contrasts and the experimental factors' main effects. The purpose of the interactions is to quantify the extent to which the main effects differ for the different types of coatings. When fitting the initial model, two of the four variance components, σ^2_δ and σ^2_λ, are estimated to zero. This leads to standard errors that are smaller than expected, and causes the Kenward-Roger degrees of freedom to be unjustifiably large compared with the skeleton analysis of variance in Table 9. After dropping some of the non-significant effects, the estimates of the variance components are positive and the degrees of freedom produced by the Kenward-Roger method are in line with those suggested by the skeleton analysis of variance. The variance component estimates for the final model are given in Table 11. The estimate for σ^2_δ indicates that there is some batch-to-batch variation, and the estimate for σ^2_λ suggests that the batch-to-batch variation was slightly different between occasions. By far the largest variance component estimate is that for σ^2_ϵ. As the separate analyses for the different coatings already showed, this is a strong indication that the three repeated observations within each experimental unit are strongly correlated.

The combined analysis confirms many of the conclusions drawn from the separate analyses in Table 8. Most main effects are highly significant. In addition, several two-factor interaction effects are significant or nearly significant at the 5% level. Also, three of the four coating contrasts appear highly significant. It should be pointed out, however, that the standard errors of these contrasts are almost certainly underestimated and that the degrees of freedom are overestimated. This is because we cannot estimate a variance component for the occasions. Therefore, some prudence is required in concluding that the three coating contrasts are significant. This is especially so for the 'Water- vs. solvent-based' contrast, C_2, which, in absolute value, has

Table 10: Mixed cumulative logit model analysis of the ASTM scores for all five coatings combined

Effect	Estimate	SE	DF	t	P-value
Intercept 1	−1.0395				
Intercept 2	0.9763				
Intercept 3	2.4495				
Intercept 4	3.6824				
Intercept 5	4.5022				
Water- vs. solvent-based (C_2)	1.2177	0.2527	58.43	4.82	$< .0001$
UV coat vs. traditional (C_3)	−1.7073	0.2517	47.92	−6.78	$< .0001$
Low-end coat vs. rest (C_4)	−4.9763	0.3127	102.10	−15.91	$< .0001$
EPDM	0.9819	0.2210	12.30	4.44	0.0008
Ethylene	1.1300	0.2196	12.62	5.15	0.0002
Talcum	1.2519	0.2676	13.46	4.68	0.0004
Mica	0.8871	0.2742	11.71	3.24	0.0074
UV	0.0559	0.2166	12.76	0.26	0.8006
Power	0.1385	0.1891	69.81	0.73	0.4663
Time	1.8894	0.1946	80.94	9.71	$< .0001$
Type of gas	2.0659	0.2545	82.14	8.12	$< .0001$
Activation gas	−0.7051	0.1885	66.58	−3.74	0.0004
Power * Type of gas	0.8238	0.2905	75.26	2.84	0.0059
EPDM * Ethylene	−0.0424	0.2163	12.07	−0.20	0.8478
Time * Time	−0.8425	0.4110	72.59	−2.05	0.0440
C_3 * UV	−0.6982	0.2442	46.58	−2.86	0.0063
C_3 * Time	−0.5631	0.2614	257.60	−2.15	0.0321
C_3 * Power	0.6464	0.2585	236.30	2.50	0.0131
C_3 * Type of gas	−0.7361	0.3454	262.30	−2.13	0.0340
C_3 * Activation gas	0.6724	0.2587	221.90	2.60	0.0100
C_3 * EPDM	0.0488	0.2448	43.44	0.20	0.8428
C_3 * Ethylene	−0.4075	0.2450	43.21	−1.66	0.1035
C_3 * EPDM * Ethylene	−0.6143	0.2446	43.71	−2.51	0.0158
C_4 * Ethylene	−0.5733	0.2940	80.93	−1.95	0.0547

Table 11: Variance component estimates corresponding to the final model for the combined analysis of the ASTM scores for the five coatings

Variance component		Estimate	SE
σ^2_δ	Batch	0.3449	0.3721
σ^2_γ	Order	0.9761	0.4384
σ^2_λ	Batch * Occasion	0.2201	0.3404
σ^2_ϵ	Run	4.4347	0.5909

the smallest point estimate of the three.

The final model does not involve the contrast 'Solvent-based coat 1 vs. 2' (C_1), which means that we found no evidence that the effects of the experimental factors differ between the two solvent-based coatings in the study. Note that it is safe to assume that this contrast is not significantly different from zero because a term that is insignificant with an underestimated standard error will also be insignificant with a correct standard error. After dropping this insignificant contrast, one degree of freedom is potentially available for estimating the variance component for the occasions. However, including random effects for the occasions in the model leads to a zero estimate for that variance component, so that the results remain unchanged.

As the contrast 'Water- vs. solvent-based' (C_2) appears to have a significant positive value, there is some indication that a good adhesion is easier to achieve for water-based than for solvent-based coatings. There is stronger evidence that it is in general harder to achieve good adhesion for UV-dried coatings and much harder for low-end coatings, as can be seen from the apparently significant negative estimates for the contrasts 'UV coat vs. traditional' (C_3) and 'Low-end coat vs. rest' (C_4).

The significance of several interaction effects involving the contrast 'UV coat vs. traditional' (C_3) implies that there is a significant difference between the UV-dried coating 5, on the one hand, and the traditional solvent-based and water-based coatings, on the other hand. The final model also contains one three-factor interaction involving the contrasts 'UV coat vs. traditional' (C_3). The significance of this interaction suggests an antagonistic interaction effect between the factors EPDM and ethylene for the UV-dried coating 5. The

interaction of the factor ethylene and the contrast 'Low-end coat vs. rest' (C_4) suggest that the factor ethylene has a smaller positive impact on the adhesion for low-end coatings than for other coatings. Note that only the standard errors and the degrees of freedom for the coating contrasts are affected by the fact that the variance component for the occasions cannot be estimated. Hence, the inference for all other effects, including the interactions with the coating contrasts, does not pose any problems.

The main advantage of the combined analysis is that it enables the experimenter to carry out formal hypothesis tests to see whether the factor effects differ from coating to coating. Moreover, if the interaction effects between the experimental factors and the coatings are insignificant, then the information in the data can be pooled across the different coatings to acquire more precise estimates of the effects and have more powerful significance tests for at least some of the remaining factor effects.

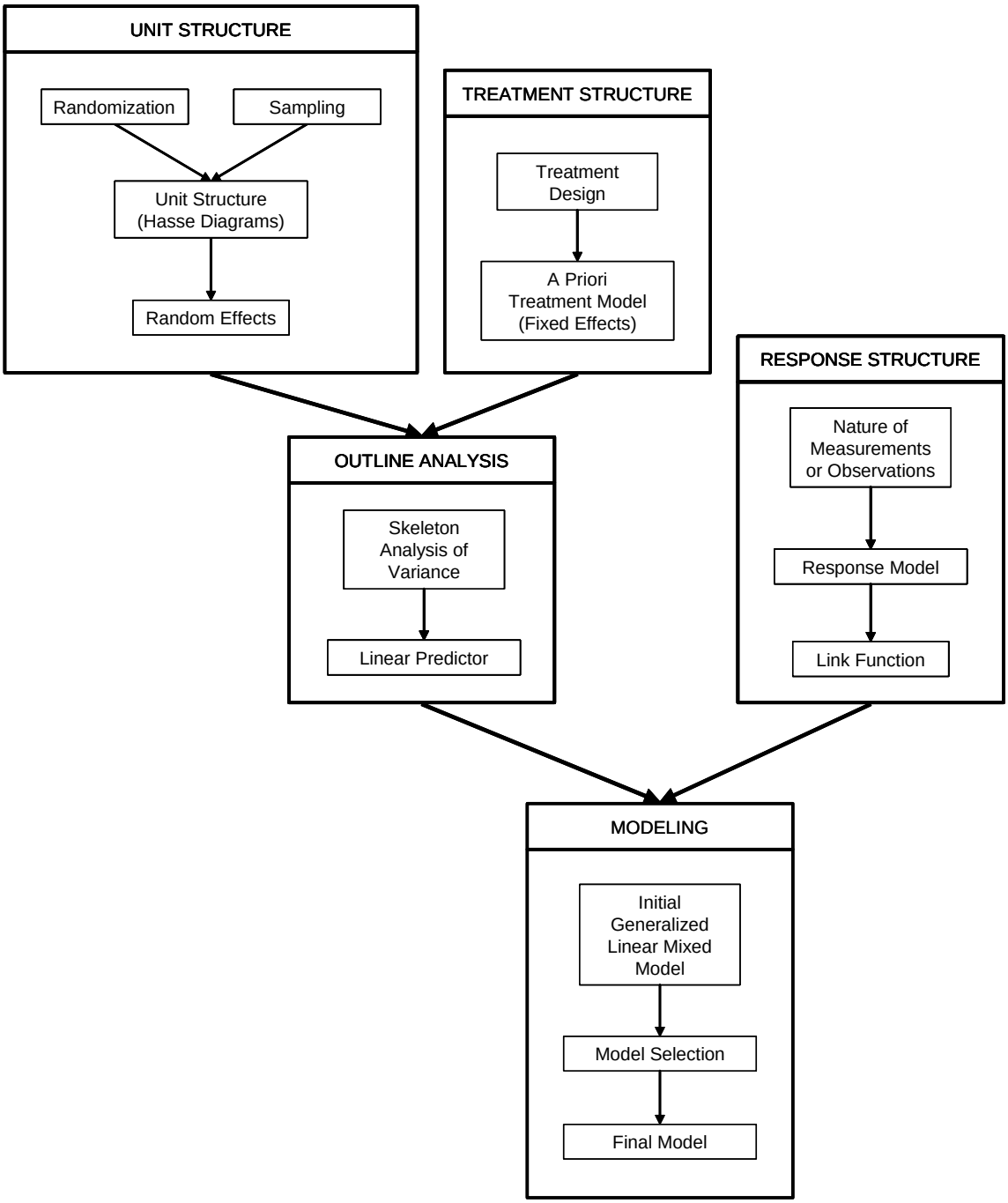
7 A General Analysis Strategy

The stages in the analysis of the polyproylene experiment have been described in some detail. In more general terms, we propose that the type of analysis strategy adopted here can be applied to the analysis of categorical response data from any multi-stratum design. Our suggested strategy is outlined in Figure 5, the different steps of which we now describe in detail.

The structure in the experimental units is determined by the randomization. The experimental units to be used and perhaps one or more blocking factors are randomized by randomly allocating labels to the factor levels, e.g. randomly assigning the labels $1, \ldots, b$ to b blocks. If a blocking factor has treatments applied to it, we assume that the design has been produced with specific treatments attached to each block label, so that the randomization automatically randomizes the treatments to the units. For example, if the main plot design has been written down with specific treatment combinations attached to each label $1, \ldots, b$ then randomly assigning these labels to the main plots achieves the appropriate randomization.

Each level of randomization and any sampling which is done within experi-

Figure 5: General analysis strategy



mental units determines a node in the Hasse diagram. Then the usual rules for Hasse diagrams are applied, i.e.:

- if factor B is nested within factor A, B appears below A, with an edge between them;
- if factors C and D are crossed, they appear at the same level and are both connected to their *supremum* (i.e. the finest grouping in which both C and D are nested), which appears above them, and their *infimum* (i.e. the coarsest grouping which is nested within both C and D, or, equivalently, the combinations of their levels), which appears below them.
- a factor called the *universe*, representing the complete experiment, appears above all other factors;
- the observational units appear below all other factors.

Then each node in the Hasse diagram represents a *stratum* in the analysis. Except for the universe and the observational units, each stratum implies a random effect in the model. The universe is represented by a fixed intercept parameter and the variation in the observational units is accounted for by the distributional assumption made.

The treatment design in the types of experiments we are discussing will have some kind of multifactorial structure, such as a fractional factorial, response surface or mixture design. Typically the design will have been chosen to efficiently fit some particular model, which we refer to as the *a priori* model in Figure 5. This model defines the fixed effects in our mixed model.

Bringing together the treatment structure and the unit structure allows us to construct the approximate skeleton analysis of variance, corresponding to a linear mixed model in a balanced design. The main effect of a treatment factor appears in the stratum defined by the blocking factor in which that treatment factor is randomized. Interactions appear in the stratum defined by the infimum of the blocking factors in which the parent treatment factors are randomized. The approximate skeleton analysis of variance allows us to check if there are too few degrees of freedom to sensibly estimate some effect and helps us to write down the linear predictor, $\boldsymbol{\eta} = \beta_0 + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}$, of the

GLMM which will be used to model our data, where $\eta_i = g(\mu_i)$ and $g(\cdot)$ is the link function.

Consideration of the nature of the response variable of interest usually leads to an obvious choice of distribution for the model as in simple GLMs. Thus, for example, dichotomous data suggest a Bernoulli distribution, counts with no upper limit suggest a Poisson distribution and polytomous data suggest a multinomial distribution. Typically, we would initially use the canonical link function (e.g. the logit link function for dichotomous data, or the log link function for count data) unless specific knowledge about how the responses arose suggests a different link function. For example, if the data are a dichotomized record of an underlying continuous variable which is believed to follow a normal distribution, then a probit link can be used. If dichotomous data represent counts of zero or greater than zero when the underlying distribution is believed to be Poisson, then a complementary log-log link is appropriate.

The assumed distribution and link function, together with the linear predictor developed, give us our initial GLMM if the responses are dichotomous or counts. For polytomous data we further have to consider the relationship between the linear predictors for the different categories, e.g. whether or not to assume proportional odds.

Now, we are finally ready to fit a model to our data. Unless the data are very rich, model selection will be an important, but difficult, issue. It is very common to find that one or more variance components are estimated to be zero and sometimes there can be convergence problems. Dealing with these to find a model which gives a simple description of the data is not entirely a prescriptive process. Knowledge of the precise objectives of the experiment and what is known about the system are important, as well some of the usual statistical tools. Things which might be considered are the standard tools of modeling, including fitting treatment models of different orders, backward elimination or forward selection starting from some specific model, merging categories in the response, changing the link function and changing the assumed distribution.

The question of how to deal with variance components which are estimated to be zero was discussed recently by Gilmour and Goos (2009). The analysis of the polypropylene data raises some other issues. Sometimes a model can

be obtained in which some treatment effect seems to be highly significant, but with variance components estimated to be zero this can be misleading. It certainly cannot be taken as strong evidence that such an effect is active. We recommend trying to drop such effects and considering the overall quality of the resulting model. Sometimes dropping some fairly highly significant effect can lead to a variance component becoming estimable, which in turn can make some previously highly significant effects become non-significant. Therefore we would recommend performing backward elimination beyond the level which would usually be done, e.g. to a significance level even lower than 1%, at least as an exploratory tool. If a plausible model can be found, which allows estimation of all variance components, then we would have a lot more confidence that such a model contains truly active effects than otherwise. Of course, other effects can be declared possibly active if their significance depends on which random effects are in the model.

8 Discussion

From our study, it should be clear that the modeling of categorical responses calls for high-quality data. The split-plot design which was used as the basis for the experiment whose data we analyzed in this paper was a D-optimal design involving 20 whole plots and 100 sub-plots for estimating main effects, a substantial number of interactions and the quadratic effects of the quantitative sub-plot factors. The treatment design was thus a high-quality design constructed using state-of-the-art methods, and the number of whole plots was substantially larger than the number of whole plot factor effects in the *a priori* model. Yet, estimating interaction effects and quadratic effects in the mixed binary and cumulative logit models we studied was often impossible due to convergence problems, even if only a few of them were included in the model.

Obviously, a large part of these problems is due to the categorical nature of the responses and to the unit structure of the experiment, which is more complex than that of a split-plot design. When planning experiments with categorical responses, any decisions regarding unit and treatment structures are thus even more important than in situations with continuous responses. We therefore suggest using the general analysis strategy outlined in Figure

5, including the use of Hasse diagrams and approximate skeleton analyses of variance, at the design stage of the experiment. The difficulties we encountered during our analysis of the polypropylene data should provide inspiration for new methodological research concerning the design of experiments with categorical response data.

In our study, we have also shown how to deal with the problem of multiple observational units within the experimental units. The dependence between repeated observations within a given experimental unit was modeled using an additional random effect in the generalized linear model. This approach is very useful for industrial experimenters who often take multiple observations, which are commonly called repeats in the literature on industrial experimental design. When repeats are taken from continuous responses, a proper statistical analysis can be done on the average responses over all repeats within an experimental unit. Obviously, this commonly used approach is not appropriate for ordered categorical outcomes such as those we have analyzed here.

Acknowledgment

This work was supported by the Royal Society International Joint Project number 2007/R2.

References

Agresti, A. (2002) *Categorical Data Analysis*, 2nd edition. New York: Wiley.

Bailey, R. A. (2007) *Design of Comparative Experiments*. Cambridge: Cambridge University Press.

Chipman, H. and Hamada, M. (1996) Bayesian analysis of ordered categorical data from industrial experiments. *Technometrics*, **38**, 1–10.

D'Ambra, L., Köksoy, O. and Simonetti, B. (2009) Cumulative correspondence analysis of ordered categorical data from industrial experiments, *Journal of Applied Statistics*, **36**, 1315–1328.

Dasgupta, T., Ma, C., Joseph, V. R., Wang, Z. L. and Wu, C. F. J. (2008) Statistical modeling and analysis for robust synthesis of nanostructures. *Journal of the American Statistical Association*, **103**, 594–603.

Gilmour, S. G. and Goos, P. (2009) Analysis of data from nonorthogonal multi-stratum designs in industrial experiments. *Applied Statistics*, **58**, 467–484.

Gilmour, S. G. and Trinca, L. A. (2006) Response surface experiments on processes with high variation. In *Response Surface Methodology* (ed. A. I. Khuri), 19–46. New York: World Scientific.

Goos, P. (2002) *The Optimal Design of Blocked and Split-Plot Experiments*. New York: Springer.

Hamada, M. and Wu, C. F. J. (1990) A critical look at accumulation analysis and related methods (with discussion). *Technometrics*, **32**, 119–162.

Jeng, S.-L., Joseph, V. R. and Wu, C. F. J. (2008) Modeling and analysis strategies for failure amplification method. *Journal of Quality Technology*, **40**, 128–139.

Jones, B. and Goos, P. (2007) A candidate-set-free algorithm for generating D-optimal split-plot designs. *Applied Statistics*, **56**, 347–364.

Jones, B. and Goos, P. (2009) D-optimal design of split-split-plot experiments. *Biometrika*, **96**, 67–82.

Joseph, V. R. and Wu, C. F. J. (2004) Failure amplification method: an information maximization approach to categorical response optimization (with discussion). *Technometrics*, **46**, 1–31.

Lee, Y. and Nelder, J. A. (2003) Robust design via generalized linear models. *Journal of Quality Technology*, **35**, 2–12.

Letsinger, J. D., Myers, R. H. and Lentner, M. (1996) Response surface methods for bi-randomization structures. *Journal of Quality Technology*, **28**, 381–397.

Lohr, S. L. (1995) Hasse diagrams in statistical consulting and teaching. *The American Statistician*, **49**, 376–381.

Myers, R. H., Montgomery, D. C. and Vining, G. G. (2002) *Generalized Linear Models: With Applications in Engineering and the Sciences*. New York: Wiley.

Nair, V. N. (1986) Testing in industrial experiments with ordered categorical data (with discussion). *Technometrics*, **28**, 283–311.

Parker, P., Kowalski, S. M. and Vining, G. G. (2007) Construction of balanced equivalent estimation second-order split-plot designs. *Technometrics*, **49**, 56–65.

Taylor, W. H. and Hilton, H. G. (1981) A structure diagram symbolization for analysis of variance. *The American Statistician*, **35**, 85–93.

Trinca, L. A. and Gilmour, S. G. (2001) Multi-stratum response surface designs. *Technometrics*, **43**, 25–33.

Vining, G. G., Kowalski, S. M. and Montgomery, D. C. (2005) Response surface designs within a split-plot structure. *Journal of Quality Technology*, **37**, 115–129.

Vivacqua, C. A. and Bisgaard, S. (2009) Post-fractionated strip-block designs. *Technometrics*, **51**, 47–55.

Wu, F. and Yeh, C. (2006) A comparative study on optimization methods for experiments with ordered categorical data. *Computers and Industrial Engineering*, **50**, 220–232.