Nonrepeatable Gauge R&R Studies Assuming Temporal or Patterned Object Variation

FRANK VAN DER MEULEN

Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands

HENK DE KONING

A.T. Kearney, B.V. Van Heuven Goedhartlaan 13, P.O. Box 22926, 1181 LE Amstelveen, 1100 DK Amsterdam, The Netherlands

JEROEN DE MAST

Institute for Business and Industrial Statistics of the University of Amsterdam (IBIS UvA), Plantage Muidergracht 12, 1018 TV Amsterdam, The Netherlands

The standard method to assess a measurement system's precision is a gauge repeatability and reproducibility (gauge R&R) study. It exploits replications to estimate variance components that are interpreted as measurement spread. For nonrepeatable measurements, it is not feasible to obtain replications because objects are destroyed when they are measured or because the object changes over time. Possible solutions are to replace replications with measurements of multiple objects or with the measurement of one object at multiple times. Subsequently, these measurements are modeled by a fixed pattern (over time or over positions). We show that the experimental design used in this type of nonrepeatable gauge R&R studies is best constructed in a way that is similar to a Latin square design. These designs have a great flexibility, can be applied in many situations encountered in practice, and have nice mathematical properties as well. We consider several examples in which this approach is applied and worked out. For the examples given, we provide the analysis and the results following the worked-out approach. Analysis of the envisaged experimental set-up is done with linear and nonlinear mixed models in which variance components are estimated by restricted maximum-likelihood estimators.

Key Words: Latin-Square Experimental Design; Measurement Error; Nonlinear Mixed Model; Restricted Maximum Likelihood; Variance Components.

The standard method to assess a measurement system's precision is a gauge repeatability and reproducibility (gauge R&R) study (see e.g., Montgomery (2005), Burdick et al. (2003)). An example of the standard layout of such a study is presented in Table 1.

Each object out of a sample of objects is measured multiple times by a number of operators. Variation within rows is measurement spread. We denote the data by y_{ijk}, where i indexes objects, j indexes operators, and k indexes replications. The data are modeled as

$$y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + \varepsilon_{ijk}. \qquad (1)$$

Dr. Van der Meulen is an Assistent Professor at Delft Institute of Applied Mathematics (Delft University of Technology). His e-mail adress is f.h.vandermeulen@tudelft.nl.

Dr. De Koning is an associate consultant with A.T. Kearney. His e-mail address is Henk.deKoning@atkearney.com.

Dr. de Mast is a principal consultant at IBIS UvA, and Associate Professor at the University of Amsterdam. He is a senior member of ASQ. His email address is j.demast@uva.nl.

TABLE 1. Standard Layout of Gauge R&R Study

	Operator					
Objects	1		2		3	
1	y_{111}	y_{112}	y_{121}	y_{122}	y_{131}	y_{132}
2	y_{211}	y_{212}	y_{221}	y_{222}	y_{231}	y_{232}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	$y_{10,1,1}$	$y_{10,1,2}$	$y_{10,2,1}$	$y_{10,2,2}$	$y_{10,3,1}$	$y_{10,3,2}$

Here μ denotes the overall average, $a_i \sim N(0, \sigma_a^2)$ are random object effects, $b_j \sim N(0, \sigma_b^2)$ are random operator effects, and $(ab)_{ij} \sim N(0, \sigma_{ab}^2)$ represent object–operator interaction. The $\varepsilon_{ijk} \sim N(0, \sigma^2)$ are error terms. All a_i, b_j, $(ab)_{ij}$, and ε_{ijk} are assumed stochastically independent. One is typically interested in the repeatability σ^2, the reproducibility $\sigma_b^2 + \sigma_{ab}^2$, and the total measurement spread $\sigma_m = \sqrt{\sigma^2 + \sigma_b^2 + \sigma_{ab}^2}$.

The standard approach exploits replications to estimate measurement spread. For some measurements, it is not feasible to obtain replications, for example, because objects are destroyed when they are measured or because the object being measured changes over time. Such measurements used to be called destructive, but are nowadays often referred to as nonrepeatable. De Mast and Trip (2005) give a precise, mathematical definition of the problem of gauge R&R studies for nonrepeatable measurements. This comes down to the following: Nonrepeatable measurements are measurements for which either of the following two conditions does not hold:

(I) *Temporal stability*, by which we mean that the real value of the object does not change in time. For example, suppose we measure the temperature of a piece of metal after heating and we wish to determine the error in this measurement. Because the metal may cool off very rapidly, the temperature is not stable in time and the condition of temporal stability is violated.

(II) *Robustness against measurement*. This condition is violated when the object is destroyed or changed significantly during measurement. An example is measuring the strength of biscuits, in which the biscuits break as a result of the measurement. Another example is measuring the rate of dissolution of a tablet. After the measurement, the tablet is (partially) dissolved, so it has changed significantly.

Performing a gauge R&R study for nonrepeatable measurements is a fundamental problem because such measurements cannot be repeated under entirely equal conditions. This problem has been a persistent problem in quality engineering. Although there is no structural solution to it, there are a number of approaches that work in some cases. De Mast and Trip (2005) give an overview of seven such approaches.

One of these approaches works with an experimental layout similar to the one in Table 1, but one in which the rows contain measurements on *different* objects instead of measurements on the same object. This experimental layout necessarily confounds measurement spread with object-to-object variation. The usual estimators for measurement spread now estimate $\sqrt{\sigma_m^2 + \sigma_a^2}$ instead of σ_m. If object-to-object variation (σ_a^2) within rows is not negligible, this approach gives an overestimation of the measurement spread. Although this is commonly the case, the approach is still useful because the bias is on the conservative side: if the estimated measurement spread is acceptable, then the true measurement spread is as well.

As suggested in De Mast and Trip (2005), this approach can be improved upon if the object-to-object spread within rows is not just noise, but has a pattern, i.e., if either of the following conditions hold:

- *Patterned temporal variation* (PTV): the variation over time of each object follows a certain pattern.
- *Patterned object variation* (POV): the variation across objects follows a certain pattern.

The idea is to fit a model for this systematic part of the within-rows-objects variation (condition POV) or temporal variation (condition PTV) and correct the data for it. This approach leads to a smaller overestimation of measurement spread. Because we correct for systematic differences between the objects within a row, the estimators for measurement spread ($\sqrt{\sigma_m^2 + \sigma_a^2}$) will be closer to σ_m.

The approaches outlined above require a more advanced experimental set-up and analysis than standard gauge studies. This paper considers three cases, which are each introduced by a practical example:

1. I is violated, but PTV holds. Objects vary over

time, but the variation over time follows a certain pattern that can be modeled.

2. II is violated, but POV holds. Objects change during measurement, but their variation follows a pattern that can be modeled.
3. In the third example, we consider the analysis of a dissolution testing gauge R&R experiment that is discussed in Gao (2007). Contrary to their analysis, we suggest taking patterned temporal variation into account.

In each case, we show the usefulness of Latin-square–related designs. The present work extends the work by De Mast and Trip (2005). First, it gives three practical case examples, showing for what kind of situations their approach can be used. Furthermore, we pay much more attention to the experimental design, explaining the rationale for using certain designs, and extend the classes of experimental designs used, increasing the flexibility of this approach. Finally, the statistical analysis of the experiments is explained in more detail than in De Mast and Trip (2005).

The remainder of this article is organized as follows. The next section introduces the three case examples, describes the experimental design, and provides the actual data for the first two examples. Subsequently, we discuss appropriate statistical models for the data and their analysis. In the final section, conclusions are drawn.

Experimental Design and Data for Three Cases

In this section, three case examples will be introduced. For each example, we will discuss the experimental set-up and data.

Measuring the Core Temperature of a Food Product: Example for PTV without I

A food product is baked until its core reaches a temperature of about 80°C. The core temperature is measured by inserting a digital thermometer into the product. Because heat is not distributed perfectly homogeneously over the product and the operators insert the thermometer by feel (aiming for the core), it is likely that random measurement error is substantial.

To estimate random measurement error, we could do a standard gauge R&R study. Each food specimen could be measured multiple times, but because the product cools down quite rapidly (about 1.0°C per minute), these repeated measurements would confound measurement spread with variation in the product's true core temperature (condition I—temporal stability—is violated).

The Constructed Design: Latin-Square–Type Designs

Let us examine how to best analyze the measurement error in this case. (Later we will see that the designs used in the other cases are very similar.) Assume that we measure at n time instances. Assume, furthermore, that $n = q \times r$, where q denotes the number of operators and r the number of times each operator measures a certain object. (This assumption is not needed, as we will point out later.) Ideally, *each* operator would measure *each* object at each time instant. However, at any given time instant, a particular food specimen can only be measured by one operator. Nonetheless, we can create an experimental design in which each operator measures at each time instant, though not always the same object. This can easily be accomplished by a Latin-square design. To set the stage, consider the case for which we have three objects ($n = 3$) and each operator measures once ($r = 1$). If we denote the operators by A, B, and C, then an example of a design satisfying our requirements is the Latin-square design given by

	t_1	t_2	t_3
object 1	A	B	C
object 2	B	C	A
object 3	C	A	B

If we want to have each object measured twice by each operator ($r = 2$), we can add measurements at 3 additional time instances t_4, t_5, and t_6. Using another Latin-square design for the measurements at these times, we obtain the following design:

	t_1	t_2	t_3	t_4	t_5	t_6
object 1	A	B	C	A	C	B
object 2	B	C	A	C	B	A
object 3	C	A	B	B	A	C

Any permutation of the columns of this design will yield a design that suits our purposes. In a permuted design still, each operator measures each object twice ($r = 2$), and still each operator measures at all time instants. This feature is what makes Latin-square designs attractive. Latin-square designs can generally

TABLE 2. Latin-Square–Type Design
(Entries Indicate Operator)

	Time					
Specimen	0	60	120	180	240	300
I	A	C	B	A	B	C
II	A	A	B	C	B	C
III	C	C	A	B	A	B
IV	C	B	A	A	C	B
V	B	A	C	B	C	A
VI	B	B	C	C	A	A

be obtained from standard Latin squares via permutation of rows, columns, and labels. (We refer to chapter 31 of Neter et al. (1985) for this.)

Typically, one wants to measure more than three objects. To obtain a design for the more general case, we can adjust the above procedure, starting with Latin squares of a different (higher) dimension. Usually, we can set up the experiment in such a way that the condition $n = q \times r$ holds, but if this condition does not hold, an appropriate design can be constructed by deleting rows or columns in a larger design (see also Cochran and Cox (1957) and chapter 31 in Neter et al. (1985) about these so called "Youden" designs).

Experimental Set-Up and Data

In the actual experiment, it was decided to select six specimens of the food product ($n = 6$). Each specimen was to be measured twice ($r = 2$) by each of three operators ($q = 3$), according to the design in Table 2. The two times three measurements were to be done with 60 seconds between successive measurements. We constructed the Latin-square–type experimental design that is shown in Table 2. The results of the experiment are shown in Table 3. The entries are the measured core temperatures (°C). Figure 1 shows a Trellis graph of the core temperature over time per food product specimen.

TABLE 3. Measurements for Food-Product Experiment

	Time					
Specimen	0	60	120	180	240	300
I	87.0	83.9	82.2	82.0	77.0	76.4
II	78.1	76.8	75.0	73.8	69.2	70.4
III	77.2	77.7	77.0	76.4	76.3	73.1
IV	74.3	72.8	73.9	70.9	70.7	69.6
V	81.6	81.9	79.9	79.3	78.3	78.1
VI	77.6	76.1	74.8	73.9	74.2	74.2

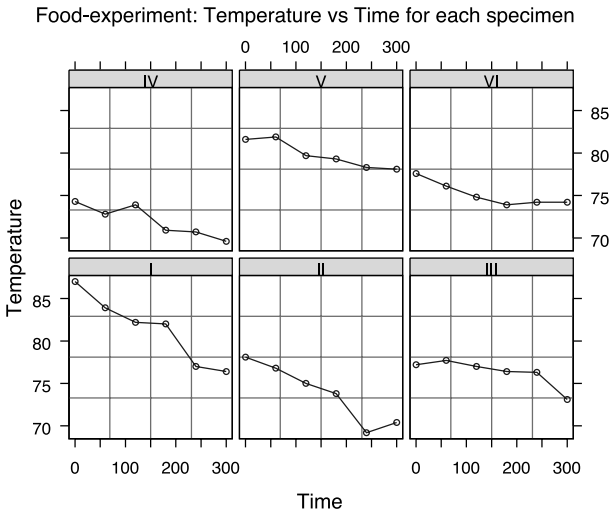

FIGURE 1. Data, Food Experiment.

Before proceeding to the analysis of this experiment, we first introduce two other examples. The statistical analysis of the three examples is given in the upcoming section.

Measuring Shrinkage of Carpet Tiles: Example for POV Without II

A company produces carpet tiles. Out of a stretch of carpet, carpet tiles are blanked. After production, the amount of shrinkage (or expansion) of the carpet tiles is measured. If the carpet tiles shrink or expand during their lifetimes, customers complain.

The challenge in this measurement procedure is to mimic the circumstances to which the carpet tiles are exposed during their lifetime. In order to stress test the carpet tile's performance with respect to shrinkage, it is exposed to extreme temperatures and moisture conditions during the measurement. The gauge R&R of such stress tests refer to the consistency of the test results that would be obtained if the test were performed multiple times on the same tile. However, these tests are irreversible and therefore the measurement procedure is nonrepeatable (vondition II—robustness against measurement—is violated here).

Experimental Set-Up and Data

In order to determine how to construct a good design, we need to know a little bit more about the process of producing carpet tiles. As we explained, carpet tiles are blanked out of a stretch of carpet. From the left to right side of the the carpet stretch, six tiles are blanked. It is assumed, however, that the amount of shrinkage varies from the left to right side of the carpet stretch. So the shrinkage follows a pattern that can be modeled, similar to the pattern over time in the first case example (in which the temperature of a food product was measured). Therefore, we choose a design that is similar to the one used in the food experiment (see Table 2). Note that, in this case, "time" should be interpreted as the position at which the carpet tile is blanked out of the stretch. Position 1 indicates the position at the utmost left, position 6 the position at the utmost right of the stretch of carpet.

The results of the experiment are shown in Table 4. Entries are measured shrinkage percentages. Figure 2 shows the shrinkage per position (from left to right) per carpet tile.

Measuring Dissolution of Tablets: Example for PTV Without I and II

The third case is based on an example given in a paper of Gao et al. (2007). They examine the process of dissolution testing of tablets. The rate at which tablets dissolve is an important aspect in pharmaceutical applications. The dissolution rate of tablets is measured by an apparatus that exposes the tablets to the flow of some liquid. Over time, measurements are done. The drug released into the medium from the tablet matrix is measured. The measurement is, of course, nonrepeatable because the tablet is partially gone after the measurement. As we will see,

TABLE 4. Measurements for Carpet-Tile Experiment

	Position					
Carpet tile	1	2	3	4	5	6
I	0.69	0.42	0.28	0.38	0.15	0.4
II	0.81	0.55	0.23	0.27	0.23	0.43
III	0.67	0.45	0.42	0.21	0.36	0.38
IV	0.53	0.28	0.32	0.22	0.12	0.23
V	0.51	0.43	0.20	0.19	0.24	0.41
VI	0.47	0.29	0.22	0.23	0.27	0.45

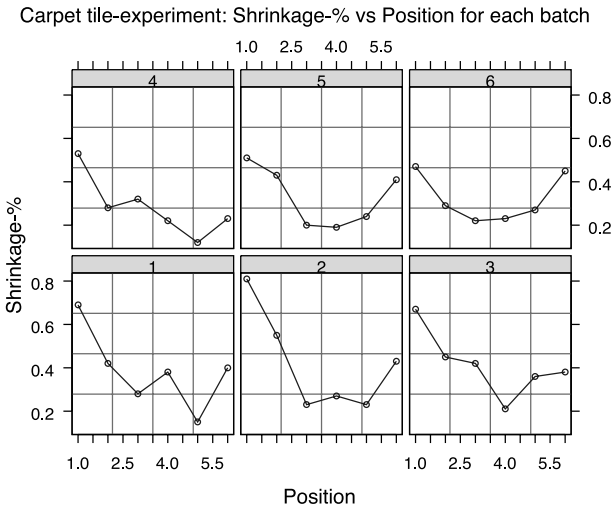

FIGURE 2. Data, Carpet-Tile Experiment.

the dissolution percentage of a single tablet follows a certain pattern in time.

Experimental Set-Up and Data

The experiment of Gao (2007) is as follows. Two operators measure the dissolution percentage of tablets on each of two apparatuses (labeled A and B). Each apparatus contains six separate vessels. This experiment is then a three-factor crossed then nested model design. The whole experiment is replicated 6 times, resulting in 6 runs. A single measurement on one tablet consists of measurements at 7 time instants ($t = 7.5, 15, 30, 45, 60, 75, 90$ minutes). Figure 3 shows these measurements for two typical tablets.

Gao (2007) based the gauge R&R analysis on the

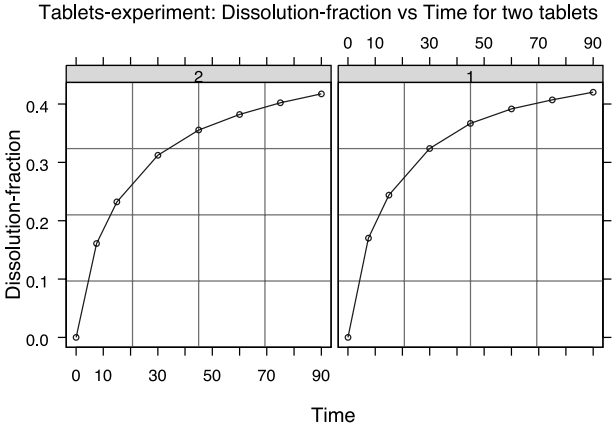

FIGURE 3. Data, Tablets Experiment (Dissolution Profile for Two Tablets).

measurements conducted at time $t = 30$ and fit a model similar to the one specified in Equation (1). Our approach in this case is twofold. First, we will model the dissolution profile in time, thereby using all data available from the experiment. Gauge R&R results are obtained from fitting a nonlinear mixed-effects model. This results in estimates of measurement error at all measurement times of the experiment. Second, we will suggest another experimental design (with fewer runs) that could have been used, resulting in reduced confounding of tablet and measurement error.

Analyzing the Three Examples

General Remarks: Estimation Method

The standard gauge-R&R model, as given in Equation (1), is an example of a linear mixed model (mixed referring to the presence of both random and fixed effects). In analyzing the first two examples, we will use linear mixed models as well. For the third example, we will fit the data by a nonlinear mixed model. In this section, we make some comments on the way these models can be estimated and our preferred method.

The traditional way to estimate variance components is the ANOVA method. Unfortunately, ANOVA has some well-established drawbacks, especially in the case of unbalanced data (which is important for our examples). The most important are (see Searle et al. (1992), pp. 35–39)

1. The possibility of negative estimates for variance components, which are not realistic from a practical viewpoint.
2. The lack of uniqueness of the choice of sums of squares. For unbalanced data, one can use, for instance, Henderson's methods I, II, and III, which are all using different sets of sums of squares. On top of this, criteria for deciding which choice for sums of squares is optimal are lacking. Therefore, the choice for a particular set of sums of squares is arbitrary.

Maximum likelihood (ML) estimation is a viable alternative for estimating variance components. Negative estimates are impossible when using ML estimation and the problem of the arbitrary nature of ANOVA is solved as well. An additional benefit of ML estimation is that the resulting estimators are asymptotically efficient. Detailed explanation of its application to variance–component estimation can be found in Searle et al. (1992). A slight drawback of ML estimation for mixed models (which includes the models under study) is that estimators for variance components depend on the fixed effects. This problem is circumvented by restricted maximum-likelihood (REML) estimation, where the fixed parameters are treated as nuisance parameters (see McCullagh and Nelder (1989), chapter 8, or Davidian and Giltiman (1995), chapters 3 and 4). The restricted likelihood is defined as the likelihood obtained by integrating out the fixed effects. By maximizing the restricted likelihood over the set of variance components, we obtain the REML estimators. Once the REML estimates are computed, they replace the variance components in the (generalized) least squares equations. From these, we can then obtain estimates for the fixed effects.

REML estimators can be justified from a Bayesian point of view as well (Searle et al. (1992), section 9.2b). In the Bayesian setup, besides the random effects, also the fixed effects are considered random. If we use a noninformative improper (flat) prior distribution on the elements of fixed effects, then it is not hard to see that the resulting posterior density is proportional to the restricted likelihood (the proportionality constant not depending on the random effects). As a consequence, the REML estimator equals the posterior mode if we use a noninformative prior for the fixed effects.

In the following, we will use the "nlme" library of `S-plus` *to perform the statistical analysis of the examples. We will report REML estimates. For a detailed explanation of the nlme package, we refer to Pinheiro and Bates (2000).*

Analysis of Examples

Measuring the Core Temperature of a Food Product

First we fit the data to the standard gauge R&R model of Equation (1). Although it is obvious from Figure 1 that the measured temperatures for each specimen are time dependent, it is instructive to compare this "naive" approach (neglecting time as covariate) to a more sophisticated one, which we will outline below. Hence, denote by y_{ijk} the temperature for specimen i, measured by operator j for the kth time. Fitting a crossed-random effect models yields the results summarized in Table 5 (Here and in the following, we report 95% confidence intervals. In the output these are denoted by [lower, upper]; est. denotes the REML estimate. Furthermore, fixed effects are aligned to the left, random effects are indented.)

TABLE 5. Results for Fitting a Crossed Random-Effects Model

	Lower	Est.	Upper
Intercept	73.36	76.43	79.47
Specimen	1.73	3.42	6.74
Operator	0.05	0.61	7.23
Specimen: operator	0.00	0.43	4521.91
Residual	1.81	2.49	3.44

The estimates for random effects are standard deviations. The width of the confidence interval for the interaction term is huge. Blind application of the standard model is therefore not sensible, as opposed to what is often done in practice. An ANOVA test shows that the interaction term can be dropped from the model. Refitting gives the results of Table 6. Therefore, the "naive" approach shows that reproducibility is a minor issue, but repeatability is a serious issue for this measurement. In fact, the estimated measurement spread equals 2.60. The proportion of measurement variance due to repeatability equals 94%.

Next, we will start with a very easy model and refine it in steps. Looking at Figure 1, it seems that, for each object, the temperature decreases linearly in time. Let $t_1, \ldots, t_6$ denote the times of measurement in the experiment. If y_{ik} denotes the temperature of specimen i at time t_k, we fit the model

$$y_{ik} = \mu + \gamma(t_k - 150) + \varepsilon_{ik}. \qquad (2)$$

Here $\{\varepsilon_{ik}\}$ denotes a sequence of independent identically distributed normal random variables. The dot plots of the residuals per specimen in Figure 4 reveal that this model is too simplistic. Because the specimens are drawn from a population of specimens, it is natural to model their effects as random effects. This is confirmed by Figure 5, which shows

TABLE 6. Results for Fitting a Random-Effects Model Without Interaction

	Lower	Est.	Upper
Intercept	73.38	76.43	79.47
Specimen	1.74	3.42	6.73
Operator	0.06	0.62	6.55
Residual	1.93	2.52	3.27

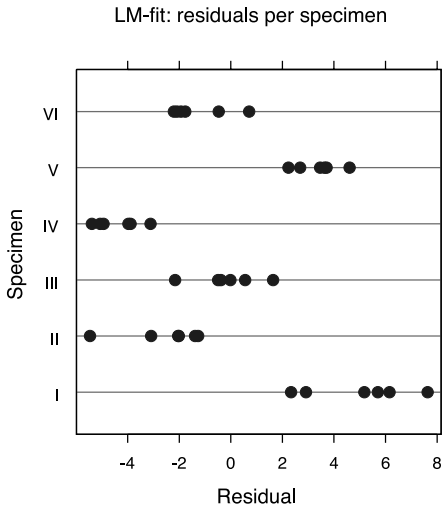

FIGURE 4. Dot Plots of Residuals for Each Specimen for Naive Model.

the fitted parameters plus confidence intervals in case we apply a linear fit for each specimen separately. The confidence intervals for the intercept parameter β_0 show little overlap. Additional drawbacks of the fixed-effects model are, first, that it does not provide an estimate of the between-specimen variability (which we are interested in) and, second, that the number of parameters in the model grows linearly with the number of specimen. Therefore, we propose the following mixed analysis of covariance model:

$$y_{ik} = \mu + a_i + \gamma(t_k - 150) + \varepsilon_{ik}. \qquad (3)$$

Here μ and γ are fixed effects and a_i is a random specimen effect. To see if an operator effect should be taken into account as well, we plot the residuals obtained by fitting model (3) in Figure 6. We made separate box plots for each operator. This figure points out a distinction between operator A and

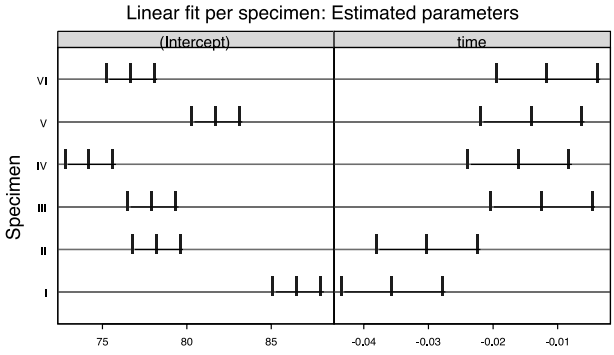

FIGURE 5. Confidence Intervals for Estimates in Simple Linear Fit for Each Specimen.

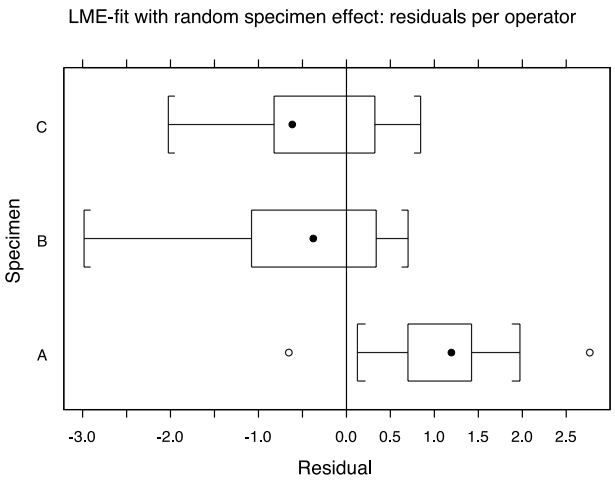

FIGURE 6. Residuals per Operator for Model (3).

operators B and C. We will assume that the three operators are drawn from a population and therefore model its effect as a random effect. This leads to the model

$$y_{ik\ell} = \mu + \alpha_i + \beta_\ell + \gamma(t_k - 150) + \varepsilon_{ik\ell}, \qquad (4)$$

where ℓ indexes operators and $\{\varepsilon_{ik\ell}$ is a sequence of independent and identically distributed normal random variables. With this model, we obtain an estimate for the between-operator spread as well (adding an additional random interaction term gives similar problems as in the naive approach). The results for model (4) are summarized in Table 7. Further residual plots did not reveal interesting patterns. We conclude that the measurement spread σ_m equals 1.42 and that 60% of the measurement variance is due to repeatability. This analysis clearly shows that the estimated measurement spread is in fact much smaller than the naive analysis pointed out.

Measuring Shrinkage of Carpets

The analysis for this example is similar to that of the food product data. We fitted both the standard gauge R&R model (without interaction term) and the following model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijk}. \qquad (5)$$

Here α, β, and ε are independent batch, operator, and repeatability random effects, respectively; μ is the overall mean and $\gamma_1, \ldots, \gamma_6$ are fixed position effects. For identifiability, we will set $\gamma_1 = 0$, which corresponds to treatment contrasts. The results for the naive analysis are in Table 8. Therefore, $\sigma_m \approx 0.16$. The proportion of measurement variance due to repeatability equals 83%.

The results for fitting model (5) are in Table 9. From this table, it follows that

$$\sigma_m \approx \sqrt{0.0766^2 + 0.0519^2} \approx 0.093.$$

The proportion of measurement variance due to repeatability equals 46%. Again, we see that the overestimation of the measurement spread is relatively large.

TABLE 7. Results for Fitting Model (4)

	Lower	Est.	Upper
Intercept	73.38	76.43	79.58
I(time − 150)	−0.023	−0.030	−0.016
Specimen	1.89	3.55	6.66
Operator	0.30	0.90	2.70
Residual	0.84	1.10	1.44

TABLE 8. Results for Fitting a Standard Gauge R&R Model Without Interaction Term

	Lower	Est.	Upper
Intercept	0.28	0.40	0.52
Batch	0.0011	0.0150	0.2134
Operator	0.0163	0.0656	0.2640
Residual	0.1148	0.1467	0.1875

Measuring Dissolution of Tablets

We now discuss the analysis of the gauge R&R experiment of Gao (2007). We are grateful to the au-

TABLE 9. Results for Fitting Model (5)

	Lower	Est.	Upper
Intercept	0.58	0.68	0.79
Position2	−0.27	−0.21	−0.15
Position3	−0.40	−0.34	−0.27
Position4	−0.42	−0.36	−0.30
Position5	−0.45	−0.39	−0.33
Position6	−0.29	−0.23	−0.17
Batch	0.0046	0.0202	0.0880
Operator	0.0277	0.0766	0.2120
Residual	0.0398	0.0519	0.0678

thors of Gao (2007) for sharing their data with us. In dissolution testing, variability arises primarily from three factors: apparatus, operator, and tablet formulation and/or manufacturing process (Gao (2007), p. 1796). Therefore, the following model could be fit for the 30-minute data:

$$y_{ijk} = \mu + b_j + c_k + (bc)_{jk} + \varepsilon_{ijk}. \quad (6)$$

Here y_{ijk} denotes the dissolution percentage after 30 minutes of tablet i, measured by operator j at apparatus k, b_j denotes a random operator effect, c_k a random apparatus effect, and $(bc)_{jk}$ a random operator $\times$ apparatus interaction effect. However, because there are only two operators (apparatuses) involved in this experiment, we doubt whether it is sensible to treat them as randomly chosen from a possibly larger population of operators (apparatuses). Instead, we suggest the following model to include their effects:

$$y_{ijk} = \mu + \varepsilon_{ijk}, \quad \text{with } \varepsilon_{ijk} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2_{jk}), \quad j = A, B,\ k = 1, 2. \quad (7)$$

Here j and k are index operator and apparatus, respectively. Model (7) is a short way to write down the model for which the measurements by operator A on apparatus 1 are assumed to follow a $N(\mu, \sigma^2_{A1})$ distribution, measurements by operator A on apparatus 2 have a $N(\mu, \sigma^2_{A2})$ distribution, measurements by operator B on apparatus 1 have a $N(\mu, \sigma^2_{B1})$ distribution, and measurements by operator B on apparatus 2 have a $N(\mu, \sigma^2_{B2})$ distribution. As pointed out on page 209 in Pinheiro and Bates (2000), `S-plus` uses a different parametrization, in which it estimates σ^2_{A1} by $\hat{\sigma}^2_{A1}$ and expresses the estimates for the other variance parameters as multiples of $\hat{\sigma}^2_{A1}$. Tablet 63 contains an exceptionally low value at time $t = 15$ and is excluded from this analysis. The results for fitting model (7) are summarized in Table 10.

TABLE 10. Each Line Shows the Results from Fitting Model (7) to the Dissolution Data at a Certain Fixed Time

Time	$10^3\sigma_{A1}$	$10^3\sigma_{A2}$	$10^3\sigma_{B1}$	$10^3\sigma_{B2}$
7.5	13.90	12.92	17.60	5.40
15	13.39	12.17	16.45	3.56
30	9.68	10.97	12.17	3.28
45	5.65	8.76	9.10	2.52
60	3.33	8.93	7.30	2.86
75	3.00	9.15	7.57	3.61
90	3.21	10.01	7.97	4.41

We will now deduce one statistical model for the data at all time instances. We start with a very simple model that specifies the dissolution profile in time. As we go along, we will include factors (similarly as in the analysis of the food-product experiment). Let $y_{i\ell}$ denote the measurement at time t_ℓ of the ith tablet. We propose to model the shape of the profile by the Weibull model,

$$y_{i\ell} = f(t_\ell; \tilde{a}, \tilde{r}, h) + \varepsilon_{i\ell},$$
$$f(t; \tilde{a}, \tilde{r}, h) = \tilde{a}\left(1 - \exp\left[-\tilde{r}\frac{t^{1-h}}{1-h}\right]\right), \quad t \geq 0. \quad (8)$$

Here $\{\varepsilon_{i\ell}\}$ is a sequence of independent and identically distributed $N(0, \sigma^2_\varepsilon)$ random variables. The parameters $\tilde{a}$ and $\tilde{r}$ are assumed to be positive and $h < 1$. Note that this modeling step is analogous to the model in Equation (2) for the food-product example. The only difference is that here we do not propose a linear function of time, but a nonlinear function. Depending on the values of $\tilde{a}$, $\tilde{r}$, and h, we will see that the graph of the mapping $t \mapsto f(t; \tilde{a}, \tilde{r}, h)$ looks like the curves observed in Figure 3.

Because $h < 1$, $\lim_{t\to\infty} f(t; \tilde{a}, \tilde{r}, h) = \tilde{a}$, the parameter $\tilde{a}$ models the final fraction of dissolved tablet. An interpretation for the other two parameters ($\tilde{r}$ and h) is given by Lánsky and Weiss (2003). We now briefly explain their approach. If we define $C(t) = f(t; \tilde{A}, \tilde{r}, h)/\tilde{a}$, then $C(t)$ is the amount of dissolved tablet at time t, divided by the amount of tablet that is going to be dissolved ultimately, as time increases. Lánsky and Weiss (2003) interpret $C(t)$ as the cumulative distribution function of a random variable T, which is defined as the dissolution time of a randomly selected molecule (or the time until a molecule of the tablet enters solution). Thus, $C(t) = P(T \leq t)$. With this probabilistic interpretation, we can give a meaning to the parameters $\tilde{r}$ and h. Suppose T has a probability density f. Define the fractional dissolution rate by

$$k(t) := \frac{f(t)}{1 - F(t)}, \quad t \geq 0. \quad (9)$$

From this definition, it is seen that the fractional dissolution rate is analogous to the hazard rate in survival analysis. Its probabilistic interpretation is as follows (see Lánsky and Weiss (2003), p. 1633): $k(t)dt$ equals the probability that a molecule in solid state will be dissolved in the interval $[t, t + dt)$ under the condition that this has not happened up to time t. For the Weibull model, the fractional dissolution

rate equals $k(t) = \tilde{r}t^{-h}$, which is determined by $\tilde{r}$ and h. Our motivation for using the Weibull model comes from its flexibility and ability to allow for both an increasing and a decreasing fractional dissolution rate (depending on the sign of h).

Because $\tilde{a}$ and $\tilde{r}$ should be positive, we use an alternative parameterization when fitting the model. We simply set $\tilde{a} = e^a$ and $\tilde{r} = e^r$, so that a and r take values in $\mathbb{R}$ (we could have reparameterized h as well, but that turned out to make no difference in the following analysis).

We fitted the model by nonlinear least squares (NLS) without the data at $t = 0$ (these are fitted without error) and found estimates for a, r,, and h equal to $\hat{a} = -0.78(0.007)$, $\hat{r} = -2.60(0.01)$, $\hat{h} = 0.28(0.01)$, respectively (standard errors are in parentheses). Note that $\tilde{a}$ is then estimated by $\exp(\hat{a}) \approx 0.46$, which seems to agree with a visual inspection of Figure 3. The residual standard deviation equals $\hat{\sigma}_\varepsilon = 0.0147$. Next, we examined the residuals. Figure 7 shows box plots of standardized residuals for each tablet. This figure clearly demonstrates a random tablet effect (as could be expected). This is further confirmed if we fit the nonlinear model for each tablet separately and plot confidence intervals for (a, r, h) for each tablet in the plane. See Figure 8. Although the confidence intervals are based on asymptotic distribution theory and should be trusted with caution, this figure points out that we may expect an improved fit if we associate with parameter r a random effect (confidence intervals for the parameter r among different tablets are disjoint for many couples of tablets).

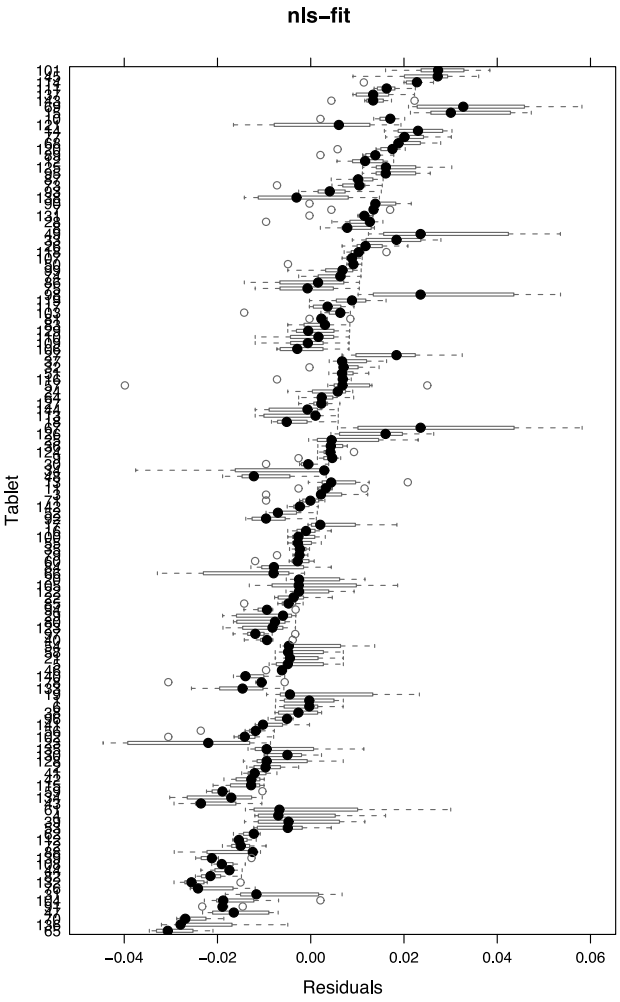

FIGURE 7. Box Plots of Residuals per Tablet after Fitting Model (8).

We henceforth propose the following model:

$$y_{i\ell} = f(t_\ell; e^a, e^{r_i}, h) + \varepsilon_{i\ell}, \quad r_i = r + \eta_i, \qquad (10)$$

where $\{\eta_i\}$ is a sequence of independent $N(0, \sigma_\eta^2)$ distributed random variables that is independent of the sequence $\{\varepsilon_{i\ell}\}$. Adding a random effect to parameter r implies that we model the fractional dissolution rate of a randomly chosen tablet by $k(t) = \tilde{r}e^\eta t^{-h}$, where η is a normal random variable with variance σ_η^2.

Fitting the model gives the following results for the fixed effets: $\hat{a} = -0.77(0.003)$, $\hat{r} = -2.61(0.01)$, $\hat{h} = 0.28(0.004)$ (standard errors are in parentheses). Estimates for the variance components are given by $\hat{\sigma}_\varepsilon = 0.0071$ ($[0.0068, 0.0074]$) and $\hat{\sigma}_\eta = 0.096$ ($[0.085, 0.108]$) (95% confidence intervals are in parentheses). Figure 9 compares the fit of the fixed-effect model (8) with the prediction from the random-effect model (10) for 15 randomly chosen tablets. The improved fit is very apparent. Residual plots for model (10) are shown in Figure 10. These plots help to check whether the model assumptions are satisfied. The rightmost picture shows a normal QQ plot for the estimated best linear unbiased predictors of the random effects r_i. The left and middle figures show standardized within-group residuals. Within-group residuals are defined as the differences between $y_{i\ell}$ and $\widehat{E(y_{i\ell} \mid r_i)}$, the latter denoting the estimated best linear unbiased prediction of $y_{i\ell}$ (thus, we plug in both the estimated random effects and the estimated fixed effects). More details on the definition of these estimated residuals and estimated random effects can be found in chapter 9 of Searle et

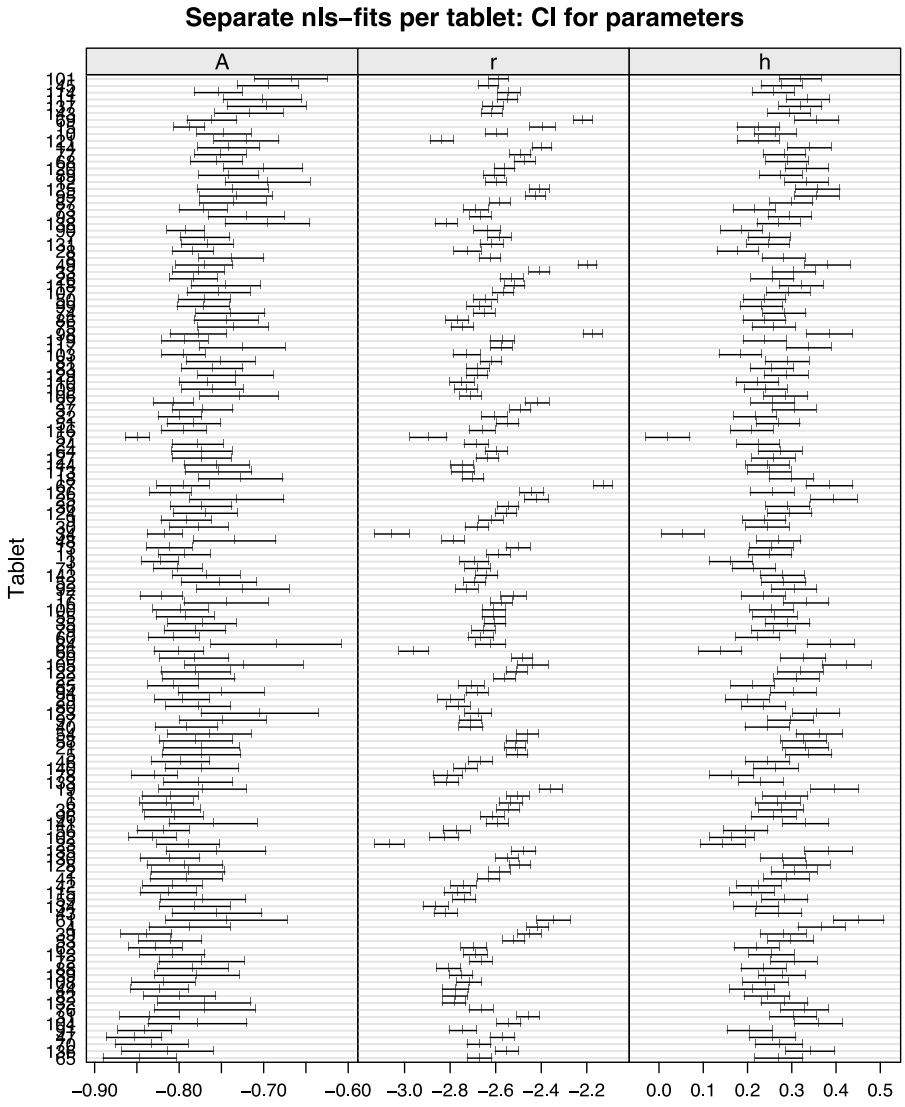

FIGURE 8. NLS Estimates for (A, r, h) According to Model (8) for Each Tablet Separately.

al. (1992). Except for a few outlying points, the residual plots in Figure 10 do not contradict adequacy of the fitted model. Because the tablets are drawn from a population of tablets, it is natural to assume that there are small differences in characteristics between tablets. With the present setup, we model these differences by differences in the fractional dissolution rate.

Examination of the residuals against covariate time indicates heteroscedasticity. Figure 11 shows box plots of raw residuals for each measurement time. A Levene test on equality of variance was rejected at the 0.001 level. We could include heteroscasticity in our model by replacing the constant residual standard deviation σ_ε by $\sigma_\varepsilon \times g(t)$, where g accounts for the differences in residual variation over time. However, to keep the exposition at an elementary level, we take a simpler route. We treat the obtained residuals at each time instance as filtered measurements, where tablet-to-tablet variation has been filtered out. Subsequently, we collect all residuals that correspond to measurements executed at time t_ℓ and fit model (7) to these residuals for each ℓ separately. The results are summarized in Table 11. From this table, we see, for example, that if operator B measures a tablet at apparatus 1 at time $t = 45$, say he/she measures $m\%$, then the unknown true value is approximately within the interval $m \pm 2 \times 0.294\%$ (with 95% confidence). A more naive approach (ig-

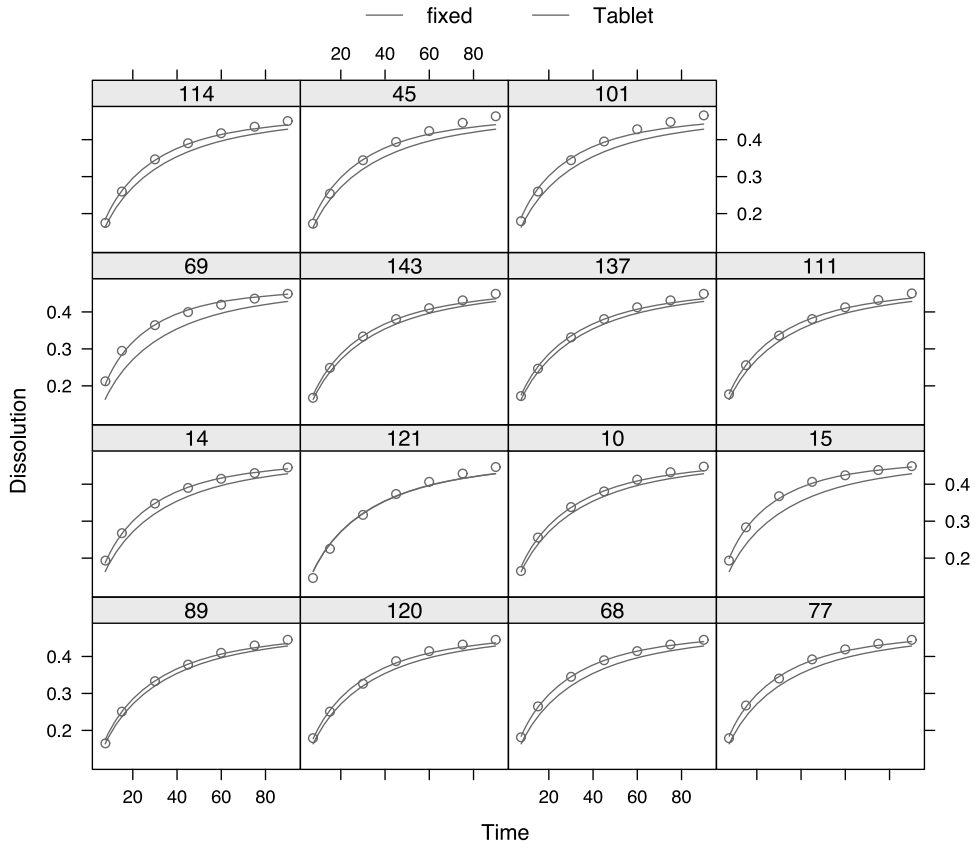

FIGURE 9. Comparison of NLS Fitted Values from Model (8) (Indexed by "Fixed") and NLME Predicted Values from Model (10) (Indexed by "Tablet") for 15 Randomly Chosen Tablets. The "tablet" curves nearly interpolate the data, indicating that Model (10) fits the data accurately.

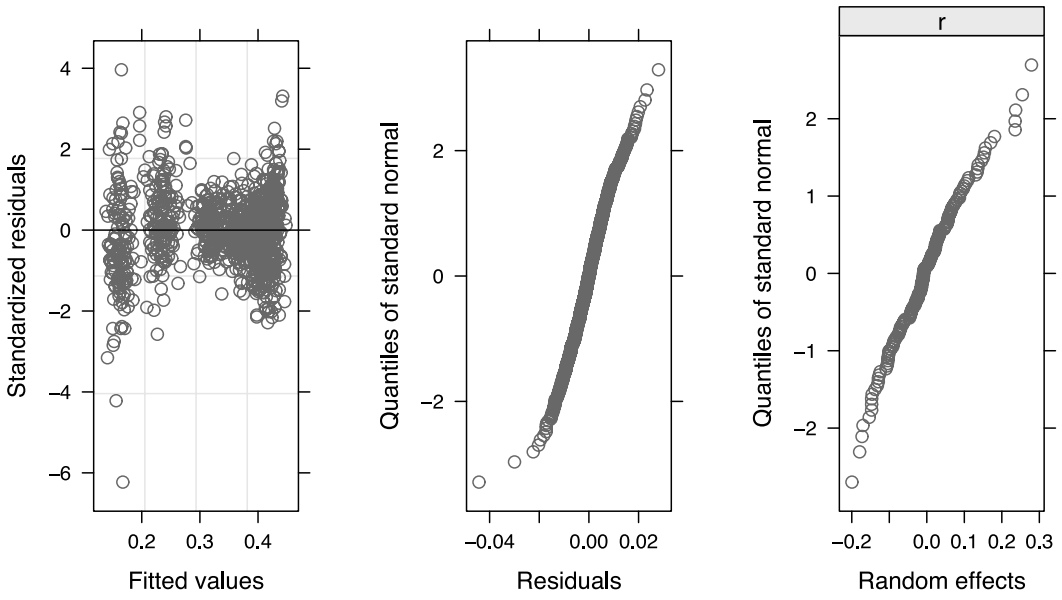

FIGURE 10. Residual Plots for Model (10).

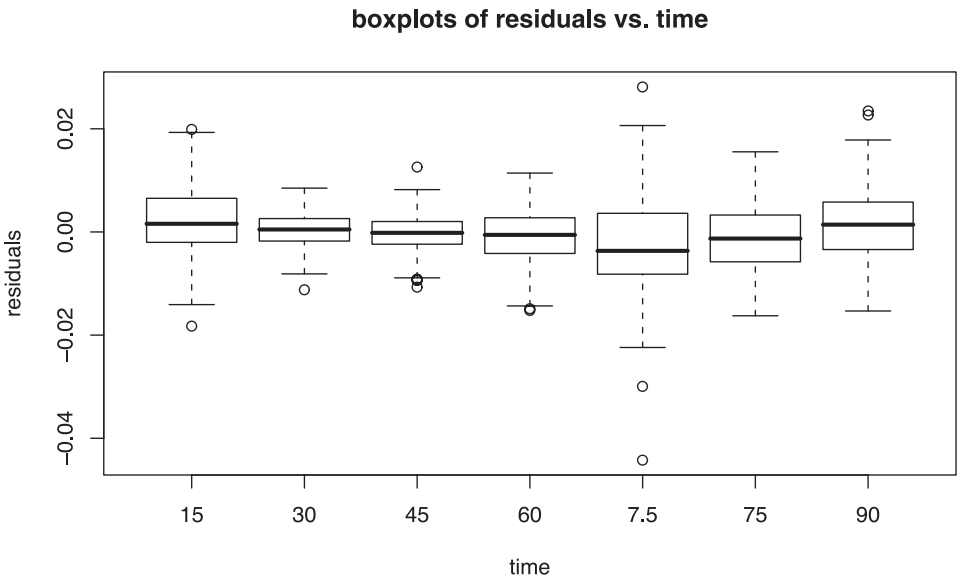

FIGURE 11. Box Plots of Raw Residuals Against Measurement Time for Model (10).

noring tablet-to-tablet variation) would give the interval $m \pm 2 \times 0.910\%$ (see Table 10).

By adjusting the design such that a single tablet is measured at various times by different operators, better randomization is achieved (reducing the confounding of tablet and measurement spread). This can be done by the methods discussed in the section on experimental design and henceforth is very similar to our food-product example.

Conclusion

In gauge R&R studies on nonrepeatable measurement system, one is often forced to replace replications with measurements of different objects, thus confounding measurement variation with between-objects within sample variation. This paper presents a methodology for reducing the resulting overestimation by exploiting patterns in both object-to-object variation and temporal variation of objects. The usefulness of the Latin-square–related design has been shown for a couple of real data examples.

The standard errors of the estimates for variance components are quite large in the first two examples, especially the errors in the estimate for reproducibility. This is due to the fact that, in the current design, only three operators are used. Three is the typical number of operators in gauge R&R studies reported in literature and expected in practice. Also in the standard gauge R&R study, this small number results in very large standard errors for the estimated variance components (Burdick and Larsen (1997)). To get better estimates one has to replicate the design, but it only helps to use the replication scheme in which the number of operators is increased.

TABLE 11. Each Line Shows the Results from Fitting Model (7) to the Residuals of Model (10) at a Certain Fixed Time

Time	$10^3\sigma_{A1}$	$10^3\sigma_{A2}$	$10^3\sigma_{B1}$	$10^3\sigma_{B2}$
7.5	8.45	7.02	10.60	3.73
15	5.23	4.83	6.40	1.64
30	0.92	3.21	1.87	1.11
45	2.70	2.01	2.94	0.87
60	3.39	4.02	3.81	1.20
75	4.49	5.66	5.11	1.79
90	4.19	6.53	5.40	2.00

Acknowledgment

We are grateful to dr. Zongming Gao for sharing the dissolution data with us and various valuable comments on earlier versions of this paper. We thank Prof. Geurt Jongbloed from Delft University of Technology for helpful discussions on the statistical analysis.

References

Burdick, R. K. and Larsen, G. A. (1997). "Confidence Intervals on Measures of Variability in R&R Studies". *Journal of Quality Technology* 39(3), pp. 261–273.

Burdick, R. K.; Borror, C. M.; and Montgomery, D. C. (2003). "A Review of Methods for Measurement Systems Capability Analysis". *Journal of Quality Technology* 35(4), pp. 342–354.

Cochran W. G. and Cox, G. M. (1957). *Experimental Designs*, 2nd edition. New York, NY: Wiley.

Davidian, M. and Giltiman, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*. Chapman & Hall.

De Mast, J. and Trip, A. (2005). "Gauge R&R Studies for Destructive Measurements". *Journal of Quality Technology* 37(1), pp. 40–49.

Gao, Z.; Moore, T. W.; Smith, A. P.; Doub, W. H.; and Westenberger, B. J. (2007). "Studies of Variability in Dissolution Testing with USP Apparatus 2". *Journal of Pharmaceutical Sciences* 96(7), pp. 1794–1801

Lánsky, P. and Weiss, M. (2003). "Classification of Dissolution Profiles in Terms of Fractional Dissolution Rate and a Novel Measure of Heterogeneity". *Journal of Pharmaceutical Sciences* 92(8), pp. 1632–1647.

Montgomery, D. C. (2005). *Design and Analysis of Experiments*, 6th edition. New York, NY: John Wiley and Sons, Inc.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edition. Chapman and Hall, London.

McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. New York, NY: John Wiley and Sons, Inc.

Neter, J.; Wasserman, W.; and Kutner, M. H. (1985). *Applied Linear Statistical Models*, 2nd edition. Richard D. Irwin, Inc.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer.

Searle, S. R.; Casella, G.; and McCulloch, C. E. (1992). *Variance Components*. New York, NY: John Wiley and Sons, Inc.