Estimating precision, repeatability, and reproducibility from Gaussian and non- Gaussian data: a mixed models approach

Peer-reviewed author version

ASSAM NKOUIBERT, Pryseley; Mintiens, Koen; Knapen, Katia; Van de Stede, Yves & MOLENBERGHS, Geert (2010) Estimating precision, repeatability, and reproducibility from Gaussian and non- Gaussian data: a mixed models approach. In: JOURNAL OF APPLIED STATISTICS, 37 (10). p. 1729-1747.

DOI: 10.1080/02664760903150706
Handle: http://hdl.handle.net/1942/11253

Estimating Precision, Repeatability, and Reproducibility From Gaussian and Non-Gaussian Data: A Mixed Models Approach

Assam Pryseley[1] **Koen Mintiens**[2] **Katia Knapen**[2] **Yves Van der Stede**[2] **Geert Molenberghs**[1]

[1] Interuniversity Institute for Biostatistics and statistical Bioinformatics. Universiteit Hasselt, Gebouw D, Agoralaan, 3590 Diepenbeek, Belgium and Katholieke Universiteit Leuven, Belgium

[2] Veterinary and Agrochemical Research Centre, VAR-CODA-CERVA, Brussels, Belgium

Abstract

Quality control relies heavily on the use of formal assessment metrics. In this paper, for the context of veterinary epidemiology, we review the main proposals, precision, repeatability, reproducibility, and intermediate precision, in agreement with ISO (**International Organization for Standardization**) practice, generalize these by placing them within the linear mixed model framework, which we then extend to the generalized linear mixed model setting, so that both Gaussian as well as non-Gaussian data can be employed. Similarities and differences are discussed between the classical ANOVA (analysis of variance) approach and the proposed mixed model settings, on the one hand, and between the Gaussian and non-Gaussian cases, on the other hand. The new proposals are applied to five studies in three diseases: Aujeszky's disease, enzootic bovine leucosis (EBL), and bovine brucellosis. The mixed-models proposals are also discussed in the light of their computational requirements.

Keywords: Accuracy; Analysis of variance; Aujeszky's disease; Bias; Bovine brucellosis; Enzootic bovine leucosis; Generalized linear mixed models; Linear mixed models; Quality control.

1 Introduction and Definitions

Whenever decisions are based on analytical results, it is important to assess the quality of the results, that is, the extent to which they can be relied on for the purpose at hand. In some sectors of analytical chemistry, it is now a formal requirement for laboratories to introduce quality assurance measures to ensure that they are capable of and are providing, data of the required quality. Let us first introduce a collection of commonly used concepts, based on the International Organization for Standardization (ISO), to be used in the remainder of the paper.

The paper is aimed predominantly at **researchers** seeking application of the methodology outlined here.

Therefore, methodological development has been kept as brief as possible. However, the interested reader is guided toward more detail through appropriate references to the literature.

1.1 Definitions

Accuracy is the closeness of agreement between a test result and an agreed upon reference for comparison, i.e., an accepted reference value. It is expressed by two components: bias and precision.

Unbiassedness is the closeness of agreement between the average value obtained from a large series of test results and an well-defined accepted reference value. It is expressed in terms of bias, i.e., the difference between the expectation of the test result and an accepted reference value. However, the focus of this paper is on the estimation of precision, not on unbiassedness. It is also referred to in some manuals as *Trueness*.

Precision is the closeness of agreement between replicated and independent test results, obtained under stipulated conditions. Precision is usually expressed in terms of *imprecision*, and computed as either, standard deviation, variance, or coefficient of variation (CV) of the test results, with less precision reflecting in large standard deviations. Often, the minimum and maximum precision estimates are of interest. Quantitative measures of precision depend critically on the stipulated conditions. These conditions depend in turn on factors affecting the variability of the results from a measurement method. *Laboratory, operator, time elapsed between measurements, calibration of equipment,* and *batch of reagents* are some of the factors affecting the variability of the results, as outlined in the ISO 5725 manual (ISO 1994, Parts 1 and 3). By varying either none, some, or all of these (measurable) factors, it is possible to estimate the *repeatability, intermediate precision(s)*, and *reproducibility* standard errors (or CV).

Repeatability is the precision obtained, under the 'same' conditions, when independent test results are obtained with the same method, on identical test items, in the same laboratory, by the same operator, using the same equipment, and within short intervals of time; these are termed repeatability conditions. Repeatability leads to an estimate of the minimum value of precision.

Reproducibility is the precision obtained, under changing conditions, when independent test results

are obtained with the same method, on identical test items, but in different laboratories, with different operators, using different (or recalibrated) equipments; these are referred to as reproducibility conditions. The result is an estimate of the maximum value of precision.

Intermediate precision(s) is the precision(s) obtained, under changed conditions, by varying at least two of the variables affecting variability of the results. For example, independent test results obtained with the same method, on identical test items, in different laboratories, by different operators, using the same equipment within short intervals of time, enables estimation of the intermediate-operator precision. Changing k factors enables the estimation of so-called $k-1$ intermediate precisions, including repeatability and reproducibility. When intermediate measures of precision are given, one must carefully state which of the factors have been allowed to vary.

The *international vocabulary of basic and general terms in metrology* (VIM) manual distinguishes between repeatability and reproducibility by referring to the former for successive measurements made under the same conditions and to the latter for measurements made under different conditions of measurements (ISO 1993b). For repeatability conditions, the VIM and ISO 3534 (ISO 1993c) definitions are almost identical. VIM's definition for reproducibility conditions, however, is more general than the ISO 3534 definition, and includes within-laboratory measurements using different principles of measurement. This more general terminology is increasingly common. For this reason, it is recommended that the conditions of measurement would be always indicated in references to reproducibility.

Accuracy is a qualitative concept and precision should not be used for accuracy (Eurachen/Citac 2000). Note that, to improve accuracy optimally, both bias has to be reduced and precision has to be increased. Improving precision alone improves accuracy but not an optimal way as the precise results may be remote from the accepted reference value. Also, reducing bias implies that the results are "close" to the accepted reference value, but this does not guarantee precision. These concepts are depicted in Figure 1.

1.2 Motivation for, Objective of, and Organization of the Paper

Users of the results of chemical quantitative analysis, particularly in those areas concerned with international trade, are coming under increasing pressure to eliminate or at least minimize the replication

of effort frequently expended in obtaining them, for reasons of cost effectiveness. Confidence in data obtained outside the user's own organization is a prerequisite to meeting this objective. In some sectors of analytical chemistry, it is now a formal, frequently also legal, requirement for laboratories to introduce quality assurance measures so as to ensure that they are capable of, and are providing data for, the required quality (ISO 1993a). Such measures include the use of validated methods of analysis, the use of well-defined internal quality control procedures, participation in proficiency testing schemes, accreditation based on ISO 17025:2005, and establishing traceability of the results of the measurements.

In analytical chemistry, there has been great emphasis on the precision of results obtained using a specified method, rather than on their traceability to a pre-specified standard or SI unit (ISO 1993a). As a consequence of these requirements, chemists are, for their part, coming under increasing pressure to demonstrate the quality of their results. This is understood to include the degree to which a result would be expected to agree with other results, i.e., precision, in principle irrespective of the analytical methods used. Thus, it is essential that laboratories use standardized methods.

The aim of this paper is to briefly review methods for estimating measures of precision, repeatability, intermediate-precision, and reproducibility. In this part, the focus is on measurements obtained on a continuous, Gaussian scale, and the basis is the ISO 5725:1994 norm. We will frame these methods within a principled modeling framework: the linear mixed model (Laird and Ware 1982, Verbeke and Molenberghs 2000). We then extend these methods to the case of non-Gaussian measurements, a non-trivial task since the modeling framework used will now be the generalized linear mixed model (Breslow and Clayton 1993, Molenberghs and Verbeke 2005), which is intrinsically non-linear.

These methods may be applied to a wide range of materials, including liquids, powders and solid objects, manufactured or naturally occurring, provided due consideration is given to any heterogeneity of the material and a nested experimental design used for the experiment (ISO 1994).

The rest of the paper is organized as follows. A basic experimental design and five motivating case studies are introduced in Section 2. A brief review of the ISO method and its framing within and extension to the mixed model framework is given in Section 3. This section also places emphasis on the relationship between precision and reliability estimates, availability of software, and the use of precision

in practice. In Section 4, we present the results obtained from applying the methods to the case studies.

2 Experimental Design and Motivating Studies

2.1 Experimental Design

In our context, a 'value' of a so-called precision experiment is considered as one particular material or specimen of the so-called measurand (ISO 1993c) such as, for example, different concentrations of a chemical solution. It is customary to perform precision experiments at many values of the measurand to investigate whether precision is constant or varies for a given range of values of the measurand. Therefore, we will refer to an experiment with x distinct specimens of the measurand as an *x-value precision experiment*. For example, an experiment involving 5 'different' concentrations of a chemical solution is a 5-value precision experiment.

Many experimental designs have been proposed for sampling and testing situations, including nested factorials (Smith and Beverly 1981), split factorials (Ankenman *et al* 2001) and assembled designs (Ankenman *et al* 2003). Fully-nested experimental designs are a common choice for precision experiments. A p fully-nested experimental design has the advantage of enabling, in one inter-laboratory study, estimation of the repeatability, reproducibility, and $p-1$ intermediate precision standard deviations, where p refers to the levels of nesting. However, this type of experiment places considerable requirements on the laboratories; k^p-1 test results are required from each laboratory for a p-factor fully-nested design with k results under each repeatability condition.

In practice, it is quite common to use completely balanced experimental designs. However, it leads to costly experiments and some variance components end up being estimated with high precision while others are estimated with poor precision (Delgado and Iyer 1999). This motivated many authors, including Bainbridge (1965), Smith and Beverly (1981), and Naik and Khattree (1998), to investigate unbalanced nested designs, called *staggered nested designs*, to spread the infomation in the experiment more equally among the variance components. Delgado and Iyer (1999) and Ankenman *et al* (2003) amongst others, developed algorithms to search for optimal designs in particular experimental settings.

This article neither introduces a new design method nor develops an algorithm for obtaining an optimal design. All motivational datasets, which will be introduced in the next section, are based on fully-nested experimental design. **Thus, methods presented for the estimation of precision are based on fully-nested designs. Implementation based on available software, such as SAS, SPLUS, R, and SPSS, is available in full generality, for both fully and staggered nested experimental designs.** Figure 2 shows a three-factor fully-nested design for a given value of the experiment, which we shall use to illustrate our estimation methods.

The subscripts i, j, and k, affixed to the data y in Figure 2, represent the three-factor fully-nested experiment. For each value ℓ, there are I laboratories, each having J days of experiment or J operators, with K replications under repeatability conditions.

It is common practice to choose between 8 and 15 laboratories. When the between-laboratory standard deviation is larger than the repeatability standard deviation, as is often the case, little is to be gained by obtaining more than 2 test results per level within each stratum of the combination of factors affecting the precision (ISO 1994). Sample size calculations for precision experiments are discussed in ISO 5725-1 (ISO 1994).

2.2 Motivating Case Studies

In this section, we present 5 motivating case studies based on 3 diseases: Aujeszky's disease, enzootic bovine leucosis (EBL), and bovine brucellosis. The rationale for considering a collection of applications, rather than a single one or at least a small number, is thatsome of the concepts and methodology to be developed hinge on replication across a range of applications. All motivational datasets are based on fully-nested experimental designs, within the levels of the precision experiment. As aforementioned in Section 2.1, using x different specimens of the measurand leads to an *x-value precision experiment*. The response is the measurement on the measurand at the different values of the experiment. Factors encompass the laboratories (for inter-laboratory trials), day of the experiment, and number of replicates.

All datasets used were provided by the Unit of Quality Care of the Veterinary and Agrochemical Research Centre VAR-CODA-CERVA, located in Brussels, Belgium. The VAR-CODA-CERVA

is the National Reference Laboratory of Belgium for veterinary diagnostic assays used within the framework of official disease monitoring and surveillance programmes. These diagnostic assays should be evaluated and validated in a standardized and consistent way.

The World Organization for Animal Health (OIE) published guidelines for the validation and certification of diagnostic assays for infectious animal diseases, with the objective of harmonizing animal disease prevention, surveillance and control, therefore, allowing the use of a 'quality label (logo)' on associated kit materials, illustrating the recognition of the status of a test as valid for the defined fitness for purpose, according to OIE parameters. For this validation procedure, repeatability and reproducibility are amongst the key assay performance characteristics required to evaluate an assay. The motivating case studies are used to evaluate assays for the 3 diseases mentioned above.

Aujeszky's disease is caused by the pseudorabies virus (PRV, Suid Herpes virus-1). In the eighties is was a zoonotic disease. Nowadays, PRV is controlled and monitored in most European countries by mass-vaccination programmes (vaccination with marker vaccines) which has been implemented to eradicate PRV since 1993. It is associated primarily with pigs. A dataset, referred to as *AUJES* in what follows, was obtained from an 8-value precision experiment. A blocking ELISA (Enzyme-Linked ImmunoSorbent Assay, a biochemical technique used mainly in immunology to detect the presence of an antibody or an antigen in a sample), by the company IDEXX laboratories, was used to measure the gE-specific antibodies. Here, gE stands for "Glycoproteine E" of the Aujeszky Virus. So the ELISA used will detect specific antibodies against the glycoprotein E of the virus in order to know that pigs are infected with the virus or not.

The experiment was performed in 3 different laboratories over an 8-day period, with 2 measurements within each day. The outcome of the experiment is the percentages of inhibition, which is assumed to be normally distributed.

Enzootic bovine leukosis (EBL) affects cattle and is caused by bovine leukemia virus (BLV). EBL occurs mainly in America, Australia, eastern Europe and Asia. European Community states such as Belgium, Ireland, Norway, and the Netherlands are free of BLV. EBL is economically significant because of prema-

ture culling or death of cattle as a result of lymphosarcoma. Other costly consequences of BLV infection can be condemnation of carcasses at slaughter and losses from export restrictions. Natural infection has also been recorded in buffaloes, sheep, and capybaras.

Two datasets, labeled *SERUM* and *MILK*, are obtained by performing precision experiments using two different EBL ELISA on pooled serum and bulk milk samples, respectively. The *SERUM* dataset, is obtained by using a blocking ELISA on pooled serum Synbiotics in a 12-value experiment, including 3 laboratories for a period of 10 days and 2 measurements daily. The outcome of the experiment is expressed as % inhibition, assumed to be normally distributed. The *MILK* dataset, stems from a 10-value experiment using indirect ELISA on pooled milk samples. Three laboratories participated in the experiment for 9 days, each obtaining 2 measurements daily. The outcome of the experiment is expressed as S/P ratio's, also assumed to be normally distributed. Here, the notation S/P is motivated by S for 'Sample' and P for 'Positive.' An S/P ratio is a (semi)quantitative value for the level of antibodies in a sample. The presence of antibodies in a sample (S) against a specific disease is determined by relating the optical density values (OD) of that sample to the positive control mean ($=$ P) of the kit by calculating the sample to positive ratio.

Brucella, particularly Brucella abortus, is the causative agent of bovine brucellosis. Abortion in cows, mostly in the first three months of pregnancy or after seven months, is the most outstanding clinical feature of the disease. Infections may occur via conjunctiva or skin, but ingestion of contaminated dairy products constitutes the main risk to the public. Several member states (including Belgium) of the EU have an official status of being free of bovine Brucellosis. In most of the other countries of the EU, eradication programs for bovine brucellosis have been implemented. Brucellosis is readily transmissible to humans and can produce serious complications in the central nervous systems.

There are two motivating datasets based on Brucellosis. The first dataset, *BRU-ELISA*, is obtained from an experiment with "SERELISA Brucella Plus Ab Mono Indirect" kit, which uses an indirect ELISA, enabling the detection of Brucella lipopolysaccharides (LPS) antibodies in individual bovine serum samples. A 10-value precision experiment involving 3 laboratories for 10 days and 2 measurements per day, was performed. The outcome of the ELISA is measured as S/P ratio's, which are assumed to be normally distributed. Another test used for detection of Brucella-specific antibodies is the serum

agglutination test according to Wright (SAW-TEST), which has been used with success for many years in surveillance and control programs for bovine brucellosis. The *BRU-SAW* dataset comes from an inter-laboratory 5-value precision experiment, involving 4 laboratories for a 10-day period and 2 measurements per day. The outcome of the experiment is a titer (with a range from titer 25, 30, 50, 100 and >100). Titers which are higher than 30 are defined as positive samples. Titers below 30 are negative samples. Subsequently, the outcome of the SAW-TEST was expressed as a binary varialble (POS/NEG) indicating presence or absence of Brucella.

S/P ratios and % inhibition are derived from raw optical density (OD) values obtained by the ELISA reader in the laboratory. The OD values are always read for all samples as well as for the internal (kit), negative and positive, controls at each run. When higher OD values are observed for the controls, higher OD values are expected for the samples as well. Thus, to adjust for uncontrollable variation, the OD values are reported in a standardized version by means of S/P ratio and or % inhibition, which depend on the instructions of the kit insert. The S/P ratio and % inhibition were calculated for each sample as follows:

$$S/P\ ratio = \frac{OD(Sample)\ -\ mean[OD(negative\ controls)]}{mean[OD(positive\ controls)]\ -\ OD(negative\ controls)},$$

$$\%\ inhibition = \frac{mean[OD(negative\ controls)]\ -\ OD(Sample)}{mean[OD(negative\ control)]} \times\ 100.$$

It should be noted that neither the S/P ratio nor the % inhibition are restricted to the unit interval. In fact, the % inhibition can attain values much higher than 100 while the S/P ratio can attain values as high as 4. Based on the opinion of the laboratory technicians and experts in VAR-CODA-CERVA, as well as previous experience with analyzing similar datasets, a normality assumption is plausible. Although in some cases the data may be highly skewed and a transformation is necessary.

3 Statistical Methodology

In this section, we first briefly review ISO's analysis of variance method, then absorb this approach within the more versatile realm of linear mixed models. Thereafter the methodology is extended to generalized

linear mixed models, allowing binary and general non-Gaussian data to be used in the estimation of precision, reproducibility, and repeatability.

3.1 Review of ISO's Analysis of Variance Method

For nested experimental designs, the time-honored analysis of variance method (Neter *et al* 1996, Mickey, Dunn, and Clark 2004) may be used to estimate measures of precision, separately for each value of a precision experiment. The mean squares together with their corresponding expected values, are used to estimate the variance components that figure in the precision measures. The motivating cases, described in the previous section, all represent fully-nested experimental design. Hence, it is sensible to restrict attention to this important design. Similar analyzes can be used, for example, for the staggered-nested experimental designs.

Consider the three-factor fully-nested experiment depicted in Figure 2. Denote the data obtained from the experiment by $y_{[\ell]ijk}$. The ANOVA model for each value, ℓ, of the experiment is as follows:

$$y_{[\ell]ijk} = \beta_0 + \beta_1 \mathsf{lab}_{[\ell]i} + \beta_2 \mathsf{day}_{[\ell]ij} + e_{[\ell]ijk},$$

where $e_{[\ell]ijk} \sim N(0, \sigma^2_{[\ell]r})$. Note that i refers to laboratory, j stands for day, and k for repeats for a given combination of laboratory and day. The total variability of the measurement process consists of components contributed by the laboratories, the days of experiment, and the error term. The total variability and its corresponding components can be estimated using the sums of squares and the formula for its expected value. The total sum of squares ($SST_{[\ell]}$) can be decomposed as:

$$SST_{[\ell]} = \sum_i \sum_j \sum_k \left(y_{[\ell]ijk} - \overline{\overline{\overline{y}}}_{[\ell]} \right)^2 = SS0_{[\ell]} + SS1_{[\ell]} + SSE_{[\ell]},$$

where

$$\begin{aligned}
SS0_{[\ell]} &= \sum_i \sum_j \sum_k \left(\overline{\overline{y}}_{[\ell]i} - \overline{\overline{\overline{y}}}_{[\ell]} \right)^2 = JK \sum_i \left(\overline{\overline{y}}^2_{[\ell]i} \right) - IJK \left(\overline{\overline{\overline{y}}}_{[\ell]} \right)^2, \\
SS1_{[\ell]} &= \sum_i \sum_j \sum_k \left(\overline{y}_{[\ell]ij} - \overline{\overline{y}}_{[\ell]i} \right)^2 = JK \sum_i \left(\overline{y}^2_{[\ell]i} \right) - IJK \left(\overline{\overline{\overline{y}}}_{[\ell]} \right)^2, \\
SSE_{[\ell]} &= \sum_i \sum_j \sum_k \left(y_{[\ell]ijk} - \overline{y}_{[\ell]ij} \right)^2.
\end{aligned}$$

Table 1: *ANOVA table for a three-factor fully-nested experiment.*

Source	Sum of squares	Degrees of freedom	Mean square	Expected mean square
Factor 0	$SS0_{[\ell]}$	$I-1$	$MS0_{[\ell]} = SS0_{[\ell]}/(I-1)$	$\sigma^2_{[\ell]r} + K\sigma^2_{[\ell](1)} + JK\sigma^2_{[\ell](0)}$
Factor 1	$SS1_{[\ell]}$	$I(J-1)$	$MS1_{[\ell]} = SS1_{[\ell]}/(I(J-1))$	$\sigma^2_{[\ell]r} + K\sigma^2_{[\ell](1)}$
Error	$SSE_{[\ell]}$	$IJ(K-1)$	$MSE_{[\ell]} = SSE_{[\ell]}/(IJ(K-1))$	$\sigma^2_{[\ell]r}$
Total		$IJK-1$		

Table 1 presents the ANOVA decomposition for a three-factor fully-nested experiment (Figure 2), for a given value ℓ. The unbiased estimates $s^2_{[\ell]r}$, $s^2_{[\ell](1)}$, and $s^2_{[\ell](0)}$ for $\sigma^2_{[\ell]r}$, $\sigma^2_{[\ell](1)}$, and $\sigma^2_{[\ell](0)}$, respectively, can be obtained from the mean squares $MS0_{[\ell]}$, $MS1_{[\ell]}$, and $MSE_{[\ell]}$, respectively, using the appropriate formulas for the expected mean squares:

$$s^2_{[\ell](0)} = \frac{MS0_{[\ell]} - MS1_{[\ell]}}{JK}, \tag{1}$$

$$s^2_{[\ell](1)} = \frac{MS1_{[\ell]} - MSE_{[\ell]}}{K}, \tag{2}$$

$$s^2_{[\ell]r} = MSE_{[\ell]}. \tag{3}$$

The values for the repeatability variance, one-factor, (i.e., factor 1), intermediate precision variance, and reproducibility variance are, for a given value ℓ:

$$s^2_{[\ell]r}\ , \tag{4}$$

$$s^2_{[\ell]I(1)} = \begin{cases} s^2_{[\ell]r} + s^2_{[\ell](1)}, & \text{if } s^2_{[\ell](1)} > 0, \\ s^2_{[\ell]r}, & \text{if } s^2_{[\ell](1)} \leq 0, \end{cases} \tag{5}$$

$$s^2_{[\ell]R} = \begin{cases} s^2_{[\ell]r} + s^2_{[\ell](1)} + s^2_{[\ell](0)}, & \text{if } s^2_{[\ell](1)} > 0 \text{ and } s^2_{[\ell](0)} > 0, \\ s^2_{[\ell]r} + s^2_{[\ell](0)}, & \text{if } s^2_{[\ell](1)} \leq 0 \text{ and } s^2_{[\ell](0)} > 0, \\ s^2_{[\ell]r} + s^2_{[\ell](1)}, & \text{if } s^2_{[\ell](1)} > 0 \text{ and } s^2_{[\ell](0)} \leq 0, \\ s^2_{[\ell]r}, & \text{if } s^2_{[\ell](1)} \leq 0 \text{ and } s^2_{[\ell](0)} \leq 0. \end{cases} \tag{6}$$

From (1) and (2), it can be seen that it is possible for $s^2_{(0)}$ and $s^2_{(1)}$ to take negative values, since they

are written as the difference between two positive numbers divided by a positive number. This may be the case when there is smaller between-day variability than variability due to replicates, and when there is less between-laboratory variability than between-day variability. In the situation where $s^2_{(0)}$ and or $s^2_{(1)}$ are negative for a specified value, their values are set to zero (ISO 5725 manual, ISO 1994, Part 2, Sec. 7.4.5.4).

There is a functional relationship between the mean of the values ($m_{[\ell]}$) and the precision of the various values ($s_{[\ell]}$), if the material heterogeneity, which forms an inseparable part of the test results' variability, is a regular function of the value means. Fitting a linear function to the pairs ($s_{[\ell]}$, $m_{[\ell]}$) is complicated by the fact that both $m_{[\ell]}$ and $s_{[\ell]}$ are estimates, and thus subject to error. As the slopes are usually small, the errors of $m_{[\ell]}$ are negligible and the errors of $s_{[\ell]}$ dominate (ISO 1994). A good estimate of the parameters requires a weighted regression because the standard error of $s_{[\ell]}$ is proportional to the predicted value ($\hat{s}_{[\ell]}$). The weighting factors have to be proportional to $1/(\hat{s}_{[\ell]})^2$. Note that the concept of relating means to variances or other precision measures is in line with a long tradition is statistics that has led, for example, to the popular coefficient of variation. It amounts to considered precision not in an absolute but rather in a relative fashion.

The average of the precision values over the values will serve as the final precision estimate in situations where no satisfactory functional relationship exist between $m_{[\ell]}$ and $s_{[\ell]}$. We shall refer to the method representing precision as a function of the value mean by *Method A* and the method representing precision as the average of the precision estimates over the different values by *Method B*. The final precision estimates for Methods A and B will be calculated as:

$$\begin{aligned} \text{Method } A \quad &: \quad \text{precision}_\ell = f(\text{mean}_\ell), \\ \text{Method } B \quad &: \quad \text{precision} \ = \frac{1}{L} \times \sum_{\ell}^{L} \text{precision}_\ell, \end{aligned}$$

where $f(\cdot)$ is a functional relationship between $m_{[\ell]}$ and $s_{[\ell]}$.

3.2 Extension Into The Realms of Mixed Models

A critical aspect in determining precision is the estimation of variance components for factors affecting the measurement process. This can be achieved by ANOVA, as reviewed above. However, ANOVA can

only be performed on Gaussian outcomes. Mixed models provide a flexible way to perform variance component analyzes to both Gaussian and non-Gaussian outcomes. We will first review the linear and then broaden the view towards the generalized linear mixed model.

3.2.1 Linear Mixed Models

Linear mixed-effects models (LMM; Laird and Ware 1982, Verbeke and Molenberghs 2000) take the form

$$\begin{aligned} Y_i &= X_i\beta + Z_i b_i + \varepsilon_i, \\ b_i &\sim N(0, D), \\ \varepsilon_i &\sim N\left(0, \Sigma_i\right), \\ b_1, \ldots, b_n, \quad & \varepsilon_1, \ldots, \varepsilon_n \text{ independent}, \end{aligned} \tag{7}$$

where Y_i is the n_i–dimensional response vector for subject i, $1 \leq i \leq N$, N is the number of subjects, X_i and Z_i are $(n_i \times p)$ and $(n_i \times q)$ dimensional matrices of known covariates, β is a p-dimensional vector containing the fixed effects, b_i is a q-dimensional vector containing the random effects for subject i, and ε_i is an n_i-dimensional vector of residual components. D is a general $q \times q$ covariance matrix with (i, j) element d_{ij}=d_{ji} and Σ_i is an $n_i \times n_i$ covariance matrix which depends on i only through its dimension n_i, i.e., the set of unknown parameters in Σ_i will not depend upon i. The resulting variance of Y_i is given by $V_i = Z_i D Z_i' + \Sigma_i$. In experiments of the type considered here, the residual variance-covariance structure will often take the form $\Sigma_i = \sigma^2 I_{n_i}$, where I_{n_i} is an n_i-dimensional identity matrix. For precision experiments, all covariates in the experiment contribute to the variance of the measurements, i.e., all covariates are admitted into the random-effects structure. Consequently, X_i becomes a $n_i \times 1$ matrix and β is a 1-dimensional vector of fixed effects. Also, q in the dimension of Z_i and b_i is the number of factors, i.e., covariates, in the precision experiment.

In the scenario of our three-factor fully-nested experiment, as displayed in Figure 2, q=2, referring to the laboratories and the days of experiment. Expression (7) can be rewritten as:

$$y_{ijk} = \beta_0 + b_{i1}\mathsf{lab}_i + b_{ij2}\mathsf{day}_{ij} + e_{ijk},$$

with the same assumptions about b_i and ε_i, and where, b_i is a 2-dimensional vector containing b_{i1} and b_{ij2}, and ε_i is the JK-dimensional vector assembling the residuals e_{ijk}. Using the maximum likelihood or restricted maximum likelihood method (Verbeke and Molenberghs 2000), the latter of which corrects for small-sample bias in maximum likelihood, the parameters in D and Σ_i, the so-called variance components, are estimated and then the corresponding covariance between measurements obtained from the V matrix, or directly from software output:

$$\begin{aligned} \mathsf{Var}(Y_{ijk}) &= \sigma_0^2 + \sigma_1^2 + \sigma_r^2, \\ \mathsf{Cov}(Y_{ijk}, Y_{ijk'}) &= \sigma_0^2 + \sigma_1^2, \\ \mathsf{Cov}(Y_{ijk}, Y_{ij'k'}) &= \sigma_0^2. \end{aligned} \tag{8}$$

The unbiased estimates s_r^2, $s_{(1)}^2$, and $s_{(0)}^2$ for σ_r^2, $\sigma_{(1)}^2$, and $\sigma_{(0)}^2$, respectively, can be obtained from the estimated variance-covariance matrix $\hat{V}$ as:

$$s_{(0)}^2 = \mathsf{Cov}(Y_{ijk}, Y_{ijk'}), \tag{9}$$

$$s_{(1)}^2 = \mathsf{Cov}(Y_{ijk}, Y_{ijk'}) - \mathsf{Cov}(Y_{ijk}, Y_{ij'k'}), \tag{10}$$

$$s_r^2 = \mathsf{Var}(Y_{ijk}) - \mathsf{Cov}(Y_{ijk}, Y_{ijk'}). \tag{11}$$

The random-effects model is fitted for each value of the precision experiment, as with analysis of variance. The estimates of repeatability variance, one-factor, i.e., factor 1, intermediate precision variance and reproducibility variance are, for a given value ℓ, obtained as in (4)–(6), respectively. Observe that these estimates, for a given value of the experiment, can be obtained directly by the following equations:

$$s_{[\ell]r}^2 = \mathsf{Var}(Y_{ijk}) - \mathsf{Cov}(Y_{ijk}, Y_{ijk'}) = \sigma_r^2, \tag{12}$$

$$s_{[\ell]I(1)}^2 = \mathsf{Var}(Y_{ijk}) - \mathsf{Cov}(Y_{ijk}, Y_{ij'k'}) = \sigma_1^2 + \sigma_r^2, \tag{13}$$

$$s_{[\ell]R}^2 = \mathsf{Var}(Y_{ijk}) = \sigma_0^2 + \sigma_1^2 + \sigma_r^2. \tag{14}$$

Of course, if any of the random components, s_r^2, $s_{(1)}^2$, and/or $s_{(0)}^2$ is less than zero, then it has to be set to zero, to preserve the hierarchical interpretation to the model which is essential for the meaning attributed to the measures studied here. Either Method A or Method B is used to obtain the final estimate for precision, depending on whether there is an adequate relationship between m_l and s_l.

3.2.2 Generalized Linear Mixed Models

An elegant feature of linear, normal-distribution based models for continuous data is that the mean and variance parameters are independent. This is no longer true for general, non-Gaussian settings, a fact that poses additional challenges. Keeping this in mind, we shall now discuss methods for precision estimation based on non-Gaussian, e.g., binary data. The generalization of the linear mixed-effects model to generalized linear mixed-effects models (Breslow and Clayton 1993, Molenberghs and Verbeke 2005) provides a unified framework to address our needs.

Within this framework, outcomes are assumed to belong to the exponential family (Agresti 2002, McCullagh and Nelder 1989). As before, y_{ijk} is the kth replicate result for the jth day of the experiment in laboratory i, $(i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K)$, and Y_i is the n_i-dimensional vector of all measurements available for cluster i. The model assumes that, conditional on a q-dimensional random effect, b_i say, assumed to be drawn independently from a $N(0, D)$ distribution, the outcomes y_{ijk} are independent with densities of the form

$$f_i\left(y_{ijk}|b_i, \beta, \varphi\right) = \exp\left[\frac{y_{ijk}\theta_{ijk} - \psi(\theta_{ijk})}{\varphi} + c(y_{ijk}, \varphi)\right],$$

with θ_{ijk} the so-called natural parameter, $\psi(\cdot)$ a function solely depending on this parameter, φ a parameter included to allow for overdispersion, and $c(\cdot)$ a function of the outcomes and possibly also of the overdispersion parameter, but not of the natural parameter. The mean μ_{ijk} of y_{ijk} follows as the first derivative of $\psi(\cdot)$ with respect to the natural parameter, and is conventionally modeled through a linear predictor, containing fixed regression parameters β as well as random-effects parameters b_i, i.e.,

$$\eta(\mu_{ijk}) = \eta[E(y_{ijk}|b_i)] = x_{ijk}\beta + z_{ijk}b_i, \tag{15}$$

for a known link function $\eta(.)$, with x_{ijk} and z_{ijk} p-dimensional and q-dimensional vectors of known covariate values, respectively, and β a p-dimensional vector of unknown fixed regression coefficients. It is natural to equate the θ-parameters to their η-function counterparts: $\theta_{ijk} = x_{ijk}\beta + z_{ijk}b_i$. The random effects b_i are assumed to be sampled from $N(0, D)$. In conventional generalized linear model terms, we can write the general model as follows:

$$Y_i = \mu_i + \varepsilon_i = h(X_i\beta + Z_ib_i) + \varepsilon_i. \tag{16}$$

Here, $h(\cdot)$ is a known response function, conventionally a one-to-one mapping between the real line and the range of the expectation of the variable Y_i. This decomposition makes explicit the decomposition of the outcome in a systematic and stochastic components, naturally allowing for consideration of the variance structure and, in the repeated case, the variance and correlation structures. A general approximate formula for the variance-covariance matrix of Y_i without any restriction, neither on the distribution of the outcome variable nor on the complexity of the model, takes the form:

$$\mathsf{Var}(Y_i) = V_i = Z_i D Z_i' \Delta_i' + \Phi_{1/2} A_i^{1/2} R_i A_i^{1/2} \Phi_{1/2}, \tag{17}$$

where $\Delta_i = \left.\frac{\partial \mu_i}{\partial \eta_i}\right|_{b_i=0}$, Φ is a diagonal matrix with the overdispersion parameters along the diagonal, R_i is the correlation matrix of the error terms, and A_i is a diagonal matrix containing the variances following from the generalized linear model specification of Y_{ijk} given the random effects $b_i = 0$, i.e., with diagonal elements $v(\mu_{ijk}|b_i = 0)$ (Molenberghs and Verbeke 2005, Vangeneugden *et al* 2006).

Expression (16) may be reduced to simpler structures by making plausible assumptions on the distribution of the outcome variable and/or the complexity of the model. If the canonical link is used, we have that $A_i = \Delta_i$. Assuming conditional independence, meaning there is no residual association between measurements other than the one generated by the random effects, reduces R_i to an identity matrix. Then, the variance-covariance matrix of Y_i, based on using a canonical link and assuming both conditional independence and no overdispersion, is given by: $V_i = \Delta_i Z_i D Z_i' \Delta_i' + \Delta_i$. Let us now switch to the special but important case of binary outcomes.

3.2.3 Binary Outcomes

For a binary response, the generalized linear mixed model, as represented in (15), reduces to:

$$\mathsf{logit}(E(y_{ijk} = 1|b_i)) = x_{ijk}\beta + z_{ijk} b_i,$$

and further, in the case of our 3-factor fully-nested experimental design, to:

$$\mathsf{logit}(E(y_{ijk} = 1|b_i)) = \beta_0 + b_{i1}\mathsf{lab}_i + b_{ij2}\mathsf{day}_{ij}. \tag{18}$$

The mean depends only on the overall intercept, leading to a constant variance for the error terms. This is not the case as soon as other fixed effects are present. In our situation, based on (18), we have

that $\Delta_i = \pi(\pi - 1)I_i$, where I_i is an identity matrix and π is the probability of success. This further simplifies the variance-covariance matrix of Y_i to

$$V_i = [\pi(1-\pi)]^2 Z_i D Z_i' + \pi(\pi - 1)I_i$$

and we then have that

$$\begin{aligned} \mathsf{Var}(Y_{ijk}) &= [\pi(1-\pi)]^2(\sigma_0^2 + \sigma_1^2) + [\pi(1-\pi)], \\ \mathsf{Cov}(Y_{ijk}, Y_{ijk'}) &= [\pi(1-\pi)]^2(\sigma_0^2 + \sigma_1^2), \\ \mathsf{Cov}(Y_{ijk}, Y_{ij'k'}) &= [\pi(1-\pi)]^2(\sigma_0^2). \end{aligned} \tag{19}$$

Using (19), in analogy with (8), and performing similar manipulations as carried out in (12)–(14) for continuous responses, the repeatability variance, one-factor intermediate precision variance, and reproducibility variance can be estimated, respectively, for each value as:

$$\begin{aligned} s^2_{[\ell]r} &= \mathsf{Var}(Y_{ijk}) - \mathsf{Cov}(Y_{ijk}, Y_{ijk'}) = \pi(1-\pi), \\ s^2_{[\ell]I(1)} &= \mathsf{Var}(Y_{ijk}) - \mathsf{Cov}(Y_{ijk}, Y_{ij'k'}) = \pi(1-\pi)[\pi(1-\pi)\sigma_1^2 + 1], \\ s^2_{[\ell]R} &= \mathsf{Var}(Y_{ijk}) = \pi(1-\pi)[\pi(1-\pi)(\sigma_1^2 + \sigma_0^2) + 1]. \end{aligned}$$

The final precision estimates for a binary variable are calculated using similar procedures as in the continuous case, either establishing an appropriate functional relationship between m_l and s_l or taking the average of the precision estimates over the values of the experiment.

3.2.4 Inference Based on Mixed Models

Standard mixed-model inferential methods are used (Verbeke and Molenberghs 2000, 2005, Pinheiro 2006), thereby taking into account such peculiarities as the positioning of null hypotheses for variance components on the border of the parameter space (Stram and Lee 1994, 1995; Raubertas, Lee, and Nordheim 1986; Shapiro 1988; Verbeke and Molenberghs 2003; Silvapulle and Silvapulle 1995), owing to the fact that these variance components typically are constrained to be non-negative. This kind of subtle issues are predominantly of interest when testing hypotheses, whereas our concern here lies primarily with parameter estimation.

3.2.5 Availability of Software

There are many statistical software packages available that may be used to perform ANOVA, such as SPLUS, R, SPSS, and SAS, to name but a few. SAS Version 9.2 was used for the analyzes conducted in this paper. Focusing on SAS, many procedures, including ANOVA, GLM, NESTED, and MIXED, may be used to perform analysis of variance. The SAS procedure GLM provides estimates for the mean squares, from which the estimates of the precision measures can be calculated using (1)–(6). The SAS procedure NESTED may be used to perform a similar analysis, calculating for each value the quantities s_r^2, $s_{(1)}^2$, and $s_{(0)}^2$.

The procedure NESTED is easier to use and computationally more efficient than the MIXED procedure. Procedure MIXED performs the same analysis as the procedure NESTED and yields identical results. However, fitting linear mixed models with the MIXED procedure has the advantage of using constrained optimization. Rather than estimating the variance components and then setting the negative values to zero, procedure MIXED estimates the variance components under the constraint that they are bounded from below by zero, obviating the need for such *ad hoc* manipulation. Note that, when the procedure MIXED is used with the 'nobound' option, then negative variance components may be returned.

Statistical packages that may be used to perform GLMM analyzes include SPLUS, R, MLwiN, and SAS. In particular, the SAS procedures NLMIXED and GLIMMIX may be used. Although NLMIXED provides more accurate approximations than GLIMMIX, it is less flexible and more computationally intensive than the latter. We opted for the SAS procedure GLIMMIX. In general, fitting GLMM in practice is even more demanding than fitting linear mixed models.

3.3 Using Precision Estimates in Practice

The coefficient of variation (CV) is commonly used in reliability theory, in particular when describing the normal distribution for positive mean values with the standard deviation significantly less than the mean. However, it breaks down theoretically, unless the distribution is known to be positive valued, since there is a non-zero probability that the distribution will assume a negative value. Furthermore, such negative values are common with laboratory measurements as a consequence of standardization

based on manufacturers' limits. Hence, we do not encourage the use of CV in interpreting precision estimates, also since it does not extend to non-Gaussian outcomes.

As mentioned in the Introduction, precision estimates may be used in a variety of decision making procedures. Depending on the application at hand, these estimates may be used differently. Let us now consider two such situations.

3.3.1 Precision Limits for Gaussian Outcomes

The procedures presented above focus on estimating the standard deviations associated with operations under repeatability or reproducibility conditions. However, in practice, differences observed between two or more test results are examined, for example to investigate the acceptability of test results from a laboratory. For this purpose, some measure similar to a critical difference is required, rather than a standard deviation. The standard deviation based on sums or differences of n independent estimates, each with standard deviation σ, is given by $\sigma\sqrt{n}$. In statistical practice, the critical difference used is often τ times the standard deviation, where the value of τ depends on the probability level to be associated with the critical difference and on the shape of the underlying distribution. For a probability level of 95% and an assumed normal distribution, $\tau = 1.96$. Thus, the critical difference for comparing the difference between two values is given by $\tau\sigma\sqrt{2}$, resulting in the value 2.77. Therefore, the repeatability and reproducibility limits are $r = 2.77\sigma_r$ and $R = 2.77\sigma_R$, respectively, where σ_r and σ_R are the repeatability and reproducibility standard deviations, respectively. In practice, when examining two single test results obtained under repeatability or reproducibility conditions, the comparison shall be made with the repeatability or reproducibility limit, respectively. The procedure to obtain precision limits for comparing more than two values is described in the ISO 5725 manual (ISO 1994, Part 6).

3.3.2 Precision and Reliability

Reliability refers to the quality of measurement. Precision has an elegant link with reliability. For example, test-retest reliability refers to applying the measurement on the same or a similar sample but under two different conditions. Classically, reliability coefficients take the form of ratios of variances:

the variance attributed to the difference among measurements divided by the total variance (Shrout and Fleiss 1979). In case of continuous data, the intraclass correlation coefficient (ICC) is used to measure reliability, although ICC-type quantities can be defined for binary and ordinal categorical data as well (Fleiss 1981). Vangeneugden *et al* (2006) extended the concept of reliability to generalizability, which encompasses the reliabilities of various types, including test-retest reliability and inter-rater agreement, amongst others, for measurements having densities in the exponential family. We then have that

$$\textit{reliability} = 1 - \left(\frac{\textit{repeatability}}{\textit{reproducibility}}\right)^2. \tag{20}$$

Reliability captures the proportion of the variability that is systematic and ranges from 0 to 1, or from 0% to 100%, and provides an alternative way of interpreting the precision estimates. This interpretation is relative, unlike the absolute interpretation of the standard errors for repeatability and reproducibility. Such a measure may be used to investigate if a measurement method has been 'properly' standardized. We recommend the use of *reliability'* defined as

$$\textit{reliability}' = \left(\frac{\textit{repeatability}}{\textit{reproducibility}}\right)^2,$$

which takes values in the same range as (20). *Reliability'* estimates the proportion of variability ascribed to random error. A properly standardized measurement method will reduce substantially the effect of factors affecting the variability of the measurement results, i.e., the values of their variance components should be smaller compared to a poorly standardized version of the measurement method. Hence, for a properly standardized measurement method, measuring the same entity in different labs or day or by different operators will induce little variability in the measurement results, relative to random errors. Thus, *reliability'* estimates may be used to measure how well the measurement method has been standardized, with values close to 1 and 0, reflecting properly and poorly standardized measurement methods, respectively. A *reliability'* of 0 is a degenerate situation occuring when repeatability is zero. Thus, it may be more sensible to look at the value of reproducibility and decide whether it is large or not.

It is often the case that correlations or ICC for binary data are restricted to a subinterval of the unit interval. A typical instance is provided by Bahadur (1961), using correlations in a Bahadur model, developed for multivariate or repeated binary data. *Reliability'*, being an ICC, is thus restricted to a

subinterval of the unit interval, for binary data. The *reliability'* estimates for binary data are thus standardized to the unit interval, using the endpoints of the restricted interval. Based on our experimental design in Figure 2,

$$\textit{reliability}' \in \left[\frac{4}{\sigma_0^2 + \sigma_1^2 + 4}, 1\right). \tag{21}$$

4 Analysis of Case Studies

Let us now apply the methods described above to the datasets introduced in Section 2.2. For all five datasets, we will calculate the repeatability, day-precision, and reproducibility standard deviations for each sample. The sample precision estimates will be combined using both Method A and Method B, described in Section 3.1. We shall also focus on how the various measures can be used in practice.

4.1 Gaussian Data

This section is dedicated to analysis of the four datasets with Gaussian outcomes. It is worth noting that SAS procedures NESTED and MIXED produce the same results in the calculation of the laboratory, day and error (replication) variance components, owing to the balanced and complete nature of the data.

The means and standard deviations of the responses, by values, are shown in Table 2. It appears that the material heterogeneity, constituting an inseparable part of the variability of the test results, is a regular, probably linear, function of the value means. This is apparently clear for the EBL Milk and BRU ELISA datasets, but remains to be confirmed using the precision estimates, which are components of the standard errors shown in Table 2. The means of the responses vary in magnitude across the datasets, probably due to the difference in units of measurement. Thus, absolute measures, such as precision standard deviations, should not be used for comparison of the performance of the various tests. Rather, standardized measures such as $reliability'$ should be used.

Indeed, it is perfectly possible that different acceptance criteria (standard deviation or CV%) are used and defined for different diagnostic assays. Therefore, the estimates for repeatability and reproducibility for one particular assay should be compared in the long term. For example, comparing the repeatability estimates for the same samples and for the same diagnostic assay

Table 2: *Means (standard deviations) of the responses, by dataset and by values.*

Value	AUJES	EBL Serum	EBL Milk	BRU ELISA
1	80.418 (3.171)	101.383 (1.506)	2.906 (0.660)	1.449 (0.156)
2	74.385 (2.508)	101.197 (1.401)	2.892 (0.625)	0.941 (0.111)
3	65.581 (3.389)	100.423 (1.468)	2.848 (0.656)	0.545 (0.078)
4	53.210 (3.905)	98.753 (2.471)	2.869 (0.642)	0.263 (0.052)
5	38.204 (4.672)	87.536 (8.918)	2.822 (0.618)	0.144 (0.030)
6	25.094 (4.950)	61.500 (13.748)	2.667 (0.516)	0.078 (0.028)
7	22.348 (23.909)	31.738 (14.438)	1.899 (0.324)	0.046 (0.022)
8	15.056 (20.189)	14.142 (11.919)	1.060 (0.219)	
9		7.000 (9.817)	0.526 (0.090)	
10		4.060 (10.698)	0.250 (0.064)	
11		2.884 (11.825)		
12		2.049 (11.960)		

but for many different batches of that assay. If a high variability is seen between the different batches, this might indicate an 'uncontrolled' production process at the producer's value.

The SAS procedure MIXED was used to estimate the laboratory, day, and error variance components by values, for the various datasets. Repeatability, day-precision, and reproducibility standard deviations for each value were then obtained using (4)–(6). It is preferable to obtain final precision estimates, over all values, using both Methods A and B. Investigators may then decide which final estimate to use, depending on whether they accept the functional relationship developed in Method A as satisfactory. Functional relationships between value means and precisions estimates are required to be simple, to ease interpretation and use. The functional relationships investigated in this paper can be written, generally, as

$$E(f(\mathit{Precision})) = \mathit{Intercept} + \mathit{Slope} \cdot g(\mathit{Means}), \tag{22}$$

where $E(\cdot)$ is the expectation, $f(\cdot)$ and $g(\cdot)$ are either the identity or natural logarithmic functions. Thus, four functional relationships were investigated for each precision estimate. The models were fitted using weighted least squares, with weights proportional to the inverse of the predicted precision estimates. A 'best' model was selected based on R^2 values. Table 3 shows the intercepts and slopes of the selected models for each precision estimate, by dataset. For the AUJES dataset, f and g are the natural logarithmic function, whereas f and g are the identity function for the other three datasets.

For repeatability in the EBL-Milk data, the R^2 value is low and the slope is not significant at the 5% level. This indicates lack of evidence for a functional relationship between repeatability and the value means for this data. However, there is evidence for a functional relationship between precision estimates and the value means for all other datasets. Thus, given the value means, the precision estimates can be obtained using (22). It should be noted that reliability and precision limits can be obtained for each value of the experiment, when using Method A. However, the use of these measures will be illustrated with Method B.

An investigator may not be satisfied with a 'best' functional relationship between precision and the value means. Also, precision experiments having small number of values, say less than 5 values, do not allow an investigator to develop such functional relationships. For these situations, Method B may be an alternative. This means that the final precision estimates are obtained by averaging precision estimates over all values of the experiment. Table 4 shows the precision standard deviations, together with the precision limits for comparing 2 measurement values from different sources, as explained in Section 3.3.1. It is obvious that these precision estimates underestimate (overestimate) the precision for values with much higher (lower) means than the overall mean, if there exists a satisfactory positive linear relationship between precision and the value means. However, these estimates are easier to interpret and use. For the gE-ELISA test used in the AUJES data, if a laboratory has 2 measurements taken on the same day with a difference greater than 17.547, the corresponding measurement process ought to be checked. Furthermore, if a reference laboratory has a difference greater than 23.953 with a measurement from another laboratory taken on different days, then the performance of the alternative laboratory should be investigated. A similar interpretation holds for the tests used in the other datasets.

Preliminary evidence of repeatability is necessary to warrant further development of an assay.

Table 3: *Functional relationships between value means and precision.*

Dataset	Precision	Intercept	Slope	R^2
AUJES	Reproducibility	6.032	-1.169	0.698
	Day-Precision	5.757	-1.133	0.689
	Repeatability	6.385	-1.357	0.765
EBL-Serum	Reproducibility	14.572	-0.091	0.455
	Day-Precision	11.922	-0.087	0.767
	Repeatability	8.297	-0.071	0.963
EBL-Milk	Reproducibility	-9.165*	0.274	0.941
	Day-Precision	5.693	0.125	0.974
	Repeatability	15.970	-0.012*	0.041
BRU-ELISA	Reproducibility	0.023	0.976	0.981
	Day-Precision	0.018	0.982	0.993
	Repeatability	0.006	0.057	0.997

$*$: *Not significant at the 5% level of significance.*

Coefficients of variation (standard deviation of replicates/mean of replicates), generally less than 20% indicates adequate repeatability. However, if evidence of excessive variation ($>$ 30%) is apparent for most samples within and/or between runs of the assay, more preliminary studies should be done to determine whether stabilization of the assay is possible, or whether the test format should be abandoned. This is important because an assay that is inherently variable has a high probability of not withstanding the rigors of day-to-day testing on samples from the targeted population of animals. Therefore, repeatability and reproducibility can be judged a priori by setting up minimal criteria such as a maximum of 15% coefficient of variation (CV) for repeatability and 30% CV for reproducibility.

Coefficients of variation based on the corresponding repeatability and reproducibility stan-

Table 4: *Precision and reliability estimates based on Method B. Column labels are: 'Repeat' for repeatability, 'DayP' for day precision, 'Repro' for reproducibility, 'RepeatL' for repeatability limit, 'ReproL' for reproducibility limit, 'Reliab' for $reliability'$, 'RepCV' and 'ReproCV' for repeatability and reproducibility CV (%) respectively.*

Dataset	Repeat	DayP	Repro	RepeatL	ReproL	Reliab	RepCV	ReproCV
AUJES	6.335	7.709	8.647	17.547	23.953	0.537	13.540	18.482
EBL-Serum	4.570	7.067	8.934	12.660	24.748	0.262	8.951	17.499
BRU ELISA	0.035	0.066	0.070	0.096	0.193	0.249	7.069	14.137
EBL-Milk	0.122	0.313	0.491	0.337	1.359	0.061	5.883	23.675

dard deviations for each motivating case study are shown in Table 4. All repeatability and reproducibility coefficients variation are less than 15% and 30% respectively. This provides motivation for further development of the various assays.

Also included in Table 4 are the $reliability'$ estimates. Unlike the standard errors, the $reliability'$ estimates are relative and indicate the proportion of the total variability attributed to the random error. Tests with low values of reliability′ indicate considerable variability induced by the different laboratories and days, which is an indication that the use of such tests can be improved. Based on Table 4, the gE-ELISA test used in the AUJES data is the 'most' reliable one, with more than 50% of the variability stemming from random error. Also, the least reliable test is the indirect-ELISA used in the EBL-Milk data, where the laboratories and days account for more than 90% of the total variability. This may be due to the plate to plate differences observed with this particular ELISA as well as the "matrix" milk.

4.2 Binary Data

In this section, we employ the methods described in Section 3.2.2 to the binary outcomes. The experiment originally had 5 values but, since 3 out of the 5 values exhibit no variation, we will focus on the remaining two values.

Obviously, Method A cannot be applied to obtain final precision values, since two remaining values are insufficient to develop a functional relationship between precision and the value means. Therefore, we present the results based on Method B, the average over both values. The repeatability, day-precision, and reproducibility standard deviations are estimated as 0.304, 0.305, and 0.307, respectively. The unadjusted reliability′ is estimated at 0.984, which belongs to the interval $[0.955, 1)$. Using (21), the adjusted reliability′ is 0.574, which lies in the unit interval. Thus, the serum agglutination test according to Wright used in the BRU-SAW data is moderately reliable.

5 Discussion

In this paper, we have reviewed the ISO method for estimating precision on Gaussian data and placed this conventional ANOVA method within the linear mixed model context. We then extended the framework to non-Gaussian data using generalized linear mixed models, which encompass the LMM framework and hence addresses the Gaussian and non-Gaussian settings simultaneously. In particular, with Gaussian outcomes, the investigator has a choice between the ANOVA and GLMM framework and it is therefore useful to clearly see the important differences that exist between them. ANOVA may be computationally somewhat more efficient, since it allows for closed-form solutions, even though for Gaussian data the difference is negligible. The computational skills required for using GLMM are a little more advanced; in particular, monitoring convergence can pose challenges. With ANOVA, some variance components may be negative, in which case they are routinely set to zero, whereas virtually all optimization routines employed for GLMM force all variance components to be non-negative. The latter is more principled in view of properly accounting for the total variability. Of course, with non-Gaussian outcomes, the ANOVA framework is no longer an option.

Once estimates of the variance components and precision estimates have been obtained by means of GLMM, depending on the number of values in the precision experiment, a satisfactory functional relationship between precision and the value-means is aimed for and, when successful, Method A can be adopted. In the reverse case, the average of the precision estimates over all experimental values is used as the final precision estimate, i.e., Method B. It is desirable that such a functional relationship be simple and easy to use, usually linear. This reduces bias of the precision estimates, especially at

values with means higher or lower than the overall mean of the measurement results. Yet, care should be exercised when using this method well outside the range of the sample used for model building. Method-B estimates may induce considerable bias in the values with means higher or lower than the overall mean of the measurement result.

Also, it is clear that material heterogeneity accounts for the high variability observed within the same assay. This indicates that an overall estimate of precision as done by method A can underestimate (overestimate) the variability for a 'particular' sample. In practice, the 'true' status of a test sample is unknown. Therefore, the choice for method A and B to evaluate the precision depends on the purpose and conclusions one wants to draw. The VAR-CODA-CERVA is planning to define, for each diagnostic assay separately, an 'acceptable' value for repeatability and reproducibility and may refine this criterion for the known positive, weak positive and negative reference samples which are used for quality monitoring within the laboratory.

Precision is commonly expressed by standard deviation. We briefly described how acceptability of measurement results can be assessed, based on precision limits obtained from the standard errors of Gaussian data. We also presented a link between precision, reliability, and reliability′, the latter being useful to gather insight about the 'degree of standardization' of a measurement method, with limiting values 1 and 0 indicating 'properly' and 'poorly' standardized measurement methods, respectively. This is based on the assumption that a high between-laboratory variability is mostly caused by some differences in the implementation of the measurement method in the different laboratories.

Five datasets motivated our research and were subsequently analyzed; they allowed us to place some emphasis on interpretation of the results. Four of them had Gaussian data. The ISO method and our method based on GLMM provided quite similar results. A satisfactorily simple relationship was obtained between precision and the value-means for all 4 datasets, except repeatability for the EBL-Milk data. Based on Method B, $reliability'$ estimates indicate that the most and least reliable of the four test used are the gE-ELISA test used in the AUJES data, and the indirect-ELISA test used in the EBL-Milk data. **Results obtained from these tests demonstrating evidence for the motivation of further development of the various assays.** Also, $reliability'$ estimates from the binary data indicates that the SAW-test used in the BRU-SAW data is reliable.

We placed some emphasis on software implementation, most importantly on SAS. Many other packages can be used, too.

Acknowledgments

The authors gratefully acknowledge support from Belgian IUAP/PAI network "Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data".

References

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

Ankenman, B.E., Aviles, A.I., and Pinheiro, J.C. (2003), "Optimal Designs for Mixed-Effects Models With Two Random Nested Factors," *Statistica Sinica*, **13**, 385–401.

Bainbridge, T.R. (1965), "Staggered nested designs for estimating variance components," *Indust. Quality Control*, **22**, 12–20.

Breslow, N.E. and Clayton, D.G. (1993), "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, **88**, 9–25.

Delgado, J. and Iyer H. (1999), "Search for Optimal Design in a Three Stage Nested Random Model," *Statist. Comput.*, **9**, 187–193.

EURACHEM/CITAC Guide CG 4 (2000), *Quantifying Uncertainty in Analytical Measurement.*

Fleiss, J.L. (1981), *Statistical Methods for Rates and Proportions*, New York: John Wiley.

ISO (1993), *Guide To The Expression Of Uncertainty In Measurement (2nd edition)*, Geneva: ISO.

ISO (1993b), International Vocabulary of Basic and General Terms in Metrology, Geneva: ISO.

ISO (1993c), *ISO 3534: Statistics—Vocabulary and Symbols*, Geneva: ISO.

ISO (1994), *ISO 5725: Accuracy (trueness and precision) of measurement methods and results*, ISO, Geneva.

ISO (1999), *ISO/IEC 17025: General Requirements for the Competence of Calibration and Testing Laboratories*, Geneva: ISO.

Laird, N.M. and Ware, J.H. (1982), "Random effects model for longitudinal data," *Biometrics*, **38**, 963–974.

McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models*, New York: Chapman & Hall.

Mickey, R.M., Dunn, O.J., and Clark, V.A. (2004), *Applied Statistics: Analysis of Variance and Regression*, New York: John Wiley.

Molenberghs, G. and Verbeke, G. (2005), *Models for Discrete Longitudinal Data,* New York: Springer.

Naik, D.N. and Khattree, R. (1998), "A computer program to estimate variance components in staggered nested designs," *J. Quality Technology*, **30**, 292–297.

Neter, J., Wasserman, W., and Kutner, M.H. (1996), *Applied Linear Statistical Models. Regression, Analysis of Variance and Experimental Designs* (5th ed.), Homewood, IL: Richard D. Irwin, Inc.

Pinheiro, J.C. (2006), "Conditional versus Marginal Covariance Representation for Linear and Nonlinear Models," *Austrian Journal of Statistics*, **35**, Number 1, 31–44.

Raubertas, R.F., Lee, C.I.C, and Nordheim, E.V. (1986), "Hypothesis tests for normal means constrained by linear inequalities," *Communications in Statistics-Theory and Methods*, **15**, 2809–2833.

Silvapulle, M.J. and Silvapulle, P. (1995), "A score test against one-sided alternatives," *Journal of the American Statistical Association*, **90**, 342–349.

Shapiro, A. (1988), "Towards a uinified theory of inequality constrained testing in multivariate analysis," *International Statistical Review*, **56** 49–62.

Shrout, P.E. and Fleiss, J.L. (1979), "Intraclass correlations: uses in assessing rater reliability," *Psychological Bulletin*, **86**, 420–428.

Smith, J.R. and Beverly, J.M. (1981), "The use and analysis of staggered nested factorial designs," *J. Quality Technology*, **13**, 166–173.

Stram, D.O. and Lee, J.W. (1994), "Variance components testing in the longitudinal mixed effects model," *Biometrics*, **50**, 1171–1177.

Stram, D.A. and Lee, J.W. (1995), "Correction to: Variance components testing in the longitudinal mixed effects model," *Biometrics*, **51**, 1196.

Vangeneugden, T., Laenen, A., Geys, H., Renard, D., and Molenberghs, G. (2006), "Marginal correlation in longitudinal binary data based on generalized linear mixed models," *Submitted for publication*.

Verbeke, G. and Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data,* New York: Springer.

Verbeke, G. and Molenberghs, G. (2003), "The use of score tests for inference on variance components," *Biometrics*, **59**, 254–262.

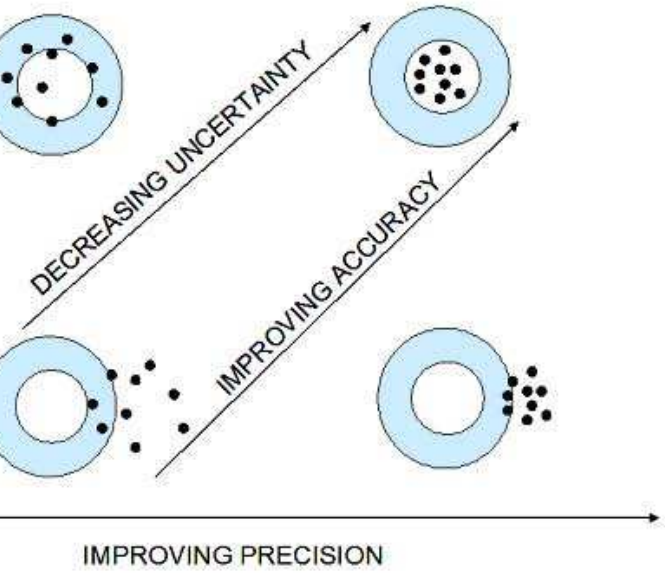

Figure 1: *A close attempt at illustrating accuracy and uncertainty. Each of these two is indicated by an arrow. A reduction in uncertainty leads to a smaller cloud of points, whereas improved accuracy leads to a cloud of points more to the center.*

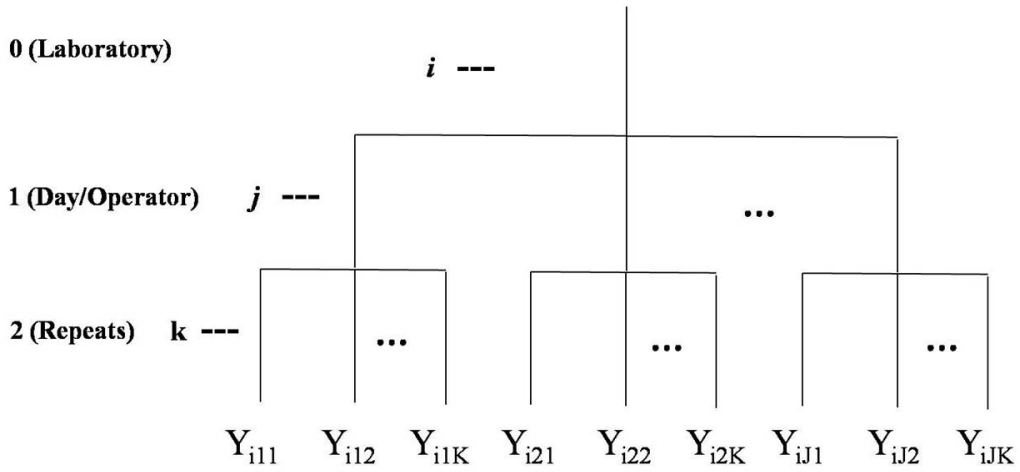

Figure 2: *Layout for three factor fully-nested experimental design (*$i = 1, dots, I; j = 1, \ldots, J; k = 1, \ldots, K$*).*