How to handle blocking

1 Introduction

It should perhaps not come as a surprise that the experimental design largely determines the complexity of the corresponding statistical model.[1] In practice, simplicity may often prevail over complexity despite the perils of simplistic statistical analyses that do not honor the experimental design.

The randomized block design is a classical choice of design, especially within agriculture related disciplines such as pesticide science. In randomized complete block designs, replicates of treatments are arranged such that all treatments occur once in all blocks. If blocks containing all treatments would become too large to be homogeneous then a balanced incomplete block design may be considered. With this design all treatment combinations will occur in the same number of blocks (this is balance), but blocks will be incomplete as they will not contain all treatment combinations.[2]

Two key components of any experimental design are the so-called treatment design and error-control design.[1] The evaluation of effects of pesticides often relies on experiments with designs that involve multiple concentrations of the pesticide of interest or multiple pesticides at specific comparable concentrations and, possibly, secondary factors of interest such as adjuvants or fertilizers. Specifically, the treatment designs may be any multi-way layouts such as two-way or three-way layouts. Such layouts may have been (partly) replicated in several blocks and/or some factors may have been (partly) replicated within blocks on several plots. This latter component of the experimental design is referred to as the error-control design.[1]

We refer to the individual replicated measurements (or simply replicates) obtained for a given treatment or treatment combination in the treatment design as observational units or simply observations. The process of obtaining replicates within the same experimental unit is referred to as (random) subsampling or pseudo replication, and it is useful for increasing the precision of estimates if the variation between observational units (within experimental units) is substantial. For instance, if the infection with a pest is known to be heterogeneous between single plants within field plots, or if repeated technical measurements from the same biological sample are known to be highly variable.

We refer to collections or groupings of observational units sharing the same treatment or treatment combination in the treatment design as experimental units (of the error-control design). Ideally experimental units should be chosen to be as homogeneous as possible. In case experimental units are not homogeneous and the experimenter does not want to restrict conclusions to a homogeneous subgroup, grouping of experimental units into more homogeneous blocks may be achieved through utilization of the actual topographical layout of the experimental design, e.g., in the field or glasshouse. This procedure is referred to as blocking and it has to be carried out before assigning treatments or treatment combinations to experimental units (through randomization). Blocking enables separating out variation between blocks from the residual within-block variation and hence it increases the sensitivity to detect differences between treatments or treatment combinations (because of the reduced residual (within-block) variation).[2] In our experience, it is not uncommon that researchers ignore the blocking structure when analyzing randomized block designs.

The aim of the present manuscript is to reinforce that it matters how blocking is dealt with in the statistical analysis of continuous responses that may be assumed to be normally distributed, possibly after a transformation. Through re-analysis of data from two experiments from pesticide science research and evaluation of results from simulated data, we explore the problems associated with the use of simplistic statistical methods and models that fail to accommodate blocking properly.

2 Experimental methods

2.1 Linear mixed models account for dependencies in the data

A key assumption of many basic statistical methods is mutual independence of all observations. This is a consequence of the assumption often described in textbooks as independence of residual error terms: After removing the differences attributable to known sources of heterogeneity (e.g., between-treatment differences or between-block differences), the remaining (residual) errors should not depend on each other.

In case experimental designs involve blocking and subsampling within the experimental units, independence of observations is no longer ensured: Even after removing the differences attributable to known sources of heterogeneity, measurements from the same experimental unit may tend to be more alike than measurements from different experimental units. In other words, measurements from the same experimental unit are not independent, but correlated, and this can be modelled by means of linear mixed models.[3] Compared to modifications of models for independent data, which were used before linear mixed models became available in the 1990's, linear mixed models are able to handle any kind of imbalance in the data (because of an unbalanced design or missing values). For specification of linear mixed models, two types of effects are distinguished: fixed effects and random effects, which may be viewed as effects pertaining to the population and block/plot level, respectively. The treatment design encapsulates the fixed effects whereas the error-control design encapsulates the random effects. Thus, a linear mixed model incorporates information on both the treatment design and on the error-control design. In particular, a linear mixed model captures how observations were randomized to treatments. The fixed effects may correspond to one or several treatment factors of interest and, possibly, also their interactions. These effects are estimated in terms of means and/or mean differences (depending on the parameterization used by the statistical software), such that hypotheses can be evaluated in the same way as for an ordinary ANOVA (where independence is assumed) by means of post hoc t-tests.[3-6]

Random effects represent different sources of variation, corresponding to the different experimental units used in the experimental design; these units are assumed to be randomly sampled from a large population of experimental units. In controlled experiments, the random effects correspond to the otherwise unexplained differences between units of the randomization procedure:[7] Years and/or locations, fields within locations, main plots or plots within blocks, single plants within plots, etc., as appropriate for a given experimental design. Random effects introduce correlations between individual measurements or observations: Two observations sharing the same level of a random effect will be correlated. The differences between levels of random effects are estimated indirectly through the so-called variance components, which reflect the variability between the observational units at the different levels in the hierarchical structure, i.e., at block and plot levels.[7] The actual degree of correlation depends on the proportions that the different variance components take up out of the total variance; the proportions will be determined from the data.

Consequently, fitting a linear mixed model may lead to different results as compared to fitting models that assume independence. Both estimated means, mean differences, and corresponding standard errors may depend on the estimated variance components of the random-effect part of the model through weighting by inverse variances. Linear mixed models may be fitted using maximum likelihood (ML) or restricted maximum likelihood (REML) estimation; the latter avoids under-estimation of standard errors in smaller datasets. Finally, it should be noted that determination of the appropriate degrees of freedom remains unresolved for linear mixed models. Several definitions have been proposed, e.g., Satterthwaite and Kenward-Roger approximations.[6,8] For larger datasets there may effectively not be much difference in results of approximate t- or F-tests obtained using different approximations and it may even be sensible simply to use the standard normal and chi-square distributions as reference distributions for the t- and F-tests, respectively.

2.2 Example: Effect of herbicide and adjuvant treatments on height in maize

In a field trial with maize the herbicide rimsulfuron was given alone or in combination with one of seven different adjuvants. The adjuvants considered were: Surfactant (0.1%), Surfactant (0.3%). Ammonium Sulphate (2%), Mineral oil (4 L ha^{-1}), Mineral oil (8 L ha^{-1}), a multinutrient fertilizer (Axan; 5kg ha^{-1}) and a cationic adjuvant (Frigate; 0.5%). The trial also contained two controls: unweeded and weed free (hand-weeded). The field trial was laid out as a randomized complete block design with ten plots (one per treatment) in each of four blocks (fields). Note that in general the use of many plots within each block is unusual and may not lead to much more homogeneity. At harvest time, several indicators of weed control ability and selectivity were recorded. For this example, we will consider crop height, which was recorded by randomly sampling of five plants from each plot (i.e., subsampling within the smallest randomization unit). A schematic picture of the randomization structure, with observational units represented by dots, is shown in Figure 1.

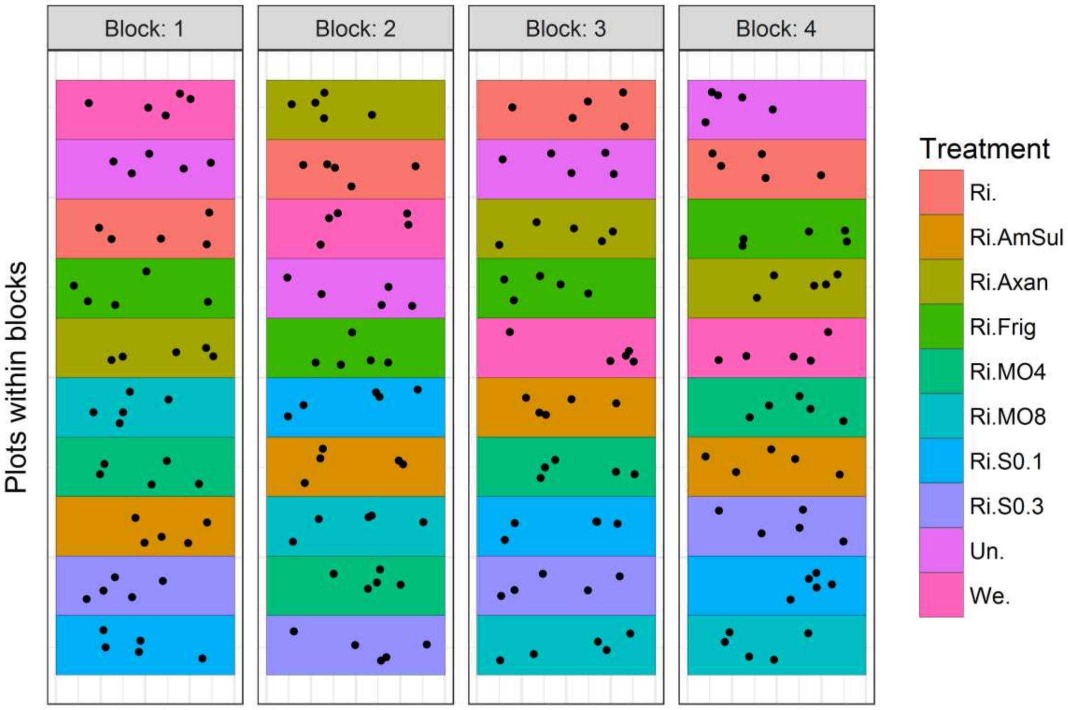

Figure 1: Experimental design for the data example on rimsulfuron and adjuvants. Within plots dots are randomly scattered to illustrate random sampling of replicates. Locations of dots, however, do not reflect the actual locations of the observational units. Ri.: rimsulfuron given alone. Ri.Amsul: rimsulfuron + Ammonium Sulphate. Ri.Axan: rimsulfuron + Axan. Ri.Frig: rimsulfuron + Frigate. Ri.MO4: rimsulfuron + Mineral oil (4 L-ha). Ri.MO8: rimsulfuron + Mineral oil (8 L-ha). Ri.SO.1: rimsulfuron + Surfactant (0.1%). Ri.SO.3: rimsulfuron + Surfactant (0.3%). Un.: Unweeded. We.: Weed-free.

We evaluated how combining rimsulfuron with different adjuvants affected the plant height, compared to the weed-free control. Specifically, seven pairwise comparisons were of interest. We considered seven different analysis strategies as detailed below.

As a first method of analysis we considered two-sample t-tests for each of the comparisons of interest, assuming homogeneous variances and implying that we completely ignored the block and plot structure and hence the dependence structure introduced by design. In this approach, the part of the variation, which could be explained by between-block and between-plot differences, became part of the residual standard error. Furthermore, for computation of standard errors and degrees of freedom each t-test involved the assumption that there were 20 independent observations per treatment group (i.e., true replicates), although there were only four replicates per treatment. Next, we analyzed data by means of one-way analysis of variance (ANOVA) where all replicates were assumed to be mutually independent measurements, also leading to inflation of the residual standard error by the between-block and plot variation. As a third approach, observations within each plot were summarized as a mean before analysis by means of a one-way ANOVA. As a fourth approach, observations within each plot were again summarized as a mean but now analyzed using a mixed model with a random block effect. Working with summary data there are no longer repeated measurements within plots, but only the latter of the two approaches takes the blocks into account. As a fifth approach, data were analyzed using a linear mixed model with a random block effect accounting for the between-block variation, but due to the subsampling in each plot, this model was still not reflecting the error-control design. As a sixth approach, data were analyzed using a linear mixed model with a random plot effect accounting for the between-plot variation, but without any random effects for the blocks (also not fully reflecting the error-control design). Finally, we analyzed data using a linear mixed model, accounting fully for the error-control design by including both random block and plot effects. This model assumed additivity with respect to the block, plot, and the treatment effects (a commonly used assumption in the analysis of data from randomized block experiments). Additional model assumptions were that the level of variation among plots was the same within each block, that the level of variation among plants was the same within every plot, and that the variability in measurements was the same for all blocks, all plots, and all plants. This model implied equal correlation between any two plots within the same block and equal correlation between measurements for any pair of plants within the same plot.

The latter four models may be thought of as one-way ANOVA models with respect to the treatment design, but incorporating additional information about the grouping of plants within blocks and plots due to the error-control design. Consequently, they may be referred to as linear mixed ANOVA models. Adequacy of model assumptions was assessed by visual inspection of residuals and estimated best linear unbiased predictors (EBLUPs) of the random effects. Confidence intervals and p-values were not multiplicity adjusted.

2.3 Example: Translocation of an insecticide seed treatment in maize

A recent study investigated the effect of a systemic insecticide, the neonicotinoid clothianidin.[9] Clothianidin was applied as a seed treatment and is translocated into plant tissue during growth providing protection from insects. The concentration of the active ingredients decreases over time within plant tissues, resulting in limited or no pest protection. Specifically, the purpose of this study was to quantify the clothianidin levels in different plant regions across the early growing season. Specifically, we considered data from two independent field experiments carried out in two successive years (2014 and 2015). Four seed treatments were evaluated: untreated seeds, fungicide treatment, low level (0.25 mg/kernel) and high level (1.25 mg/kernel). The latter two treatments were also supplemented with the same fungicides as the fungicide treatment. In a randomized complete block design, each treatment was replicated four times with four plots in each of four blocks. To minimize the contamination in the untreated plots (untreated and fungicide treatment) samples were collected from the central rows of each plot. However, some clothianidin contamination was expected for the untreated plots due to subsurface flow and the proximity of untreated plots to treated plots. We only analyzed measurements taken approximately 20 days after planting (day 19 and 20 in experiment 1 and 2, respectively). Measurements taken at other time points could be analyzed similarly. These separate analyses per time points could eventually be combined into a more comprehensive analysis of changes over time;[10] alternatively a simultaneous model for all time points could be considered. However, these analyses were beyond the scope of this article.

Ten randomly selected maize plants were removed intact from each plot and stored at -20°C until processing. From these, up to three individual samples per plot were split into seed, root and shoot prior to chemical analysis of the clothianidin level. A few samples were lost before the homogenization step leaving a total of 48 and 30 measured clothianidin levels in shoots and 42 and 30 measured levels in roots in 2014 and 2015, respectively. Figure 2 shows the experimental design for the root data, dots represent single measurements, i.e., subsampling of individual plots. In 2014 one to three samples were taken per plot, and no plot dropped completely out of analysis. In 2015 two subsamples were always taken per plot, and one entire plot had to be left out completely from analysis. Hence, this experimental design has four hierarchical layers of randomization and sampling (year/block/plot/single plant) and it is unbalanced at several of these layers.

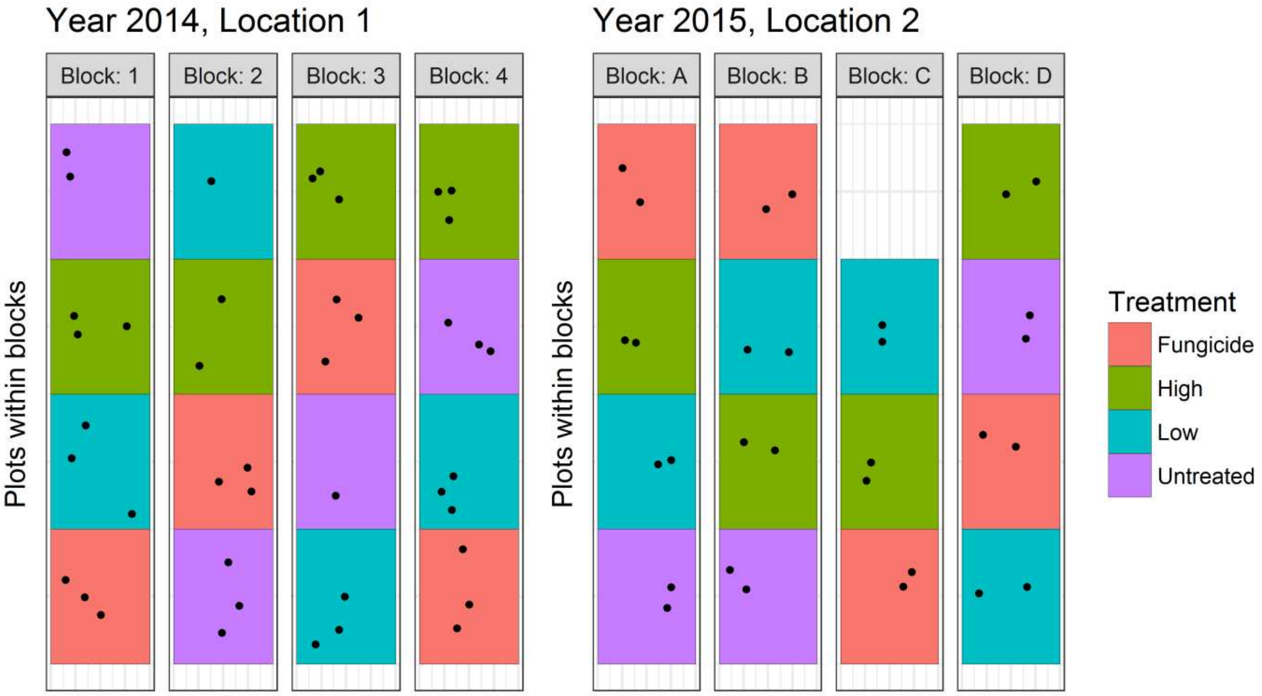

Figure 2: Experimental design for the root data from the example with translocation of insecticide seed treatment. Within plots dots are randomly scattered to illustrate random sampling of replicates. Locations of dots, however, do not reflect the actual locations of the observational units.

Data were analyzed by year and combined and three different analysis strategies were considered for both. First, for each year, an ordinary one-way ANOVA (assuming independence between all observations) was fitted, only including the treatment effect. Second, for each year, observations (of log clothianidin) within plots were summarized as a mean and analyzed using a linear mixed ANOVA model with treatment as fixed effect and block as random effect. Finally, a linear mixed ANOVA model with treatment as fixed-effects factor and plot and block as random-effects factors was fitted for each year separately.

As a joint analysis strategy, first an ordinary one-way ANOVA (assuming independence between all observations) was fitted, only including the treatment effect. This approach is not reflecting the error-control design because there are 1-3 replicates per experimental unit. Furthermore, due to the unbalanced design, variation between years and blocks may unnecessarily inflate the residual standard error and lead to confounded estimates of the treatment effects. Second, observations within each plot were summarized as a mean and analyzed using a linear mixed one-way ANOVA model with year and block as random-effects factors. Finally, a linear mixed ANOVA model with treatment as fixed-effects factor and plot, block, and year as random-effects factors was fitted, fully reflecting the error-control design. For all approaches, adequacy of model assumptions was assessed by visual inspection of residuals and EBLUPs. Clothianidin levels were log-transformed before analysis to ensure variance homogeneity. Estimates and 95% confidence intervals were obtained on the transformed scale and back-transformed. No adjustment for multiplicity was applied. As contamination was unavoidable in untreated plants, the protection from seed treatments was considered expired if clothianidin levels were similar in treated and untreated plants.

2.4 Statistical software

All analyses were carried out using the statistical programming environment R version 3.4.0[11] and, in particular, using the extension packages *lme4*[12] and *multcomp*[5] for fitting linear mixed models and obtaining pairwise comparisons, respectively.

3 Results

Both data examples had a one-way factorial treatment design. The error-control design of the first example was an RCBD with balanced subsampling. The error-control design of the second example was planned as an RCBD with unbalanced subsampling, but it became incomplete in the course of the experiment.

3.1 Example: Effect of herbicide and adjuvant treatments on height in maize

Figure 3 shows the raw data on the single plant level (grey symbols) along with the means per plot (black symbols) in four different blocks. Plots in block 1 mostly had the smallest or at least below average heights, whereas block 3 mostly had the largest plants. Therefore, visually, there was a clear block effect. For a considerable number of plots, all plants were clearly smaller than those in the remaining plots with the same treatment, illustrating the presence of additional variation between plots within blocks. Additionally, the five replicates within plots showed considerable variation in all plots.

Results of the seven pairwise comparisons for all seven approaches are shown in Table 1. The approach of applying t-tests identified a significant decrease in height when adding surfactant (0.3%), ammonium sulphate, or mineral oil at a dose of 4 L ha^{-1} to rimsulfuron as compared to the weed-free control.

The ordinary ANOVA resulted in wider confidence intervals than the t-tests and significant effects on height of rimsulfuron combined with ammonium sulphate and mineral oil at a dose of 4 L ha^{-1} compared to the weed-free control. ANOVA on summary measures and the linear mixed model with plot as random effect resulted in identical and the widest confidence intervals among all methods. No significant treatment differences were observed with these analysis strategies. Analyzing data using the linear mixed model only including block as random effect resulted in the narrowest confidence intervals almost resembling those confidence intervals obtained from the t-tests. Four significant treatment differences in height were found: rimsulfuron combined with surfactant (0.3%), ammonium sulphate, mineral oil at a dose of 4 L ha^{-1} or 8 L ha^{-1} compared with the weed-free control. Finally, the linear mixed model analysis fully accounting for the experimental design resulted in a single significant difference, the combination of rimsulfuron and ammonium sulphate compared to the weed-free control. Identical estimates and confidence intervals were observed from the linear mixed ANOVA model with block as random-effect factor. This was a consequence of the balanced design with the same number of plants sampled from each plot. The confidence intervals from these models were much wider than those obtained from t-tests and the ANOVA. As seen from the estimated variance components the random effects for block and plot explained a large part of the total variation in data. Note that all methods resulted in the same estimated means because of the balanced design.

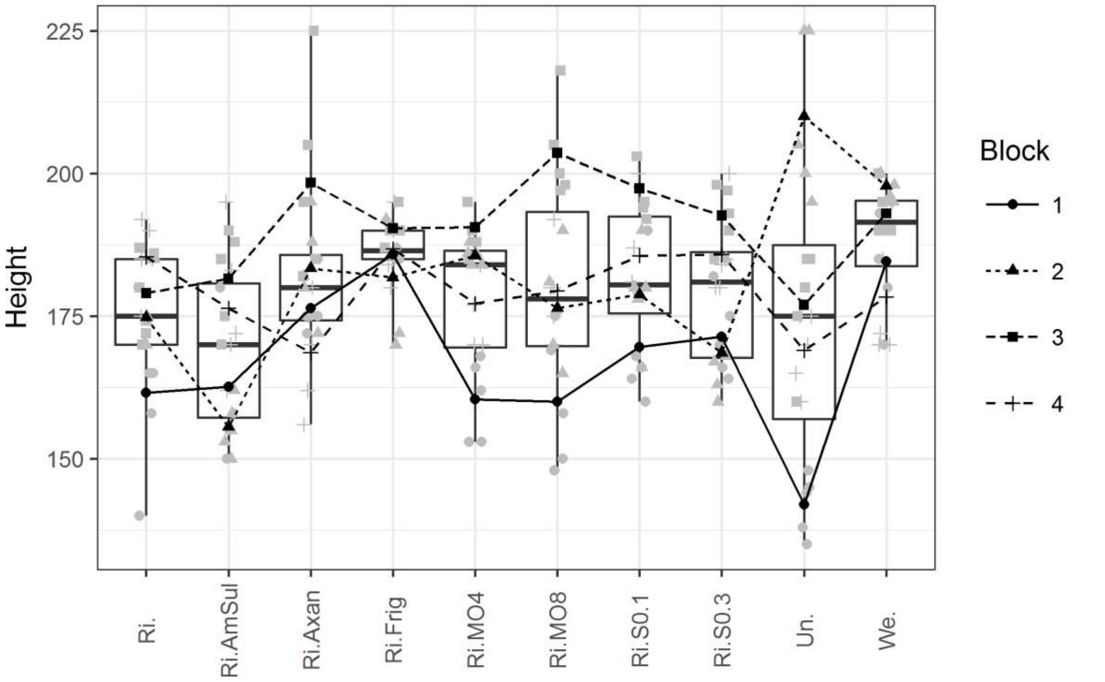

Figure 3: Mean values (black symbols) and raw data (grey symbols) for each treatment and block from the data example on rimsulfuron and adjuvants. Ri.: rimsulfuron given alone. Ri.Amsul: rimsulfuron + Ammonium Sulphate. Ri.Axan: rimsulfuron + Axan. Ri.Frig: rimsulfuron + Frigate. Ri.MO4: rimsulfuron + Mineral oil (4 L-ha). Ri.MO8: rimsulfuron + Mineral oil (8 L-ha). Ri.SO.1: rimsulfuron + Surfactant (0.1%). Ri.SO.3: rimsulfuron + Surfactant (0.3%). Un.: Unweeded. We.: Weed-free.

3.2 Example: Translocation of an insecticide seed treatment in maize

The estimated clothianidin levels in roots and shoots are shown in Table 2 for the two years separately. Treatment differences are presented in Table 3. In 2014 different numbers of plants were sampled in different plots causing different variances for the different sampling means when computed for individual plots. As a consequence, the three analysis strategies resulted in different estimated clothianidin levels, treatment differences and corresponding confidence intervals. The linear mixed ANOVA model analyzing data from all individual plants, including random effects for block and plot, was an appropriate analysis strategy as it does the proper weighting of the plot means to deal with the unequal sample sizes in the plots. In 2015, the same number of plants was sampled from each plot (except one where all data were missing). As a result, the two mixed models resulted in the same estimated treatment means, treatment differences and confidence intervals. While accounting for the block effects, the two linear mixed models also deals with the missing plot when estimating the treatment differences. Overall, the three different analysis strategies resulted in similar conclusions: In 2014, the clothianidin level in the roots did not differ significantly between the untreated and the low level clothianidin treated plots but significant differences were found for all other treatment comparisons. The same conclusions could be drawn for the shoots for all analysis strategies except for the linear mixed ANOVA on summary data, which resulted in significant differences in the clothianidin level in the shoots between all treatments. In 2015 all treatments resulted in significantly different clothianidin levels in both roots and shoots. More importantly, the estimated treatment means and treatment differences varied noticeably between the two years. A joint analysis may therefore be of great help to reach an overall conclusion.

Table 4 shows the estimated clothianidin level for each treatment in combined analysis. The estimated means differed between models, being consistently smaller for the ordinary ANOVA than for the mixed ANOVA model, as the former did not account for data missing from a block (block C) with an overall high clothianidin level (results not shown). Confidence intervals were notably narrower for the ordinary ANOVA than for the mixed ANOVA model. Results of post hoc treatment comparisons are shown in Table 5. As a consequence, of the smaller estimated standard errors in the ordinary ANOVA, the post hoc treatment comparisons also resulted in narrower confidence intervals and smaller p-values. However, in this example the overall conclusions remained the same. For both approaches, we concluded that the fungicide and untreated treatment resulted in the same low clothianidin level in both roots and shoots, followed by a somewhat larger and significantly different level in the plants from the low-level treatment. The high-level treatment also had the highest clothianidin level in both roots and shoots after 20 days.

On logarithmic scale, the residual standard error for the ANOVA models for roots and shoots were 0.94 and 0.96, respectively. For the linear mixed model on summary data, the residual errors were reduced to 0.81 and 0.62 for roots and shoots respectively and for the linear mixed model, the residual errors were reduced to 0.50 and 0.60 for the roots and shoots models, respectively, indicating that the random effects explained more than one third of the total variation. In the linear mixed model assessing the clothianidin levels in the shoots, the year effect explained most of the variation (σ_{YEAR}=0.75) as compared to the block and plot effects (σ_{BLOCK}=0.00, σ_{PLOT}=0.23). In the model for the roots, plots (σ_{PLOT}=0.55) and years (σ_{YEAR}=0.45) also explained most variation, with a negligible contribution from the block effects (σ_{BLOCK}=0.00). For both root and shoot no parameter estimates changed when the models were refitted without block, indicating that the four blocks did not contribute additional information to the model than what was already explained by plot and year. In the linear mixed models fitted to the summary data the year effect again explained most of the variation σ_{YEAR}=0.36 and σ_{YEAR}=0.66 for roots and shoots respectively, compared to the blocks (σ_{BLOCK}=0.00 for both).

4 Simulations

To substantiate and generalized the conclusions from the above examples, we performed a simulation study, i.e., we evaluated the behavior of the different methods and models for a large number of simulated datasets. Specifically, data were simulated from the final model fitted in the first example, i.e., a mixed ANOVA with block and plot as random effects. Three different designs were considered: First, data were simulated from a randomized complete block design with 10 treatments randomized without repetitions in each of 4 blocks with 5 pseudo replicates within each plot resulting in a data set with a total of N = 200 observations. Second, data were simulated from a randomized complete block design with 10 treatments randomized without repetitions within each of 4 blocks and 25 pseudo replicates within each plot resulting in a data set with N=1000 observations. Finally, data were simulated from a randomized complete block design with 5 repetitions of each treatment within each of 4 blocks and 5 pseudo replicates within each plot resulting in a data set with N=1000 observations. For all three experimental designs we also considered data where one treatment (the control) was observed without repetitions in block 4 and in block 3 two other treatments, treatment 1 and 3, were not observed at all. The lost data resulted in unbalanced data sets with a total of N = 186, 926 or 925 observations, respectively.

For each data set we estimated the treatment mean for each of the 10 treatments and the treatment differences comparing treatment 1 to 7 with the control treatment (treatment 10), as in example 1. The following models were fitted: 1) A mixed ANOVA with block and plot as random effects using REML, 2) A mixed ANOVA with block and plot as random effects using ML, 3) A mixed ANOVA with block and plot as random effects using REML and Kenward-Roger's approximation of degrees of freedom, 4) A mixed ANOVA with block as random effect using REML, 5) A mixed ANOVA with block as random effect using ML, 6) A mixed ANOVA with plot as random effect using REML, 7) A mixed ANOVA with plot as random effect using ML, 8) ANOVA on all observations, 9) ANOVA on summary measures, repetitions within each plot were summarized as a mean before analysis, 10) Mixed ANOVA on summary measures with block as random effect using REML, repetitions within each plot were summarized as a mean before analysis, 11) t-tests. The t-tests were only used for assessing treatment differences, while the mixed ANOVA using the Kenward-Roger approximation of degrees of freedom were only used in the scenarios with unbalanced data. Results were based on 1000 simulated data sets and were presented as mean percentage bias, coverage of Wald-type 95% confidence intervals and width of Wald-type 95% confidence bands summarized over all estimated means or treatment differences. For estimated treatment differences, results were furthermore presented as type 1 error rate, and power. The type 1 error rate was calculated for the treatment difference between treatment 1 and the control and power was calculated for finding a treatment difference between treatment 3 and the control. Simulations and analyses were made using the statistical programming environment R version 3.4.0[11] and in particular the extension packages *lme4*[12] and *lmerTest.*[13]

Results from the simulations are summarized in Table S2 and Table S3 for estimated means and in Table 6 and Table 7 for estimated treatment differences. For balanced data (Table S2 and Table 6) all analysis strategies resulted in the same estimated means and treatment differences and accordingly the same percentage bias. A larger bias was observed for estimated treatment differences compared to estimated means but overall bias was decreasing with increasing sample size. For the unbalanced data (Table S3 and Table 7) the different analysis strategies no longer resulted in the same estimated treatment means and differences as reflected in the dissimilar percentage biases.

Overall, the mixed ANOVA including both block and plot as random effects repeatedly resulted in coverage probabilities very close to the nominal level for treatment means and treatment differences, when estimated using REML. This model also consistently resulted in type 1 error rates closest to the nominal level. For ML estimation the coverage probabilities were consistently lower and the Type 1 error rates consistently higher. For the unbalanced data using an adjustment of the degrees of freedom (Kenward-Roger approximation) resulted in coverage probabilities closest to the nominal level but no improvement in the type 1 error rate and a lower power compared to the mixed model using the standard normal distribution approximation.

In the present simulations, ignoring the correlation between pseudo replications within plots resulted in low coverage probabilities mainly caused by more narrow confidence bands and corresponding high type 1 error rate. For the scenarios with only one replication of each treatment in each block, ANOVA on summary measures and the mixed model with plot as random effect resulted in appropriate coverage probabilities and type 1 error rates, but to a less degree for unbalanced data. Results from the mixed ANOVA on summary measure with block as random effect were identical to the results from the mixed ANOVA with plot and block as random effects (both using REML) for balanced data. For unbalanced data, coverage was in general somewhat lower and the type 1 error higher. For the scenario with treatment repetitions within each block coverage probabilities became far too low and type 1 error rates to high. t-tests and ordinary ANOVA overall resulted in the worst properties including critically low coverage and very high type 1 error rate.

Analysis strategies resulting in a low type 1 error also resulted in a low power and vice versa. However, the mixed model fully accounting for the error-control design resulted in higher power compared to other analysis strategies and comparable type 1 error rate.

5 Discussion

In the present paper, we have shown what happens if blocking and, possibly, subsampling are not appropriately incorporated in the statistical analyses. Specifically, we have shown that results from sub-optimal approaches (two-sample t-tests and ordinary ANOVA) may be quantitatively and qualitatively different from the results obtained using an appropriate mixed model.

As seen in the first example for randomized complete balanced block designs with no missing values, two-sample t-tests, ANOVA, and a linear mixed model will provide the exact same estimated treatment means. However, estimated standard errors and confidence intervals may differ substantially, depending on the chosen method. A two-sample t-test will typically not encompass the entire range of variation, because it is only based on a subset of the data. This was also supported by the simulations. ANOVA based on all measurements and linear mixed ANOVA ignoring subsampling within plots will tend to result in overly precise estimates, i.e., too small standard errors, if pseudo-replication is treated as if it were true replication. Using a linear mixed model (based on all measurements and fully reflecting both the treatment design and the error-control design) will strike the right balance between replication and pseudo-replication through a weighing step involving the estimated variance components. All statistical methods and models not fully respecting the experimental design resulted in a higher number of statistical significant treatment differences, some of which were false positive results are seen from the simulations.

In the second example we showed that a sub-optimal analysis not only influenced the estimated standard errors but also lead to different estimated treatment means. This was a consequence of the unbalanced design. It was also seen in the simulation results. For the ordinary ANOVA treatment means were estimated as raw averages whereas the mixed ANOVA accommodated the unbalanced design and produced weighted averages. Particularly, for markedly unbalanced designs such as incomplete block designs, it is well known that treatment comparisons based on simplistic arithmetic means (as used in the two-sample t-test or ordinary one-way ANOVA) yields incorrect estimates because they are confounded by block effects.[14]

A simplistic way to meet the shortcoming related to correlated observations within the experimental units is to summarize data, e.g., in terms of a mean, and then do the analysis on the summary measures, as was done in both examples and the simulations. Taking averages will remove some variation, but at the same time the degrees of freedom for the residual standard error will be reduced because of fewer observations underlying the analyses. For balanced designs, this approach lead to results resembling those from a linear mixed model if no other levels of correlation are ignored in the analysis. For unbalanced designs the different weighing of individual observations according to the different analysis strategies may result in different estimated treatment means as well as standard errors.[15]

Blocking is a special case of a hierarchical sampling structure. In general, hierarchical (or nested) sampling or randomization structures occur if there are more measurements or observations planned than there are experimental units for a treatment or treatment combination. For instance, two or more treatment factors are investigated in the same experiment; it may be technically impossible, very expensive, or practically unfeasible to randomly assign the levels of one factor to the smaller experimental units although it may be feasible for the remaining treatment factor(s). For example, different infection levels may only be randomly assigned to large subunits or strips in a field, while the remaining factors can be assigned to individual plots within the large subunits or even to single plants, or tissue samples within plots or single plants.

We considered data from field trial where different sources of variation were associated with space: blocking is useful to account for spatial gradients in the field. In highly controlled laboratory experiments, spatial arrangement may be far less important than in the field. Instead, variation of experimental conditions may be associated with time, biological or chemical raw material, technical devices, or the persons involved in the handling of the experiment. The expected value of measurements as well as the expected efficacy of treatments may change over time due to changing climatic conditions, changing quality of plant or pathogen material, changing lots of chemicals or other raw material, unobserved differences between technical devices or errors committed by experimenters. If such sources of variation exist, blocking against time (or other sources of error) paired with an appropriate linear mixed model analysis accounting for the introduced block effects may be used to account for the heterogeneity of experimental conditions over time.

A different situation where we might want to collect multiple measurements within each experimental unit is in case interest lies in change over time, e.g., determination of the dissipation kinetic of one xenobiotic over time. Multiple measurements are then not simply taken to explore the variability within each experimental unit; on the contrary, the experimenter is specifically interested in the observed within-plot differences over time. In this case, we do not usually talk about subsampling or pseudo-replication but about repeated measurements. However, analysis of repeated measurements, which also requires modelling correlation over time, is beyond the scope of this paper.

Instead of including blocks as random effects in a linear mixed model, an alternative may be to include blocks as fixed effect in an ordinary ANOVA (assuming all observations are independent). As with the summary measures, this approach may again result in bias in the estimated treatment means and too large or too small standard errors caused by the lack of appropriate weighing. Besides possibly interfering with the conclusions, the choice of considering blocks as fixed or random effects will also influence the interpretation of the results. When block is included as a fixed effect it should be because it is of interest to compare blocks or because it may be considered a confounder. Strictly speaking, results from such a study may only be interpreted for the specific blocks used in the experiment. On the contrary, when blocks are of no particular interest and they may be assumed to be sampled from a large pool of units, they should be included as random effects in a linear mixed model. In this case results may be interpreted within the range of that wider pool of experimental units, i.e., results may be generalized beyond the blocks that were actually used in the experiment. Moreover, in case of imbalance in the data, linear mixed models allow borrowing strength from blocks with more measurements to blocks with fewer measurements.

Finally, we only considered continuous responses that were normally distributed on some scale. Non-normally distributed responses would need to be modelled under other distributional assumptions.[16-18] It should be noted that the same problems of aligning the statistical model with the experimental design exist and the same type of solutions are available.[14,19,20]

In conclusion, for experiments in which treatments are randomly assigned to plots within blocks and observations are taken on individual units sampled from the plots, always include random block and plot effects in the analysis. This is most easily done using a mixed model approach that contains fixed effects for all treatments of interest and a set of random effects for block and plots within blocks. Leaving out either the random plot or block effects, or both, results in incorrect p-values for tests involving the treatment means and incorrect confidence intervals for differences in treatment means. In special balanced cases, a more simple analysis based on the sample means for each plot, but still including a random block effect, will result in the same results as the analysis based on the mixed effects model including both plot and block as random effects. For unbalanced data, this strategy will not suffice. We recommend that author guidelines should explicitly point out that the authors need to indicate how the statistical analysis reflects the experimental design, i.e., the treatment design as well as the error-control design. Additionally, it could be mentioned that, based on the experimental design, the statistical analysis could almost entirely be specified prior to execution of the experiment.

References

1. Hinkelmann K, Kempthorne O, Design and analysis of experiments, Volume 1: Introduction to experimental design, 2nd edition. John Wiley & Sons, Inc., Hoboken, NJ, USA (2008).
2. Oehlert GW, A first course in design and analysis of experiments, WH Freeman New York, New York (2000).
3. Pinheiro JC, Bates DM, Mixed-effects models in S and S-PLUS, Springer, New York (2000).
4. Spilke J, Piepho HP, Hu X, A simulation study on tests of hypotheses and confidence intervals for fixed effects in mixed models for blocked experiments with missing data. *J Agric Biol Environ Stat* **10**:374-389 (2005).
5. Hothorn T, Bretz F, Westfall P, Simultaneous inference in general parametric models. *Biom J* **50**:346-363 (2008).
6. Lenth RV, Least-Squares Means: The R package lsmeans. *J Stat Softw* **69**:1-33 (2016).
7. Piepho HP, Büchse A, Emrich K, A hitchhiker's guide to mixed models for randomized experiments. *J Agron Crop Sci* **189**:310-322 (2003).
8. Luke SG, Evaluating significance in linear mixed-effects models in R. *Behav Res Methods* 49:1494-1502 (2017).
9. Alford A, Krupke CH, Translocation of the neonicotinoid seed treatment clothianidin in maize. *PLoS One* **12**:e0173836 (2017).
10. Pallmann P, Pretorius M, Ritz C, Simultaneous comparisons of treatments at multiple time points: Combined marginal models versus joint modeling. *Stat Methods Med Res.* 10.1177/0962280215603743 (2015)
11. R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (2017). URL https://www.R-project.org/
12. Bates A, Maechler M, Bolker B, Walker S, Fitting linear mixed-effects models using lme4. *J Stat Softw* **67**:1-48 (2015).

13. Kuznetsova A, Brockhoff PB, Christensen RHB, lmerTest: Tests in linear mixed effects models. R package version 2.0-33. (2016). https://CRAN.R-project.org/package=lmerTest.
14. Littell RC, Analysis of unbalanced mixed model data: A case study comparison of ANOVA versus REML/GLS. *J Agric Biol Envir Stat*, **7**:472-490 (2002).
15. Dean A, Voss D, Design and analysis of experiments, Springer, New York (1999).
16. Szöcs E, Schäfer RB, Ecotoxicology is not normal. *Environ Sci Pollut Res* **22**:13990-13999 (2015).
17. Stroup WW, Rethinking the analysis of non-normal data in plant and soil science. *Agron J* **107**:811-827 (2015).
18. Warton DI, Lyons M, Stoklosa J, Ives AR, Three points to consider when choosing a LM or GLM test for count data. *Methods in Ecology and Evolution* **7**:882-890 (2016).
19. Piepho HP, Analysing disease incidence data from designed experiments by generalized linear mixed models. *Plant Pathol* **48**:668-674 (1999).
20. Madden LV, Turechek W, Nita M, Evaluation of generalized linear mixed models for analyzing disease incidence data obtained in designed experiments. *Plant Dis* **86**:316-325 (2002).

Table 1: Estimated differences in height between maize treated with rimsulfuron and one of different adjuvants and a weed-free control. Estimated variance components are also shown.

	t-tests		ANOVA		ANOVA[2]		ANOVA[2] with block		ANOVA[3] with plot		ANOVA[3] with block		ANOVA[3] with block and plot	
Adjuvant[1]	Est. diff (95%-CI)	p-value	Est. diff (95%-CI)	p-value	Est. diff (95%-CI)	p-value	Est. diff (95%-CI)	p-value	Est. diff (95%-CI)	p-value	Est. diff (95%-CI)	p-value	Est. diff (95%-CI)	p-value
Surfactant	-5.60		-5.60		-5.60		-5.60		-5.60		-5.60		-5.60	
(0.1%)	(-13.26-2.06)	0.1472	(-14.93- 3.73)	0.2409	(-25.42-14.22)	0.5839	(-21.31-10.11)	0.4849	(-25.42-14.22)	0.5898	(-13.41- 2.21)	0.1601	(-21.31-10.11)	0.4849
Surfactant	**-8.85**		-8.85		-8.85		-8.85		-8.85		**-8.85**		-8.85	
(0.3%)	**(-16.26- -1.44)**	**0.0205**	(-18.18- 0.48)	0.0646	(-28.67-10.97)	0.3885	(-24.56- 6.86)	0.2697	(-28.67-10.97)	0.3815	**(-16.66- -1.04)**	**0.0264**	(-24.56- 6.86)	0.2697
Ammonium sulphate	**-19.40**		**-19.40**		-19.40		**-19.40**		-19.40		**-19.40**		**-19.40**	
(2%)	**(-27.41- -11.39)**	**<0.0001**	**(-28.73- -10.07)**	**0.0001**	(-39.22-0.42)	0.0646	**(-35.11--3.69)**	**0.0155**	(39.22-0.42)	0.0551	**(-27.21--11.59)**	**<0.0001**	**(-35.11--3.69)**	**0.0155**
Mineral oil	**-10.00**		**-10.00**		-10.00		-10.00		-10.00		**-10.00**		-10.00	
(4 L ha^{-1})	**(-17.54- -2.46)**	**0.0107**	**(-19.33- -0.67)**	**0.0370**	(-29.82-9.82)	0.3307	(-25.71- 5.71)	0.2123	(-29.82-9.82)	0.3228	**(-17.81- -2.19)**	**0.0121**	(-25.71- 5.71)	0.2123
Mineral oil	-8.60		-8.60		-8.60		-8.60		-8.60		**-8.60**		-8.60	
(8 L ha^{-1})	(-18.16-0.96)	0.0766	(-17.93- 0.73)	0.0724	(-28.42-11.22)	0.4019	(-24.31- 7.11)	0.2834	(-28.42-11.22)	0.3951	**(-16.41- -0.79)**	**0.0310**	(-24.31- 7.11)	0.2834
Multinutrient fertilizer	-6.75		-6.75		-6.75		-6.75		-6.75		-6.75		-6.75	
(Axan; 5 kg ha^{-1})	(-15.14-1.64)	0.1116	(-16.08- 2.58)	0.1578	(-26.57-13.07)	0.5096	(-22.46- 8.96)	0.3998	(-26.57-13.07)	0.5045	(-14.56- 1.06)	0.0904	(-22.46- 8.96)	0.3998
Cationic adjuvant	-2.25		-2.25		-2.25		-2.25		-2.25		-2.25		-2.25	
(Frigate; 0.5%)	(-7.85-3.35)	0.4213	(-11.58- 7.08)	0.6370	(-22.07-17.57)	0.8254	(-17.96-13.46)	0.7790	(-22.07-17.57)	0.8239	(-10.06- 5.56)	0.5725	(-17.96-13.46)	0.7790
σ_{BLOCK}							8.72				9.26		8.72	
σ_{PLOT}									13.75				10.63	
Σ			15.05		14.3		11.34		8.79		12.61		8.79	

[1]Compared to weed-free control

[2]Model fitted to summary data. Repetitions within each plot were summarized as a mean before analysis.

[3]Mixed model fitted with REML estimation.

Table 2: Estimated mg clothianidin per g of plant tissue with corresponding 95% confidence intervals in roots and shoots by year 20 days after planting for three models: a one-way ANOVA assuming independent observations, a one-way ANOVA mixed model on summary data with block as random-effect variable, and a one-way ANOVA mixed model with year, block, and plot as random-effects variables.

		2014						2015					
		ANOVA		Mixed ANOVA on summary data[1]		Mixed ANOVA[1]		ANOVA		Mixed ANOVA on summary data[1]		Mixed ANOVA[1]	
Plant tissue	Treatment	Estimate	95%-CI	Estimate	95%-CI	Estimate	95%-CI	Estimate	95%-CI	Estimate	95%-CI	Estimate	95%-CI
Roots	Untreated	0.12	(0.07 - 0.20)	0.11	(0.06 - 0.19)	0.11	(0.06 - 0.21)	0.03	(0.02 - 0.06)	0.03	(0.01 - 0.06)	0.03	(0.01 - 0.06)
	Fungicide	0.05	(0.03 - 0.07)	0.05	(0.03 - 0.08)	0.05	(0.03 - 0.08)	0.12	(0.07 - 0.19)	0.12	(0.06 - 0.23)	0.12	(0.06 - 0.23)
	0.25 mg clothianidin/kernel + fungicide	0.14	(0.09 - 0.24)	0.12	(0.07 - 0.22)	0.14	(0.08 - 0.25)	0.59	(0.36 - 0.98)	0.59	(0.31 - 1.14)	0.59	(0.31 - 1.14)
	1.25 mg clothianidin/kernel + fungicide	1.73	(1.07 - 2.80)	1.84	(1.03 - 3.27)	1.76	(1.01 - 3.04)	4.32	(2.61 - 7.15)	4.32	(2.24 - 8.33)	4.32	(2.24 - 8.33)
Shoots	Untreated	0.04	(0.02 - 0.06)	0.04	(0.02 - 0.05)	0.04	(0.02 - 0.06)	0.02	(0.01 - 0.03)	0.02	(0.01 - 0.03)	0.02	(0.01 - 0.03)
	Fungicide	0.02	(0.01 - 0.02)	0.02	(0.01 - 0.02)	0.02	(0.01 - 0.02)	0.06	(0.04 - 0.09)	0.06	(0.04 - 0.09)	0.06	(0.04 - 0.09)
	0.25 mg clothianidin/kernel + fungicide	0.06	(0.04 - 0.10)	0.06	(0.04 - 0.09)	0.06	(0.04 - 0.10)	0.40	(0.26 - 0.61)	0.40	(0.25 - 0.63)	0.40	(0.25 - 0.63)
	1.25 mg clothianidin/kernel + fungicide	0.93	(0.60 - 1.46)	0.93	(0.63 - 1.38)	0.93	(0.59 - 1.46)	2.64	(1.73 - 4.02)	2.64	(1.67 - 4.17)	2.64	(1.67 - 4.17)

Table 3: Estimated treatment differences in log mg clothianidin per g of plant tissue in maize with corresponding 95% confidence intervals in roots and shoots by year 20 days after planting for three models: a one-way ANOVA assuming independent observations, a one-way ANOVA mixed model on summary data with block as random-effect variable, and a one-way ANOVA mixed model with year, block, and plot as random-effects variables.

		2014						2015					
		ANOVA		Mixed ANOVA on summary data[1,2]		Mixed ANOVA[1]		ANOVA		Mixed ANOVA on summary data[1,2]		Mixed ANOVA[1]	
Plant tissue	Treatment difference[3]	Est (95%-CI)	p-value	Est (95%-CI)	p-value	Est (95%-CI)	p-value	Est (95%-CI)	p-value			Est (95%-CI)	p-value
Roots	Untreated vs Fungicide	0.90 (0.19 - 1.60)	0.0176	0.83 (0.02 - 1.65)	0.0451	0.89 (0.09 - 1.69)	0.0290	-1.29 (-2.07 - -0.52)	0.0029	-1.43 (-2.18 - -0-68)	0.0002	-1.43 (-2.18 - -0-68)	0.0002
	Untreated vs 0.25 mg clothianidin/kernel	-0.21 (-0.94 - 0.53)	0.5861	-0.13 (-0.95 - 0.69)	0.7556	-0.19 (-1.02 - 0.64)	0.6531	-2.91 (-3.68 - -2.14)	<0.0001	-3.05 (-3.80 - -2.30)	<0.0001	-3.05 (-3.80 - -2.30)	<0.0001
	Untreated vs 1.25 mg clothianidin/kernel + fungicide	-2.70 (-3.43 - -1.98)	<0.0001	-2.83 (-3.64 - -2.01)	<0.0001	-2.73 (-3.54 - -1.92)	<0.0001	-4.90 (-5.67 - -4.13)	<0.0001	-5.04 (-5.79 - -4.29)	<0.0001	-5.04 (-5.79 - -4.29)	<0.0001
	0.25 mg clothianidin/kernel + fungicide vs Fungicide	1.10 (0.42 - 1.79)	0.0032	0.96 (0.15 - 1.78)	0.0206	1.08 (0.30 - 1.86)	0.0068	1.61 (0.90 - 2.33)	0.0001	1.61 (0.93 - 2.30)	<0.0001	1.61 (0.93 - 2.30)	<0.0001
	0.25 vs 1.25 mg clothianidin/kernel + fungicide	-2.50 (-3.20 - -1.80)	<0.0001	-2.70 (-3.51 - -1.88)	<0.0001	-2.54 (-3.33 - -1.74)	<0.0001	-1.99 (-2.70 - -1.28)	<0.0001	-1.99 (-2.67 - -1.31)	<0.0001	-1.99 (-2.67 - -1.31)	<0.0001
	1.25 mg clothianidin/kernel + fungicide vs Fungicide	3.60 (2.93 - 4.27)	<0.0001	3.66 (2.85 - 4.48)	<0.0001	3.62 (2.85 - 4.38)	<0.0001	3.61 (2.89 - 4.32)	<0.0001	3.61 (2.92 - 4.29)	<0.0001	3.61 (2.92 - 4.29)	<0.0001
Shoots	Untreated vs Fungicide	0.83 (0.20 - 1.46)	0.0137	0.83 (0.28 - 1.38)	0.0032	0.83 (0.20 - 1.46)	0.0120	-1.05 (-1.70 - -0.41)	0.0036	-1.05 (-1.75 - -0.35)	0.0031	-1.05 (-1.75 - -0.35)	0.0031
	Untreated vs 0.25 mg clothianidin/kernel + fungicide	-0.60 (-1.23 - 0.04)	0.0709	-0.60 (-1.15 - -0.05)	0.0336	-0.60 (-1.23 - 0.04)	0.0642	-3.00 (-3.65 - -2.36)	<0.0001	-3.00 (-3.70 - -2.30)	<0.0001	-3.00 (-3.70 - -2.30)	<0.0001
	Untreated vs 1.25 mg clothianidin/kernel + fungicide	-3.28 (-3.91 - -2.65)	<0.0001	-3.28 (-3.83 - -2.73)	<0.0001	-3.28 (-3.91 - -2.65)	<0.0001	-4.89 (-5.54 - -4.25)	<0.0001	-4.89 (-5.59 - -4.19)	<0.0001	-4.89 (-5.59 - -4.19)	<0.0001
	0.25 mg clothianidin/kernel + fungicide vs Fungicide	1.43 (0.79 -2.06)	<0.0001	1.43 (0.87 - 1.98)	<0.0001	1.43 (0.79 - 2.06)	<0.0001	1.95 (1.35 - 2.54)	<0.0001	1.95 (1.30- 2.59)	<0.0001	1.95 (1.30 - 2.59)	<0.0001
	0.25 vs 1.25 mg clothianidin/kernel + fungicide	-2.69 (-3.32 - -2.05)	<0.0001	-2.69 (-3.24 - -2.14)	<0.0001	-2.69 (-3.32 - -2.05)	<0.0001	-1.89 (-2.49 - -1.29)	<0.0001	-1.89 (-2.54 - -1.24)	<0.0001	-1.89 (-2.54 - -1.24)	<0.0001
	1.25 mg clothianidin/kernel + fungicide vs Fungicide	4.11 (3.48 - 4.74)	<0.0001	4.11 (3.56 - 4.66)	<0.0001	4.11 (3.48 - 4.74)	<0.0001	3.84 (3.24 - 4.43)	<0.0001	3.84 (3.19 - 4.48)	<0.0001	3.84 (3.19 - 4.48)	<0.0001

[1] Model estimated with REML.

[2] Model fitted to summary data. Repetitions within each plot were summarized as a mean before analysis.

[3] Treatment difference on log scale

Table 4: Estimated mg clothianidin per g of plant tissue with corresponding 95% confidence intervals in roots and shoots 20 days after planting for three models: a one-way ANOVA assuming independent observations, a one-way ANOVA mixed model on summary data with year and block as random-effects variables, and a one-way ANOVA mixed model with year, block, and plot as random-effects variables.

		ANOVA		Mixed ANOVA on summary data[1,2]		Mixed ANOVA[1]	
Plant tissue	Treatment	Estimate	95%-CI	Estimate	95%-CI	Estimate	95%-CI
Roots	Untreated	0.07	(0.04 - 0.11)	0.07	(0.03 - 0.15)	0.07	(0.03 - 0.15)
	Fungicide	0.07	(0.04 - 0.10)	0.07	(0.04 - 0.16)	0.07	(0.04 - 0.15)
	0.25 mg clothianidin/kernel + fungicide	0.27	(0.17 - 0.42)	0.27	(0.13 - 0.57)	0.28	(0.13 - 0.58)
	1.25 mg clothianidin/kernel + fungicide	2.54	(1.65 - 3.90)	2.82	(1.33 - 5.98)	2.76	(1.32 - 5.77)
Shoots	Untreated	0.03	(0.02 - 0.05)	0.03	(0.01 - 0.08)	0.03	(0.01 - 0.09)
	Fungicide	0.03	(0.02 - 0.04)	0.03	(0.01 - 0.08)	0.03	(0.01 - 0.08)
	0.25 mg clothianidin/kernel + fungicide	0.13	(0.09 - 0.20)	0.16	(0.06 - 0.44)	0.15	(0.05 - 0.41)
	1.25 mg clothianidin/kernel + fungicide	1.41	(0.93 - 2.15)	1.57	(0.57 - 4.30)	1.56	(0.57 - 4.28)

[1] Model estimated with REML.

Table 5: Estimated treatment differences in log mg clothianidin per g of plant tissue in maize with corresponding 95% confidence intervals in roots and shoots 20 days after planting for three models: a one-way ANOVA assuming independent observations, a one-way ANOVA mixed model on summary data with year and block as random-effects variables, and a one-way ANOVA mixed model with year, block, and plot as random-effects variables.

		ANOVA		Mixed ANOVA on summary data[1,2]		Mixed ANOVA[1]	
Plant tissue	Treatment difference[3]	Est (95%-conf.)	p-value	Est (95%-conf.)	p-value	Est (95%-conf.)	p-value
Roots	Untreated vs Fungicide	0.02 (-0.62 - 0.66)	0.9515	-0.11 (-0.93 - 0.71)	0.7904	-0.074 (-0.88 - 0.73)	0.8566
	Untreated vs 0.25 mg clothianidin/kernel	-1.35 (-2.01 - -0.70)	0.0001	-1.40 (-2.22 - -0.58)	0.0008	-1.403 (-2.22 - -0.59)	0.0007
	Untreated vs 1.25 mg clothianidin/kernel + fungicide	-3.60 (-4.25 - -2.96)	<0.0001	-3.74 (-4.56 - -2.93)	<0.0001	-3.702 (-4.51 - -2.89)	<0.0001
	0.25 mg clothianidin/kernel + fungicide vs Fungicide	1.37 (0.76 - 1.98)	<0.0001	1.29 (0.50 - 2.08)	0.0014	1.329 (0.56 - 2.10)	0.0007
	0.25 vs 1.25 mg clothianidin/kernel + fungicide	-2.25 (-2.87 - - 1.64)	<0.0001	-2.34 (-3.13 - -1.55)	<0.0001	-2.299 (-3.07 - -1.52)	<0.0001
	1.25 mg clothianidin/kernel + fungicide vs Fungicide	3.62 (3.02 - 4.22)	<0.0001	3.63 (2.84 - 4.42)	<0.0001	3.628 (2.86 - 4.39)	<0.0001
Shoots	Untreated vs Fungicide	0.11 (-0.50 - 0.72)	0.7160	-0.01 (-0.64 - 0.63)	0.9824	0.11 (-0.51 - 0.73)	0.7307
	Untreated vs 0.25 mg clothianidin/kernel + fungicide	-1.52 (-2.13 - -0.91)	<0.0001	-1.69 (-2.33 - -1.06)	<0.0001	-1.55 (-2.16 - -0.93)	<0.0001
	Untreated vs 1.25 mg clothianidin/kernel + fungicide	-3.89 (-4.50 - -3.28)	<0.0001	-3.98 (-4.61 - -3.35)	<0.0001	-3.88 (-4.50 - -3.27)	<0.0001
	0.25 mg clothianidin/kernel + fungicide vs Fungicide	1.63 (1.04 - 2.23)	<0.0001	1.69 (1.08 - 2.30)	<0.0001	1.65 (1.06 - 2.25)	<0.0001
	0.25 vs 1.25 mg clothianidin/kernel + fungicide	-2.37 (-2.96 - -1.77)	<0.0001	-2.29 (-2.90 - -1.68)	<0.0001	-2.34 (-2.94 - -1.74)	<0.0001
	1.25 mg clothianidin/kernel + fungicide vs Fungicide	4.00 (3.41- 4.60)	<0.0001	3.97 (3.36 - 4.58)	<0.0001	3.99 (3.39- 4.59)	<0.0001

[1] Model estimated with REML.

[2] Model fitted to summary data. Repetitions within each plot were summarized as a mean before analysis.

[3] Treatment difference on log scale

Table 6: Results of simulation study for balanced data. Percentage bias and coverage were averages across seven treatment comparisons to the control. Type-1 error rate were obtained using a single treatment difference, treatment 1 compared to the control. Power were obtained for the treatment difference, treatment 3 compared to the control.   Data were simulated from a randomized complete block design with 10 treatments and 4 blocks with 1) 1 repetition of treatment within block and 5 samples taken within each plot, 2) 5 repetitions of treatment within block and 5 samples taken within each plot, or 3) 1 repetition of treatment within block and 25 samples taken within each plot.

			1 rep within block 5 samples within plot					1 rep within block 25 samples within plot					5 rep within block 5 samples within plot				
Fixed-effects Design	Random-effects design	Est. Proc.	% Bias	Coverage	Width 95% CI	Type 1 error	Power	% Bias	Coverage	Width 95% CI	Type 1 error	Power	% Bias	Coverage	Width 95% CI	Type 1 error	Power
One-way ANOVA	Block and plot	REML	-2.3	0.94	31.10	0.07	0.66	-0.7	0.94	29.60	0.06	0.71	-0.2	0.95	14.03	0.04	>0.99
		ML	-2.3	0.90	26.93	0.10	0.76	-0.7	0.90	25.64	0.11	0.79	-0.2	0.94	13.70	0.05	>0.99
	Block	REML	-2.3	0.66	15.58	0.30	0.92	-0.7	0.34	6.88	0.65	0.97	-0.2	0.71	7.52	0.28	>0.99
		ML	-2.3	0.65	15.21	0.32	0.93	-0.7	0.34	6.85	0.65	0.97	-0.2	0.70	7.49	0.28	>0.99
	Plot	REML	-2.3	0.98	39.23	0.02	0.48	-0.7	0.98	37.95	0.02	0.50	-0.2	0.98	16.96	0.02	>0.99
		ML	-2.3	0.95	33.97	0.05	0.58	-0.7	0.95	32.87	0.05	0.62	-0.2	0.97	16.53	0.03	>0.99
One-way ANOVA			-2.3	0.75	18.58	0.23	0.89	-0.7	0.41	8.20	0.58	0.97	-0.2	0.77	8.64	0.21	>0.99
One-way ANOVA[1]			-2.3	0.98	39.23	0.02	0.45	-0.7	0.98	37.95	0.02	0.47	-0.2	0.98	16.96	0.02	>0.99
	Block	REML	-2.3	0.94	31.10	0.07	0.66	-0.7	0.94	29.60	0.06	0.71	-0.2	0.95	14.03	0.04	>0.99
t-tests			-2.3	0.75	18.99	0.24	0.88	-0.7	0.40	8.16	0.59	0.97	-0.2	0.77	8.66	0.22	>0.99

[1] Model fitted to summary data. Repetitions within each plot were summarized as a mean before analysis.

Table 7: Results of simulation study for unbalanced data. Results of simulation study for balanced data. Percentage bias and coverage were averages across seven treatment comparisons to the control. Type-1 error rate were obtained using a single treatment difference, treatment 1 compared to the control. Power were obtained for the treatment difference, treatment 3 compared to the control. Data were simulated from a randomized complete block design with 10 treatments and 4 blocks with 1) 1 repetition of treatment within block and 5 samples taken within each plot, 2) 5 repetitions of treatment within block and 5 samples taken within each plot, or 3) 1 repetition of treatment within block and 25 samples taken within each plot. In block 3 all samples for treatment 1 and 3 were considered lost, while all but one samples for the control treatment were considered lost in block 4.

			1 rep within block 5 samples within plot					1 rep within block 25 samples within plot					5 rep within block 5 samples within plot				
Fixed-effects design	Random-effects design	Est. Proc.	% Bias	Coverage	Width 95% CI	Type 1 error	Power	% Bias	Coverage	Width 95% CI	Type 1 error	Power	% Bias	Coverage	Width 95% CI	Type 1 error	Power
One-way ANOVA	Block and plot	REML	-2.1	0.94	32.54	0.05	0.59	0.5	0.94	31.10	0.05	0.64	0.8	0.95	15.39	0.06	>0.99
		ML	-2.0	0.90	28.00	0.09	0.68	0.6	0.89	26.77	0.10	0.72	0.8	0.94	15.00	0.07	>0.99
		K-R[1]	-2.1	0.95	34.10	0.04	0.55	0.5	0.95	32.59	0.04	0.59	0.8	0.95	15.49	0.06	>0.99
	Block	REML	-1.7	0.67	16.86	0.32	0.85	1.4	0.34	7.56	0.65	0.96	0.7	0.71	8.33	0.29	>0.99
		ML	-1.7	0.65	16.44	0.33	0.86	1.4	0.34	7.52	0.65	0.96	0.7	0.70	8.29	0.30	>0.99
	Plot	REML	-1.9	0.98	40.69	0.03	0.41	0.4	0.98	39.43	0.02	0.45	1.6	0.95	18.38	0.09	0.96
		ML	-1.8	0.95	35.01	0.05	0.55	0.5	0.95	33.94	0.05	0.57	1.6	0.94	17.89	0.10	0.96
One-way ANOVA			-1.7	0.73	19.96	0.27	0.81	1.2	0.37	8.91	0.63	0.94	1.7	0.67	9.45	0.40	>0.99
One-way ANOVA[2]			-2.4	0.97	40.27	0.03	0.38	-0.2	0.97	38.93	0.03	0.42	1.6	0.95	18.33	0.08	0.96
	Block	REML	-2.5	0.93	32.07	0.06	0.61	-0.3	0.93	30.54	0.07	0.64	0.8	0.94	15.31	0.07	>0.99
t-tests			-1.7	0.73	20.37	0.28	0.80	1.2	0.37	8.72	0.64	0.94	1.7	0.66	9.42	0.40	>0.99

[1] Mixed ANOVA with block and plot as random effect estimated using REML and degrees of freedom calculated using the Kenward-Roger (K-R) approximation.

[2] Model fitted to summary data. Repetitions within each plot were summarized as a mean before analysis.

Table S2: Results of simulation study for balanced data. Mean percentage bias and coverage were reported across all 10 estimated means estimated by different modelling strategies. Data was simulated from a complete randomized block design with 10 treatments and 4 blocks with 1) 1 repetition of treatment within block and 5 samples taken within each plot, 2) 5 repetitions of treatment within block and 5 samples taken within each plot, or 3) 1 repetition of treatment within block and 25 samples taken within each plot.

			1 rep. within block 5 samples within plot			1 rep. within block 25 samples within plot			5 rep. within block 5 samples within plot		
Model	Random effects	Est. proc	% Bias	Coverage	Length 95% CI	% Bias	Coverage	Length 95% CI	% Bias	Coverage	Length 95% CI
One-way ANOVA	Block and plot	REML	0.1	0.94	27.74	0.1	0.94	26.84	-0.1	0.92	19.14
		ML	0.1	0.90	24.02	0.1	0.90	23.24	-0.1	0.89	17.16
	Block	REML	0.1	0.82	20.61	0.1	0.76	17.85	-0.1	0.87	17.18
		ML	0.1	0.79	18.51	0.1	0.71	15.65	-0.1	0.83	15.12
	Plot	REML	0.1	0.94	27.74	0.1	0.94	26.84	-0.1	0.78	11.99
		ML	0.1	0.90	24.02	0.1	0.90	23.24	-0.1	0.76	11.69
ANOVA			0.1	0.64	13.14	0.1	0.33	5.80	-0.1	0.47	6.11
ANOVA[1]			0.1	0.94	27.74	0.1	0.94	26.84	-0.1	0.78	11.99
	Block		0.1	0.94	27.74	0.1	0.94	26.84	-0.1	0.92	19.14

[1] Model fitted to summary measures. Repetitions within each plot were summarized as a mean before analysis

Table S3: Results of simulation study for unbalanced data. Mean percentage bias and coverage were reported across all 10 estimated means estimated by different modelling strategies. Data was simulated from a complete randomized block design with 10 treatments and 4 blocks with 1) 1 repetition of treatment within block and 5 samples taken within each plot, 2) 5 repetitions of treatment within block and 5 samples taken within each plot, or 3) 1 repetition of treatment within block and 25 samples taken within each plot. In block 3 all samples for treatment 1 and 3 were considered lost, while all but one samples for the control treatment were considered lost in block 4.

			1 rep. within block			1 rep. within block			5 rep. within block		
			5 samples within plot			25 samples within plot			5 samples within plot		
Model	Random effects	Est. proc	% Bias	Coverage	Length 95% CI	% Bias	Coverage	Length 95% CI	% Bias	Coverage	Length 95% CI
One-way ANOVA	Block and plot	REML	0.1	0.94	28.41	0.1	0.94	27.45	-0.1	0.92	19.45
		ML	0.1	0.89	24.43	0.1	0.89	23.61	-0.1	0.89	17.45
		K-R[1]	0.1	0.96	31.62	0.1	0.95	30.73	-0.1	0.95	25.11
	Block	REML	0.1	0.82	20.93	0.1	0.75	18.01	-0.1	0.86	17.34
		ML	0.1	0.78	18.81	0.1	0.70	15.80	-0.1	0.83	15.27
	Plot	REML	0.1	0.94	28.65	0.1	0.94	27.70	-0.1	0.77	12.48
		ML	0.1	0.89	24.58	0.1	0.89	23.76	-0.1	0.76	12.14
ANOVA			0.1	0.64	13.60	0.1	0.32	6.00	-0.1	0.47	6.37
ANOVA[2]			0.1	0.94	28.70	0.1	0.94	27.74	-0.1	0.77	12.48
	Block	REML	0.1	0.94	28.46	0.1	0.94	27.50	-0.1	0.92	19.45

[1] Mixed ANOVA with block and plot as random effect estimated using REML and degrees of freedom calculated using the Kenward-Roger (K-R) approximation.

[2] Model fitted to summary data. Repetitions within each plot were summarized as a mean before analysis.