Annex II

Analysis of data of Annex I with SAS/STAT® 15.1.

Final germination percentages (FGP) from Gianinetti et al. (2018) are analysed (see Annex I for the whole dataset) as an example. It should be noted that germination data can be arranged as either binary data (individual Bernoulli outcomes, with every response being either an event or a nonevent), that is, the response of every individual seed is kept as a separate record (i.e., a line in the dataset), or, more frequently, data can be arranged as means of the binomial distribution across clusters (e.g., plates), that is, means of a dependent (response) variable that represent proportions from aggregated binary responses (in this respect, the Bernoulli distribution is a special case of the binomial distribution where a single trial is conducted for each replicate). Binomial data can even be aggregated into a single mean across all the seeds tested for a given condition (that is, their clustering into plates may be ignored and, for every mean, the overall, across-seeds, distribution of the proportion of successes is considered as a single binomial aggregation of Bernoulli trials). In any case the error distribution is binomial, but slightly different statistical approaches suit to the two arrangements of data (i.e. binary or aggregate binomial). Besides, binomial data can be recorded as either counts (number of germinated seeds for each cluster), proportions (number of germinated seeds for each cluster divided by the total number of live seeds in the cluster) or percentages (proportions times 100). In the exemplary dataset (Annex I), each plate is an aggregate of 20 Bernoulli trials. The comments to the statistical analysis presented here are based on the SAS/STAT® 15.1 User's Guide (2018), Littell et al. (2006), Stroup (2015) and Stroup et al. (2018), to which the reader should refer for a more in-depth exposition of the matter.

Some testing of the error distributions is initially done. Counts are used in these preliminary tests, but proportions/percentages work as well. The file with the data is referred to as 'reffile' and it does not need to be specified in the procedure statement if it is the current input (the last file opened in the current SAS session).

First, homogeneity of variances is tested (by means of both Levene's and Brown-Forsythe tests).

Note that, as the GLM procedure allows homogeneity of variance testing for simple one-way models only, the model must include the highest interaction of fixed-effects alone (that is, the fixed effect of highest degree, wherein the error distribution can be considered around means; in other words, the cross-factor cell means displayed in the table of results in the main text).

```
ODS graphics on; /*This activates the graphical function, if it is not enabled by default*/
proc GLM data=reffile plot=diagnostics;
class grain trt;
model germ = grain*trt;
means grain*trt / hovtest=Levene hovtest=bf;
output out=resids r=residual; /*This provides the input file of residuals for the subsequent analysis*/
run;
ODS graphics off; /*This closes the graphical function, if necessary*/
```

RESULTS (excerpts):

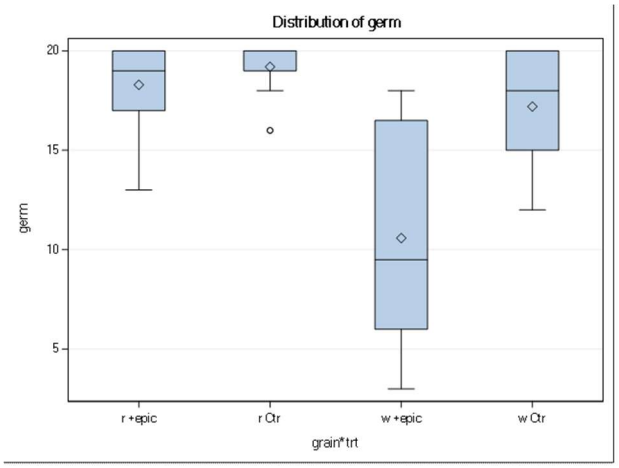

Figure 1.

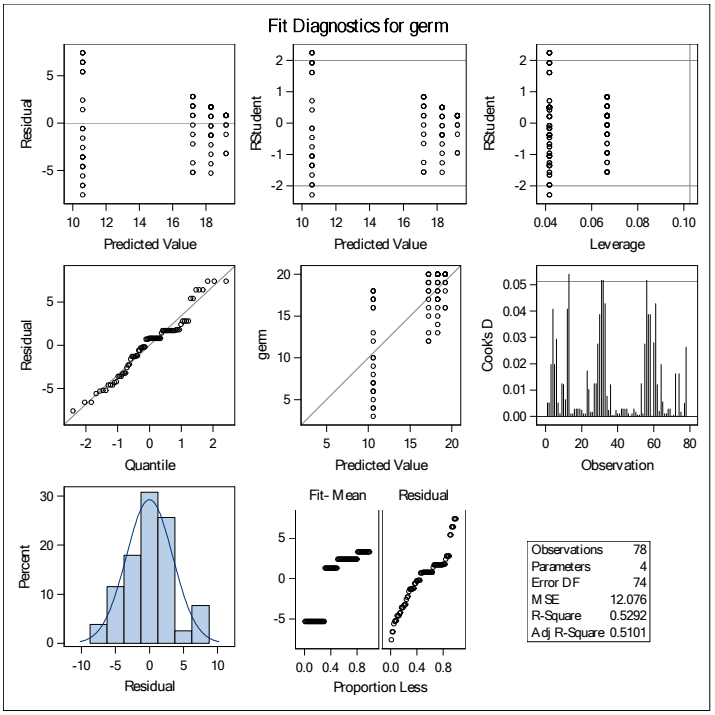

Figure 2.

Table 1.

Levene's Test for Homogeneity of germ Variance ANOVA of Squared Deviations from Group Means					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
grain*trt	3	8389.2	2796.4	19.04	<.0001
Error	74	10868.6	146.9		

Table 2.

Brown and Forsythe's Test for Homogeneity of germ Variance ANOVA of Absolute Deviations from Group Medians					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
grain*trt	3	166.7	55.5694	13.76	<.0001
Error	74	298.8	4.0377		

The spread of the error distributions is clearly diverse for the different group (cell) means (Figures 1 and 2), and homogeneity tests (Tables 1 and 2) confirm that, as expected, the assumption of equal variances is untenable. The plot of Cook's D statistic indicates that observations in all the groups are similarly influential on the analysis of group means. The Fit-Mean/Residual spread (Proportion less) plot (lower central plot) shows that the spread in the residuals is wider that the spread in the centred fit. This indicates that the fitted model accounts only for part of the variation in the data, as expected. Residual data display a staircase pattern with plateaus and gaps (in the Residual vs Quantile, and Proportion less plots) because they are discrete (germinated seeds out of 20 tested seeds).

The normality of errors can then be tested (based on the residuals obtained from the previous analysis) with:

```
proc UNIVARIATE data=resids normal;
var residual;
run;
```

RESULTS (excerpt):

Table 3.

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.960317	**Pr < W**	0.0159
Kolmogorov-Smirnov	**D**	0.13364	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.226104	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	1.255363	**Pr > A-Sq**	<0.0050

Table 3 indicates that errors of germination data are not normal. This was also evident from Figure 1, though, in the 'Residual' plot on the lower-left of Figure 2, the overall residuals appear to be approximately normal. This apparent incongruence occurs because a normal distribution of the overall residuals does not necessarily ensure that each individual mean error is normal as well (individual slightly skewed error variances can be hidden by overlying the different heterogeneous error variances, thus that the overall residuals deceptively appear to be normal). In the present case, data for 'r' grain (first two samples in the boxplot of Figure 1) are close to the highest possible value (20 seeds germinated out of 20 seeds sown) and their error distributions are clearly lopsided.

Given the binomial nature of germination data (with non-normal error distributions and heterogenous variances), GzLMM are specifically indicated for their analysis (Stroup, 2015). The GLIMMIX procedure has been used for this purpose. Data were first analysed considering all the factors as fixed.

```
proc GLIMMIX data=reffile;
class grain trt cv;
model germ/n = grain trt grain*trt cv(grain);
run;
```

The events/trials syntax ('germ/n', in this case) in the 'model' statement indicates that the response variable is binomial, each observation is a subpopulation (a cluster of responses) and the response means are sample proportions of Bernoulli outcomes: the first variable, 'events', is the number of germinated seeds in each plate, and the value of the second variable, 'trials', gives the total number of seeds in each plate (i.e., the number of Bernoulli trials). Because of the events/trials syntax, the GLIMMIX procedure defaults to the binomial distribution and to its default link, the logit.

As there are no random effects, thus that Maximum Likelihood is used as estimation technique.

RESULTS (excerpts):

The iterative fitting process converges to a solution of the model, whose fit is shown in Table 4. The ANOVA table is given in Table 5.

Table 4.

Fit Statistics	
-2 Log Likelihood	401.92
AIC (smaller is better)	417.92
AICC (smaller is better)	420.00
BIC (smaller is better)	436.77
CAIC (smaller is better)	444.77
HQIC (smaller is better)	425.46
Pearson Chi-Square	243.92
Pearson Chi-Square / DF	3.48

Table 5.

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
grain	1	70	80.51	<.0001
trt	1	70	43.86	<.0001
grain*trt	1	70	5.89	0.0178
cv(grain)	4	70	10.72	<.0001

Table 4 refers the model fitting statistics. The –2 Log Likelihood value is specifically useful for comparing models that are syntactically nested, that is, models that are hierarchically contained, one in the other, with

respect to fixed effects and/or the covariance parameters. Table 5 shows that all the effects are significant ($P \leq 0.05$).

An important parameter is given at the end of Table 4: the Pearson Chi-Square / DF. This ratio measures the residual variation in the fit, once the binomial variance has been accounted for. For data with a binomial (or Poisson) error distribution, such ratio is useful to assess overdispersion of the model. In practice, overdispersion occurs when Pearson Chi-Square / DF is notably higher than 1. Thus, the Pearson Chi-Square / DF is the overdispersion parameter, which estimates the scale parameter, by which the binomial variance must be multiplied to obtain the actual variance observed in the data, which is inclusive of the variance that is not already explained by the binomial variance. In other words, overdispersion occurs when data appear more dispersed than is expected for distributions that have mean-variance relationship (for which the expectation is Pearson Chi-Square / DF = 1). Overdispersion, in fact, can be measured only if the variability a model can capture is constrained because of a functional relationship between mean and variance, whereas in a model for Gaussian data overdispersion cannot be assessed. Specifically, in the present case of germination data, variability exceeds what predicted by the binomial distribution (Pearson Chi-Square / DF = 3.48; Table 4). This might happen because the observed distribution is a mixture of different distributions, wherein the binomial error distribution is superimposed on another variance distribution due to some unmodeled variable (note that, anyway, the scale parameter is multiplicative with respect to the binomial variance, that is, even the unmodeled variance is automatically managed as heterogeneous in relation to the mean). For germination data, this can be due to a misspecified, or incomplete, model; to heterogeneity in the observational units; or to an incorrect specification of the covariance structure. In the first case, often some source of random variation is not accounted for, which equates to a positive correlation among corresponding observations. Correlation within clusters of observations is linked to overdispersion because the correlation is due to some unmodeled factor that affects the observational units and causes clustering of responses; the effect of such factor therefore adds to the theoretical random distribution of residuals. Blocks are modelled arrangements that, if properly identified, superimpose on the clusters affected by the unmodeled factor and are therefore able to capture the relative additional variance it introduces, so that it can be statistically accounted for. Random factors usually have a Gaussian distribution and are therefore modelled as such (Littell et al., 2006). The addition of this further, Gaussian, variability in the residuals can produce a deviation from the expected distribution of residuals, which are modelled as binomial, thereby negatively affecting the power of the statistical tests (which are based on the residual variance) and, therefore, the resulting significance. Hence, type I errors are inflated by overdispersion, which makes *F* tests on the remaining effects too conservative.

Prior to considering overdispersion, however, the probit link function is introduced (it is therefore explicitly specified) in place of the logit, because the former is more suitable to threshold models, which represent the best theoretical background for modelling germination/dormancy data (Roberts, 1961; Bradford, 1990; Gianinetti and Cohn, 2007; Hardegree et al., 2015). The two functions, anyway, are very close.

```
proc GLIMMIX;
class grain trt cv;
model germ/n = grain trt grain*trt cv(grain) / link=probit;
run;
```

RESULTS (excerpts):

The model fit is shown in Table 6 and the ANOVA table is given in Table 7.

Table 6.

Fit Statistics	
-2 Log Likelihood	405.77
AIC (smaller is better)	421.77
AICC (smaller is better)	423.86
BIC (smaller is better)	440.62
CAIC (smaller is better)	448.62
HQIC (smaller is better)	429.32
Pearson Chi-Square	243.82
Pearson Chi-Square / DF	3.48

Table 7.

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
grain	1	70	108.81	<.0001
trt	1	70	54.11	<.0001
grain*trt	1	70	9.77	0.0026
cv(grain)	4	70	10.27	<.0001

The fit is slightly worsened (Table 6), but the significance of the interaction has improved (Table 7).

Overdispersion is still a problem (Table 6), but before considering overdispersion, however, another, even more important, modification must be made to the model: as the target of the experiment was to ascertain whether the grain colour type has a significant effect on the germination response in presence of a fungal infection, it is clear that the intent of the analysis was to generalize the findings to any cultivar with either red or white grain, not simply to test the specific cultivars used in the experiment. This kind of generalization implies that the main genetic effect is the grain colour and the specific genetic backgrounds of the used cultivars represent random effects. In other words, in the experimental design, cultivars are considered as subjects possessing a common chief feature whose effect was under investigation. In general, nested effects usually represent random effects within a fixed-effects structure.

The shift of 'cv' from being a fixed to a random factor may, however, have a heavy effect on significances (Quinn and Keough, 2002; Gbur et al., 2012), because the *F* test of the 'grain' factor effect is thereby based on the effect of the nested factor, 'cv(grain)', even though only the degrees of freedom are changed, because in mixed models all the *F* ratios use the same denominator, namely the penalized residual sum of squares divided by the REML degrees of freedom. In fact, the degrees of freedom in the denominator of the *F* ratio drop from the remainder degrees to the degrees corresponding to the 'cv(grain)' effect (DF=4, as can be seen in Table 7). The updated model is:

```
proc GLIMMIX;
class grain trt cv;
model germ/n = grain trt grain*trt / link=probit;
random cv(grain);
run;
```

Although every cultivar is uniquely identified across the whole experiment, nesting is made explicit because its specification is necessary to trigger the use of this random effect in the denominator of the *F* ratio for testing the significance of the 'grain' factor: 'cv(grain)' is a random effect that syntactically contains 'grain' and it is therefore automatically used to compute the denominator degrees of freedom .

As a random factor is present, the Residual Pseudo-Likelihood (REPL) is now used as estimation technique, and the model that now includes a random factor (i.e., a so-called G-side random effect, where G is the matrix used to model random factors) is called a conditional model.

RESULTS (excerpts):

Table 8.

Fit Statistics	
-2 Res Log Pseudo-Likelihood	258.68
Generalized Chi-Square	245.49
Gener. Chi-Square / DF	3.32

Table 9.

Covariance Parameter Estimates		
Cov Parm	**Estimate**	**Standard Error**
cv(grain)	0.1067	0.08542

Table 10.

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
grain	1	4	13.43	0.0215
trt	1	70	53.95	<.0001
grain*trt	1	70	9.96	0.0024

The model fit statistics (Table 8) have changed (because the introduction of a random effect elicited the use of mixed model statistics) and therefore they cannot be compared with those of the previous model. In particular, the Generalized Chi-Square / DF (Table 8) has replaced the Pearson Chi-Square / DF as a measure of overdispersion. It ought to be noted that the Generalized Chi-Square / DF is a rough, albeit dimly indicative, measure of overdispersion (whereas the Pearson Chi-Square / DF is an exact measure of overdispersion). Table 9 shows the estimated variance (0.1067 on the linked scale, here the probit scale) that the 'cv(grain)' random effect contributes to the total model variance (random effects are usually considered relevant only in relation to the random variance that they can account for, thus that their modelling can improve the statistical power of the analysis). It can be noted that the estimate of the variance of 'cv(grain)' is of similar size as its standard error, suggesting this variance might not to be different from zero. Table 10 shows that, as expected, the significance of the 'grain' effect has been strongly reduced (the *P* of the data given the null hypothesis, i.e. no effect, has noticeably increased), though it still supports a significant role of the grain colour on the germination response. The generalization of the model comes at a price.

As previously said, overdispersion occurs when Pearson Chi-Square / DF is notably higher than 1. In this case, the Generalized Chi-Square / DF parameter is 3.32, not much departing from the value of 3.48 found for the Pearson Chi-Square / DF overdispersion parameter prior to introducing the 'cv(grain)' random factor. Thus, there is still indication of ovedispersion. As said, a first reason for this could be an incomplete specification of the model; in fact, an effect is missing, namely, the plate effect. Plates represent replicated blocks (clusters) whose averaged value is used in the model. The binomial variance is then modelled on this average, but the replicated plates represent a random effect with a Gaussian distribution around the cell mean. This between plates variability contributes to the dispersion of data and must be considered into the model, otherwise it manifests itself as overdispersion.

It may be noted that, whereas Gaussian models inherently account for unmodeled residual variance components by performing an estimate of the overall residual variance, in the case of binomial models, once the mean has been estimated also its variance is known (Stroup, 2015). Binomial models, therefore, cannot automatically accommodate additional, unmodeled variance components. This leads to overdispersion, that is, the model fails to account for all the variability in the data (Stroup, 2015; Stroup et al., 2018). On the other hand, the fact that overdispersion can be assessed in binomial (and Poisson) models is a convenient way to ascertain the capability of the model to account for the variability existent in the data, and therefore its fit.

The 'plate' random effect is then introduced (in the class statement too). Plates represent replicated (aggregate) measures of the cell means predicted within combinations of the fixed effects; differences between plate values represent, therefore, random fluctuations around the seed population mean response. The plate effects are thus modelled as Gaussian random deviations from the predicted response means of the cells relative to the interaction of highest level present in the model, i.e. 'grain*trt'. Note that, in GzLMM, variances of random factors are estimated on the linked scale. This is because they are modelled as random deviations (hence, with mean zero) from the intercept of the linear regression, and the model is assumed to be linear on the scale determined by the link function. Accordingly, the variance of each random factor is assumed to be Gaussian and identical throughout the levels of the fixed factors model, i.e. for all the means, on the linear (linked) scale. On the scale of the data, however, the random response deviations among plates (clusters) are constrained to the inherent binomial distribution, that is, they are smaller toward the lower and upper boundaries and become skewed as the latter are approached. The model becomes:

```
proc GLIMMIX /*method=Laplace*/; /*the 'method=Laplace' option provides more exact fit statistics*/
class grain trt cv plate;
model germ/n = grain trt grain*trt / link=probit;
random cv(grain);
random plate / group=grain*trt;
```

```
covtest homogeneity; /*This tests homogeneity of variance parameters across groups*/
covtest zeroG; /*It tests whether (G-side) random effects have a significant effect on the model*/
covtest 0; /*It tests whether 'cv(grain)' variance is significantly different from zero*/
run;
```

The 'group=' option is used here to specify the factor combination within which plates represent random fluctuations around the cell means because it allows to evaluate whether variances parameters vary with respect to what predicted by the superimposition of a Gaussian random variability among plates on the binomial distribution. This evaluation is done by means of 'covtest' statements.

Note that every plate is uniquely identified across the whole experiment, thus that no confusion between diverse plates can occur, and, therefore, it is not necessary to specify that they are nested within the other factors. In this kind of hierarchical (aka multilevel) models, there is one intercept (random deviation) for each (within-factor) level of every hierarchical-level factor, and hierarchically-structured model levels correspond to the superimposed Gaussian distributions of the corresponding random effects (assumed to have mean zero and independent variances).

The 'covtest 0' statement actually tests whether the first parameter in the Covariance Parameter Estimates table, which in the present case indeed is the 'cv(grain)' variance, is significantly different from zero.

RESULTS (excerpts):

Table 11.

Fit Statistics	
-2 Res Log Pseudo-Likelihood	179.19
Generalized Chi-Square	70.54
Gener. Chi-Square / DF	0.95

Table 12.

Covariance Parameter Estimates			
Cov Parm	**Group**	**Estimate**	**Standard Error**
cv(grain)		0.1329	0.1319
plate	grain*trt r +epic	0.2978	0.1519
plate	grain*trt r Ctr	0.2477	0.2024
plate	grain*trt w +epic	0.2665	0.1291
plate	grain*trt w Ctr	0.7109	0.3887

Table 13.

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
grain	1	4	9.19	0.0387
trt	1	70	18.26	<.0001
grain*trt	1	70	4.38	0.0399

Table 14.

Tests of Covariance Parameters Based on the Residual Pseudo-Likelihood					
Label	**DF**	**-2 Res Log P-Like**	**ChiSq**	**Pr > ChiSq**	**Note**
Homogeneity	3	180.93	1.75	0.6270	DF
No G-side effects	5	292.01	112.82	<.0001	MI
Parameter list	1	182.84	3.65	0.0280	MI

DF: P-value based on a chi-square with DF degrees of freedom.
MI: P-value based on a mixture of chi-squares.

Table 11 would suggest that fitting (specifically, -2 Res Log Pseudo-Likelihood) has improved by taking into account the between-plates effect. Unfortunately, pseudo-likelihoods cannot be compared across different models, even if the models are syntactically nested with respect to fixed and/or random factors. Nevertheless, exact fit statistics can be obtained by using the Laplace approximation method (or other integral approximation, such as Gaussian quadrature, which would, however, require a re-arrangement of the syntax), with comparable results in the present instance (not shown; this will be discussed later). Generalized Chi-Square / DF (which, as already seen, is the overdispersion parameter when REPL is used) is now just below 1, corroborating that overdispersion was due to the random effect of plates. Table 12 confirms that the estimate of the variance of 'cv(grain)' (on the probit scale) is of a very similar size as its standard error. Even if it should not be significantly different from zero, 'cv(grain)' ought to be retained in the model because it is a noticeable factor of the experimental design (Stroup, 2015), and it is also necessary to explicitly provide a correct *F* test for the grain effect (assuming cultivar as a random factor). Table 12 also shows estimates for variance/covariance parameters for each level of the 'grain*trt' factor (on the probit scale). It deserves to be mentioned that, as 'grain*trt' is a fixed factor whose response error terms are modelled according to a binomial distribution, the estimates of Table 12 represent between-plates Gaussian variances that superimpose onto (i.e., overlap and, if data are not underdispersed, include) the theoretical binomial variances inherent to the fixed means. Table 13 indicates significant ($P \leq 0.05$) effects for all the factors. Although Table 12 also shows that estimates of the between-plates random variances within the different levels of the 'grain*trt' factor appear similar but for the variance of the untreated white grain cultivars ('w Ctr'), Table 14 indicates that homogeneity of the between-plates Gaussian variances cannot be rejected. The standard error of the between-plates variance of the untreated white-grained cultivars ('w Ctr') is, indeed, higher too (Table 12), and its contrast with the other between-plates variances has thereby lost power. These variances had already been shown to be heterogenous on the data scale (see Tables 1 and 2), but such heterogeneity is inherent to the binomial distribution, and it is therefore accounted for by the link

function. In fact, on the probit scale (the presently chosen link function, on which all random-factor variances are evaluated) they no longer result heterogeneous. Table 14, finally, suggests that the inclusion of (G-side) random effects in the model is highly significant for the fit of the model to the data, and, specifically, that the 'cv(grain)' variance is significantly different from zero. It should be mentioned that tests of significance for covariance parameters, requested with 'covtest' statements, are based on likelihood ratios, which are exact only if true likelihood, not pseudo-likelihood, is used (here, Residual Pseudo-Likelihood was utilized; Table 14). These tests provide, nonetheless, useful indicative clues about Chi-square probabilities. Again, exact tests can be obtained by using the Laplace approximation method, with comparable results in the present instance, though the 'cv(grain)' variance would thereby be demonstrated to be different from zero only with P = 0.1269 (not shown). This random effect is anyway retained into the model because it is a structural element of the experimental design.

Now that the model has been fully specified, residuals can be displayed to diagnostic fitting problems:

```
proc GLIMMIX plots=(residualpanel(ilink marginal) studentpanel(conditional));
class grain trt cv plate;
model germ/n = grain trt grain*trt / link=probit;
random cv(grain);
random plate / group=grain*trt;
run;
```

RESULTS (excerpts):

Figure 3.

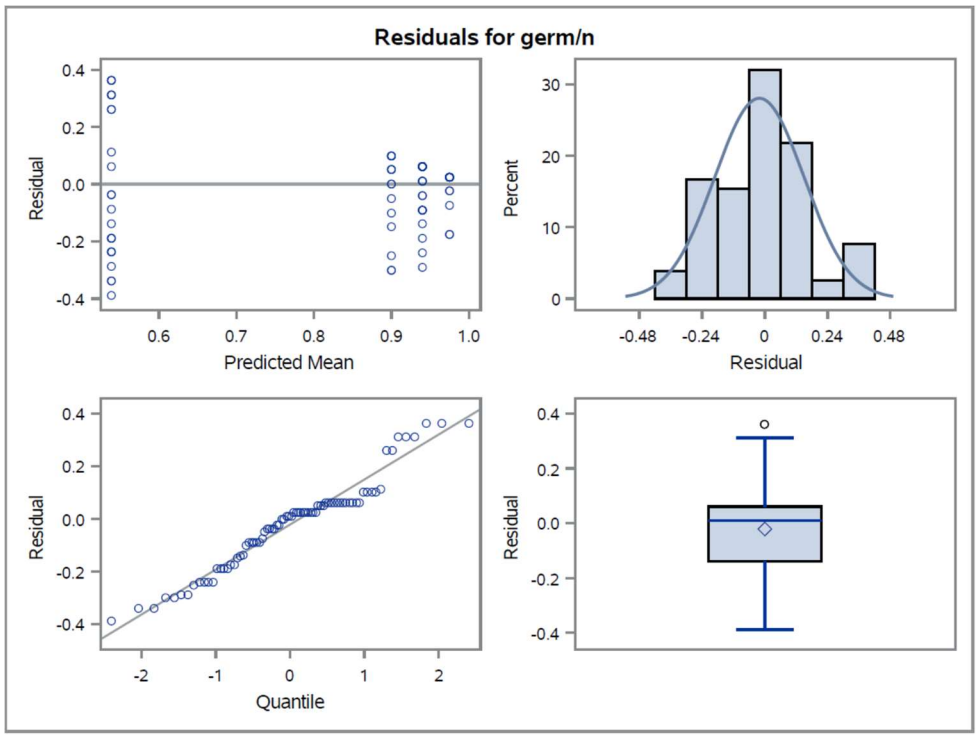

Figure 3 displays the GLIMMIX graphical analysis of marginal (i.e. deviations from the overall mean) raw residuals on the data scale ('ilink' indicates the inverse linked scale, i.e. the original data scale), corresponding to the residual plot of the fit diagnostics panel initially obtained with the GLM procedure (Figure 2). The upper-left plot of the panel in Figure 3 shows the error distributions for every (predicted) mean of germination proportion (there are four distinct values for the predicted means, corresponding to the four

combinations of the 'grain' and 'trt' fixed factors). As already remarked for Figure 2, error distributions sharply decline from about 0.5 to near 1 mean values. Table 14 demonstrates that these residuals essentially correspond to what expected for binomial data, as no additional heterogeneity of the between-plates Gaussian variances was detected across these means (on the linear scale). The two plots on the right indicate that the overall residuals (i.e. the pooling, or superimposition, of all the error terms) are distributed quite normally, but this is not necessarily the case for binomial data: if means were concentrated close to the two boundaries (0 and 1), the histogram plot (upper right) could turn out to be bimodal in the upper right plot, as the two error distributions would be oppositely skewed. As expected, the mean of the residuals (diamond in the boxplot, lower right) is reasonably close to zero. From the Normal Quantile plot, on the lower left, it can be appreciated that the quantiles of the residuals roughly match the theoretical normally-distributed quantiles (because, as H_0 hypothesis, residuals are supposed to be normally distributed). In the present instance, data still display a staircase pattern because they are discrete (germination events out of 20 Bernoulli trials), and BLUPs for individual plates are not considered since these are marginal residuals. Note that, in the Quantile plot, the number of quantiles is established as to match the number of data (residuals), so that the theoretical reference distribution is spliced in as many slices as the number of residuals in the studied dataset, each slice corresponding to an equal area of the theoretical reference distribution; thereby, each quantile corresponds to the same probability, and is therefore expected to include, on average, one residual datum, which is plotted against the midpoint of its corresponding theoretical quantile. In the Normal Quantile plot, the theoretical quantiles are represented on the x-axis in terms of standard deviations (probits or *z*-scores): more extreme residual values correspond to wider standard deviations from the zeroed mean (residuals are assumed to sum up to zero) and have lower probabilities. In this way, quantile width is lowest around the mode (in the middle, for a theoretical normal distribution), where probability density is highest, and decreases toward the tails.

Figure 4.

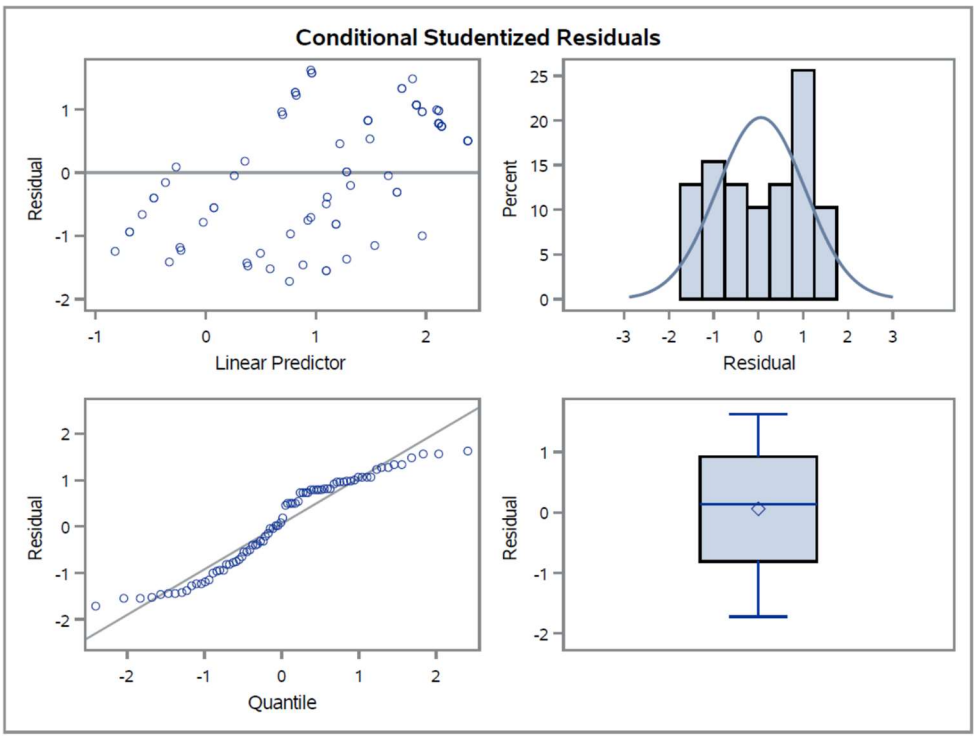

Figure 4 illustrates the GLIMMIX graphical analysis of conditional (i.e. measuring deviations from the conditional mean, that is, adjusted according to the BLUPs of the random effects) studentized residuals (which are adjusted for variance heterogeneity, i.e. standardized; thereby providing a neater picture of

residuals). Conditional studentized residuals are especially good at capturing the nuances of binomial data. The upper-left plot of the panel in Figure 4 shows the scatter plot of the conditional studentized residuals versus the linear predictor. Heteroscedasticity is no longer an issue on the probit scale (or, more properly in the case of these residuals, on a studentized scale), nor a whatsoever trend is apparent, thereby supporting randomness. Although residuals are reasonably centred on a mean of zero (lower right plot), they no longer appear Gaussian, but rather asymmetrically bimodal with truncated tails (upper right plot). This is due to the greater sharpness offered by conditional studentized residuals, which allow to focalize and discriminate the symmetric error distribution for the response close to 50%, i.e. for white-kernelled seeds artificially infected, from the other responses, which are closer to the upper limit (100%) and therefore skewed (studentized residuals are adjusted for variance heterogeneity, but, not being modelled on the linear scale, they maintain the skewedness of the original data). The bimodal residual distribution reflects into a sigmoidal relationship of the conditional studentized residuals versus the theoretical normally-distributed quantiles (Quantile plot on the lower left). The departures from the diagonal at the two extremes correspond to extreme residuals occurring closer to the mean than expected for a Gaussian distribution (which, indeed, is not the expected distribution for binomial data, but works well as H_0 hypothesis). In other words, data have fewer and/or less extreme values than would be expected if they truly came from a Gaussian distribution, that is, the distribution of residuals is platykurtic (short tailed). It is platykurtic because it is multimodal (which, as seen, results from uneven skewness across error terms). The error distribution of germination data is, indeed, binomial, and such distribution asymptotically approximates a Gaussian distribution as the number of binomial clusters (that is, plates) and, largely more effective, the number of Bernoulli trials in every binomial cluster (that is, seeds per plate) increase. Only when there are several binomial clusters (plates) and very many Bernoulli trials in every binomial cluster (seeds per plate), binomial residuals closely approximate ordinary Gaussian residuals. If, as often is the case, there are not several binomial clusters, and, above all, the number of Bernoulli trials in every binomial cluster is not large (>50), the studentized binomial residuals behave intermediately with respect to the Gaussian and the Bernoulli distributions, and the latter is strongly platykurtic. In the present case, 20 seeds per plate and 15-24 plates (across three genotypes) for each mean are evidently far from being enough to approximate a Gaussian distribution of conditional studentized residuals. Although this could be a trouble for ordinary ANOVA and GLM, GzLMM properly deal with the binomial error.

Altogether, residuals plots correspond to what expected for this kind of data, and they do not evidence anomalous outliers.

The statistical model can then be considered satisfactory (although further small improvements could be considered, some of which will be subsequently mentioned). The means of interest can then be compared, and their confidence intervals calculated:

```
/*conditional model*/
proc GLIMMIX plots=boxplot(group observed);
class grain trt cv plate;
model germ/n = grain trt grain*trt / link=probit;
random cv(grain);
random plate / group=grain*trt;
lsmeans grain*trt / cl adjust=Tukey ilink plot=meanplot;
run;
```

The ‘plots=’ option provides plots of observed values and Pearson conditional residuals for the levels of the ‘grain*trt’ effect (the ‘group’ effect referred to across the random statements). The ’lsmeans’ statement prescribes that: least-squares means are estimated for the levels of the ‘grain*trt’ effect (on the linked scale), their confidence limits (cl) are calculated, and multiple contrasts are performed with Tukey’s adjustment for

multiplicity; whereas the 'ilink' option requests that the estimated means and confidence limits are also reported on the scale of the data, and the 'plot=' option ensures that a plot of the requested estimates (on the linked scale) is displayed.

RESULTS (excerpts):

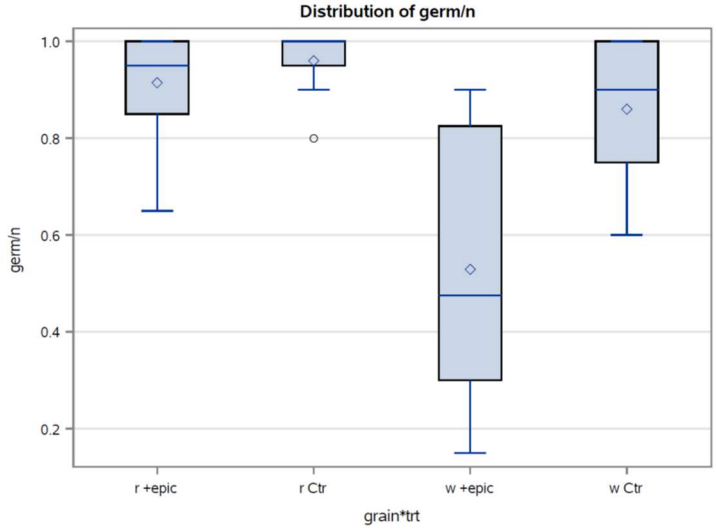

Figure 5.

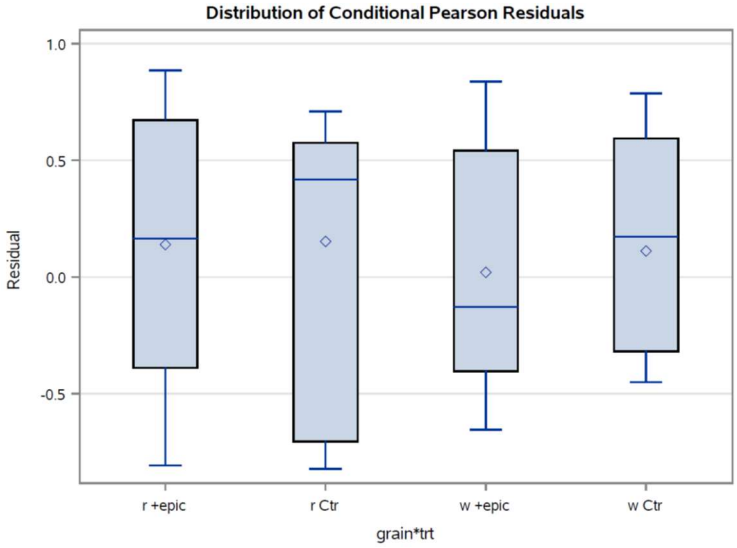

Figure 6.

Table 15.

grain*trt Least Squares Means													
grain	trt	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Mean	Standard Error Mean	Lower Mean	Upper Mean
r	+epic	1.5516	0.2575	70	6.03	<.0001	0.05	1.0380	2.0651	0.9396	0.03083	0.8504	0.9805
r	Ctr	1.9571	0.2944	70	6.65	<.0001	0.05	1.3698	2.5443	0.9748	0.01731	0.9146	0.9945
w	+epic	0.09623	0.2433	70	0.40	0.6936	0.05	-0.3889	0.5814	0.5383	0.09660	0.3487	0.7195
w	Ctr	1.2806	0.3229	70	3.97	0.0002	0.05	0.6365	1.9247	0.8998	0.05675	0.7378	0.9729

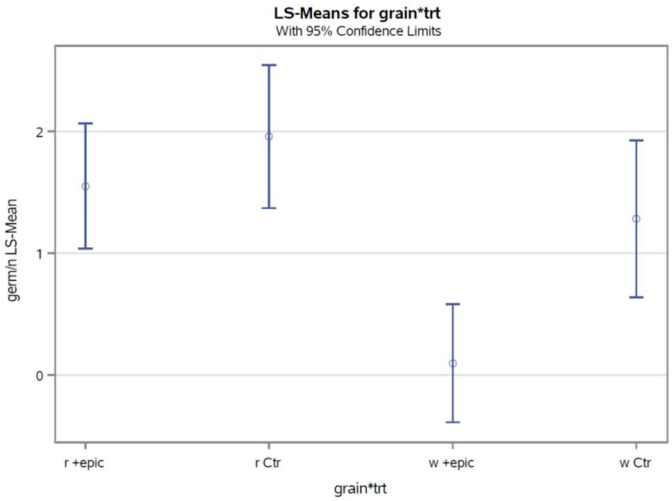

Figure 7.

Table 16.

Differences of grain*trt Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer														
grain	**trt**	**grain**	**trt**	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Adj P**	**Alpha**	**Lower**	**Upper**	**Adj Lower**	**Adj Upper**
r	+epic	r	Ctr	-0.4055	0.2521	70	-1.61	0.1122	0.3806	0.05	-0.9083	0.09729	-1.0690	0.2580
r	+epic	w	+epic	1.4553	0.3542	70	4.11	0.0001	0.0006	0.05	0.7488	2.1618	0.5231	2.3876
r	+epic	w	Ctr	0.2710	0.4130	70	0.66	0.5140	0.9131	0.05	-0.5528	1.0947	-0.8161	1.3580
r	Ctr	w	+epic	1.8608	0.3819	70	4.87	<.0001	<.0001	0.05	1.0991	2.6225	0.8557	2.8660
r	Ctr	w	Ctr	0.6765	0.4370	70	1.55	0.1262	0.4149	0.05	-0.1951	1.5481	-0.4737	1.8266
w	+epic	w	Ctr	-1.1844	0.2736	70	-4.33	<.0001	0.0003	0.05	-1.7300	-0.6387	-1.9044	-0.4644

Figure 5 basically furnishes the same plot as Figure 1, which had been obtained with the GLM procedure, but, in the GLIMMIX plot, data are reported as germination proportions (on the scale of observed data). Figure 6 displays the Pearson conditional residuals for the four means, that is, the error terms of the means are standardized according to their variances. The error terms do not appear to be closely centred around zero, but the overall Studentized conditional residuals (Figure 4, lower right plot) provide a better overall evaluation of this assumption, also given Figure 6 uses means of the original data (just like Figure 1) not the means estimated according to the GzLMM procedure (thus that the plot of Figure 6, which is automatically generated with the previous one, is not really informative, in the present instance). Table 15 shows the estimated means and confidence limits on the linked scale ('Estimate', 'Lower' and 'Upper') as well as on the data scale ('Mean', 'Lower Mean' and 'Upper Mean'; back-transformed by applying the inverse link transform to the estimates). 'Pr > |t|' is the test of the null hypothesis that the mean equals zero (on the linked scale), which, on the probit scale, equates to testing whether a germination mean differs from 50% (not really informative, at least in the present instance). Figure 7 gives a graphical representation of the least squares means with their confidence intervals on the linked scale. Table 16 displays the multiple comparisons between means (on the linked scale). 'Adj P' is the probability, adjusted for multiplicity of tests of the null hypothesis that the difference between the contrasted means equals zero. This adjustment is shown only for the sake of completeness, as it is not necessary in the present instance, given it is typically recommendable for experimental designs wherein many items (e.g., genotypes) are compared in the absence of relevant interaction effects, but it can be avoided when planned conditions are compared in structured, i.e. multifactor, treatment designs. Overall, these values demonstrate that the mean of the white rices treated with *Epicoccum* ('w +epic') is different from any other mean, and no other difference is statistically significant ($P \leq 0.05$). Notice that the conditional model makes fully explicit the hierarchical, multilevel, structure of

these data, wherein the random plate factor is nested within the random cultivar factor, which, in turn is nested within the fixed grain type (colour) factor. The model is partly nested (Quinn and Keough, 2002) because of the between-subjects 'trt' factor.

The statistical analysis can be also performed formulating the model as quasi-marginal (i.e. comprising both a G-side random factor and an R-side random effect of residuals), that is, modelling the random between-plates effect not as a random factor, but, rather, as a residual effect:

```
/*quasi-marginal model*/
proc GLIMMIX;
class grain trt cv plate;
model germ/n = grain trt grain*trt / link=probit;
random cv(grain);
random residual / group=grain*trt;
lsmeans grain*trt / cl adjust=Tukey ilink plot=meanplot;
covtest homogeneity; /*This tests homogeneity of variance parameters across groups*/
covtest zeroG; /*It tests whether (G-side) random effects have a significant effect on the model*/
run;
```

The 'random residual' statement now models the error terms, corresponding to the between-plates effect as no other residual effect is present, as an R-side variance/covariance structure. The only effect of this statement is allowing for different variances among levels of the 'grain*trt' interaction.

RESULTS (excerpts):

Table 17.

Dimensions	
G-side Cov. Parameters	1
R-side Cov. Parameters	4
Columns in X	9
Columns in Z	6
Subjects (Blocks in V)	1
Max Obs per Subject	78

Table 18.

Optimization Information	
Optimization Technique	Dual Quasi-Newton
Parameters in Optimization	5
Lower Boundaries	5
Upper Boundaries	0
Fixed Effects	Profiled
Starting From	Data

Table 19.

Fit Statistics	
-2 Res Log Pseudo-Likelihood	167.99
Generalized Chi-Square	74.00
Gener. Chi-Square / DF	1.00

Table 20.

Covariance Parameter Estimates			
Cov Parm	**Group**	**Estimate**	**Standard Error**
cv(grain)		0.1191	0.1164
Residual (VC)	grain*trt r +epic	2.5313	0.7651
Residual (VC)	grain*trt r Ctr	1.7614	0.6917
Residual (VC)	grain*trt w +epic	3.6726	1.3229
Residual (VC)	grain*trt w Ctr	6.2302	2.6837

Table 21.

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
grain	1	4	9.81	0.0351
trt	1	70	17.56	<.0001
grain*trt	1	70	3.28	0.0746

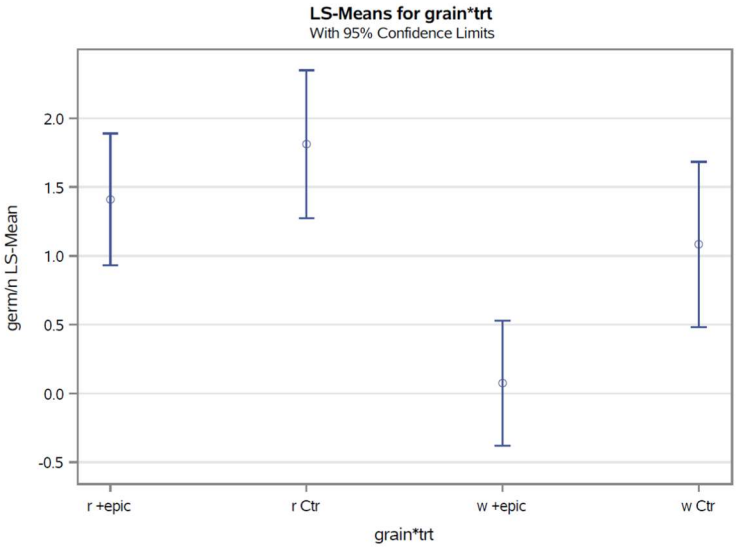

Figure 8.

Table 22.

Differences of grain*trt Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer														
grain	**trt**	**grain**	**trt**	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Adj P**	**Alpha**	**Lower**	**Upper**	**Adj Lower**	**Adj Upper**
r	+epic	r	Ctr	-0.4005	0.2225	70	-1.80	0.0762	0.2821	0.05	-0.8442	0.04331	-0.9861	0.1851
r	+epic	w	+epic	1.3363	0.3315	70	4.03	0.0001	0.0008	0.05	0.6752	1.9973	0.4639	2.2086
r	+epic	w	Ctr	0.3270	0.3858	70	0.85	0.3995	0.8315	0.05	-0.4424	1.0964	-0.6883	1.3423
r	Ctr	w	+epic	1.7367	0.3533	70	4.92	<.0001	<.0001	0.05	1.0321	2.4414	0.8069	2.6666
r	Ctr	w	Ctr	0.7275	0.4047	70	1.80	0.0766	0.2832	0.05	-0.07968	1.5346	-0.3376	1.7926
w	+epic	w	Ctr	-1.0093	0.2523	70	-4.00	0.0002	0.0009	0.05	-1.5124	-0.5061	-1.6733	-0.3453

Table 23.

Tests of Covariance Parameters Based on the Residual Pseudo-Likelihood					
Label	**DF**	**-2 Res Log P-Like**	**ChiSq**	**Pr > ChiSq**	**Note**
Homogeneity	3	173.79	5.79	0.1221	DF
No G-side effects	1	172.11	4.12	0.0212	MI

DF: P-value based on a chi-square with DF degrees of freedom.
MI: P-value based on a mixture of chi-squares.

Table 17 shows that now there are five parameters in optimization: one G-side parameter, the variance of 'cv(grain)', and four R-side parameters, namely, the between-plates variances of the four levels of the 'grain*trt' interaction. Fixed effects are 'profiled' from the analysis (Table 18), meaning they are not optimized by the likelihood function, but are modelled by a linear model working on pseudo-data obtained by transposing the original data onto the linked scale (this is why the estimation technique is called pseudo-likelihood). The Generalized Chi-Square / DF parameter is estimated as 1 (Table 19), suggesting there is no relevant overdispersion (rather, it was previously demonstrated that platykurtosis exists), and therefore no overly mistaken model specification is apparent. The 'cv(grain)' variance estimate (Table 20) is not much different from that obtained with the conditional model (see Table 12), but the between-plates variances are now computed as normalized R-side 'working' variances (a sort of scale parameters), not amenable to interpretation outside the same table (Stroup et al., 2018). It may, anyway, be noted that larger variances were obtained for white-grained cultivars, owing to greater susceptibility of these caryopses to stochastic infection by occasional microbial contaminants given the conditions of this experiment. The two fixed factors were significant (at $P \leq 0.05$), but not their interaction, though $P = 0.0746$ is only slightly above the significance threshold (Table 21). LS-means are illustrated in Figure 8, and inference about them is practically the same as in the conditional model (Table 22). The significant effect of the G-side variance is confirmed, as is homogeneity of between-plates residuals (Table 23), even though their heterogeneity no longer appears to be highly improbable. Thus, inference provided by the quasi-marginal model comes close to that obtained with the conditional model. Once properly specified, both the conditional and quasi-marginal approaches to modelling seed germination appear valid alternatives. If nesting of cultivars within grain types were not present, a marginal model would evidently be suitable as alternative to a conditional one.

It may be worth noting that the above-mentioned platykurtosis implies that residuals are less spread (into distribution tails) than would expected for a normal distribution (Figure 4). This is equivalent to say that the data appear to be underdispersed. Underdispersion can be measured by means of the Pearson Chi-Square / DF ratio (i.e. the overdispersion parameter), whereas the Generalized Chi-Square / DF ratio (which is the approximative overdispersion parameter when REPL is used) is not as good in detecting underdispersion. If the Laplace method is used to approximate the marginal distribution of GzLMM by maximum likelihood (ML), the Pearson Chi-Square / DF ratio can be measured even in presence of (G-side only) random effects (that is, for the conditional model only). For the present data, this approach (not shown) provides a Pearson Chi-Square / DF ratio = 0.23, which would seem to evidence underdispersion. Although underdispersion can take place because of model overfitting, an apparent underdispersion (according to the Pearson Chi-Square / DF ratio) is inherent to binomial data (Stroup et al., 2018). It is, therefore, not a trouble here. For binomial data, overdispersion is, most commonly, the real issue.

A straightforward difference between classical ANOVA models and GzLMM is that, for non-Gaussian response variables, specifically for response variables whose distribution is a one-parameter member of the exponential family, there is no residual variance component to estimate, and, therefore, all the random factors and their interactions should be included into the model, even because they could be needed for proper *F* tests (Gbur et al., 2012). However, hierarchical models, like the present one, are characterized by nested effects, which, in this kind of designs wherein the levels of the nested factor are different within each level of the main factor, represent entirely unbalanced interaction effects; therefore, the nested effect is completely confounded with its interaction (Quinn and Keough, 2002). Separate specification of nested effects and their interactions would thus be redundant, and must be omitted, at least in the case of plates. Besides, when passing from a 'random' (G-side) statement to a 'random residual' (R-side) one, the variance/covariance structure corresponding to the 'variance components' structure (which is used by default, if not otherwise specified) is the 'compound symmetry' structure, which includes a scale parameter to make up for the variance of the random (block) factor (namely, the plate effect), not directly modelled by the R matrix (Gbur et al., 2012). However, the modelling of the effect associated with the interaction of the random effect is already embedded into the R-side covariance structure. Thus, as germination data typically represent hierarchical models wherein the plate effect is completely confounded with its interaction, an R-side 'compound symmetry' structure would introduce redundant effects, and was therefore excluded.

Cultivars are nested within grain type, but, differently from the case of plates, a cross interaction could be envisaged between cultivars and treatment. A random coefficients model, with both a random intercept (for the random effect) and a random slope (that is, the random x fixed factor interaction) across treatments for each cultivar, could then be utilized (Littell et al., 2006). However, nesting of different cultivars within grain types baffles a proper disentanglement of the, partially overlapping, effects of the two random coefficients, and, in fact, a random coefficients model would incur into computational troubles because of a zero-variance component estimate for the cultivar effect once its interaction has been evaluated (not shown). Since the 'cv(grain)' effect must be retained, as previously discussed, its interaction is rather neglected. This, anyway, does not cause any overdispersion, as seen. Indeed, in random coefficients models, negative and zero variance component estimates can be artefacts due to negligibly unequal random slopes, and such issue often disappears once the random slope term is removed from the model (Gbur et al., 2012). When the interaction effect, which is evaluated first in type III ANOVA, is very small or null, and it partially overlaps with the random effect, boundary estimation within the same matrix can result in assigning all the significance to the interaction effect. A purely nested, and perfectly balanced, design would be more properly evaluated using type II ANOVA, which is more powerful if there is no significant interaction effect; but experimental designs of germination studies almost never are purely nested. When some interaction effect is expected, in fact, the experimental design is partly nested (Quinn and Keough, 2002).

It might also be useful mentioning that an analysis of the conditional model using the (computationally more exact) Laplace method would indicate a less significant effect of the interaction (P = 0.1753, not shown). Although Figure 5 would suggest that the interaction effect is large on the data scale, when the means are considered on the linear scale (Figure 7), where the model is tested, the interaction effect looks, indeed, much smaller. Nevertheless, the finding that only a mean, the 'w +epic' one, is significantly different from all the others (Table 16) is confirmed even if the Laplace method is used (not shown). Using the Laplace method is generally recommended (Stroup, 2015; Stroup et al., 2018) for FGP; but for basic, planned germination experiments, the improvements appear moderate. Unplanned experiments, on the other hand, greatly benefit from integral approximation because it allows proper evaluation of the model.

Finally, it must be noticed that establishing the right degrees of freedom is an important task in GzLMM to have appropriate *F* tests and confidence limits. In this respect, it deserves to be highlighted that for small-sized studies and, even more important, when designs with heavily unbalanced replication are utilized in the presence of random factors, the Satterthwaite correction should be used to compute the denominator degrees of freedom for tests of fixed effects. In the present case, this adjustment would slightly change significance probabilities, leaving anyway unchanged the conclusion of the analysis (not shown). In fact, the containment method for determining denominator degrees of freedom, which is used by default, can fare well even in presence of nesting and moderate unbalancing of replications, as occurs for the present data. The Satterthwaite correction is, however, recommended (Stroup et al., 2018) for more strongly unbalanced designs.

References

Bradford K.J. (1990). A water relations analysis of seed germination rates. *Plant Physiology* 94: 840-849.

Gbur E.E., Stroup W.W., McCarter K.S., Durham S., Young L.J., Christman M., West M. and Kramer M. (2012). *Analysis of Generalized Linear Mixed Models in the Agricultural and Natural Resources Sciences*. American Society of Agronomy: Madison, WI, USA.

Gianinetti A. and Cohn M.A. (2007). Seed dormancy in red rice. XII: Population-based analysis of dry-afterripening with a hydrotime model. *Seed Science Research* 17: 253-271.

Gianinetti A., Finocchiaro F., Maisenti F., Kouongni Satsap D., Morcia C., Ghizzoni R. and Terzi V. (2018). The caryopsis of red-grained rice has enhanced resistance to fungal attack. *Journal of Fungi* 4: 71.

Hardegree S.P., Walters C.T., Boehm A.R., Olsoy P.J., Clark P.E. and Pierson F.B. (2015). Hydrothermal germination models: comparison of two data-fitting approaches with probit optimization. *Crop Science* 55: 2276-2290.

Littell R.C., Milliken G.A., Stroup W.W., Wolfinger R.D. and Schabenberger O. (2006). *SAS® for Mixed Models, Second Edition*. SAS Institute Inc.: Cary, NC, USA.

Quinn G.P. and Keough M.J. (2002). *Experimental Design and Data Analysis for Biologists*. Cambridge University Press: Cambridge, UK.

Roberts E.H. (1961). Dormancy of rice seed: I. The distribution of dormancy periods. *Journal of Experimental Botany* 12: 319-329.

Stroup W.W. (2015). Rethinking the analysis of non-normal data in plant and soil science. *Agronomy Journal* 107: 811-827.

Stroup W.W., Milliken G.A., Claassen E.A. and Wolfinger R.D. (2018). *SAS® for Mixed Models: Introduction and Basic Applications*. SAS Institute Inc.: Cary, NC, USA.