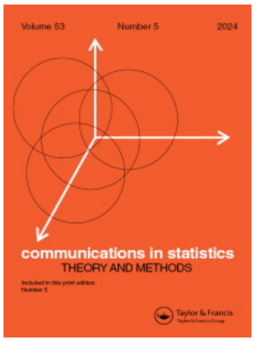

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/lsta20

Testing components of two-way interaction in multi-environment trials

Johannes Forkman, Waqas Ahmed Malik, Steffen Hadasch & Hans-Peter Piepho

To cite this article: Johannes Forkman, Waqas Ahmed Malik, Steffen Hadasch & Hans-Peter Piepho (2024) Testing components of two-way interaction in multi-environment trials, Communications in Statistics - Theory and Methods, 53:5, 1716-1735, DOI: 10.1080/03610926.2022.2108058

To link to this article: https://doi.org/10.1080/03610926.2022.2108058

Published online: 10 Aug 2022.

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
2024, VOL. 53, NO. 5, 1716–1735
https://doi.org/10.1080/03610926.2022.2108058

OPEN ACCESS

Testing components of two-way interaction in multi-environment trials

Johannes Forkman[a], Waqas Ahmed Malik[b], Steffen Hadasch[b], and Hans-Peter Piepho[b]

[a]Department of Crop Production Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden; [b]Institute of Crop Science, University of Hohenheim, Stuttgart, Germany

ABSTRACT

Experiments with two factors are commonly analyzed using two-way analysis of variance, where testing significance of interaction is straightforward. However, using bilinear models, interaction can be analyzed further. The additive main effects and multiplicative interaction (AMMI) model uses singular value decomposition for partitioning interaction into multiplicative terms, such that the first terms typically account for a large portion of the sum of squares, whereas the last terms are of minor importance. This model is used extensively for analysis of genotype-by-environment interaction in multi-environment trials. A recurring question is how to determine the number of terms to retain in the model. If data is replicated, which is usually the case, the F_R test can be used for this purpose. The simple parametric bootstrap method is another option, although this test was developed for unreplicated data. Since both of these tests of significance may be applied in cases with replication, researchers need advice on which of the methods to use. We discuss several statistical models and show that the two methods address different questions.

ARTICLE HISTORY
Received 29 October 2021
Accepted 26 July 2022

KEYWORDS
AMMI model; bilinear model; genotype-environment interaction; mixed-interaction model

1. Introduction

Multi-environment trials compare genotypes in varying environments, i.e., locations. Since relative performance of genotypes depends on environments, the focus is typically not only on average performance of genotypes over environments, but also on genotype-by-environment interaction. Estimates of this interaction determine which genotypes farmers choose to grow in their special environments. The *additive main effects and multiplicative interaction* (AMMI) *model* (Gauch 1988) is a popular method for the analysis. The AMMI model includes additive main effects of genotypes, additive main effects of environments, and multiplicative effects of genotype-by-environment interaction. The multiplicative effects are estimated using singular value decomposition. Usually just one or two multiplicative terms are retained in the model, although it is possible to include more. The question about how many multiplicative terms to include in the AMMI model for a specific multi-environment trial is crucial for the conclusions

CONTACT Johannes Forkman johannes.forkman@slu.se Department of Crop Production Ecology, Swedish University of Agricultural Sciences, Box 7043, 750 07 Uppsala, Sweden.

from the analysis about how different genotypes perform in varying environments, and thus decisive for which genotypes are actually cultivated.

Many methods have been proposed for testing the significance of the multiplicative terms of the AMMI model. No exact tests exist, but under the common assumption of normally distributed observations with homogeneous variance, the F_R test (Piepho 1995) and the simple parametric bootstrap (*SPB*) test (Forkman and Piepho 2014) are useful options.

Using the *SPB* test, only genotype-by-environment means must be known, since information about variance within genotype-by-environment combinations is not utilized. The *SPB* test is also applicable when there are no replicates. The F_R test, on the other hand, requires information about variance within genotype-by-environment combinations, since this information is needed for computation of the F_R statistic. Thus, the F_R test cannot be applied if there are no replicates. Most commonly in practice, the researcher has access to replicates and then faces the choice of using either the *SPB* test or the F_R test. Some researchers have used both tests and found that in practice they often give different results (Plavšin et al. 2021). In the present article, we shall examine why the tests typically give different results, as this has not been investigated before.

The F_R test and the *SPB* test were proposed for different statistical models. In both models, the response variable is the genotype-by-environment means. Both models include an intercept, fixed effects of genotypes and environments, and a sum of multiplicative interaction terms. In addition, both models include normally distributed residual error terms. However, different assumptions were made for the variance of this error term. The model for which the F_R test was proposed uses an error variance that is inversely proportional to the number of replicates within genotype-by-environment combinations. This is the pure within-environment error variance of genotype means. The model for which the *SPB* test was proposed, by contrast, uses a larger error variance, comprising both the pure error variance and residual, unexplained, interaction variance.

In this article, we derive a unified model, which includes both models as special cases. This enables a comparison between the two tests, with regard to which hypotheses they test and how they work under different scenarios. Furthermore, we show how the variance components of this model can be estimated.

The new model includes an intercept, fixed effects of genotypes and environments, a sum of fixed multiplicative interaction terms, random residual interaction, and random replication error. Thus, the residual terms of the models for the F_R test and the *SPB* test are replaced by two terms: a random residual interaction and a random replication error. Since the model includes both fixed and random effects of interaction, we call this model a *mixed-interaction model.*

Our mixed-interaction model is related to several early-proposed models that are still much used for analysis of genotype-by-environment interaction. Specifically, there is a strand of work on stability analysis, starting with Yates and Cochran (1938) and further developed by Finlay and Wilkinson (1963), who regressed genotype-by-environment means on environment means. Since regression cannot explain all interaction, residuals from the regression lines must comprise interaction, which is modeled as random. Eberhart and Russell (1966) extended this work by fitting genotype-specific residual variance components. Shukla (1972) provided a framework that explicitly modeled replicated data and separated plot error from residual interaction. However, in their regression approach, not only plot error and genotype-by-environment interaction effects, but also environmental main effects, are random. Replacing environment means by random effects in the Finlay and Wilkinson (1963) regression leads to factor-analytic models (Gogel, Cullis, and Verbyla 1995; Piepho 1997), which are mixed-effects model extensions of fixed-effects AMMI models. Factor-analytic models are fitted using iterative residual maximum likelihood procedures, which may sometimes not converge. Singular value decomposition is more convenient in this regard, since it is easily performed using fast and stable computer routines, and has by construction no convergence issues.

The Finlay and Wilkinson (1963) model is a linear regression model, which includes the assumption of random deviations from the line. The AMMI model generalizes the Finlay and Wilkinson (1963) model from simple regression to multiple regression. The regressors of the AMMI model are latent environmental variables. In our mixed-interaction model, the random residual interaction term is the random deviation from the latent multiple regression model. Gauch (1988) proposed a similar model, with residual interaction in addition to random replication error, but did not specify residual interaction as fixed or random.

Malik, Forkman, and Piepho (2019) compared the *SPB* test and the F_R test under the assumption of a model without any random interaction, i.e., using the model for which the F_R test was proposed. However, that comparison did not explain the fundamental difference between the two tests with regard to which hypotheses they are testing and the results are limited to the special case of no random residual interaction.

The purpose of this article is to show that the *SPB* test and the F_R test aim at different null hypotheses. This fact, which has not been noticed before, is clarified by the introduction of the mixed-interaction model. The performance of the tests is investigated through simulation. Finally, the article aims to provide recommendations on which test should be used.

Sections 2.1–2.3 discuss models proposed earlier for analysis of multi-environment trials. Section 2.4 introduces the mixed-interaction model. Section 2.5 proposes a method for estimating the variance components of this model. Section 2.6 presents the F_R test and the *SPB* test in the framework of the mixed-interaction effects model. Section 2.7 highlights the difference between the null hypotheses of the two tests. Section 3 provides an example, which illustrates the use of the two tests. Section 4 presents a simulation study, which shows how the two tests perform under different mixed-interaction model scenarios. Section 5 discusses practical consequences and gives advice on which test to use depending on the aim of the analysis.

2. Models and methods

2.1. Linear fixed-effects models

In a multi-environment experiment for comparison of I genotypes in J environments, with R replicates per environment, let y_{ijr} denote the yield in the rth replicate of the of the ith genotype in the jth environment. We will assume that, at each environment, a randomized compete block design is used. Furthermore, we will assume that all genotypes are tested in all environments, which is a common situation in official crop variety testing. Such multi-environment experiment data can be analyzed using the linear fixed-effects model

$$y_{ijr} = \mu + \alpha_i + \xi_j + \eta_{jr} + \theta_{ij} + e_{ijr} \tag{1}$$

where μ is an intercept, α_i is a fixed effect of the ith genotype, ξ_j is a fixed effect of the jth environment, η_{jr} is a fixed effect of the rth complete block in the jth environment, θ_{ij} is a fixed effect of interaction between the ith genotype and the jth environment, and e_{ijr} is a random residual error assumed to be normally distributed: $e_{ijr} \sim N(0, \sigma_E^2)$. Let $\bar{y}_{ij\cdot} = \sum_{r=1}^{R} y_{ijr}/R$. Then,

$$\bar{y}_{ij\cdot} = \mu + \alpha_i + \beta_j + \theta_{ij} + \bar{e}_{ij\cdot} \tag{2}$$

where $\beta_j = \xi_j + \sum_{r=1}^{R} \eta_{jr}/R$ and $\bar{e}_{ij\cdot} = \sum_{r=1}^{R} e_{ijr}/R$.

2.2. Linear mixed-effects models

Multi-environment trials can also be analyzed using linear mixed-effects models. With fixed effects of genotypes and random effects of environments (Shukla 1972; Patterson 1978), the model can be written

$$\bar{y}_{ij\cdot} = \mu + \alpha_i + b_j + s_{ij} + \bar{e}_{ij\cdot} \tag{3}$$

where $b_j \sim N(0,\ \sigma_B^2)$, $s_{ij} \sim N\left(0,\ \sigma_S^2\right)$ and all other terms are defined as in (2). The interaction is random because one of the main effects is random (Piepho, Büchse, and Emrich 2003). Alternatively, effects of genotypes and environments may be modeled as random and fixed, respectively (Smith, Cullis, and Gilmour 2001):

$$\bar{y}_{ij\cdot} = \mu + a_i + \beta_j + s_{ij} + \bar{e}_{ij\cdot} \tag{4}$$

where $a_i \sim N(0,\ \sigma_A^2)$ and all other terms are defined as in (2) and (3). This is a popular model used for genomic selection in breeding. However, in that application, effects of genotypes as well as effects of interaction are correlated according to an observed genomic relationship matrix (Montesinos-López et al. 2018). Such information is not usually available in official variety testing.

It is also possible to use a model with fixed main effects for environments and genotypes, but random effects for their interaction:

$$\bar{y}_{ij\cdot} = \mu + \alpha_i + \beta_j + s_{ij} + \bar{e}_{ij\cdot} \tag{5}$$

where all terms are defined as in (3) and (4). If some genotypes are missing in some environments, Model (3) may be preferable to Model (5), because by Model (3), inter-environment information about the differences between the genotypes can be recovered. However, this information is often small, as the variance between environments is usually large (Piepho and Möhring 2006). When the dataset is balanced, Models (3) and (5) give exactly the same results as regards differences between estimated marginal genotype means. Two-way models with fixed main effects of treatments and experiments, and random treatment-by-experiment interaction, i.e., similar to Model (5), are commonly used in meta-analysis (Piepho, Williams, and Madden 2012).

2.3. Bilinear models

The Finlay and Wilkinson (1963) regression model for analysis of genotype-by-environment interaction can be written as

$$\bar{y}_{ij\cdot} = \mu + \alpha_i + \phi_i w_j + s_{ij} + \bar{e}_{ij\cdot} \tag{6}$$

where μ, α_i, s_{ij} and $\bar{e}_{ij\cdot}$ are defined as in (5), and ϕ_i is the sensitivity of the ith genotype to a latent environmental variable w_j (Piepho 1999). In practice, environment means are often used as estimates of w_j, even though these are not the least-squares estimates (Digby 1979; Mandel 1995). Models with multiplicative terms, such as $\phi_i w_j$, are known as bilinear models (Gabriel 1978).

Mandel (1971) proposed a partitioning of the interaction into a sum of multiplicative terms. For the application of multi-environment trials, this model is known as the AMMI model (Gauch 1988). The AMMI model for I cultivars observed in J environments can be written as

$$\bar{y}_{ij\cdot} = \mu + \alpha_i + \beta_j + \sum_{m=1}^{M+1} \gamma_{im}\lambda_m\delta_{jm} + p_{ij} \tag{7}$$

where μ, α_i and β_j are defined as in (5), $p_{ij} \sim N(0, \sigma_P^2)$ and $\sum_{m=1}^{M+1} \gamma_{im}\lambda_m\delta_{jm}$ is the singular value decomposition of the $I \times J$ matrix $\mathbf{\Theta} = \{\bar{y}_{ij\cdot} - \mu - \alpha_i - \beta_j - p_{ij}\}$. Specifically, γ_{im} is the ith element of the mth left-singular vector of $\mathbf{\Theta}$, δ_{jm} is the jth element of the mth right-singular vector of $\mathbf{\Theta}$, λ_m is the mth singular value of $\mathbf{\Theta}$, and $M = \min(I-1,\ J-1)$. The rank of the matrix with elements $\{\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots}\}$ is at most M, because of rows and columns being centered. The rank of $\mathbf{\Theta}$, however, is at most $M+1$. Singular values are assumed to be sorted in decreasing order, i.e., $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{M+1} \geq 0$. Note that some or all of these singular values can be 0. If main effects of genotypes, α_i, are omitted from (7), then the resulting model is the genotype main effects and genotype-by-environment interaction effects (GGE) model (Yan et al. 2000), which is another popular model for analysis of multi-environment trials.

Let $\bar{y}_{\cdots} = \sum_{i=1}^{I}\sum_{j=1}^{J}\bar{y}_{ij\cdot}/(IJ)$ denote the grand mean, $\bar{y}_{i\cdot\cdot} = \sum_{j=1}^{J}\bar{y}_{ij\cdot}/J$ the mean for the ith genotype, and $\bar{y}_{\cdot j\cdot} = \sum_{i=1}^{I}\bar{y}_{ij\cdot}/I$ the mean in the jth environment. The parameters of Model (7) can be fitted in two steps using the method of least squares (Gabriel 1978). In the first step, μ is estimated as $\bar{y}_{\cdots}$, α_i as $\bar{y}_{i\cdot} - \bar{y}_{\cdots}$, and β_j as $\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots}$. These are the unique least-squares estimates given the constraints $\sum_{i=1}^{I}\alpha_i = 0$ and $\sum_{j=1}^{J}\beta_j = 0$. In the second step, the $I \times J$ matrix $\hat{\mathbf{\Theta}} = \{\bar{y}_{ij\cdot} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots}\}$ is subjected to singular value decomposition, which yields M positive singular values $\hat{\lambda}_1$, $\hat{\lambda}_2$, ..., $\hat{\lambda}_M$, such that $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_M \geq \hat{\lambda}_{M+1} = 0$ The $(M+1)$th singular value, $\hat{\lambda}_{M+1}$, is 0 since the rank of $\hat{\mathbf{\Theta}}$ is M.

2.4. The mixed-interaction model

When there are R replicates per environment, i.e., when $\bar{y}_{ij\cdot} = \sum_{r=1}^{R} y_{ijr}/R$, we may assume that $p_{ij} = s_{ij} + \sum_{r=1}^{R} e_{ijr}/R$, where $s_{ij} \sim N(0, \sigma_S^2)$ and $e_{ijr} \sim N(0, \sigma_E^2)$. Then $\sigma_P^2 = \sigma_S^2 + \sigma_E^2/R$ and

$$\bar{y}_{ij\cdot} = \mu + \alpha_i + \beta_j + \theta_{ij} + s_{ij} + \bar{e}_{ij\cdot} \tag{8}$$

where $\theta_{ij} = \sum_{m=1}^{M+1} \gamma_{im}\lambda_m\delta_{jm}$ and $\bar{e}_{ij\cdot} = \sum_{r=1}^{R} e_{ijr}/R$. Model (8) is essentially the AMMI model for replicated data (Gauch 1988), with the addition of the explicit assumptions that $s_{ij} \sim N(0, \sigma_S^2)$ and $\bar{e}_{ij\cdot} \sim N(0, \ \sigma_E^2/R)$. In Model (8), genotype-by-environment interaction has been decomposed into a fixed part, θ_{ij}, and a random part, s_{ij}. We refer to Model (8) as the *mixed-interaction model.*

Through this decomposition, the AMMI model can be viewed as a generalization of the Finlay and Wilkinson (1963) regression model (6), which also includes two random terms: s_{ij}, which is the departure from the regression line, and $\bar{e}_{ij.}$, which is the experimental error term. Whereas Model (6) describes the fixed-effects interaction with a single multiplicative term, Model (8) uses at most $M + 1$ non-zero multiplicative terms. These can be regarded as regressions on latent predictor variables. Consequently, the random interaction term, s_{ij}, in (8), is still a deviation from regression. In addition, Model (8) includes main effects, β_j, of environments.

When $\sigma_S^2 > 0$, the sum of the multiplicative terms and the random interaction can, through singular value decomposition, be written $\theta_{ij} + s_{ij} = \sum_{m=1}^{M+1} \gamma'_{im}\lambda'_m\delta'_{jm}$. With this notation, Model (8) becomes

$$\bar{y}_{ij\cdot} = \mu + \alpha_i + \beta_j + \sum_{m=1}^{M+1} \gamma'_{im}\lambda'_m\delta'_{jm} + \bar{e}_{ij\cdot} \tag{9}$$

when $\sigma_S^2 > 0$. In the other event, when $\sigma_S^2 = 0$, Model (8) is

$$\bar{y}_{ij\cdot} = \mu + \alpha_i + \beta_j + \sum_{m=1}^{M+1} \gamma_{im}\lambda_m\delta_{jm} + \bar{e}_{ij\cdot} \tag{10}$$

Model (8) is an extension of Models (5) and (10), since Model (5) is the special case that $\sum_{m=1}^{M+1} \gamma_{im}\lambda_m\delta_{jm} = 0$, which occurs when $\lambda_1 = 0$, and Model (10) is the special case that $\sigma_S^2 = 0$. When both these special cases occur, there is no genotype-by-environment interaction at all. In the following, we will assume the mixed-interaction-effects model (8), since this model covers all cases.

2.5. Estimation of variance components in the mixed-interaction model

The variance component σ_E^2 of (8) is readily estimated as the error mean square (MSE), defined as

$$MSE = \frac{SSE}{DF} \tag{11}$$

where SSE and DF are the error sum of squares and error degrees of freedom, respectively. Using Model (1), $SSE = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{r=1}^{R} \left(y_{ijr} - \bar{y}_{ij.} - \bar{y}_{.jr} + \bar{y}_{.j.}\right)^2$ and $DF =$

$J(I-1)(R-1)$. When the design does not include blocks but is completely randomized, the block effects, η_{jr}, must be omitted from (1). In that case, $SSE = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{r=1}^{R}\left(y_{ijr}-\bar{y}_{ij.}\right)^2$ and $DF = IJ(R-1)$.

It is more challenging to estimate σ_S^2 of (8) when $\sum_{m=1}^{M+1}\gamma_{im}\lambda_m\delta_{jm} \neq 0$. However, when $\lambda_1, \lambda_{2,} \ldots, \lambda_\kappa$ of (8) are large as related to σ_P and $\lambda_{\kappa+1} = \lambda_{\kappa+2} = \ldots = \lambda_{M+1} = 0$, then the joint distribution of $\hat{\lambda}_{\kappa+1}^2/\sigma_P^2$, $\hat{\lambda}_{\kappa+2}^2/\sigma_P^2$, $\ldots, \hat{\lambda}_M^2/\sigma_P^2$ is approximately central Wishart distributed (Muirhead 1978) such that the expected value of $\sum_{m=\kappa+1}^{M}\hat{\lambda}_m^2$ is approximately $(I-1-\kappa)(J-1-\kappa)\sigma_P^2$. Under this condition, σ_P^2 can be estimated as the residual interaction mean square (MSR) divided by R, where

$$MSR = \frac{R\sum_{m=\kappa+1}^{M}\hat{\lambda}_m^2}{(I-1-\kappa)(J-1-\kappa)} \tag{12}$$

Thus, σ_S^2 can be estimated as

$$\hat{\sigma}_S^2 = \frac{MSR - MSE}{R}, \quad \kappa = 0,\ 1,\ \ldots M-1 \tag{13}$$

In the special case of Model (5), $MSR = R\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\bar{y}_{ij.}-\bar{y}_{i.}-\bar{y}_{.j.}+\bar{y}_{...}\right)^2/\left((I-1-\kappa)(J-1-\kappa)\right)$, at which (13) is an unbiased estimator of σ_S^2.

2.6. Tests for significance of multiplicative terms

Under the same condition as required for estimation of σ_S^2 using (13), the SPB method can be used for sequentially testing the hypotheses $H_0 : \lambda_{\kappa+1} = 0$ in Model (7), where $\kappa = 0,\ 1,\ \ldots, M-2$. The SPB test uses as test statistic $T = \hat{\lambda}_{\kappa+1}^2/\sum_{m=\kappa+1}^{M}\hat{\lambda}_m^2$. The distribution for this test statistic, under the null hypothesis, is simulated through repeated sampling of $(I-1-\kappa)\times(J-1-\kappa)$ matrices of standard normally distributed values. For each random sample, T_b is computed as the ratio of the first squared singular value to the sum of all squared singular values. The distribution of T_b thus obtained, when $b = 1,\ 2,\ \ldots,\ B$, where B is large, is the simulated distribution of the test statistic T under the null hypothesis.

The F_R test was proposed for testing the hypotheses $H_0 : \lambda_{\kappa+1} = 0$ in Model (10), where $\kappa = 0,\ 1,\ \ldots, M-1$. Thus, with the F_R test, it is possible to test one null hypothesis more than with the the SPB test. The F_R test uses as test statistic $F_R = MSR/MSE$, where MSE and MSR are defined as in (11) and (12), respectively. The F_R test statistic should be compared with an F distribution with $(I-1-\kappa)(J-1-\kappa)$ and DF degrees of freedom, where DF is the error degrees of freedom, defined as in (11).

Assuming Model (8), with $p_{ij} =\ s_{ij} + \bar{e}_{ij}.$, where $\sigma_S^2 > 0$, the SPB method is readily applicable. However, Model (8) is not on the form required for the F_R test as specified by Piepho (1995), since Model (8) includes not just a single random error term, but two random terms, s_{ij} and $\bar{e}_{ij}.$. Still, if there are replicates, i.e., $R > 1$, the F_R test can be used for testing the hypotheses $H_0 : \lambda_{\kappa+1}^{'} = 0$, in Model (9), where $\kappa = 0,\ 1,\ \ldots, M-1$.

The *SPB* test and the proposed estimator (13) of the interaction variance, σ_S^2, utilize the above mentioned Muirhead (1978, 23) approximation requiring the first κ singular values to be large at testing $H_0 : \lambda_{\kappa+1} = 0$. If this condition is not fulfilled, i.e., if λ_κ is small or moderate, then in our experience $\sum_{m=\kappa+1}^{M} \hat{\lambda}_m^2$ tends to be smaller than $(I-1-\kappa)(J-1-\kappa)\sigma_P^2$. Since also $\hat{\lambda}_{\kappa+1}^2$ tends to be smaller when the condition is not fulfilled, the effect on the test statistic $T = \hat{\lambda}_{\kappa+1}^2 / \sum_{m=\kappa+1}^{M} \hat{\lambda}_m^2$ is not clear. However, the effect on the F_R test statistic, as written $F_R = R\sum_{m=\kappa+1}^{M} \hat{\lambda}_m^2 / \left((I-1-\kappa)(J-1-\kappa)MSE\right)$ is obvious. This test statistic will be smaller if the condition is not fulfilled, giving significant results less often. In the simulation study of Section 3, these properties of the *SPB* and F_R tests will be explored.

2.7. The *SPB* and the F_R tests aim at different hypotheses

Models (8) and (9) can be written together as

$$\bar{y}_{ij\cdot} = \mu + \alpha_i + \beta_j + \sum_{m=1}^{M+1} \gamma_{im}\lambda_m\delta_{jm} + s_{ij} + \bar{e}_{ij\cdot} = \mu + \alpha_i + \beta_j + \sum_{m=1}^{M+1} \gamma'_{im}\lambda'_m\delta'_{jm} + \bar{e}_{ij\cdot}$$

The null hypotheses of the *SPB* test is

$$H_0 : \lambda_{\kappa+1} = 0$$

whereas the null hypothesis of the F_R test is

$$H_0 : \lambda'_{\kappa+1} = 0$$

By generalizing the Finlay-Wilkinson regression model (6) into an AMMI model that includes a random departure from regression, s_{ij}, we have in this article uncovered the difference in intention between the two tests. Since we now know that the two methods test different null hypotheses, it is no longer surprising that they usually give different results with regard to how many multiplicative terms are significant.

Because $\lambda'_{\kappa+1} \geq \lambda_{\kappa+1}$ and $\sum_{i=1}^{I}\sum_{j=1}^{J} \bar{e}_{ij\cdot}^2 \leq \sum_{i=1}^{I}\sum_{j=1}^{J} (s_{ij} + \bar{e}_{ij\cdot})^2$, the F_R test typically gives more significant results than the *SPB* test. This is illustrated by the following example.

3. Example

Shafii and Price (1998) studied an AMMI analysis of a multi-environment trial in winter rapeseed. The data is easily accessible, since it is included as the `shafii.rapeseed` dataset of the `agridat` package (Wright 2021) of R. We shall use the part of the experiment that was carried out in 1989. This subset comprises $I = 6$ genotypes investigated in $J = 9$ environments, with $R = 4$ observations per genotype and environment. Genotypes 1–6 are Dwarf, Jet, Cascade, Bridger, Glacier, and Bienvenu, respectively. Environments 1–9 are TGA, NC, GGA, SC, VA, TN, NY, WA, and ID, respectively. Observations are yield (tonnes/ha).

An initial two-way analysis of variance using Model (1) shows that there are significant main effects of genotypes ($F = 2.92,\ p = 0.016$), significant main effects of environments ($F = 219.62,\ p < 0.001$) and a significant genotype-by-environment interaction ($F = 6.42,\ p < 0.001$). Since the interaction is significant, this is analyzed further.

With main effects removed, the transpose of the interaction matrix is

$$\hat{\boldsymbol{\Theta}}^{\mathrm{T}} = \begin{pmatrix} -0.68 & -0.33 & 0.59 & 0.83 & -0.24 & -0.17 \\ 0.17 & -0.45 & -0.29 & 0.18 & 0.12 & 0.26 \\ -0.11 & 0.07 & -0.08 & -0.57 & 0.40 & 0.30 \\ -0.87 & -0.81 & 0.82 & 0.86 & -0.14 & 0.14 \\ -0.02 & 0.23 & -0.35 & -0.21 & 0.21 & 0.14 \\ 0.22 & 0.41 & -0.67 & -0.24 & -0.29 & 0.56 \\ 0.58 & 0.20 & -0.68 & -0.19 & 0.15 & -0.07 \\ 0.07 & 0.10 & 0.69 & 0.75 & -0.35 & -1.26 \\ 0.64 & 0.58 & -0.04 & -1.42 & 0.15 & 0.09 \end{pmatrix}$$

The singular values of $\hat{\boldsymbol{\Theta}}$ are: $\hat{\lambda}_1 = 3.00$, $\hat{\lambda}_2 = 1.57$, $\hat{\lambda}_3 = 1.14$, $\hat{\lambda}_4 = 0.68$, and $\hat{\lambda}_5 = 0.46$.

Using the F_R test, three multiplicative terms are significant, but using the simple parametric bootstrap test, only the first multiplicative term is significant, as reported in Table 1. The choice of test may have practical consequences. In a model with three multiplicative terms retained, Bienvenu is the best genotype for environment GGA (Georgia). In a model with a single multiplicative term retained, however, Dwarf is the best genotype for environment GGA. In the Discussion, we argue that the F_R test is preferable for this question, and thus recommend Bienvenu for environment GGA.

4. Simulation study

4.1. Design

A simulation study was performed to illustrate the main differences between the F_R and *SPB* tests. Observations were repeatedly simulated following the model

$$y_{ijr} = \gamma_{i1}\lambda_1\delta_{j1} + \gamma_{i2}\lambda_2\delta_{j2} + s_{ij} + e_{ijr} \tag{14}$$

where $i = 1,\ 2,\ \dots, 15$; $j = 1,\ 2,\ \dots, 10$; and $r = 1,\ 2,\ 3,\ 4,$ thus simulating multi-environment trials comparing fifteen genotypes in ten environments with four replicates per genotype and environment. Eight specific combinations of parameter values were simulated. In all cases, $\sigma_E^2 = 1$. Table 2 lists the settings for the other three parameters: λ_1, λ_2 and σ_S^2. In Cases 1 and 2, there were no multiplicative terms ($\lambda_0 = 0$). In Cases 3 and 4, a mild first multiplicative term was present ($\lambda_1 = 2$), and in Cases 5–8, a strong first multiplicative term ($\lambda_1 = 10$). In Cases 5 and 6, the second multiplicative term was zero, whereas in Cases 7 and 8, a mild second multiplicative term was assumed ($\lambda_2 = 2$). Cases with and without random interaction ($\sigma_S^2 = 0$ and 0.04, respectively) were explored, as listed in Table 2. When the interaction variance, σ_S^2, is 0.04, the interaction standard deviation, σ_S is 0.2, i.e., one fifth of the error standard deviation, σ_E. This value of σ_S was chosen because it proved to illustrate the differences between the

Table 1. Test statistics and *p*-values when testing multiplicative terms in the Shafii and Price (1998) example, using the F_R test and the simple parametric bootstrap test with $B = 100\,000$ bootstrap samples.

	F_R test		*SPB* test	
Multiplicative term	F_R	p-value	T	p-value
1	6.42	<0.001	0.669	0.006
2	3.04	<0.001	0.556	0.377
3	2.10	0.009		
4	1.30	0.237		

Table 2. Average estimates of the variance σ_S^2 and frequencies (freq.) of significant results when testing at significance level 0.05.

Parameter settings				Mean	F_R test		*SPB* test	
Case	λ_1	λ_2	σ_S^2	$\hat{\sigma}_S^2$	Null hypothesis	Freq.	Null hypothesis	Freq.
1	0	0	0	0.000	$H_0:\ \lambda_1^{'} = 0$	0.051[a]	$H_0:\ \lambda_1 = 0$	0.051[a]
2	0	0	0.04	0.040	$H_0:\ \lambda_1^{'} = 0$	0.281[b]	$H_0:\ \lambda_1 = 0$	0.050[a]
3	2	0	0	−0.019	$H_0:\ \lambda_1^{'} = 0$	0.181[b]	$H_0:\ \lambda_1 = 0$	0.131[b]
4	2	0	0.04	0.016	$H_0:\ \lambda_1^{'} = 0$	0.517[b]	$H_0:\ \lambda_1 = 0$	0.109[b]
5	10	0	0	−0.001	$H_0:\ \lambda_2^{'} = 0$	0.047[a]	$H_0:\ \lambda_2 = 0$	0.050[a]
6	10	0	0.04	0.039	$H_0:\ \lambda_2^{'} = 0$	0.248[b]	$H_0:\ \lambda_2 = 0$	0.050[a]
7	10	2	0	−0.020	$H_0:\ \lambda_2^{'} = 0$	0.194[b]	$H_0:\ \lambda_2 = 0$	0.138[b]
8	10	2	0.04	0.013	$H_0:\ \lambda_2^{'} = 0$	0.503[b]	$H_0:\ \lambda_2 = 0$	0.115[b]

[a)]Frequency of Type I error,
[b)]Power.

tests well. For each of the eight cases, 100,000 datasets of $10 \cdot 15 \cdot 4 = 600$ observations were randomly generated following Model (14).

For each case, multiplicative terms were generated as follows. In the qth simulation, $q = 1,\ 2,\ \ldots,\ 100\ 000$, a 15×10 matrix of random standard normally distributed values, $z_{ij}^{(q)}$, was generated and subjected to singular value decomposition: $z_{ij}^{(q)} = \sum_{m=1}^{10} \gamma_{im}^{(q)} \lambda_m^{(q)} \delta_{jm}^{(q)}$. The first and second multiplicative terms were set to $\gamma_{i1}^{(q)} \lambda_1 \delta_{j1}^{(q)}$ and $\gamma_{i2}^{(q)} \lambda_2 \delta_{j2}^{(q)}$, respectively, where λ_1 and λ_2 were selected as specified in Table 2.

For the analyses, Model (1) without block effects was assumed. Consequently MSE was estimated as $\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{r=1}^{R} \left(y_{ijr} - \overline{y}_{ij.}\right)^2 / (IJ(R-1))$. The interaction variance σ_S^2 was estimated using (13), employing $\kappa = 0$ in Cases 1 and 2, $\kappa = 1$ in Cases 3–6, and $\kappa = 2$ in Cases 7 and 8. In Cases 1–4, the significance of the first singular value was tested. Precisely, the null hypothesis $H_0:\ \lambda_1^{'} = 0$ was tested using the F_R test applied to Model (9), and the null hypothesis $H_0:\ \lambda_1 = 0$ was tested using the *SPB* test, with $B = 1000$ bootstrap samples, applied to Model (7). In Cases 5–8, the significance of the second singular value was tested, i.e., the null hypotheses $H_0:\ \lambda_2^{'} = 0$ and $H_0:\ \lambda_2 = 0$ were tested using the F_R test and the *SPB* test ($B = 1000$), respectively. Note that in cases with no random interaction, i.e., in Cases 1, 3, 5 and 7, the two tests aimed at the same null hypothesis, since in these cases $\lambda_1 = \lambda_1^{'}$ and $\lambda_2 = \lambda_2^{'}$. In addition, the hypotheses $H_0 : \lambda_2 = 0$ and $H_0 : \lambda_1 = 0$ were tested in Cases 3 and 5, respectively.

4.2. Results

Table 2 presents the average estimate of the interaction variance and the observed frequencies of significant results when testing these hypotheses at level 0.05. On average, the estimator (13) of the interaction variance σ_S^2 performed well in Cases 1, 2, 5 and 6. In Cases 1 and 2, this estimator is unbiased. In Cases 5 and 6, the agreement between $\hat{\sigma}_S^2$ and σ_S^2 was good as consequence of the highest positive singular value, $\lambda_1,$ being large, such that the necessary condition for the Wishart approximation was met. However, in Cases 3 and 4, with $\lambda_1 = 2,$ and in Cases 7 and 8, with $\lambda_2 = 2,$ the highest positive singular value was not sufficiently dominating, thus resulting in less good estimates of the interaction variance.

In case of no multiplicative terms and no random interaction (Case 1), both tests showed Type I error rates close to the nominal level 0.05 (Table 2). Figures 1a and b, for the F_R test and the *SPB* test, respectively, show that these tests are exact when testing the significance of the first singular value, since the quantiles of the simulated p-values agrees perfectly with the quantiles of a uniform distribution on the interval from 0 to 1.

In Case 2, which included no multiplicative terms but a random interaction, using the F_R test, the first singular value was significant at level 0.05 in 28.1% of the cases (power), whereas the frequency of significant results of the *SPB* test was the same as the nominal level, 5.0% (Type I error). The bend in Figure 1c of the curve below the 45° reference line shows that the F_R test resulted in more significant results than would a random draw from a uniform distribution, regardless of the chosen level of significance. The addition of the random interaction, as compared to Case 1, increased the probability of a large first singular value, which resulted in a raised frequency of significant F_R tests. Since in Case 2, interaction is present, the result 0.281 (Table 2) is the estimated power of the F_R test, when performed at significance level 0.05. If σ_S^2 had been larger than 0.04, an estimated power larger than 28.1% would have been expected, because the larger this variance, the larger the expected effects of interaction. The *SPB* test yielded the same frequency of significant results at all levels of significance (Figure 1d). This result illustrates the property of the *SPB* test that it does not test the significance of random interaction. On the contrary, the *SPB* test is designed for detecting patterns in the data that is not random, but systematic.

Cases 3 and 4 both investigate power, since in these cases $\lambda_1 = 2,$ which implies that also $\lambda_1^{'} > 0.$ Both in case of no random interaction (Case 3) and in case of random interaction (Case 4), the F_R test was more powerful than the *SPB* test (Table 2). The difference between the methods was larger in Case 4 than in Case 3, since the *SPB* test does not test for random interaction, but the F_R test does. As shown in Figure 2, these conclusions are valid also for other levels of significance than 0.05, where Case 3 is presented in Figures 2a and b, and Case 4 in Figures 2c and d. The curves are more bended in Figures 2a and c, for the F_R test, than in Figures 2b and d, which show results for the *SPB* test. Note that in Case 3, since $\lambda_1^{'} = \lambda_1,$ the two methods test the same hypothesis, while in Case 4, since $\lambda_1^{'} \neq \lambda_1,$ different hypotheses are tested.

In Cases 5–8, the second singular value, $\lambda_2,$ was tested, assuming the presence of a first multiplicative term with a strong singular value, $\lambda_1 = 10.$ Simulation results for these cases were very similar to those for Cases 1–4, where Case 5 corresponds to Case 1, and so on (Table 2, Figures 3–4). The *SPB* test maintains the correct type I error rate

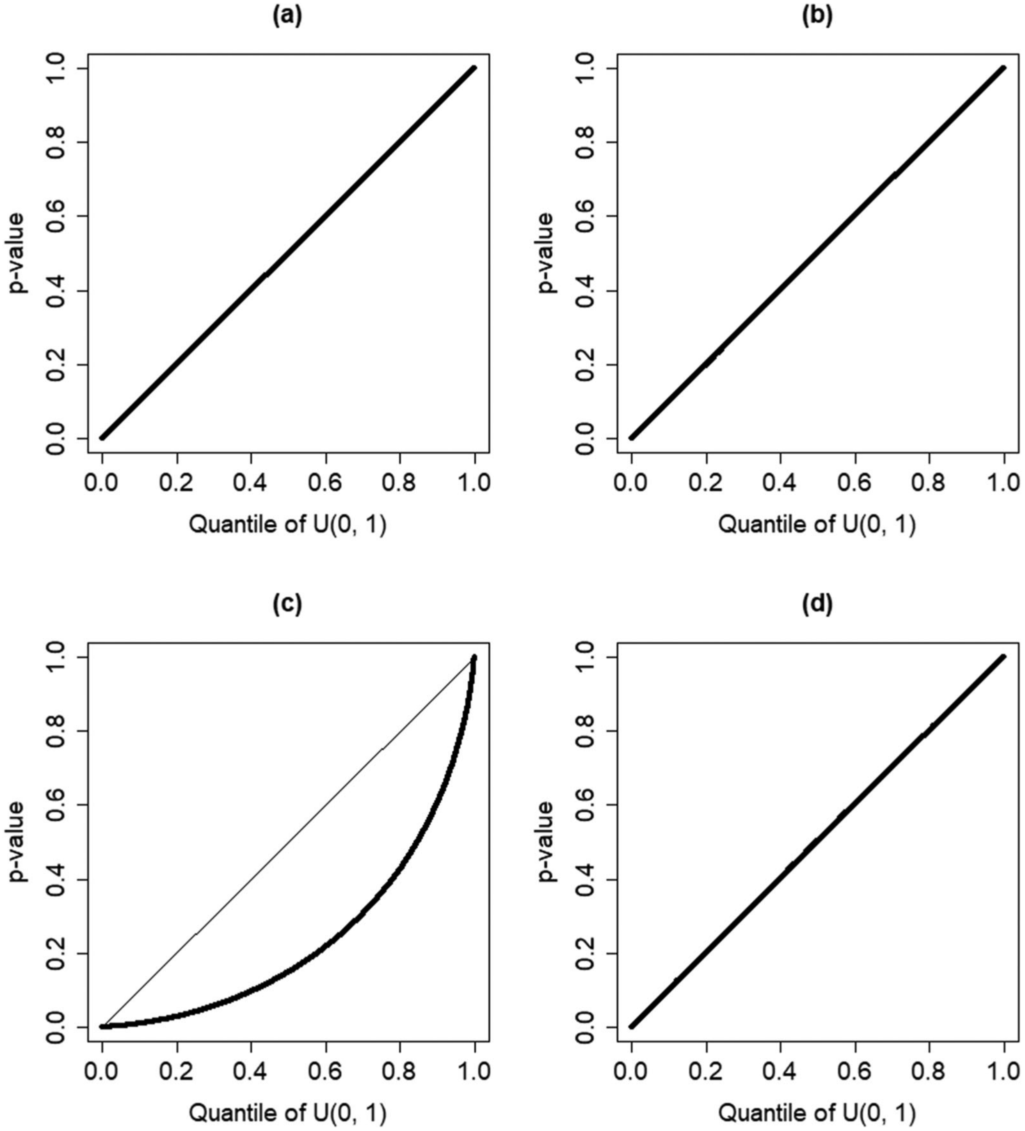

Figure 1. Quantiles of observed *p*-values, at testing $H_0: \lambda_1 = 0$, versus quantiles of the uniform distribution. Results are based on 100 000 simulated datasets following model (14) with $\lambda_1 = \lambda_2 = 0$. (a) F_R test, $\sigma_S^2 = 0$; (b) SPB test, $\sigma_S^2 = 0$; (c) F_R test, $\sigma_S^2 = 0.04$; d) SPB test, $\sigma_S^2 = 0.04$.

although a random interaction is added, as seen by comparing Case 6 with Case 5 in Table 2.

In order to investigate the importance of the Muirhead (1978, 23) condition for the performance of the F_R and SPB tests, the hypothesis $H_0 : \lambda_2 = 0$ was tested in Case 3, where λ_1 was not large. The observed frequencies of significant results were 0.013 and 0.021, for F_R and SPB, respectively.

One might falsely get the impression from Table 2 that neither the F_R test nor the SPB test are powerful methods, because the largest estimated power was only 0.517. However, power depends on effect size. As an example, in Case 5, the frequency of

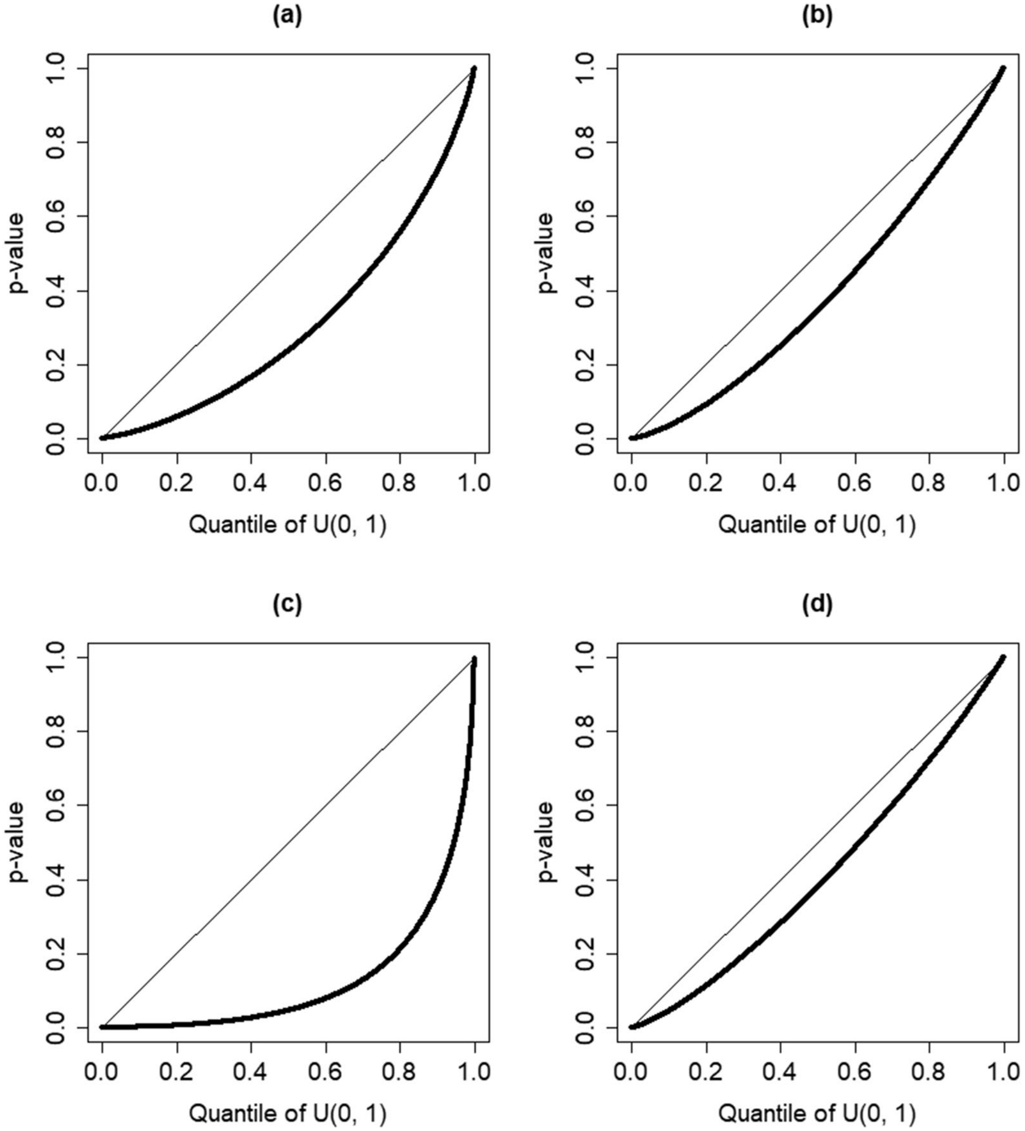

Figure 2. Quantiles of observed p-values, at testing $H_0: \lambda_1 = 0$, versus quantiles of the uniform distribution. Results are based on 100 000 simulated datasets following Model (14) with $\lambda_1 = 2$ and $\lambda_2 = 0$. (a) F_R test, $\sigma_S^2 = 0$; (b) *SPB* test, $\sigma_S^2 = 0$; (c) F_R test, $\sigma_S^2 = 0.04$; d) *SPB* test, $\sigma_S^2 = 0.04$.

significant results at testing $H_0 : \lambda_1 = 0$ instead of $H_0 : \lambda_2 = 0$ was 1.000 for both the F_R test and the *SPB* test, although this result was not included in Table 1. We refer to the original articles (Piepho 1995; Forkman and Piepho 2014) for more information about the power of these tests.

5. Discussion

The *SPB* test investigates null hypotheses regarding the fixed-effects part of the interaction, whereas the F_R test examines null hypotheses regarding the sum of the fixed and

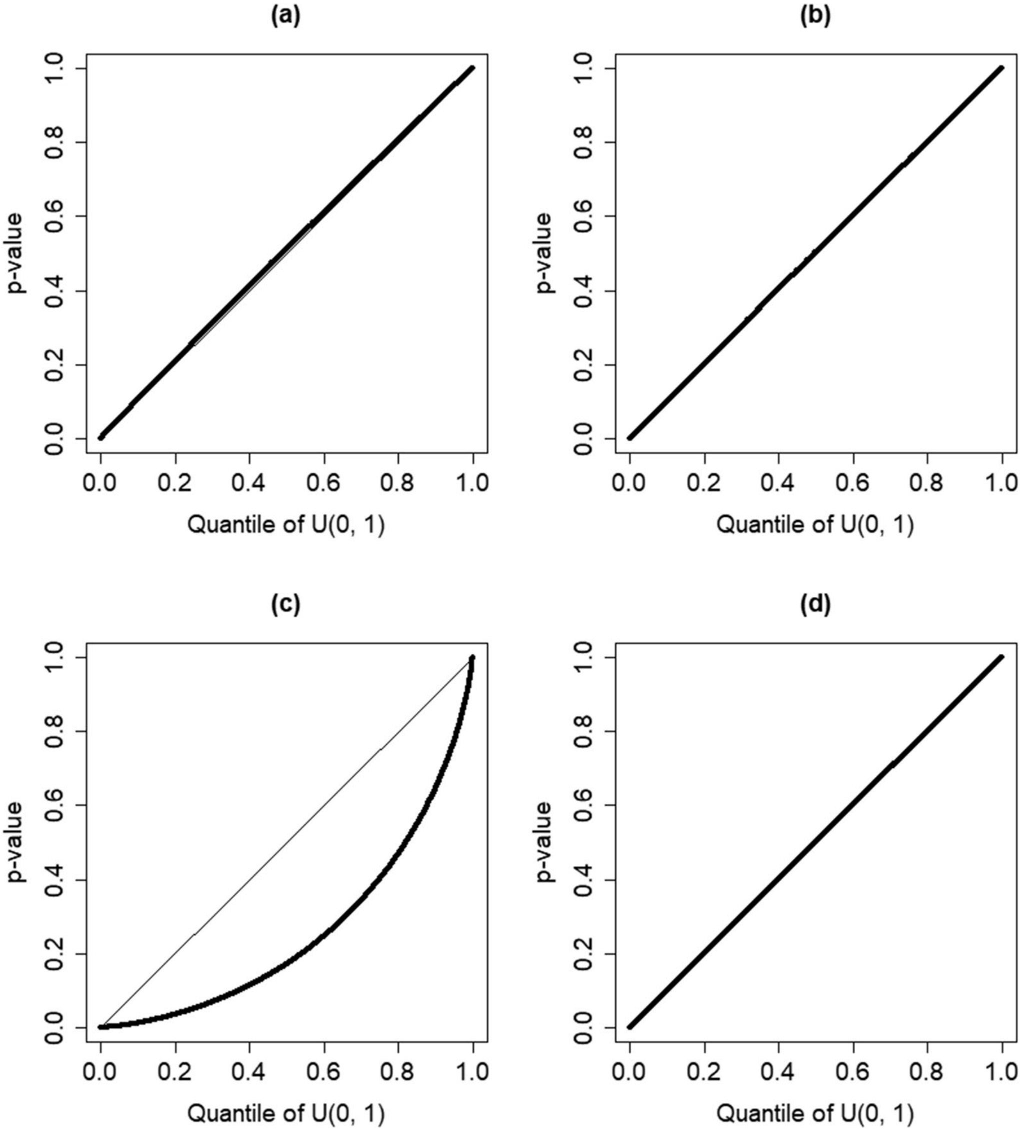

Figure 3. Quantiles of observed *p*-values, at testing $H_0: \lambda_2 = 0$, versus quantiles of the uniform distribution. Results are based on 100 000 simulated datasets following Model (14) with $\lambda_1 = 10$ and $\lambda_2 = 0$. (a) F_R test, $\sigma_S^2 = 0$; (b) *SPB* test, $\sigma_S^2 = 0$; (c) F_R test, $\sigma_S^2 = 0.04$; d) *SPB* test, $\sigma_S^2 = 0.04$.

the random interaction. The choice between the two tests depends on the aim of the analysis. If the aim is to find patterns in the interaction that are larger than would be expected by random genotype-by-environment interaction, then the *SPB* test is the right choice. If, however, research aims at finding out whether or not there is any interaction at all in the data, regardless of this interaction being fixed or random, and how many multiplicative terms are needed for describing this total interaction, then the F_R test should be chosen.

In reality, there is almost certainly some genotype-by-environment interaction in a multi-environment trial. If so, the null hypothesis $H_0: \lambda_1' = 0$ is always rejected by the

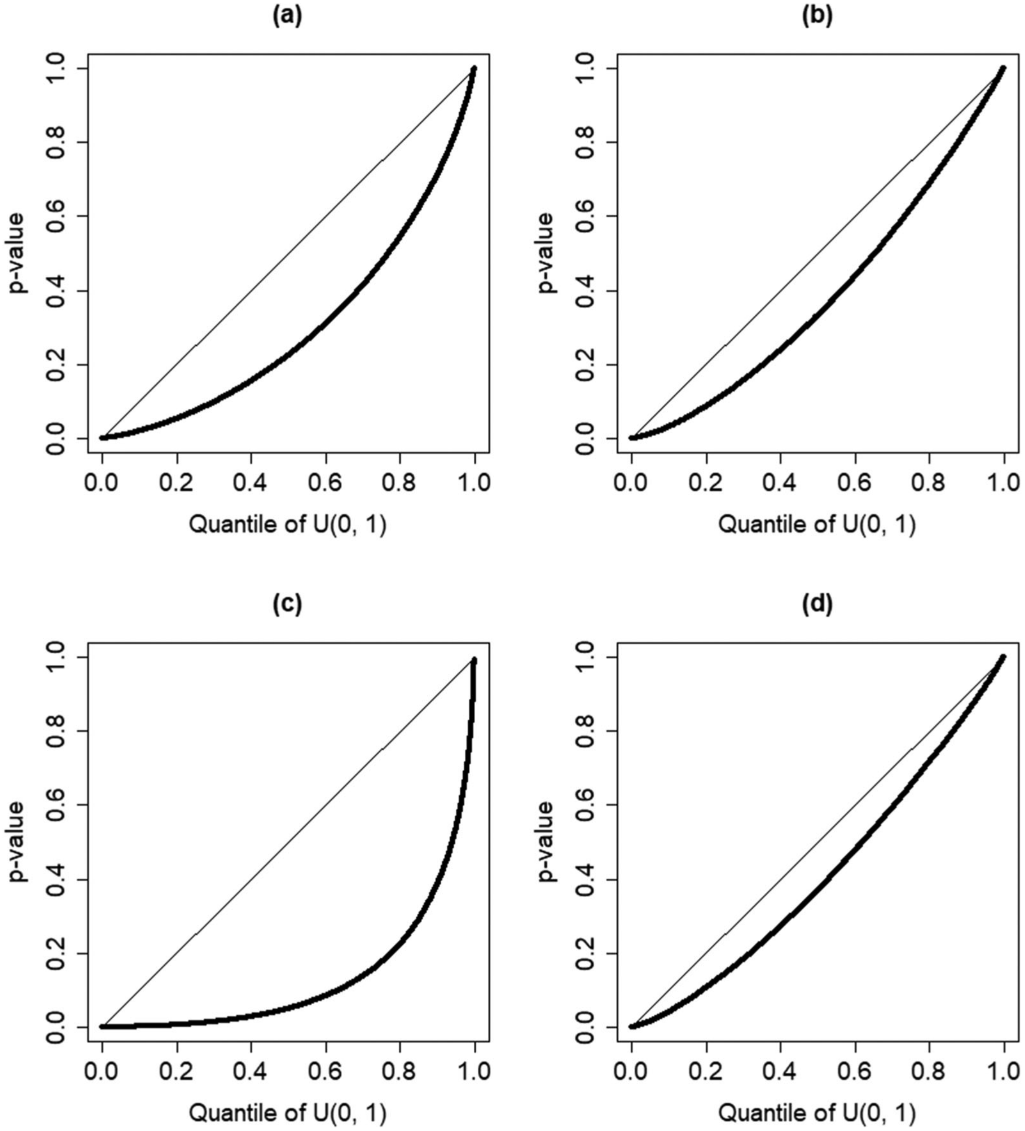

Figure 4. Quantiles of observed *p*-values, at testing $H_0: \lambda_2 = 0$, versus quantiles of the uniform distribution. Results are based on 100 000 simulated datasets following Model (14) with $\lambda_1 = 10$ and $\lambda_2 = 2$. (a) F_R test, $\sigma_S^2 = 0$; (b) *SPB* test, $\sigma_S^2 = 0$; (c) F_R test, $\sigma_S^2 = 0.04$; d) *SPB* test, $\sigma_S^2 = 0.04$.

F_R test, provided there are sufficiently many replicates. This is in contrast to the *SPB* test, which may give non-significant results also in studies with very large numbers of replicates, specifically when genotype-by-environment interaction is purely random. The F_R test is used for determining if the residual interaction mean square is significantly larger than the error mean square. Using the F_R test, the aim is to separate the part of the estimated genotype-by-environment interaction that is not a result of insufficient replication from the part that might be.

In cases with no random genotype-by-environment interaction at all, but just variance between replicates, hypotheses coincide. Based on the results of this study, as long as the two methods perform similarly with regard to Type I error, the F_R test is preferable to the SPB test when there is no random genotype-by-environment interaction, since the F_R test is more powerful than the SPB test in this situation. The F_R test of the first singular value is equivalent to a test of interaction in analysis of variance. The results of the simulation showed that it is possible to use the F_R test of the first singular value for testing the entire genotype-by-environment interaction also if this interaction comprises random components.

The choice between the two methods is a choice regarding s_{ij} in the equation for Model (8) as signal or noise. If s_{ij} is regarded as signal, i.e., if interest is in interaction including this term, irrespective of it being random or not, then this term should be included in the sum of multiplicative terms, as in Model (9), for which the F_R test can be used. If, on the other hand, s_{ij} is regarded as noise, then the SPB test can be employed for testing the significance of the singular values of Model (8), where s_{ij} is random.

AMMI analyses can be aimed at many questions. For determining which genotype performs best in some specific environment, we recommend the F_R test, because here the s_{ij} term is regarded signal. For this question, it does not matter if interaction is fixed or random. Similarly, for the question of which environment is best for a specific genotype, the F_R test is the preferred method. For the question of how environments should be grouped into mega-environments, however, we recommend the SPB test. Here, the purpose is to divide the environments into groups of environments that are similar with respect to how genotypes perform. These mega-environments will probably be used in the future, for new genotypes, i.e., for others than just those included in the experiment. If such a division is to be sustainable, it is required that the interaction is systematic, not just random. In this case, the s_{ij} term is regarded noise. Similarly, in order to find out which genotypes respond similarly under different environmental conditions, the SPB test is the more relevant of the two. This is so if we want to be able to generalize the conclusions to a population of environments from which the studied environments can be considered as a sample. In practice, the environments are usually experimental stations, and we want to generalize the results to the entire region to which the experimental stations belong. In summary, we recommend the F_R test for narrow inference, which is confined to the investigated genotypes and environments, and we recommend the SPB test for broad inference, which extends beyond the experiment.

As discussed in Section 2, not only the SPB test, but also the F_R test is conditioned on the singular values of the null model being large. This property of both tests calls for using a sequential testing procedure, starting with testing the significance of the first singular value, and proceeding with testing the significance of the following singular values, one at a time, until a non-significant result is obtained.

The F_R test is comparatively robust against nonnormality and heteroscedasticity, especially for testing the first singular value (Piepho 1995). The SPB test, however, is not (Forkman and Piepho 2015). Malik et al. (2018) proposed tests that are similar to the SPB test, but more robust. Thus, if data is nonnormal or heteroscedastic, and the aim of the research is to discern multiplicative terms in the data that have larger singular values than could be expected by random genotype-by-environment interaction, then those tests can be recommended rather than the F_R and SPB tests. If the aim is rather to discern multiplicative terms in the data that have larger singular values than could be expected by random variance between replicates, the resampling-based methods proposed by Malik, Forkman, and Piepho (2019) is, due to its better properties as regard robustness, an even better choice than the F_R test.

Singular value decomposition requires complete matrices, i.e., every genotype must occur in every environment. In multi-environment crop breeding trials, this requirement is often not fulfilled. In official crop variety testing, however, data is usually balanced within years. For analysis of multi-environment trials using singular value decomposition, Gauch and Zobel (1990), Paderewski (2013), García-Peña et al. (2016), Arciniegas-Alarcón et al. (2014), Arciniegas-Alarcón, García-Peña, and Krzanowski (2016), and Arciniegas-Alarcón, García-Peña, and Rodrigues (2020) have proposed methods for imputation of missing values. As regards the F_R statistic, imputation is needed for computation of *MSR*. In case the numbers of replicates varies between the observed combinations of genotypes and environments, an approximate value of R must be used in (12). The error mean square, *MSE*, can be computed although some values are missing. Once values have been imputed, it is possible to use the F_R test as well as the *SPB* test, although it is unclear what effects the imputation may have on the tests' performance, depending on the extent of missing values. Linear mixed models, on the other hand, can usually be fitted even if not all genotypes are included in all environments. For correct inference, however, the missing values must be missing at random (Piepho and Möhring 2006).

The F_R test was introduced in 1995 when the randomized complete block design was more commonly used in crop variety testing. Nowadays more advanced design, such as alpha design (Patterson and Williams 1976), partially replicated design (Cullis, Smith, and Coombes 2006), and row-column design with or without spatial balance (Piepho, Williams, and Michel 2015, Piepho, Michel, and Williams 2018), are probably more common than at that time. With these more advanced designs, it is not obvious which mean square should be used in the denominator of the F_R test. The *SPB* test operates on means and does not need an estimate of the pure within-environment error variance. However, the assumptions of independence and homoscedasticity may be violated when the experimental design is complex. In this case, the AMMI model should be fitted using generalized least squares instead of ordinary least squares (Hadasch et al. 2018). More research is needed on how to perform a *SPB* test in this situation. The fact that the test works well with balanced data suggests that it would be worthwhile to explore options for extensions to the unbalanced case.

In plant breeding, genomic selection utilizes marker information for prediction. A commonly used model is our model (4) with fixed effects of environments and random effects of genotypes and genotype-by-environment interaction, but assuming genetically correlated random effects. The genomic covariance matrix is usually constructed using marker information rather than ancestral information (VanRaden 2008, Montesinos-López et al. 2018). In recent times, machine-learning methods have been explored for prediction in multi-environment trials. This novel research has shown that yield and other traits in multi-environment trials can be successfully predicted using deep learning methods (Montesinos-López et al. 2018; Khaki and Wang 2019), or using deep kernel or Gaussian kernel methods for genomic selection (Crossa et al. 2019). When marker-information is incorporated into a factor-analytic model (Burgueño et al. 2007), the same question about the number of multiplicative terms arises. Information criteria can be used to select the model order (Verbyla 2019). Alternatively, for the purpose of significance testing, one may revert to a fixed-effects model and then, having selected the model order, switch back to the mixed-effects model.

In summary, when the multi-environment trial includes replicates within environments, the researcher may use either the *SPB* test or the F_R test for assessing significance of multiplicative terms of genotype-by-environment interaction in AMMI models. The *SPB* test investigates whether there are patterns in the interaction, i.e., multiplicative terms that are larger than expected by random interaction. The F_R test explores which estimates of multiplicative terms are larger than one might expect due to variance between replicates within genotype-by-environment combinations. The *SPB* test investigates the fixed part of the interaction, whereas the F_R test investigates the total interaction.

ORCID

Johannes Forkman http://orcid.org/0000-0002-5796-0710
Waqas Ahmed Malik http://orcid.org/0000-0001-6455-5353
Hans-Peter Piepho http://orcid.org/0000-0001-7813-2992

References

Arciniegas-Alarcón, S., M. García-Peña, and P. C. Rodrigues. 2020. New multiple imputation methods for genotype-by-environment data that combine singular value decomposition and Jackknife resampling or weighting schemes. *Computers and Electronics in Agriculture* 176:105617. doi:10.1016/j.compag.2020.105617.

Arciniegas-Alarcón, S., M. García-Peña, and W. Krzanowski. 2016. Missing value imputation in multi-environment trials: Reconsidering the Krzanowski method. *Crop Breeding and Applied Biotechnology* 16 (2):77–85. doi:10.1590/1984-70332016v16n2a13.

Arciniegas-Alarcón, S., M. García-Peña, W. J. Krzanowski, and C. d Santos Dias. 2014. An alternative methodology for imputing missing data in trials with genotype-by-environment interaction: Some new aspects. *Biometrical Letters* 51 (2):75–88. doi:10.2478/bile-2014-0006.

Burgueño, J., J. Crossa, P. L. Cornelius, R. Trethowan, G. McLaren, and A. Krishnamachari. 2007. Modeling additive × environment and additive × additive × environment using genetic covariances of relatives of wheat genotypes. *Crop Science* 47 (1):311–20. doi:10.2135/cropsci2006.09.0564.

Crossa, J., J. W. R. Martini, D. Gianola, P. Pérez-Rodríguez, D. Jarquin, P. Juliana, O. Montesinos-López, and J. Cuevas. 2019. Deep kernel and deep learning for genome-based prediction of single traits in multienvironment breeding trials. *Frontiers in Genetics* 10:1168. doi:10.3389/fgene.2019.01168.

Cullis, B. R., A. B. Smith, and N. E. Coombes. 2006. On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological, and Environmental Statistics* 11 (4):381–93. doi:10.1198/108571106X154443.

Digby, P. G. N. 1979. Modified joint regression analysis for incomplete variety x environment data. *The Journal of Agricultural Science* 93 (1):81–6. doi:10.1017/S0021859600086159.

Eberhart, S. A., and W. A. Russell. 1966. Stability parameters for comparing varieties. *Crop Science* 6 (1):36–40. doi:10.2135/cropsci1966.0011183X000600010011x.

Finlay, K. W., and G. N. Wilkinson. 1963. The analysis of adaptation in a plant-breeding programme. *Australian Journal of Agricultural Research* 14 (6):742–54. doi:10.1071/AR9630742.

Forkman, J., and H. P. Piepho. 2014. Parametric bootstrap methods for testing multiplicative terms in GGE and AMMI models. *Biometrics* 70 (3):639–47. doi:10.1111/biom.12162.

Forkman, J., and H. P. Piepho. 2015. Robustness of the simple parametric bootstrap method for the additive main effects and multiplicative interaction (AMMI) model. *Biuletyn Oceny Odmian (Cultivar Testing Bulletin)* 34:11–8.

Gabriel, K. R. 1978. Least squares approximation of matrices by additive and multiplicative models. *Journal of the Royal Statistical Society: Series B* 40:186–96.

Gauch, H. G. 1988. Model selection and validation for yield trials with interaction. *Biometrics* 44 (3):705–15. doi:10.2307/2531585.

Gauch, H. G., and R. W. Zobel. 1990. Imputing missing yield trial data. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik* 79 (6):753–61. doi:10.1007/BF00224240.

Gogel, B. J., B. R. Cullis, and A. P. Verbyla. 1995. REML estimation of multiplicative effects in multienvironment variety trails. *Biometrics* 51 (2):744–9. doi:10.2307/2532960.

García-Peña, M., Arciniegas-Alarcón, S., Krzanowski, W., and Barbin, D. 2016. Multiple imputation procedures using the GabrielEigen algorithm. *Communications in Biometry and Crop Science* 11 (2):149–63.

Hadasch, S., J. Forkman, W. A. Malik, and H. P. Piepho. 2018. Weighted estimation of AMMI and GGE models. *Journal of Agricultural, Biological and Environmental Statistics* 23 (2): 255–75. doi:10.1007/s13253-018-0323-z.

Khaki S., and Wang, L. 2019. Crop yield prediction using deep neural networks. *Frontiers in Plant Science* 10:621. doi:10.3389/fpls.2019.00621.

Malik, W. A., J. Forkman, and H.-P. Piepho. 2019. Testing multiplicative terms in AMMI and GGE models for multienvironment trials with replicates. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik* 132 (7):2087–96. doi:10.1007/s00122-019-03339-8.

Malik, W. A., S. Hadasch, J. Forkman, and H. P. Piepho. 2018. Nonparametric resampling methods for testing multiplicative terms in AMMI and GGE models for multienvironment trials. *Crop Science* 58 (2):752–61. doi:10.2135/cropsci2017.10.0615.

Mandel, J. 1971. A new analysis of variance model for non-additive data. *Technometrics* 13 (1): 1–18. doi:10.1080/00401706.1971.10488751.

Mandel, J. 1995. *Analysis of two-way layouts*. New York: Chapman & Hall.

Montesinos-López, A., O. A. Montesinos-López, D. Gianola, J. Crossa, and C. M. Hernández-Suárez. 2018. Multi-environment genomic prediction of plant traits using deep learners with dense architecture. *G3 (Bethesda, Md.)* 8 (12):3813–28. doi:10.1534/g3.118.200740.

Muirhead, R. J. 1978. Latent roots and matrix variates: A review of some asymptotic results. *The Annals of Statistics* 6 (1):5–33. doi:10.1214/aos/1176344063.

Paderewski, J. 2013. An R function for imputation of missing cells in two-way data sets by EM-AMMI algorithm. *Communications in Biometry and Crop Science* 8:60–9.

Patterson, H. D. 1978. Routine least squares estimation of variety means in incomplete tables. *Journal of the National Institute of Agricultural Botany* 14:401–13.

Patterson, H. D., and E. R. Williams. 1976. A new class of resolvable incomplete block designs. *Biometrika* 63 (1):83–92. doi:10.1093/biomet/63.1.83.

Piepho, H. P. 1995. Robustness of statistical tests for multiplicative terms in the additive main effects and multiplicative interaction model for cultivar trials. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik* 90 (3–4):438–43. doi:10.1007/BF00221987.

Piepho, H. P. 1999. Stability analysis using the SAS system. *Agronomy Journal* 91 (1):154–60. doi: 10.2134/agronj1999.00021962009100010024x.

Piepho, H.-P. 1997. Analyzing genotype-environment data by mixed models with multiplicative terms. *Biometrics* 53 (2):761. doi:10.2307/2533976.

Piepho, H. P., A. Büchse, and K. Emrich. 2003. A hitchhiker's guide to mixed models for randomized experiments. *Journal of Agronomy and Crop Science* 189 (5):310–22. doi:10.1046/j.1439-037X.2003.00049.x.

Piepho, H. P., V. Michel, and E. R. Williams. 2018. Neighbor balance and evenness of distribution of treatment replications in row-column designs. *Biometrical Journal. Biometrische Zeitschrift* 60 (6):1172–89. doi:10.1002/bimj.201800013.

Piepho, H. P., and J. Möhring. 2006. Selection in cultivar trials – is it ignorable? *Crop Science* 46 (1):192–201. doi:10.2135/cropsci2005.04-0038.

Piepho, H. P., E. R. Williams, and L. V. Madden. 2012. The use of two-way linear mixed models in multitreatment meta-analysis. *Biometrics* 68 (4):1269–77. doi:10.1111/j.1541-0420.2012.01786.x.

Piepho, H. P., E. R. Williams, and V. Michel. 2015. Beyond Latin squares A brief tour of row-column designs. *Agronomy Journal* 107 (6):2263–70. doi:10.2134/agronj15.0144.

Plavšin, I., J. Gunjača, R. Šimek, and D. Novoselović. 2021. Capturing GEI patterns for quality traits in biparental wheat populations. *Agronomy* 11 (6):1022. doi:10.3390/agronomy11061022.

Shafii, B., and W. J. Price. 1998. Analysis of genotype-by-environment interaction using the additive main effects and multiplicative interaction model and stability estimates. *Journal of Agricultural, Biological, and Environmental Statistics* 3 (3):335–45. doi:10.2307/1400587.

Shukla, G. 1972. Some statistical aspects of partitioning genotype-environmental components of variability. *Heredity* 29 (2):237–45. doi:10.1038/hdy.1972.87.

Smith, A., B. Cullis, and A. Gilmour. 2001. The analysis of crop variety evaluation data in Australia. *Australian & New Zealand Journal of Statistics* 43 (2):129–45. doi:10.1111/1467-842X.00163.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. *Journal of Dairy Science* 91 (11):4414–23. doi:10.3168/jds.2007-0980.

Verbyla, A. P. 2019. A note on model selection using information criteria for general linear models estimated using REML. *Australian & New Zealand Journal of Statistics* 61 (1):39–50. doi:10.1111/anzs.12254.

Wright, K. 2021. agridat: Agricultural Datasets. R package.

Yan, W., L. A. Hunt, Q. Sheng, and Z. Szlavnics. 2000. Cultivar evaluation and mega-environment investigation based on the GGE biplot. *Crop Science* 40 (3):597–605. doi:10.2135/cropsci2000.403597x.

Yates, F., and W. G. Cochran. 1938. The analysis of groups of experiments. *The Journal of Agricultural Science* 28 (4):556–80. doi:10.1017/S0021859600050978.