Published online June 21, 2018

RESEARCH

More, Larger, Simpler: How Comparable Are On-Farm and On-Station Trials for Cultivar Evaluation?

P. Schmidt, J. Möhring, R. J. Koch, and H.-P. Piepho*

ABSTRACT

Traditionally, cultivar evaluation trials have been conducted as replicated small-plot, on-station trials at multiple locations and years. To this day, this is the method of choice for cultivar registration trials conducted by official federal institutes. Given a different purpose (e.g., marketing), cultivar evaluation may also be done as on-farm trials with single replicates and fewer plots laid out as large strips. Such trials are often conducted at a larger number of locations. It is not clear how comparable these two trial systems are. Our objective therefore was to compare the precision and accuracy of these two systems using yield data from both on-farm trials and from official on-station trials for winter oilseed rape (*Brassica napus* L.) across 8 yr. We set up multivariate mixed models to analyze the combined dataset of both trial systems and estimate heterogeneous variance components. Furthermore, based on 23 hybrid genotypes common to both datasets, we investigated the genetic correlation between systems and tested for genotype × system interaction effects. The results suggest that on-farm trials are comparable with on-station trials in terms of precision of a single plot, but that there are genotype × system interaction effects prohibiting the comparison of yield estimates for genotypes between systems. One potential explanation for this difference was identified as the system-specific group effect of semidwarf vs. long-strawed genotypes.

P. Schmidt, J. Möhring, and H.-P. Piepho, Biostatistics Unit, Institute of Crop Science, Univ. of Hohenheim, Fruwirthstrasse 23, 70599 Stuttgart, Germany; R.J. Koch, Pioneer Hi-Bred Northern Europe Sales Division, Apensener St. 198, 21614 Buxtehude, Germany. Received 14 Sept. 2017. Accepted 6 Mar. 2018. *Corresponding author (piepho@uni-hohenheim.de). Assigned to Associate Editor Lucía Gutiérrez.

Abbreviations: BLUE, best linear unbiased estimator; BLUP, best linear unbiased predictor; BSA, Bundessortenamt (Federal Plant Variety Office); MET, multi-environment trial; OF, on-farm; OS, on-station; RCBD, randomized complete block design; TPE, target population of environments; VC, variance component; VCU, value for cultivation and use.

Once a new crop cultivar has been bred, its value for cultivation and use (VCU) must be evaluated in cultivar evaluation trials. These trials should be designed to be as efficient as possible and yield valid results. In this context, "valid" means that differences between cultivars (genotypes) found in the trials should represent the actual differences in performance in the farmers' fields.

In many cases (e.g., for German official VCU trials), the method of choice can be described as replicated multi-environment, small-plot, on-station (OS) trials. This trial system usually comprises identical sets of genotypes that are tested at multiple locations and/or across several years, making it a multi-environment trial (MET), where a year × location combination is referred to as an environment. Furthermore, they are OS trials, as all field trials are planned and executed by trained personnel following established protocols and using field trial technology. A major reason for using specialized technology is the relatively small plot size of, for example, ~10 m^2 for winter oilseed rape (*Brassica napus* L.) trials (Bundessortenamt, 2000). German VCU trials are usually laid out as randomized complete block designs (RCBD) with three to four replicates (Fig. 1), as α-designs with three replicates, or as split-plot design with fertilizer treatments

Published in Crop Sci. 58:1–11 (2018).
doi: 10.2135/cropsci2017.09.0555

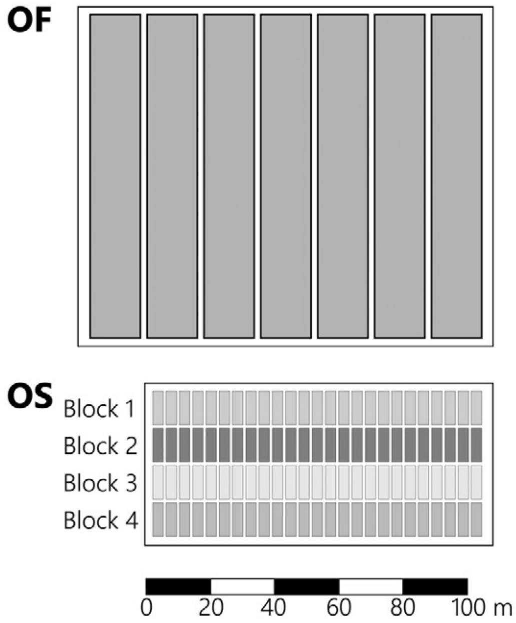

Fig. 1. Schematic layout of an on-farm (OF) trial with seven plots (top) and an on-station (OS) trial with four blocks, each with 25 plots (bottom), in a single environment.

randomized on main plots within complete blocks and genotypes randomized to subplots according to a RCBD (Laidig et al., 2008). Data within years are balanced with respect to genotypes and locations. Every year, the evaluation of a new set of genotypes starts. This set is tested in a 3-yr-cycle, during which some genotypes are discarded each year. These OS trials can be seen as the traditional approach, and they have been widely used and accepted for roughly a century now (Smith et al., 2005).

A more recent development is the use of on-farm (OF) trials (Bradley et al., 1988; Troyer, 1996; Troyer and Wellin, 2009; Yan et al., 2002). For a review of design and analysis of OF trials, see Piepho et al. (2011). Here, many aspects of the experimental design are substantially different from traditional approaches: plots are laid out in strips that can be hundreds of times larger than OS plots, and the number of plots per field site is often confined to <10 (Fig. 1). Therefore, the number of genotypes per environment becomes smaller.

Another difference from OS trials is the reduction to a single replication per environment. Thus, instead of replicating plots per environment, resources in OF trials are shifted towards increasing the plot size and the number of tested environments. In place of field trial technology, the farmer's machinery is used, setting its working width as plot widths for the respective environment. Typically, more genotypes are under investigation than can be grown at single OF field sites. This is handled by distributing genotypes across environments, with only subsets of genotypes tested in any one environment. Thus, the obtained datasets are not only unreplicated within environments, but they are also highly unbalanced with respect to the location × year and genotype × environment classifications.

Given the substantial differences in design between OF and OS trials, it is of interest to compare these two trial systems in terms of precision and accuracy. Yan et al. (2002) summarized it well in their article, which investigated analogous winter wheat (*Triticum aestivum* L.) trial systems in Canada: "Understanding the relationship between the two systems could have a significant impact on cultivar evaluation strategy. If they are highly correlated, either system would be sufficient; if they are complementary, both systems would be helpful; and if they are mutually exclusive, a decision must be made on which system is appropriate for cultivar evaluation."

For this purpose, we here consider German MET data on winter oilseed rape. The Pioneer Accurate Crop Testing System (PACTS) constitutes an OF trial system, which each year comprises ~20 genotypes in ~100 locations with single replicates in Germany. Data from these OF trials were compared with OS trials from official cultivar evaluations conducted by the German Federal Plant Variety Office (Bundessortenamt, Hannover; BSA). These OS trials comprise ~20 locations laid out as RCBD with three to four replicates in Germany each year and at the beginning of a cycle, including ~90 genotypes.

The aim of this article is to evaluate the precision, accuracy, and hence the overall usefulness of OF trials compared with OS trials regarding cultivar evaluation in winter oilseed rape.

MATERIALS AND METHODS

Data

Both METs (OF and OS) were conducted in Germany, and their datasets cover a period of 8 yr (2007–2014).

On-Farm Data

The OF data were obtained in trials overseen and conducted by Du Pont Pioneer in cooperation with local farmers. Their intention was on the one hand to assess the performance of a relatively small number of genotypes and on the other hand to present these genotypes for marketing purposes. All trials were sown with standard drilling technology either by Du Pont Pioneer staff or by the farmers themselves. The sown plot was at least 3 m wider than the harvested area of the plot. Crop husbandry measures such as fertilization and spraying were applied by the farmers according to their local practices. The farmers harvested each strip with their combine separately, whereas the respective yield was assessed and documented by Du Pont Pioneer staff with the company's technology as grain yield corrected for 9% moisture in 10^{-1} t ha^{-1}. This was done via three different weighing methods that all used electronic scales: (i) bagging scales with a weighing bag attached to an aluminum frame on top of a trailer, (ii) weighing plates that the trailer stands on, and (iii) trailers with an integrated weighing system. Within each environment, only a single weighing system and a single combine were used. The dataset comprises

6680 plot records, arising from a median of 100.5 environments per year and 801 environments in total (Table 1). There were 4 to 18 (median = 8) plots per environment, and genotypes were never replicated within an environment. Within each year up until 2012, one out of four possible genotype sets was tested at each environment. Although all four sets included a core set of seven genotypes with greatest commercial relevance in the respective year, three sets also comprised two to five additional and generally new genotypes. Starting in 2013, a fifth genotype set was added, exclusively containing genotypes with a herbicide tolerance. Accordingly, this set did not share genotypes with the other sets and was tested at environments treated with the corresponding herbicide. Finally, and for each environment individually, it was common practice to add one or two genotypes that were of interest to the respective farmer. These are subsequently referred to as external genotypes. Tall and semidwarf genotypes were present in every environment, and randomization plans kept them separate to improve comparisons within growth types. During the first 2 yr, no randomization within those groups was applied. For all the following years, one out of four randomization plans was used to lay out each trial. The median plot size was 710 m^2, with a minimum of 27 m^2 and a maximum of 3360 m^2. On a yearly basis, new genotypes were included, while others were excluded due to selection so that the number of years a genotype was tested ranged from 1 to 7 yr (median = 2). In total, the dataset comprises 110 genotypes (37 Pioneer, including 13 semidwarf; and 73 anonymized external genotypes). The number of environments in which a genotype was tested ranged from 1 to 618 (median = 76) for Pioneer and 1 to 402 (median = 3) for external genotypes, leading to 87.5% of all observations coming from Pioneer genotypes. On average, ~31 genotypes were tested each year (19 when excluding genotypes with only a single observation).

On-Station Data

The OS data were obtained in official trials run on behalf of and supervised by the BSA. Their results serve as the basis for variety registration decisions. Sowing and harvest were done using standard field trial technology. Trial plots were separated from neighboring plots by borders of the same genotype either via the plot-in-plot or the double-plot approach. A procedure where plants are physically separated (i.e., *Scheiteln*) was implemented ~2 wk before harvest. *Scheiteln* is necessary, because rapeseed plants of adjacent plots intertwine towards the end of a season. Concerning crop husbandry, it was attempted to recreate the farmers' local practices with the exception of fungicides, which generally were not applied. Semidwarf and tall genotypes were always grouped together. Whenever a semidwarf plot was adjacent to a tall genotype plot, an extra semidwarf plot was planted in between. The latter was not harvested but only served as a buffer. Grain yield was harvested and weighed according to BSA protocols (Bundessortenamt, 2014).

Table 1. Summary of multi-environment trial data for on-farm and on-station datasets. Numbers in parentheses represent medians.

Parameter	On-farm data	On-station data
Years	2007–2014	2007–2014
Data points	6,680	39,246
Total no. of genotypes	110	727
Median plot size (m^2)	710	10
Environments per year	76–121 (100.5)	17–29 (20.5)
Replicates per environment	1	3–4 (3)
Genotypes per environment	4–18 (8)	25–150 (70)
Environments per genotype	1–618 (4)	6–171 (11)

The OS data comprises 39,246 plot records coming from a median of 20.5 environments per year and 171 environments in total (Table 1). The BSA implements a procedure where genotypes are tested and selected in a 3-yr evaluation cycle. Since new genotypes are introduced on a yearly basis, stages of different evaluation cycles overlap and all stages (i.e., first year, second year, and third year) are present each year. Although genotype sets from different evaluation stages may be tested at the same environments, they are tested separately in adjacent trials. All trials were laid out as RCBD with either three or four replicates. The number of tested genotypes per environment ranges from 25 to 150 with a median of 70. In total, the OS dataset contains information on 727 genotypes with a median of 149.5 genotypes tested each year. The number of environments in which a genotype was tested ranged from 6 to 171 (median = 11).

Dataset Comparability

In summary, we have data of winter oilseed rape genotypes obtained within two different cultivar evaluation systems. Grain yield was used as the response, as it is the most important trait. It is important to note that the two datasets do not share a single location, yet both are the result of a cultivar evaluation in recent cultivation periods in the growing region of Germany. Therefore, the target population of environments (TPE) of both systems is comparable. Out of the total of 814 genotypes (801 tall, 13 semidwarf), we found that both datasets share a set of 23 (15 tall, 8 semidwarf) Pioneer genotypes (Fig. 2). External genotypes were anonymized in the OF data. Hence, any further shared genotypes among the external ones could not be identified, and the 23 genotypes will be referred to as the "identified shared" set. All Pioneer genotypes, but not all other genotypes, are hybrids.

Analysis

The OF and OS data were combined into a single dataset and then analyzed in a two-stage analysis (Möhring and Piepho, 2009; Piepho et al., 2012, 2016; Schulz-Streeck et al., 2013). In the first stage, OS data of each environment were analyzed separately with a linear model. Because the OF trials are unreplicated, their environments cannot be analyzed individually, and it is not possible to apply an analogous first-stage analysis. To resolve this problem, an alternative approach (described below) that allows for a joint analysis in the second stage was chosen. In the second stage, two different bivariate mixed models were fitted to analyze data across environments (Möhring and Piepho, 2009; Piepho et al., 2012, 2016; Schulz-Streeck et al., 2013). All analyses were done with ASREML-R (Gilmour et al., 2009) and SAS software 9.4 (SAS Institute, 2013).

First Stage and Weighting

In the first stage of our two-stage analysis, data per environment were analyzed by modeling effects that account for the respective design. As a result, adjusted means (i.e., best linear unbiased

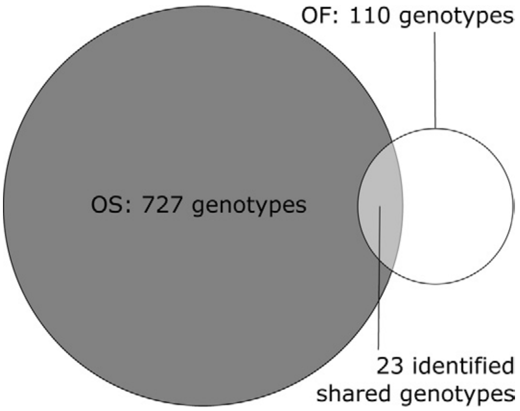

Fig. 2. Venn diagram of genotypes exclusive to and shared between the on-farm (OF) and on-station (OS) dataset.

estimators [BLUEs]) for the genotypes were obtained and could be used as the response in the second stage to perform an analysis across environments. The analysis for each year × location combination was performed for the OS data using the model

$$\mathbf{y} = G + T/R \qquad [1]$$

where $\mathbf{y}$ is the vector of observed plot yields, G represents the genotypes, T represents the trials, and R represents the replicates. The plot error associated with observation $\mathbf{y}$ is not listed in the model for simplicity but is part of the fitted model and was fitted as heterogeneous between trials. Here, we use the notation described in Piepho et al. (2003), where the dot operator (·) defines crossed effects ($A \cdot$ B), the crossing operator (×) defines a full factorial model ($A \times B = A + B + A \cdot B$), and the nesting operator (/) indicates that a factor B is nested within another factor A ($A/B = A + A \cdot B$). The colon (:) is used to separate fixed (first) from random effects (last). Our first-stage model (Eq. [1]) takes all factors (except the plot error) as fixed. G was taken as fixed to obtain a vector of adjusted genotype means ($\bar{\mathbf{y}}_1$), which are submitted to the second stage. T and R were also taken as fixed. When submitting estimates from the first to the second stage, a weighting method was used to allow for variances of and covariances between adjusted genotype means, which can slightly improve precision (Möhring and Piepho, 2009). In this study, we used Smith's weights (Smith et al., 2001, 2005; Damesa et al., 2017), which can be obtained as the diagonal elements of the inverse of the variance–covariance matrix of the adjusted genotype means from the first stage.

For the OF data, each data point for a genotype in an environment is simply taken as the genotype's adjusted mean (i.e., $\mathbf{y} = \bar{\mathbf{y}}_1$). The effect for the main plot in the OF data caused by separating tall and semidwarf genotypes was fitted in the second stage.

Second Stage

One may set up a full model for analyzing MET data across environments in a single system and then extend this to cover both systems using a bivariate approach (Piepho and Möhring, 2011), allowing for heterogeneity of variance between systems for the same type of effect and for correlations between systems for any effect observed for both systems (Przystalski et al., 2008). Below, we will develop different models, starting from the case of a single system and then extending this to two systems. The general approach taken for this extension can be described in three steps:

1. Cross each effect of the single-system model (including the intercept) with a factor S for system, which in our case has two levels (OF and OS).
2. Assume heterogeneity of variance between systems for all effects.
3. For any effect observed for more than one system, allow for a covariance between systems.

In our case, a covariance is needed only when the effect does not involve locations because the systems do not share a single location. This is because locations are nested within systems and any effect crossed with location can occur in only one system. The factor S will be taken as fixed throughout because our interest is in comparing systems.

Analysis Ignoring Growth Types. The standard model for analyzing MET data with the addition of a random main-plot effect across environments in a single system is

$$\bar{\mathbf{y}}_1 = \mu + Y \times L \times G + Y \cdot L \cdot M \qquad [2]$$

where $\bar{\mathbf{y}}_1$ is the vector of adjusted genotype × environment means computed in the first stage, μ is the overall intercept, Y represents the years, L represents the locations, G represents the genotypes, and M represents the main plot (Table 2). Note that $Y \cdot L \cdot M$ is only fitted for the OF trials, where semidwarf and tall genotypes are allocated to two different main plots per trial. In this model, all effects (except μ) are taken as independent and identically distributed random variables with constant variance components (VCs).

As the OF trials involve no replication, we cannot perform a first-stage analysis, meaning the observed plot data is transferred as is from the first to the second stage, formally treating them as the adjusted means obtained from OS trials. We therefore denote these data as pseudoadjusted means. Moreover, we cannot dissect plot errors from the highest order interaction effect with OF data. To deal with this problem at the second stage of analysis, we fix the error variance of the OF data to a tiny value and estimate only the VC for the highest order interaction. This is accomplished by giving all the pseudoadjusted means $\bar{\mathbf{y}}_1$ from the first stage a very large weight (i.e., 100,000). The resulting variance estimate will confound the error variance and the actual interaction variance (see Discussion). Also, the analysis effectively assigns the same weight to each OF trial because a trial-specific error variance cannot be estimated.

Applying Steps 1 to 3 to Eq. [2], we obtain the corresponding model for multiple systems:

$$\bar{\mathbf{y}}_1 = S + S \cdot \left(Y \times L \times G + Y \cdot L \cdot M\right) \qquad [3]$$

where S represents the systems and all other terms are defined as for Eq. [2]. For all random effects, we assume heterogeneity of variance between systems. Additionally, a covariance between systems is allowed for $S \cdot Y$, $S \cdot G$, and $S \cdot Y \cdot G$ (Table 2).

Note that by estimating a covariance for, for example, $S \cdot G$ effects, we are assuming a correlation between $S \cdot G$ effects for the same genotype in the two different systems. The variance–covariance matrix of $S \cdot G$ for a single genotype is

Table 2. Full display of all models used in the second stage of the analysis in the notation described in Piepho et al. (2003).

Model	Full model†
Eq. [2]	$\bar{\mathbf{y}}_1 = \mu + Y \times L \times G + Y \cdot L \cdot M$ $= \mu : Y + L + Y \cdot L + G + Y \cdot G + L \cdot G + Y \cdot L \cdot G + Y \cdot L \cdot M$
Eq. [3]	$\bar{\mathbf{y}}_1 = S + S \cdot (Y \times L \times G + Y \cdot L \cdot M)$ $= S : \underline{S \cdot Y} + S \cdot L + S \cdot Y \cdot L + \underline{S \cdot G} + \underline{S \cdot Y \cdot G} + S \cdot L \cdot G + S \cdot Y \cdot L \cdot G + S \cdot Y \cdot L \cdot M$
Eq. [4]	$\bar{\mathbf{y}}_1 = \mu + Y \times L \times (D/G)$ $= \mu + D : Y + L + Y \cdot L + Y \cdot D + L \cdot D + Y \cdot L \cdot D + D \cdot G + Y \cdot D \cdot G + L \cdot D \cdot G + Y \cdot L \cdot D \cdot G$
Eq. [5]	$\bar{\mathbf{y}}_1 = S + S \cdot [Y \times L \times (D/G)]$ $= S \cdot (\mu + D + Y + L + Y \cdot D + L \cdot D + Y \cdot L \cdot D + D \cdot G + Y \cdot D \cdot G + L \cdot D \cdot G + Y \cdot L \cdot D \cdot G)$ $= S + S \cdot D : \underline{S \cdot Y} + S \cdot L + S \cdot Y \cdot L + \underline{S \cdot Y \cdot D} + S \cdot L \cdot D + S \cdot Y \cdot L \cdot D + \underline{S \cdot D \cdot G} + \underline{S \cdot Y \cdot D \cdot G} + S \cdot L \cdot D \cdot G + S \cdot Y \cdot L \cdot D \cdot G$

† In models, the dot operator (·) defines crossed effects ($A \cdot B$), the crossing operator (×) defines a full factorial model ($A \times B = A + B + A \cdot B$), and the nesting operator (/) indicates that a factor B is nested within another factor A ($A/B = A + A \cdot B$). The colon (:) is used to separate fixed (first) from random effects (last). $\bar{\mathbf{y}}_1$ is the vector of adjusted genotype × environment means computed in the first stage of the analysis, μ is the overall intercept, Y represents the years, L represents the locations, G represents the genotypes, M represents the main plot, S represents the systems and D represents the growth type. For all random effects crossed with the system factor, system-specific variances were fitted. Effects for which an additional covariance between systems is allowed are underscored.

$$\begin{pmatrix} \sigma^2_{S \cdot G-\mathrm{OF}} & \sigma_{S \cdot G-\mathrm{cov}} \\ \sigma_{S \cdot G-\mathrm{cov}} & \sigma^2_{S \cdot G-\mathrm{OS}} \end{pmatrix}$$

Accordingly, we can compute the correlation between the two systems as $\rho = \sigma_{\mathrm{cov}} \big/ \sqrt{\sigma^2_{\mathrm{OS}} \cdot \sigma^2_{\mathrm{OF}}}$. In the extreme, this correlation could be unity if the effects were identical in the two systems, apart from scaling differences reflected by heterogeneity of variance. The same reasoning applies for the corresponding two correlated effects of the other two terms $S \cdot Y$ and $S \cdot Y \cdot G$. Since the correlation is a standardized measure that can be interpreted more intuitively, we always present the correlation coefficients instead of their corresponding covariance estimates. Note that by applying Eq. [3], we obtain estimates for the same variances as if we had analyzed the datasets separately with the standard Eq. [2]. In addition, we obtain three covariance–correlation estimates.

Analysis Accounting for Growth Types. After investigating the two systems' accuracies in the first analysis, we found a systematic difference between the estimates per system for the shared genotypes identified (see Results). This corroborated a previous conjecture of a discrepancy between the two systems regarding the relative performance of different growth types. We therefore conducted a second analysis where we considered a grouping of genotypes (G) into two categories, tall and semidwarf, represented by a fixed factor for growth type (D). For a single system, Eq. [2] extends to

$$\bar{\mathbf{y}}_1 = \mu + Y \times L \times (D/G) \qquad [4]$$

where all terms except D are defined as in Eq. [2] (Table 2). Note that since a main plot in the OF data groups genotypes of the same growth type, $Y \cdot L \cdot M$ is completely confounded with $Y \cdot L \cdot D$, which is why we only use the latter in this analysis. Applying Steps 1 to 3 to Eq. [4], we obtain the corresponding model for multiple systems:

$$\bar{\mathbf{y}}_1 = S + S \cdot [Y \times L \times (D/G)] \qquad [5]$$

where all terms are defined as in Eq. [2–4] (Table 2).

We applied Eq. [3] and [5] to estimate VCs and their SE, best linear unbiased predictors (BLUPs), and adjusted means (BLUEs), as well as to test fixed effects via Wald χ^2 tests. For visual comparisons of the two systems, we plotted BLUPs for $S \cdot G$ from Eq. [3] and for $S \cdot D \cdot G$ from Eq. [5]. Based on $S \cdot D$ and $S \cdot D \cdot G$, which are present in Eq. [5] only, we were able to add growth-type-specific lines: for a given growth type (semidwarf or tall), we fitted the slope for a line in the plot based on Eq. [5] as the first eigenvector obtained from a spectral decomposition of the variance–covariance matrix of $S \cdot D \cdot G$ (Jackson and Dunlevy, 1988). Finally, we computed adjusted means for $S \cdot D$. Thus, for a given growth type, we obtained two means, one for each system. These two means define a point through which the line was fitted in the plot.

Evaluation Criteria

Precision

To allow for a meaningful comparison of precision, VCs and their SE were estimated via Eq. [3] for each system. The main focus of our analysis was then on the size of the plot error variances. As stated above, the error variances in both second-stage models of this article were fixed to estimates from the first stage. Estimates for the overall plot error VC were obtained as follows: (i) for the OS data, the mean of the plot error variances across all environments was estimated by assuming a γ distribution and applying the generalized linear model

$$\log\left[E\left(\sigma^2_{\mathrm{plot}_i}\right)\right] = \lambda \qquad [6]$$

where $\sigma^2_{\mathrm{plot}_i}$ is the plot error variance for the ith environment, E denotes the marginal expectation, and λ is an intercept. The log function was chosen as the link function to link the expected value of the γ-distributed plot error variance estimates to the linear predictor (Cullis et al., 1996; Frensham et al., 1998). Via Eq. [6], we obtained an estimate for the mean plot error variance across environments ($\bar{\sigma}^2_{\mathrm{plot}}$), as well as its SE. (ii) As explained before, for the OF data, no first-stage analysis was conducted and no plot error variances per environment could be estimated. Hence, in the second-stage analysis, we obtained only a single VC that confounds the variances of error and the highest interaction term. We present this estimate as an upper bound to the error variance, denoting it as error variance for simplicity, and display no variance for the highest interaction effect. Additionally, a main plot effect is fitted for the OF data, whose VC represents a second plot error variance. For the overall comparison of error variances, we sum up the confounded OF plot error variance with the OF main plot error variance. This sum then serves as an upper limit of the true

OF error variance and is contrasted with the mean plot error variance obtained for the OS data.

Accuracy

As a measure to compare the accuracies of both systems, we first estimated the genetic correlation ρ_G (i.e., the pairwise correlation of the same genotype in the two systems). It was estimated in three different ways: (i) fitting Eq. [3] with $S \cdot G$ set as random to the full dataset to estimate ρ_{G1}; (ii) fitting Eq. [5] with $S \cdot D \cdot G$ set as random to the full dataset to estimate ρ_{G2}; and (iii) fitting Eq. [3], assuming $S \cdot G$ as random, to a reduced dataset where all data points coming from semidwarf genotypes were excluded to estimate ρ_{G3}. Note that the main plot effects cannot be estimated for this reduced dataset.

Second, to test for $S \cdot G$ and $S \cdot D$ interactions, we slightly modified Eq. [3] by taking S, G, and $S \cdot G$ as fixed and Eq. [5] by taking S, D, and $S \cdot D$ as fixed. Due to the computational burden, in this analysis, it was not possible to use the Kenward–Roger approximation (Kenward and Roger, 1997); instead, the number of denominator df was set to infinity, and thus a Wald χ^2 test was performed (Rao, 1973; Butler et al., 2009, p. 91).

RESULTS

Precision

Estimates for all VCs (and their correlations) are presented in Table 3 and Supplemental Fig. S1. In general, estimates are similar for both systems: the VC for location × year interactions dominates those of years and locations. Likewise, the sizes of the error VC estimates tend to be in between the relatively large environmental VC (Y, L, and $Y \cdot L$) and those of the rather small VC for genotype main and interaction effects (G, $G \cdot Y$, $G \cdot L$, and $G \cdot Y \cdot L$). This overall relationship between VC sizes corroborates findings by Laidig et al. (2008), who estimated VCs with a comparable model for winter oilseed rape BSA data in the period of 1991 to 2006.

Examining the differences, it can be seen that the OS/OF ratio of VCs for Y, L, and $Y \cdot L$ are around or below one, whereas those for G, $G \cdot Y$, and $G \cdot L$ are all >1.3 (Table 3).

Table 3. Variance component (VC) estimates with their SE for grain yield, VC ratios and correlations between VC of winter oilseed rape from German on-station and on-farm trial data in the period 2007–2014 obtained via model (3).

	VC estimate ± SE		On-station/	
Factor†	On-farm	On-station	on-farm ratio	Correlation
	10^{-2} t^2 ha^{-2}			
Y	11.1 ± 6.2	12.0 ± 7.3	1.09	0.51 ± 0.31
L	16.1 ± 2.7	8.2 ± 3.5	0.51	
$Y \cdot L$	32.4 ± 2.1	24.9 ± 3.6	0.77	
G	2.0 ± 0.5	3.1 ± 0.2	1.58	0.66 ± 0.19
$Y \cdot G$	0.7 ± 0.2	1.0 ± 0.1	1.37	0.60 ± 0.27
$L \cdot G$	0.3 ± 0.1	0.8 ± 0.2	2.59	
$Y \cdot L \cdot G$	–‡	3.3 ± 0.2		
$Y \cdot L \cdot M$	3.6 ± 0.2			
Error	5.4 ± 0.1‡	6.7 ± 0.2		

† Y = year, L = location, G = genotype, and M = main plot.

‡ Highest interaction term and error variance are confounded and presented as error variance.

Accuracy

Genetic Correlation

When applying Eq. [3] to the full dataset, ρ_{G1} was estimated to be a moderate 0.66 with a SE of 0.19 (Table 4). In contrast, the estimate for ρ_{G3} is very high (0.97), and the estimate for ρ_{G2} is even approaching unity.

Test of Interaction Effects

As can be seen in Table 5, for both $S \cdot G$ and $S \cdot D$, interaction effects were found to be significant.

DISCUSSION

Precision

Why We Chose the Plot Error Variance as an Evaluation Criterion for Precision

When deciding on a measure to compare the precision of OF and OS trials, several choices are available. One option is half the variance of a difference between two genotype means. For balanced data, this is given by (Talbot, 1984):

$$V_{\mathrm{gm}} = \frac{\sigma^2_{YG}}{n_Y} + \frac{\sigma^2_{LG}}{n_L} + \frac{\sigma^2_{YLG}}{n_Y n_L} + \frac{\sigma^2_{\mathrm{plot}}}{n_Y n_L n_R} \qquad [7]$$

where V_{gm} is the variance of a genotype mean, n_Y, n_L, and n_R are the numbers of years, locations and replicates, respectively, σ^2_{GY}, σ^2_{GL}, and σ^2_{GYL} are VCs for $Y \cdot G$, $L \cdot G$, and $Y \cdot L \cdot G$, respectively, and σ^2_{plot} is the plot error variance. This formula holds true only for balanced data, but it is useful also when VCs were estimated from unbalanced data, because it shows the impact of design variables, such as the number of environments. Thus, comparing V_{gm} estimates from different METs would only be fair if the same amount of temporal and financial resources was available to conduct the trials for the same set of genotypes in the same cultivation area. In other words, irrespective of the VC, it must be clear that n_Y, n_L, and n_R, which are essentially limited by available resources, have a crucial impact on V_{gm}, which is why one would have to include resource information to put V_{gm} from different trials into perspective.

Moreover, it can be shown that using heritability as an indicator for the comparison is limited in a similar manner. This becomes clear when using the ad hoc measure of heritability (Holland et al., 2003; Piepho and Möhring, 2007), given as

$$\bar{H}^2_{\mathrm{Piepho}} = \frac{\sigma^2_{\mathrm{G}}}{\sigma^2_{\mathrm{G}} + 0.5\bar{v}_{\mathrm{gd}}} \qquad [8]$$

where σ^2_{G} is the VC for the genotype main effect and $\bar{v}_{\mathrm{gd}}$ is the mean variance of a difference of two adjusted genotype means. This latter quantity, and hence the heritability in Eq.

Table 4. Genetic correlation (ρ_G) estimates from Eq. [3] and [5] using the full and a reduced dataset.

				Genetic correlation	
Parameter	Genotypes in dataset	Model	Effect†	Estimate	SE
ρ_{G1}	All	Eq. [3]	$S \cdot G$	0.66	0.19
ρ_{G2}	All	Eq. [5]	$S \cdot D \cdot G$	1.00‡	–‡
ρ_{G3}	Semidwarf excluded	Eq. [3]	$S \cdot G$	0.97	0.06

† *S*, system; *G*, genotype; *D*, growth type.

‡ Fixed to 1 in the sense that the non-negativity constraints on variance components led to one of the heterogeneous variances of $S \cdot D \cdot G$ to be fixed to zero, which resulted in the correlation being one, given the nonzero covariance.

[8], can also be computed for unbalanced data. Note that $\bar{H}^2_{\text{Piepho}}$ coincides with the standard broad-sense heritability H^2 in the case of balanced data. Moreover, for balanced data, $0.5\bar{v}_{\text{gd}}$ coincides with V_{gm} in Eq. [5].

Furthermore, the VCs themselves may not allow for a meaningful comparison either. This is because purely environmental VCs are assumed to be similar in both analyses, since they were estimated in the same TPE (i.e., recent cultivation periods in Germany). Moreover, they are irrelevant for genotype mean comparisons, which only depend on VCs comprising the genotypic factor. These latter VCs, on the other hand, do depend on the set of genotypes tested, which cannot necessarily be assumed to be identical for the two datasets. The OF data used here mainly contain genotypes bred by a single breeding company, whereas the OS trials evaluate virtually all genotypes intended for registration in Germany and thus represent a much broader gene pool. This is also a reasonable explanation why the OS/OF ratios for the genotypic VCs are all larger than one. In other words, concerning our pursued comparison, we have purely environmental VCs on one hand, which are assumed to be similar and not directly relevant for cultivar evaluation. Genotypic VCs, on the other hand, are indeed relevant for yield performance comparisons, but contrasting them with the chosen dataset is not necessarily meaningful due to the different composition of tested genotype sets.

Conversely, the plot error variance is an appropriate indicator of a trial's precision, as it provides a measure for the trial's precision for individual plots. Restricted maximum likelihood estimators (Searle et al., 1992) of VCs typically have little bias, and in the present case, because of the large size of the datasets and the small number of fixed effects, the bias is not expected to strongly depend on the number of observations and replicates. Thus, this comparison among the two trial systems is feasible despite the differences in MET design.

Table 5. *F*-tests of fixed effects for Eq. [3] taking system (*S*), genotype (*G*) and $S \cdot G$ as fixed and for Eq. [5] taking *S*, growth type (*D*) and $S \cdot D$ as fixed. The number of denominator df was set to infinity, and thus a Wald χ^2 test was performed.

Model	Effect	df	*F*-value	*p*-value
Eq. [3]	S	2	2375.1	<0.001
	G	813	3883.4	<0.001
	$S \cdot G$	22	40.0	0.011
Eq. [5]	S	2	989.8	<0.001
	D	1	0.4	0.556
	$S \cdot D$	1	14.0	<0.001

Comparing Plot Error Variances

As stated above, for unreplicated trials it is impossible to estimate separate effects for the error and the highest interaction in the respective model. Instead, we obtain a single confounded effect and a corresponding confounded VC. Additionally, the VC for the main plot effect essentially represents a second error variance. This VC is also confounded because of the experimental design, more specifically due to the fact that an environment's main plot always groups all genotypes of the same growth type. Since the focus lies on the error variance, we can paraphrase this situation by saying that for Eq. [3], this means that (i) the OF error variance is potentially inflated by the VC of $S \cdot Y \cdot L \cdot G$, and (ii) the OF main plot VC also includes the VC of $S \cdot Y \cdot L \cdot D$. This implies that by taking the sum of the plot error variance ($5.4\ 10^{-2}$ $\text{t}^2\ \text{ha}^{-2}$) and the main plot VC ($3.6\ 10^{-2}$ $\text{t}^2\ \text{ha}^{-2}$) for OF data, we obtain an upper limit for the true error variance ($9.0\ 10^{-2}$ $\text{t}^2\ \text{ha}^{-2}$). This upper limit is 34% larger than the estimated plot error variance for OS trials ($6.7\ 10^{-2}$ $\text{t}^2\ \text{ha}^{-2}$). Note, however, that since the former is an upper limit, it would only apply if the respective confounded VCs (i.e., $S \cdot Y \cdot L \cdot G$ and $S \cdot Y \cdot L \cdot D$) were zero. This seems particularly unlikely for Eq. [3], where the corresponding $S \cdot Y \cdot L \cdot G$ VC estimate for OS data was $3.3 \pm 0.2\ 10^{-2}$ $\text{t}^2\ \text{ha}^{-2}$.

Therefore, the results indicate that these two systems have comparable precision regarding a single plot. Keep in mind, however, that (i) as long as the OF trials only use a single replicate, the confounded VC will inevitably be included in the residual variance, which is expected to be 34% larger than that of OS trials, and (ii) OS trials have three to four replicates and are therefore more precise for a single trial.

Accuracy

Although comparing the precision of both trial systems is useful, this comparison misses out on the question whether there are systematic differences between their genotypic evaluations, which would adversely affect the accuracy. Therefore, an accuracy comparison serves to assess whether the two systems evaluate genotype performances in the same manner or not. Note, however, that all results for the accuracy comparisons are based on—and thus are limited to—the 23 identified shared genotypes, which are all Pioneer hybrids (15 tall hybrids and 8 semidwarf hybrids).

As stated above, a genetic correlation estimate close to one would indicate that the evaluation of genotypes in both systems leads to equivalent outcomes. The estimate of the genetic correlation ρ_{G1} from Eq. [3], however, is only 0.66 (± 0.19). Combined with the significant test results for the $S \cdot G$ interaction effects (again using Eq. [3], but taking S, G, and $S \cdot G$ as fixed), this suggests that the two systems do not generally evaluate genotypes in the same way.

It is very striking by comparison, however, that the estimate for ρ_{G2} from Eq. [5] is very close to unity, and furthermore, that the $S \cdot D$ interaction effects in the modified Eq. [5] were found to be significant as well. To verify this considerably larger correlation estimate, we reparameterized the variance–covariance structure of $S \cdot D \cdot G$ by omitting its covariance and adding a random $D \cdot G$ main effect instead. As expected, the VCs for $S \cdot D \cdot G$ were estimated as zero for both systems, which confirmed the high genetic correlation. Furthermore, when fitting Eq. [3] to a dataset where all semidwarf genotypes were excluded, the estimate for ρ_{G3} was also close to one.

Note that by creating Eq. [5] according to Steps 1 to 3, we allowed for heterogeneous variance and a covariance between genotypes of different systems, but not between different growth types. We did consider including both variance structures by defining the variance–covariance structure for genotypes as the Kronecker product of unstructured correlation matrices for D and S, respectively. We found, however, that the Akaike information criterion of this model (−61491.99) is larger than that of our presented Eq. [5] (−61502.82), which led us to the decision against the former. Nevertheless, we would like to point out that this disregarded model provided estimates for the genetic correlation between systems for each growing type, respectively, both being >0.98. Furthermore, it could be argued that heterogeneous genetic variances between environments should be allowed (e.g., by applying factor-analytic variance structures). Due to the large number of environments in our dataset (>900), however, such an approach would lead to just as many additional VCs and to a relatively complex model that may be difficult to fit. Thus, we decided against such an approach to keep the focus on the heterogeneity between systems.

Altogether, this suggests that both systems do not evaluate the identified shared genotypes in a similar manner; there are notable interaction effects between the systems and growth types (Fig. 3). There can be multiple causes for this difference, as each system is a composition of a large number of measures and the two systems differ in several of them—not only regarding plot sizes, number of environments, and machinery, but also, for example, that fungicides are often applied in OF trials but never in OS trials. Thus, it is principally impossible to identify one or even multiple measures as the underlying biological cause for this interaction. This suggests that additional investigations and/or experiments on the issue may be worthwhile. Furthermore, it is not clear which system (if any) estimates the correct yield. On one hand, it is possible that OS trials underestimated semidwarf or overestimated tall hybrid genotypes compared with OF trials. On the other hand, OF trials could have overestimated the semidwarf hybrids or underestimated the tall hybrid genotypes compared with OS trials. Furthermore, note that the ambiguity of these statements is due to fact that we cannot know which of the two systems is closer to the truth, simply because we do not know the true genotypic performances. That being said, it might just as well be that the truth lies in between the findings of the two systems, but this issue is not the main concern here. Note also that only OS trials include nonhybrid tall genotypes, and that these nonhybrid tall genotypes were always separated from hybrid tall genotypes (until 2009 via border plots, as well as via the plot-in-plot or double-plot approach; starting in 2009, via plot-in-plot or double-plot).

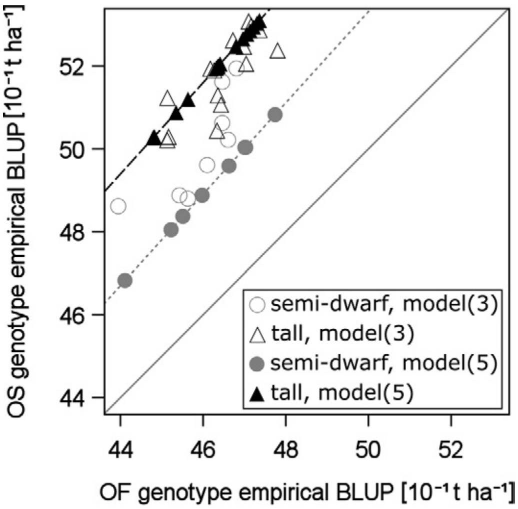

Fig. 3. System-specific best linear unbiased predictors (BLUPs) of winter oilseed rape grain yield for 23 identified shared genotypes estimated via Eq. [3] (unfilled symbols) and via Eq. [5] (filled symbols). Different symbols indicate whether a genotype is tall (triangles) or a semidwarf (circles). The solid reference line passes through the origin and has a slope of one. The grey dotted (semidwarf) and black dashed (tall) lines each pass through a point given by the adjusted system × growth type means and have a slope defined by the first eigenvector calculated from the estimated variance–covariance matrix of $S \cdot D \cdot G$ in Eq. [5]. OS, on-station; OF, on-farm.

The quintessence from these results is that there are strong indicators of a systematic difference between the two systems when collectively evaluating semidwarf and tall hybrid genotypes, yet it must also be realized that when either a system-specific group effect accounting for these differences in the evaluation of semidwarf and tall genotypes was implemented (ρ_{G2}) or tall genotypes were analyzed separately (ρ_{G3}), the genotypic correlation estimates were almost one and thus hybrid genotype evaluations of both systems were very similar. This suggests that one cause for the difference between the systems is related to a systematic factor, the growth type.

Experimental Design Comparison

It can be argued that—metaphorically speaking—the overall intention of this article is to compare apples and oranges, because the two systems serve different purposes and there are several important differences between them. We understand and partly agree with this criticism, since there are indeed multiple differences between OF and OS trial systems that are not only considerable, but conceptual. As a result, any attempt to compare the two systems may shift the question of how these systems should be compared to whether they should be compared at all. At this point, however, we would like to argue that even though the comparison may not always be straightforward, it is nevertheless justified, valid, and worthwhile. To give deeper insight into some of the major differences between the systems' concepts, they are discussed individually below.

Unreplicated vs. Replicated Data

On-station trials are and have been the method of choice for the major part of agricultural field trials worldwide, and there are several reasons supporting this tradition. Probably the biggest influence on how to conduct single field trials came from the work of Fisher (1926), which is often quoted and taught as the very base of field trial design. For a detailed exposition of intent and implementation of the three main principles of experimental design (i.e., randomization, replication, and blocking), see Casler (2015). When considering METs rather than single-location trials, however, it is known that in terms of efficiency one should maximize the number of environments instead of the number of replicates per environment (Talbot, 1984; Casler, 2015); this follows directly from inspection of Eq. [7] and maximization for a fixed total number of plots. Nevertheless, Piepho et al. (2011) suggested that replication "should generally be adhered in OF experiments." The underlying reason for this advice is that, without replication, crucial consequences ensue. First of all, analyzing single environments becomes virtually impossible. Instead, environments are now treated as random blocks drawn from a larger TPE, and thus the data can indeed only be analyzed to provide genotypic means across environments. Second, losing a single environment's plot due to, for instance, weather damage means losing all information for the respective genotype in that environment. This, however, is not as devastating as in OS trials, since the number of tested environments is usually larger for OF trials. A third consequence is the confounding of the highest order interaction term with the plot error, which complicates the interpretation of VC estimates and is thus at the heart of this article. Furthermore, in case the VC that is confounded with the error variance is not zero, the analysis inevitably becomes less precise than it would be if they were not confounded. None of these consequences are genuinely desirable, yet it must be clear that they are not necessarily fatal.

In fact, by accepting the restrictions and further endorsing the shift towards more environments and higher practicability, head-to-head comparisons as an unreplicated field trial design approach became popular in plant breeding and genotype testing. With sometimes even the minimum of only two plots per environment comparing two genotypes, head-to-head comparisons can be seen as the extreme case in OF trials and have been applied by plant breeders for >30 yr now (Bradley et al., 1988).

Balanced vs. Unbalanced Data

In OS trials, usually all genotypes available in a given year are tested together in the same trials. By contrast, the OF trials considered in this paper often tested only a subset of all genotypes in each trial, because the capacity of each trial was limited. Using the same set of genotypes and locations each year would yield perfectly balanced data and allow for a conventional analysis via ANOVA techniques, which were indeed favored in the early times (Smith et al., 2005). With today's mixed model approaches and computational possibilities, however, the unbalance of data itself need not be avoided only for the sake of a conventional statistical analysis. As long as it can be assumed that data are missing at random or missing completely at random, mixed models can be used for data analyses (Piepho and Möhring, 2006; Little and Rubin, 2014). At the same time, it must be acknowledged that environments become incomplete blocks in these OF trials, and for a given set of test environments with fixed test capacity, it is useful to optimize the allocation of genotypes to environments regarding environments as incomplete blocks (Piepho et al., 2011). As a result, however, computation of means across environments entails substantial adjustments for environmental effects, which adversely affects the precision of genotype mean comparisons compared with a complete genotype × environment classification. Hence, from a design perspective, it is desirable to test as many genotypes as possible together in the same environments. It should be pointed out that, for OF trials, the number of genotypes tested at a single environment is more limited than for OS trials.

Strip Plot vs. Small Plot

The ability to use the farmers' machinery implies an increased practical relevance of OF trials (Piepho et al., 2011); arguably, compared with OS trials, OF trials often show a broader validity than those drawn solely from "the narrow confines of a research institute setting" (University of Reading, 1998). Another crucial argument is the occurrence of border effects arising from competition (Büchse, 2002). This is especially true when plants in adjacent plots cannot be expected to be similarly competitive (e.g., due to different inherent growing sizes, root formation etc.). As stated above, however, it is common practice in the OS trials to use buffer plots between plots with semidwarf and tall genotypes to overcome this problem. Besides border effects due to competition between plants, there is also the possibility of front border effects. Front borders refer to plot borders adjacent to paths between plots. It can be argued that tall hybrids have a bigger photosynthetically active canopy by being able to lean into paths further than semidwarfs. Naturally, larger plots lead to borders taking up a smaller fraction of the plot area and accordingly diminish the influence of border effects.

Another critical aspect is the harvest of small plots with field trial technology and thus not conventional technology, especially when *Scheiteln* was applied. *Scheiteln* itself poses an unwanted but inevitable manipulation of the plants if not conducted with required care (UFOP, 2016). Moreover, the crop gets pushed down so that relatively low cutting needs to be done for harvest, which can lead to excessive strain on the straw walkers. In contrast, farmers always try to cut oilseed rape as high as possible to efficiently harvest oilseed grains.

Therefore, both differences in the purpose of cultivar evaluation trials and differences in plot sizes require differences between the two systems in the applied methods and technology. It seems probable that these changes play an important role in the observed difference in yield estimates.

CONCLUSION

In this article, we used bivariate mixed model analyses to compare unreplicated OF strip trials with traditional, replicated OS small-plot trials regarding their precision and accuracy for cultivar evaluation. Results indicate a comparable precision for a single plot of the two systems' yield estimates for winter oilseed rape genotypes in Germany. By contrast, we identified systematic growth type × system interaction effects for yield estimations. These interactions could not be attributed to a single cause, yet we found the growth type to be a factor that allowed us to model these interactions. After all, the two systems can currently only be expected to yield equivalent results within but not across growth types.

Conflict of Interest

The authors declare that there is no conflict of interest.

Supplemental Material Available

Supplemental material for this article is available online.

Acknowledgments

We thank Pioneer Hi-Bred Northern Europe for providing the dataset used in this study. This research was funded by the German Research Foundation (DFG Grant PI 377/18-1).

References

Bradley, J.P., K.H. Knittle, and A.F. Troyer. 1988. Statistical methods in seed corn product selection. J. Prod. Agric. 1:34–38. doi:10.2134/jpa1988.0034

Büchse, A. 2002. Optimierung der Versuchstechnik bei Winterraps. UFOP-Schriften 18. Union zur Förderung von Oel-un Proteinpflanzen, Bonn, Germany.

Bundessortenamt. 2000. Richtlinien für die Durchführung von landwirtschaftlichen Wertprüfungen und Sortenversuchen. Landbuch Verlag, Hannover, Germany. http://www.bundessortenamt.de/internet30/fileadmin/Files/PDF/Richtlinie_LW2000.pdf (accessed 9 June 2016).

Bundessortenamt. 2014. Richtlinien für die Durchführung von landwirtschaftlichen Wertprüfungen und Sortenversuchen. Bundessortenamt, Hannover, Germany.

Butler, D.G., B.R. Cullis, A.R. Gilmour, and B.J. Gogel. 2009. ASREML-R reference manual. Release 3.0. Tech. Rep. Queensland Dep. Prim. Ind., Brisbane, QLD.

Casler, M.D. 2015. Fundamentals of experimental design: Guidelines for designing successful experiments. Agron. J. 107:692–705. doi:10.2134/agronj2013.0114

Cullis, B.R., F.M. Thomson, J.A. Fisher, A.R. Gilmour, and R. Thompson. 1996. The analysis of the NSW wheat variety database. I. Modelling trial error variance. TAG. Theor. Appl. Genet. 92:21–27. doi:10.1007/BF00222947

Damesa, T., J. Möhring, M. Worku, and H.-P. Piepho. 2017. One step at a time: Stage-wise analysis of series of experiments. Agron. J. 109:845–857. doi:10.2134/agronj2016.07.0395

Fisher, R.A. 1926. The arrangement of field experiments. J. Min. Agric. Great Britain 33:503–513.

Frensham, A.B., A.R. Barr, B.R. Cullis, and S.D. Pelham. 1998. A mixed model analysis of 10 years of oat evaluation data: Use of agronomic information to explain genotype by environment interaction. Euphytica 99:43–56. doi:10.1023/A:1018395731621

Gilmour, A.R., B.J. Gogel, B.R. Cullis, and R. Thompson. 2009. ASReml user guide. Release 3.0. VSN Int., Hemel Hempstead, UK.

Holland, J.B., W.E. Nyquist, and C.T. Cervantes-Martínez. 2003. Estimating and interpreting heritability for plant breeding: An update. Plant Breed. Rev. 2003:9–112. doi:10.1002/9780470650202.ch2

Jackson, J.D., and J.A. Dunlevy. 1988. Orthogonal least squares and the interchangeability of alternative proxy variables in the social sciences. Statistician 37:7–14. doi:10.2307/2348374

Kenward, M.G., and J.H. Roger. 1997. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 53:983–997. doi:10.2307/2533558

Laidig, F., T. Drobek, and U. Meyer. 2008. Genotypic and environmental variability of yield for cultivars from 30 different crops in German official variety trials. Plant Breed. 127:541–547. doi:10.1111/j.1439-0523.2008.01564.x

Little, R.J.A., and D.B. Rubin. 2014. Statistical analysis with missing data. 2nd ed. Ser. Prob. Stat. Wiley, Hoboken, NJ.

Möhring, J., and H.-P. Piepho. 2009. Comparison of weighting in two-stage analysis of plant breeding trials. Crop Sci. 49:1977–1988. doi:10.2135/cropsci2009.02.0083

Piepho, H.-P., A. Büchse, and K. Emrich. 2003. A hitchhiker's guide to mixed models for randomized experiments. J. Agron. Crop Sci. 189:310–322. doi:10.1046/j.1439-037X.2003.00049.x

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

Piepho, H.-P., and J. Möhring. 2007. Computing heritability and selection response from unbalanced plant breeding trials. Genetics 177:1881–1888. doi:10.1534/genetics.107.074229

Piepho, H.-P., and J. Möhring. 2011. On estimation of genotypic correlations and their standard errors by multivariate REML using the MIXED procedure of the SAS system. Crop Sci. 51:2449–2454. doi:10.2135/cropsci2011.02.0088

Piepho, H.-P., J. Möhring, T. Schulz-Streeck, and J.O. Ogutu. 2012. A stage-wise approach for the analysis of multi-environment trials. Biom. J. 54:844–860. doi:10.1002/bimj.201100219

Piepho, H.-P., M.F. Nazir, M. Qamar, A. Rattu, Riaz-ud-Din, M. Hussain, et al. 2016. Stability analysis for a countrywide series of wheat trials in Pakistan. Crop Sci. 56:2465–2475. doi:10.2135/cropsci2015.12.0743

Piepho, H.-P., C. Richter, J. Spilke, K. Hartung, A. Kunick, and H. Thöle. 2011. Statistical aspects of on-farm experimentation. Crop Pasture Sci. 62:721–735. doi:10.1071/CP11175

Przystalski, M., A. Osman, E.M. Thiemt, B. Rolland, L. Ericson, H. Østergård, et al. 2008. Comparing the performance of cereal varieties in organic and non-organic cropping systems in different European countries. Euphytica 163:417–433. doi:10.1007/s10681-008-9715-4

Rao, C.R., editor. 1973. Linear statistical inference and its applications. John Wiley & Sons, Hoboken, NJ. doi:10.1002/9780470316436

SAS Institute. 2013. Base SAS 9.4 procedures guide: Statistical procedures. 2nd ed. SAS Inst., Cary, NC.

Schulz-Streeck, T., J.O. Ogutu, and H.-P. Piepho. 2013. Comparisons of single-stage and two-stage approaches to genomic selection. TAG. Theor. Appl. Genet. 126:69–82. doi:10.1007/s00122-012-1960-1

Searle, S.R., G. Casella, and C.E. McCulloch. 1992. Variance components. Wiley, New York. doi:10.1002/9780470316856

Smith, A., B. Cullis, and A. Gilmour. 2001. The analysis of crop variety evaluation data in Australia. Aust. N. Z. J. Stat. 43:129–145. doi:10.1111/1467-842X.00163

Smith, A.B., B.R. Cullis, and R. Thompson. 2005. The analysis of crop cultivar breeding and evaluation trials: An overview of current mixed model approaches. J. Agric. Sci. 143:449–462. doi:10.1017/S0021859605005587

Talbot, M. 1984. Yield variability of crop varieties in the U.K. J. Agric. Sci. 102:315–321. doi:10.1017/S0021859600042635

Troyer, A.F. 1996. Breeding widely adapted, popular maize hybrids. Euphytica 92:163–174. doi:10.1007/BF00022842

Troyer, A.F., and E.J. Wellin. 2009. Heterosis decreasing in hybrids: Yield test inbreds. Crop Sci. 49:1969–1976. doi:10.2135/cropsci2009.04.0170

UFOP. 2016. Beiträge zum Sortenprüfwesen bei Öl- und Eiweißpflanzen für die deutsche Landwirtschaft. Union zur Förderung von Öl-und Proteinpflanzen, Bonn, Germany. http://www.ufop.de/agrar-info/erzeuger-info/raps/beitraege-zum-sortenpruefwesen-bei-oel-und-eiweisspflanzen-fuer-die-deutsche-landwirtschaft/ (accessed 8 June 2016).

University of Reading. 1998. On-farm trials: Some biometric guidelines. Stat. Serv. Ctr., Univ. of Reading, Reading, UK.

Yan, W., L.A. Hunt, P. Johnson, G. Stewart, and X. Lu. 2002. On-farm strip trials vs. replicated performance trials for cultivar evaluation. Crop Sci. 42:385–392. doi:10.2135/cropsci2002.0385