Food and Chemical Toxicology 156 (2021) 112517

Contents lists available at ScienceDirect

Food and Chemical Toxicology

journal homepage: www.elsevier.com/locate/foodchemtox

Equivalence tests for safety assessment of genetically modified crops using plant composition data

Jasper Engel [*], Hilko van der Voet

Biometris, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB, Wageningen, the Netherlands

A R T I C L E I N F O

Handling editor: Dr. Jose Luis Domingo

Keywords:
Food safety
Equivalence test
Distribution-wise equivalence
Linear mixed model
Generalized pivotal quantity

A B S T R A C T

The evaluation of compositional characteristics of plants harvested from field trials is an important step in the safety assessment of a genetically modified crop and its derived products for food and feed. The European Food Safety Authority (EFSA) evaluates safety by testing for equivalence between the GM genotype and other genotypes, typically with a history of safe use. Here, a new equivalence test is proposed, which addresses issues with the EFSA test. The method is motivated by a recently proposed equivalence test for analysis of data from animal feeding trials. In order to be suitable for practical safety assessment, the new method has a statistical power set to a desired value, e.g. 95%, by construction. In addition, we assess distributions rather than average values. This way, equivalence limits can also be established when there is limited genotypic variation. The original EFSA equivalence test breaks down in this case. The method is illustrated by its application to data from a field study on maize grain. Simulation studies indicate that the proposed test has appropriate performance characteristics and is competitive with respect to recently proposed alternatives, including the EFSA/EU equivalence test.

1. Introduction

Food safety is a core value in modern societies, and many countries and regions have established regulations to ensure a high level of consumer protection against health hazards related to food (EUGFL, 2002; FQPA, 1996). Foods produced from genetically modified (GM) crops have been introduced on the market since 1994 and are therefore relatively new. Consequently, food safety authorities have installed procedures to assess the safety of new GM crops (Codex, 2008; EC, 2013; EFSA, 2011; FDA, 2020).

The evaluation of the compositional characteristics of plants in field trials has become a standard part of GM crop safety assessment. This evaluation serves as a general screening method against unintended effects of the genetic modification. For this purpose, the plant composition data from the test (T) genotype are compared to those from other genotypes that are not genetically modified, with a history of safe use, typically commercial varieties of the crop. Two types of other genotypes are commonly included in the field trials. Firstly, if available, a genetically close comparator is included as control (C) to estimate as precisely as possible the changes induced by the new genetic trait. Secondly, a range of reference (R) genotypes is included as a background for estimating the effects of existing genetic variation in the crop.

In Europe, a specific protocol for GM crop safety assessment was published by the European Food Safety Authority (EFSA) in 2011 (EFSA, 2011), that subsequently became part of EU law in 2013 (EC, 2013). This protocol specifies the minimum number of reference genotype and sites (environments) that must be included in a field trial. Furthermore, the protocol requires specific statistical methods to be used for the analysis of the compositional data. For investigating the trait effect, the differences between T and C should be tested for significance (difference tests). For investigating the equivalence of T to the R varieties, a specific form of equivalence testing should be performed, where the equivalence limits are in a first step estimated from the field trial data (van der Voet et al., 2011). The results of both types of test lead to seven outcome types, but for practical use EFSA summarises the equivalence results into four categories: I. equivalent; II. equivalent more likely than not; III. non-equivalent more likely than not; IV. non-equivalent. In practice, results in category II are accepted by EFSA, and the distinction between categories III and IV is of less interest. The distinction between categories III and IV is in fact the result of a test for non-equivalence rather than a test for equivalence, which is of no direct regulatory relevance because further investigation is needed anyway.

The focus of this paper is the T-R equivalence test. Since its introduction in 2011, the EFSA method has been used successfully in practice

---

[*] Corresponding author.
E-mail address: jasper.engel@wur.nl (J. Engel).

https://doi.org/10.1016/j.fct.2021.112517
Received 23 April 2021; Received in revised form 3 August 2021; Accepted 15 August 2021

Table 1

Equivalence test methods for plant compositional data that are compared in this paper, or (DesPow0) was the basis for the DesPow test that is introduced in this paper. Symbols and abbreviations: PC = plant composition in field trial; AF – animal feeding study; T = Test genotype; R = Reference genotype (or genotype group); R’ = alternative reference genotype; C= Control genotype; DWE = Distribution-Wise Equivalence; DesPow = desired power; Δ_{TC} = difference between T and C means (at appropriate scale, e.g. log); σ_R = standard deviation of R distribution; σ_E = standard deviation within genotypes; α = significance level of equivalence test; $1 - \beta$ = desired power of equivalence test; n_0 = minimum sample size per genotype.

Acronym	Study type	Description	Characteristics	References
EFSA	PC	• Compare genotype averages of T and R • Two-step approach: 1. Estimate equivalence limits, 2. Use limits for test • Based on fitting two related linear mixed models • Define equivalence limit as ‘outer’ confidence limit of estimated percentile of the distribution of reference genotypes	• Method required by EFSA and EU • Results shown together with T vs. C difference test on a T-C scale • Regulatory choices: field design, α	• (EFSA, 2010, 2011) (EC, 2013) • van der Voet et al. (2011) • Implemented in https://www.efsa.europa.eu/en/applications/gmo/tools
Perc	PC	• Compare distributions of T-R and R-R′ • One-step approach, with fixed value of the equivalence limit • Based on generalized inference • Define equivalence limit as percentile of the distribution of reference genotypes	• Results shown on a T-R scale • Regulatory choices: field design, α • Assumes $\sigma_R \gg \sigma_E$ when setting equivalence limit	• (Vahl and Kang, 2016) • (Kang and Vahl, 2014)
DesPow0	AF	• Compare distributions of T-C and R-R′ • One-step approach • Based on generalized inference • Define equivalence limit by desired power of test, based on simulation for cases without R variation	• Results shown together with T vs. C difference test on a T-C scale • Regulatory choices: field design, α, $1 - \beta$ • Power defined for a simplified test case: $\Delta_{TC} = 0$ and $\sigma_R = 0$	• (van der Voet et al., 2017)
DesPow	PC	• Compare distributions of T-R and R-R′ • Two-step approach: 1. Estimate equivalence limits, 2. Use Limits for test • Based on fitting one linear mixed model • Based on generalized inference • Define equivalence limit by desired power of test, based on simulation where test cases are “just another reference”	• Results shown together with T vs. R difference test on a T-R scale • Regulatory choices: field design, α, $1 - \beta$ • Power defined for tests T that are “just another reference”	• this paper

in scientific opinions from the EFSA GMO Panel.[1] However, regarding the statistical modelling and equivalence testing suggestions for other approaches have been made. An approach based on generalized inference and a fixed equivalence limit was proposed to address the two-step nature of the EFSA method (Kang and Vahl, 2014) and to avoid breakdown of the method in cases without discernible variation between the R varieties (Vahl and Kang, 2016). However, the suggested approaches have some problems of their own, as we will outline in the Results and Discussion sections. Based on the suggested approaches, we here propose an update for the EFSA approach for compositional equivalence testing to integrate the best elements of the method in use and the proposals made. Most importantly, we assess distributions rather than average values, to avoid situations where a lack of genotypic variation between the R varieties would block the calculation of equivalence limits. In order to be suitable for practical safety assessment, we also propose to set equivalence limits based on a principle of ‘desired power’, i.e. such that we control the power of showing equivalence in a simplified scenario where T is “just another reference” genotype.

The newly proposed equivalence test based on desired power (and therefore designated here as DesPow) can be compared to related tests, as summarized in Table 1. DesPow is derived from another desired-power equivalence test for animal feeding studies (van der Voet et al., 2017). This method is abbreviated DesPow0 in Table 1. In animal feeding studies it is not possible to include many references, and therefore this test focused on a comparison of T and C in the light of typical variation between references as observed in previous studies. DesPow0 is therefore a T-C equivalence test, while for compositional data from field studies the current standard EFSA method is a T-R test. Consequently, the new DesPow test is also a T-R equivalence test. In addition, DesPow0 uses a fixed equivalence limit that follows from a specific assumption regarding the amount of variation between the reference genotypes. DesPow omits this assumption and consequently uses an equivalence limit that is estimated from the reference data. In this paper, DesPow is compared to the EFSA method and to a method by Vahl and Kang (Vahl and Kang, 2016), here designated as the Perc method (Table 1). The reason to include the Perc method is that it has many similarities to the new method (in fact it inspired our work), but also has a very different assumption regarding the amount of variation between the reference genotypes when specifying the equivalence limit.

2. Method

2.1. Models

The DesPow equivalence test presented in this paper can be used to compare the Test genotype to the Reference genotypes. The aim of the T-R equivalence test is to test if the dissimilarity in the plant composition values of the genotypes of interest is smaller than an acceptable limit, the equivalence limit. The value for this limit could be based on expert knowledge. In practice, however, such an explicitly specified equivalence limit is typically not available. Therefore, DesPow estimates limits from the plant composition data itself. The outcome of the resulting equivalence test is presented according to an earlier proposed graphical format (van der Voet et al., 2019; van der Voet et al., 2017). An example is shown in Fig. 2 that will be further discussed in the Results section. For each analyte (variable), a significant difference is shown when the confidence interval for the T-R dissimilarity (horizontal black line) does not contain zero. In contrast, equivalence is shown when the confidence interval falls completely inside the equivalence region (green area). In case of an equivalent outcome, it is typically assumed that, with respect to this analyte, any unintended effect is small enough not to be a safety concern. Note that exceeding the estimated equivalence limit does not necessarily imply biological harm. Therefore, the limits should be regarded as screening thresholds and equivalence tests are to be considered as a screening tool. Variables with uncertain equivalence might be considered to need further exposure, hazard or risk characterization.

---

[1] See the collection of Scientific Opinions from the EFSA GMO Panel at https://www.efsa.europa.eu/en/publications/?f%5B0%5D=im_field_subject%3A61906&f%5B1%5D=sm_field_so_type%3Aopinion.

2.1.1. Statistical model

A typical field experiment for safety evaluation consists of comparing T to a chosen direct comparator C and/or a set of commercial R genotypes at several sites. Within each site the varieties are planted together according to a randomized block design. The T and C are included at each site but each R may only be included at a subset of sites. Typical designs include balanced complete block designs, balanced incomplete block designs, partially balanced incomplete block designs and unbalanced incomplete block designs (Kang and Vahl, 2014). For the method proposed in this paper, the direct comparator C does not need to be present, but we include it in the model description below for completeness.

Compositional data are positive (or at least non-negative) concentration values. This may lead to skew distributions if the range of values is relatively close to zero. In order to have data that conform better to standard linear model assumptions such as additivity of effects, normal error distributions and homoscedasticity, it is convenient to work with log-transformed values (note that log-transformation is usually helpful if ranges are relatively close to zero, and do not have much influence if ranges are relatively far from zero, so it is convenient to apply the transformation in all cases; this also allows the same interpretation of results for all cases, e.g. as effects on a ratio scale). This is also recommended by EFSA (EFSA, 2010). As in all data analyses, checks on the linear model assumptions remain necessary also after log-transformation, and in case of severe violations the methodology described hereafter should not be used or only in adapted form.

Let y_{ijl} be the (log-transformed) response of genotype i in block l at site j. Let $n_R,\ n_s$ and n_B denote the number of reference genotypes, sites, and blocks, respectively. Let $n_{R,\ tot},\ n_{T,tot}$, and $n_{C,tot}$ denote the total number of experimental units corresponding to an R, T or C genotype, respectively. The linear mixed model (LMM) $y_{ijl} = \mu + GF_i + B_{lj} + E_{ijl}$is used to analyse the compositional data, where random effects of (reference) genotype and error are represented by $GF_i \sim N(0, \sigma_R^2)$ for $i \leq n_R$, and $E_{ijl} \sim N(0, \sigma_E^2)$, respectively. As usual, these random effects are assumed to be mutually independent. Note that the genotype factor considered here also has fixed levels ($i = n_R + 1,\ n_R + 2$). Therefore, intercept μ represents the mean of the R genotypes. The average differences between T or C on the one hand and R genotypes on the other hand are given by GF_{n_R+1} or GF_{n_R+2}, respectively. The fixed effect of block j at site l is indicated by B_{lj}. Note that the linear mixed model in the EFSA and Perc methods contains random site and block within site effects instead (Kang and Vahl, 2014; van der Voet et al., 2011). However, the number of sites is often limited, which makes it difficult to estimate the variance between sites with sufficient precision. An additional advantage of fixed block effects is that the mathematics of the resulting equivalence test are more straightforward. Therefore, these effects are taken as fixed here, which forces the analysis to use intrablock information only when comparing T to R. As will be shown later, our equivalence criterion also only uses intrablock information. The equivalence test with random site and block effects is discussed in supplementary material 4. Note that analyses based on fixed or random site and block effects are the same for balanced designs since there is no interblock information to recover.

The structure of the LMM is quite akin to the model used by van der Voet et al. (2017) for equivalence testing using data from animal studies. This can be seen by partitioning the data such that the LMM may be written in terms of three models, two with the fixed levels of the genotype factor and the other one with its random levels:

$$y_{ijl} = \begin{cases} \mu_R + R_i + B_{lj} + E_{ijl}, & i = 1, \ldots, n_R \\ \mu_T + B_{lj} + E_{ijl}, & i = n_R + 1 \\ \mu_C + B_{lj} + E_{ijl}, & i = n_R + 2 \end{cases} \tag{1}$$

Fixed effects of the three genotype groups, namely R (all references), T and C are represented by $\mu_R = \mu$, $\mu_T = \mu + GF_{n_R+1}$, and $\mu_C = \mu + GF_{n_R+2}$, respectively. Below, this set of three means will be referred to as the factor genotype-group with symbol G. The random effect corresponding to the R genotypes is represented by $R_i \sim N(0, \sigma_R^2)$. As indicated above, it is assumed to be independent from the residual random effect.

2.1.2. Equivalence testing

Vahl and Kang discuss several equivalence criteria for GM crop safety assessment (Vahl and Kang, 2016). Here, the so-called distribution wise equivalence (DWE) criterion is used to measure the dissimilarity between T and R. DWE studies the differences between these genotype-groups relative to typical differences between any pair of commercial reference genotypes (R and R'), i.e. ordinary acceptable variation. As mentioned earlier, the current EFSA method breaks down when there is a lack of variation between the references, i.e. σ_R^2 is zero. This issue is circumvented when DWE is used to show 'conditional' equivalence (Vahl and Kang, 2016). This means that the differences between genotypes are only assessed within blocks. Similar to the DWE criterion proposed for animal feeding studies (van der Voet et al., 2017), we define the conditional DWE criterion (θ) as follows:

$$\theta = \frac{E\left(y_{Tjl} - y_{Rjl}\right)^2}{E\left(y_{R'jl} - y_{R''jl}\right)^2} = \frac{\Delta_{TR}^2 + \sigma_R^2 + 2\sigma_E^2}{2\sigma_R^2 + 2\sigma_E^2} \tag{2}$$

with indices T, R, R', and R'' representing the test, and any three reference genotypes, respectively. $E(.)^2$ corresponds to the expected squared difference at the plant-level indicated in its argument. The difference between the T and R means is given by $\Delta_{TR} = \mu_T - \mu_R$. Expression 2, is closely related to the criterion labelled 'DWE-C' of Vahl and Kang (Vahl and Kang, 2016). It is also similar to the DWE criterion proposed for animal feeding studies (van der Voet et al., 2017). More details regarding DWE are provided in (Vahl and Kang, 2016).

Large values for θ may indicate a lack of equivalence. An equivalence test based on this DWE-criterion can therefore be expressed as:

$$H_0 : \theta \,/\, EL_\theta \geq 1 \ H_1 : \theta / EL_\theta < 1 \tag{3}$$

where EL_θ corresponds to the equivalence limit.

Following the procedure for equivalence testing for animal feeding studies (van der Voet et al., 2017), we propose to apply equivalence test (3) following the scheme shown in Fig. 1, namely (1) analyse the data using the mixed model, (2) obtain an estimate $\widehat{EL}_\theta$ for EL_θ using simulated upper confidence limits for θ, (3) divide the upper confidence limit θ_{upp} for θ estimated from the data by $\widehat{EL}_\theta$ and test for equivalence by comparing this ratio to 1, and (4) visualization of results. Note that on the θ-scale shown in Fig. 1 no distinction can be made between categories III (non-equivalent more likely than not) and IV (non-equivalent) from EFSA for interpretation of equivalence (van der Voet et al., 2011). For visualization, the output of the equivalence test is expressed on another scale where a distinction between all four categories can be made. This also allows for simultaneous visualization of the outcome of the equivalence test and an associated T-R difference test.

1 Data analysis: Obtain an estimate and one-sided upper confidence limit for $\boldsymbol{\theta}$

The generalized pivotal procedure has been introduced by Weerahandi as a straightforward approach of constructing confidence intervals for a complicated function of model parameters, with frequentist coverage close to the nominal level (Krishnamoorthy and Mathew, 2009; Meeker et al., 2017; Roy and Bose, 2009). The confidence interval follows from appropriate quantiles of the so-called generalized pivotal quantity (GPQ) of the (function of) parameter(s) of interest. The approach has been used successfully to obtain confidence intervals for numerous complex problems, including other safety assessment procedures (Chiu et al., 2013; Kang and Vahl, 2014, 2016; Krishnamoorthy and Mathew, 2009; McNally et al., 2003; Meeker et al., 2017; van der Voet et al., 2017).

The GPQ procedure employs Henderson's method III (for variance estimation) in concert with generalized least squares (for subsequent estimation of fixed effects) for estimation of the parameters in linear mixed model (1) (Searle et al., 1992). The estimators and GPQs of the corresponding parameters are shown in Table 2. Details are provided in supplementary material 1 and 2.

An estimate for θ is readily obtained by plugging in estimates for Δ_{TR}, σ_R^2, and σ_E^2 in expression 2. Similarly, a GPQ for θ is obtained by plugging the GPQs of these parameters into expression 2, i.e.:

$$GPQ_\theta = \frac{\left[GPQ_{\Delta_{TR}}\right]^2 + GPQ_{\sigma_R^2} + 2GPQ_{\sigma_E^2}}{2GPQ_{\sigma_R^2} + 2GPQ_{\sigma_E^2}} \tag{4}$$

Typically, the distributions of the GPQs cannot be expressed analytically. Instead, the empirical distribution of the GPQs is obtained by independently sampling $U_R \sim \mathrm{X}^2_{df_{R|G,B}}$, $U_E \sim \chi^2_{df_E}$, and $Z \sim N(0,1)$ a large number, let's say 10000, of times. We will refer to this number as the number of GPQ samples. For each sampling, the value for GPQ_θ is computed. The accumulated GPQ-values yield the empirical distribution for GPQ_θ. The upper limit, θ_{upp}, of the $100(1-\alpha)\%$ confidence interval for θ is obtained as the upper $100(1-\alpha)$ percentile of the empirical distribution of GPQ_θ. Typically, $\alpha = 0.05$. Supplementary material 6 shows that the coverage of the confidence interval for θ is acceptable for DesPow.

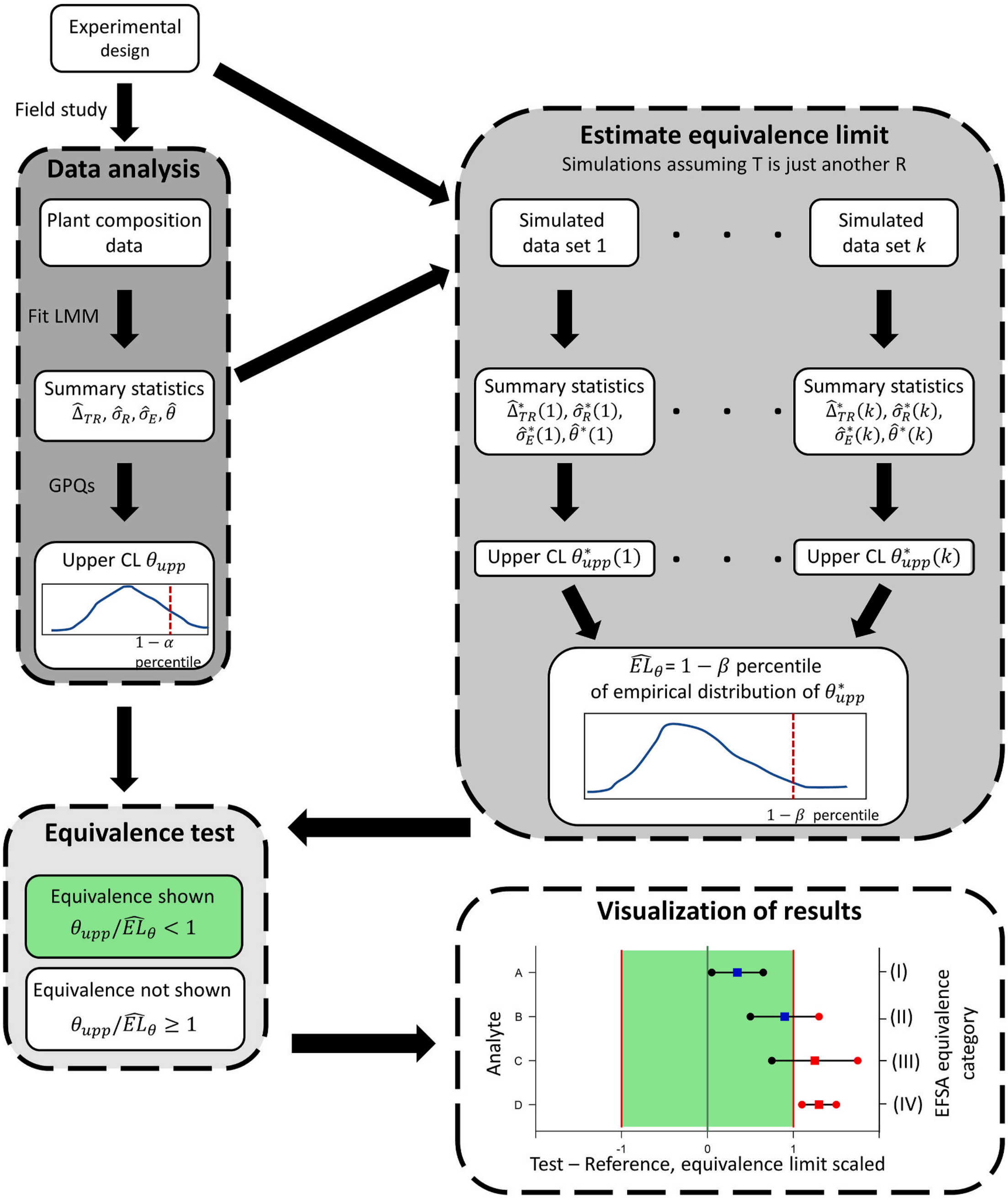
Experimental design
Field study
Data analysis
Plant composition data
Fit LMM
Summary statistics $\widehat{\Delta}_{TR}, \hat{\sigma}_R, \hat{\sigma}_E, \hat{\theta}$
GPQs
Upper CL θ_{upp}
$1-\alpha$ percentile
Estimate equivalence limit
Simulations assuming T is just another R
Simulated data set 1
Simulated data set k
Summary statistics $\widehat{\Delta}^*_{TR}(1), \hat{\sigma}^*_R(1), \hat{\sigma}^*_E(1), \hat{\theta}^*(1)$
Summary statistics $\widehat{\Delta}^*_{TR}(k), \hat{\sigma}^*_R(k), \hat{\sigma}^*_E(k), \hat{\theta}^*(k)$
Upper CL $\theta^*_{upp}(1)$
Upper CL $\theta^*_{upp}(k)$
$\widehat{EL}_\theta = 1-\beta$ percentile of empirical distribution of θ^*_{upp}
$1-\beta$ percentile
Equivalence test
Equivalence shown $\theta_{upp}/\widehat{EL}_\theta < 1$
Equivalence not shown $\theta_{upp}/\widehat{EL}_\theta \geq 1$
Visualization of results
Analyte
A
B
C
D
(I)
(II)
(III)
(IV)
EFSA equivalence category
-1
0
1
Test – Reference, equivalence limit scaled

Fig. 1. *Schematic overview of the DesPow method. The equivalence test is carried out by (1) analysing the data using the mixed model and computing the one-sided upper confidence limit for θ (θ_{upp}) using the generalized pivotal quantity (GPQ) of θ, (2) estimating the equivalence limit EL_θ using upper confidence limits for θ from a large number of simulated data sets,(3) dividing the upper confidence limit θ_{upp} by $\widehat{EL}_\theta$ and test for equivalence, and (4) visualization of results. Note that the diagonal arrows indicate that DesPow uses the experimental design and the estimates of the variance components $\sigma_R = \hat{\sigma}_R$ and $\sigma_E = \hat{\sigma}_E$ when simulating data to set the equivalence limit.*

2 estimating equivalence limits

The estimate of the equivalence limit is based on a principle of 'desired power', i.e. the limit is chosen such that the power of showing equivalence is controlled for a scenario with a predefined similarity between T and R. Given the predefined similarity between T and R, many data sets are simulated according to model (1) following the experimental design of the study. The following values are used for the fixed effects and variance components in (1). Firstly, for a T-R equivalence test, we consider 'safe cases' which are simulated by drawing new T and individual R means from the same distribution in each iteration. In other words, T is 'just another reference' and its mean in the simulation is given by $\mu_T = \mu_R + R_i$. For each data set that is simulated under model 1 a new draw from the distribution of the (safe) test mean is considered, i.e. $\mu_T(k) \sim N(\mu_R, \sigma_R^2)$ for simulation k. Without loss of generality we take $\mu_R = 0$. Finally, the values for the variance components are set to the estimates from part 1, i.e. $\widehat{\sigma}_R^2$ and $\widehat{\sigma}_E^2$ are used.

Typically, the number of simulated data sets M that are simulated according to model (1) is 1000 or 10000. For each simulated data set, an upper confidence limit for θ is obtained using the GPQ approach outlined in part 1. Let $\theta^*_{upp}(m)$ denote the upper confidence limit obtained for the mth simulated data set. An estimate for EL_θ in the DesPow method is obtained as the $1-\beta$ percentile of the confidence limits from the simulated data:

$$\widehat{EL}_\theta = P_{1-\beta}\left(\left\{\theta^*_{upp}(1),\ \theta^*_{upp}(2), \ldots, \theta^*_{upp}(M)\right\}\right) \tag{5}$$

The desired power approach outlined here is similar to the T-C equivalence test for animal feeding trials (van der Voet et al., 2017), but a crucial difference is that in the latter $\mu_T = \mu_C$ was chosen and that no estimates of the variance components were needed in the simulations.

3 Equivalence tests

Equivalence is shown when $\theta_{upp}/\widehat{EL}_\theta < 1$. Thus, the 'desired power' approach for setting the equivalence limit guarantees that equivalence can be shown with probability $1-\beta$ under the scenario of interest. The probability $1-\beta$ is the power of the test for showing equivalence under the prespecified scenario (T is 'just another reference').

4 Visualization of results

For each variable the outcome of the equivalence test is visualized as equivalence limit scaled difference (ELSD) according to the procedure introduced for safety assessment of animal feeding trials (van der Voet et al., 2017; van der Voet and Paoletti, 2019) and generalized to other cases (van der Voet et al., 2019). Briefly, the scale of the DWE-criterion (θ) is not easy understandable. For example, no distinction can be made between positive and negative differences between the genotype-groups. Therefore, the outcome of the equivalence test is re-expressed on the scale of Δ_{TR} according to the strategy of developed for the DesPow0 test (van der Voet et al., 2017). See supplementary material 3 for details. Subsequently, an equivalence standardization step is carried out resulting in the following reformulation of hypothesis (3):

$$H_0 : \Delta_{\mathrm{TR}} \,/\, |EL_{\Delta_{\mathrm{TR}}}| \leq -1 \text{ or } \Delta \,/\, |EL_{\Delta_{\mathrm{TR}}}| \geq 1 \;vs.\; H_1 : -1 < \Delta_{\mathrm{TR}} \,/\, |EL_{\Delta_{\mathrm{TR}}}| < 1 \tag{6}$$

On this standardized scale, which we refer to as ELSD, the equivalence limits are always equal to -1 and $+1$. Note that the standardization requires $|EL_{\Delta_{TR}}| > 0$. Equivalence is shown when an appropriate two-sided confidence interval, derived from GPQ_{ELSD} lies completely within the interval $(-1, +1)$, see supplementary material 3. Note that this interval is symmetric around zero since equivalence test 3 can make no distinction between positive and negative differences.

A difference test comparing T and R can also be also be carried out on the ELSD scale:

$$H_0 : \Delta_{\mathrm{TR}} \,/\, |EL_{\Delta_{\mathrm{TR}}}| = 0 \;vs.\; H_1 : \Delta_{\mathrm{TR}} \,/\, |EL_{\Delta_{\mathrm{TR}}}| \neq 0 \tag{7}$$

A two-sided confidence interval which is given by the $\alpha/2$ and $(1-\alpha/2)$ percentile points of GPQ_{ELSD}, and, hence, not necessarily symmetric around 0, can be used for this purpose: a significant difference is shown if this interval does not include zero. This GPQ-based difference test is very similar to a t-test carried out on the estimates from the LMM. To avoid confusion, we note that the T-R difference test applied here is another test than the T-C difference test used in EFSA (2011).

For simultaneous visualization of the outcome of the equivalence and difference test, their intervals are combined into a single interval, see supplementary material 3. An example, to be discussed in more detail later is shown in Fig. 2. Hypothesis (7) of no difference is rejected when the combined interval does not contain zero. Hypothesis (6) of no-equivalence is rejected when the interval lies fully inside $(-1, +1)$. The combined interval can also be used to summarize the equivalence results in the four categories specified by EFSA. The outcome is equivalent (cat. I) if null hypothesis (6) is rejected. Equivalence (cat. II) is more likely than not if part of the interval, as well as the point estimate, lie inside $(-1, +1)$. Note that the median of GPQ_{ELSD} is used as point estimate. Non-equivalence (cat. III) is more likely than not if part of the combined interval, but not the point estimate, lies inside (−1,+1). Finally, the outcome is non-equivalent if the combined interval fully lies outside (−1,+1) (cat. IV). Note that the null hypothesis (6) of no-equivalence is not rejected in categories II – IV. The null hypothesis (7) of no difference is rejected in category IV, and possibly rejected in categories I-III.

2.2. *Case study: safety assessment of a maize variety*

Application of the DesPow equivalence test for comparative assessment for GM safety is illustrated by analysis of a maize composition data set used before (EFSA, 2010, 2011; van der Voet et al., 2011). To the best of our knowledge this is the only publicly available dataset for GM risk assessment. Briefly, the study involved 13 reference varieties, a GM

Table 2

Estimators and generalized pivotal quantities of the parameters in the conditional DWE criterion (3). Details are provided in supplementary material 1 and 2. Symbols: $ms_{R|G,B}$ = *Henderson's method III mean squares for differences between reference varieties (given block and genotype group);* ms_E = *the error (residual) mean squares;* $df_{R|G,B}$ *and* df_E = *the accompanying degrees of freedom;* n_{eff} = *effective number of replications;* X = *the design matrix of the fixed effects in model (1);* V = *an estimate of the variance-covariance matrix of* y; C_{TR} = *a vector specifying the contrast between genotype-groups T and R;* U_R, U_E, *and* Z = *random variables distributed as* $U_E \sim \chi^2_{df_E}$, $U_R \sim \chi^2_{df_{R|G,B}}$, *and* $Z \sim N(0,1)$; *finally, the variance of the estimator for* Δ_{TR} *is expressed as a linear function of the expected values of the mean squares with coefficients* h_1 *and* h_2.

Parameter	Estimator	Generalized pivotal quantity
σ_E^2	$\widehat{\sigma}_E^2 = ms_E$	$GPQ_{\sigma_E^2} = GPQ_{E(ms_E)} = (ms_E * df_E)/U_E$
σ_R^2	$\widehat{\sigma}_R^2 = \max[0, (ms_{R\|G,B} - ms_E)/n_{eff}]$	$GPQ_{\sigma_R^2} = \max[0, (GPQ_{E(MS_{R\|G,B})} - GPQ_{E(ms_E)})/n_{eff}]$ $GPQ_{E(ms_{R\|G,B})} = (ms_{R\|G,B} * df_{R\|G,B})/U_R$
$\Delta_{TR} = \mu_T - \mu_R$	$\widehat{\Delta}_{TR} = C_{TR}(X^TV^{-1}X)^{-1}X^TV^{-1}y$	$GPQ_{\Delta_{TR}} = \widehat{\Delta}_{TR} + Z\sqrt{h_1 GPQ_{E(ms_{R\|G,B})} + h_2 GPQ_{E(ms_E)}}$

variety and a control variety which were planted in one year at four sites according to a randomized block design with three blocks per site. The study protocol specified that each site was to have been planted with six maize varieties, namely the GM (test), the conventional counterpart (control) and four references. More specifically, three reference varieties were planted at two sites, but the other varieties were planted at one site. Most varieties were replicated three times at each site, but some twice or once. The control was replicated twice at two sites and three times at the other sites. The GM variety was replicated three times at each site.

Analysis focused on 68 analytes in maize grain. Like (van der Voet et al., 2011), fifteen of these analytes, namely 13 fatty acids, furfural and sodium, were discarded because they were not detected in the references, test and control. Following (EFSA, 2010), visual inspection of the log-transformed values of the remaining 53 analytes showed single outliers in four analytes, namely 18:0 Stearic acid, 18:2 Linoleic acid, Copper, and Ferulic acid. In addition, six non-detects (imputed with 0.5*LOD) in 16:1 Palmitoleic acid and one for Phytic acid were marked as outlying. All outliers were omitted from equivalence testing. The log-transformed values of the processed data were analysed by the EFSA, Perc, and DesPow equivalence tests. Following the original procedures, the EFSA and Perc methods were applied using random site and block within site effects. These effects are taken as fixed in DesPow, see model (1). The significance level α of the tests was 5%, the desired power $1-\beta$ was 0.95, the number of GPQ samples was 10000 and M, the number of simulated datasets for calculation of EL_θ, was set to 10000 as well. The results of DesPow, Perc and EFSA were visualized as ELSD according to the procedure described earlier. For the EFSA approach this involved plotting the ratio between the confidence interval for Δ_{TR} and the equivalence limit. In the notation of (van der Voet et al., 2011), pp 16, this corresponds to the ratio between $m_G - m_R \pm lsd(GR;1;97.5)$ and $lsd(GR;2;97.5)$.

2.3. Simulation study

Simulation studies were conducted to investigate the statistical properties of DesPow and to compare its power to the EFSA and Perc method. Data was simulated according to model (1) using some aspects of the experimental design of the case study example. More specifically, a balanced design with 4 sites and 3 blocks per site was used. The number of references n_R was set at 10, 13, 25, 50 or 100, unless mentioned otherwise. The reference variance component σ_R^2 was varied from 0 to 1000. In addition the parameter settings $\mu_R = 0$, $\mu_C = \mu_R + R_i$, and $\sigma_E^2 = 1$ were taken, with R_i indicating a random value from $N(0, \sigma_R^2)$. The settings for μ_T are described below for each simulation study. Block effects were simulated by $B_{lj} = S_j + B_{l(j)}$, with $S_j \sim N(0,4)$ and $B_{l(j)} \sim N(0, 0.25)$. A thousand data sets were simulated for each combination of parameter settings. These were analysed by DesPow using, unless mentioned otherwise, $\alpha = 0.05$ and $1-\beta = 0.95$.EFSA and Perc were also applied with $\alpha = 0.05$.

The first simulation study focused on the DesPow estimate of the equivalence limit. Here, the Test was seen as “just another reference”, i.e. $\mu_T = \mu_R + R_i \sim N(\mu_R, \sigma_R^2)$. For each simulated data set, 10000 GPQ samples and $M = 25000$ were used. In addition, “optimal” values of the equivalence limits were obtained by repeating the procedure to set the equivalence limit shown in Fig. 1 using the true simulation settings for σ_R and σ_E rather than their estimates. Next, the same setting for μ_T was used to assess whether the DesPow approach to set the equivalence limit indeed controlled the desired power of the equivalence test at the desired level $1-\beta$. Here, the number of GPQ samples and the number of data sets to estimate EL_θ (M) were set to 10000.

Finally, the proportion of null hypothesis of non-equivalence rejected by EFSA, Perc, and DesPow was compared for $\sigma_R = \{0, 1, 3, 10\}$ and varying values for Δ_{TR}. The number of references n_R was set at 10, 13, 25 or 50. For values of $\sigma_R = 0$ or 1 the test mean μ_T was varied from 0 to 4 in steps of 0.25. For $\sigma_R = 3$ the mean μ_T was varied from 0 to 10 in steps of 1. Parameter μ_T was varied from 0 to 50 in steps of 5 for the case $\sigma_R = 10$. The number of GPQ samples and number of data sets to set the equivalence limit was set to 10000. The Perc test was also based on 10000 GPQ samples and the 2.5th and 97.5th percentiles of a standard normal distribution were used to set the equivalence limit (Vahl and Kang, 2016). The EFSA test also used 95% equivalence limits (van der Voet et al., 2011).

The number of sites (4) and blocks in site (3) in all simulations was somewhat low. For a limited number of configurations, simulations were repeated for a larger number of sites. Similar results were obtained (not shown).

3. Results

3.1. Case study: analysis of maize compositional data

The results for DesPow T-R comparisons on the ELSD scale are shown in Fig. 2c. Table 3 and Supplementary Table S3 – S5 (supplementary material 5) present additional details. In the Figure, the confidence intervals for the ELSD between the Test and Reference means are indicated by the black horizontal lines. Their intersection with the dark green vertical line at 0 indicates non-significance of a traditional two-sided difference test at the 5% significance level. As many as 36 of the 53 analytes (blue and red squares) showed significant differences. For example, the Carbohydrates concentration is seen to be significantly larger in the Test group than in the Reference group, while the Zinc concentration is significantly lower. The vertical red lines at -1 and $+1$ correspond to the equivalence limits. The area within the equivalence limits is stressed by the green background color. The results of DesPow can be summarized as one of the four EFSA equivalence categories by comparing the confidence interval to the equivalence region (see also Fig. 1). Equivalence is shown (cat. I) when the confidence interval falls completely inside the equivalence region. Equivalence is shown for 42 analytes, including a large number for which a significant difference was detected such as Carbohydrates. The result that equivalence was more likely than not (cat. II) was found for 8 analytes, such as Niacin: less than half the confidence interval is outside the equivalence region. For three analytes, 16:0 Palmitic, Phytic acid and Potassium, non-equivalence was more likely than not (cat. III): more than half the confidence interval is outside the equivalence region. Category IV, non-equivalent, was not observed. In practice, the analytes from categories III and IV are further evaluated for possible concerns.

For further interpretation of the DesPow results, examples of boxplots of the data are given in Fig. 3. Note that these boxplots ignore the block effects in the DesPow model and therefore do not present an exact representation of the model. First, it is interesting to note that Test and Control tend to deviate in the same direction from the References. For 16:0 Palmitic the Test (and the Control) are clearly higher than the range of Reference values and the outcome of non-equivalent more likely than not is unsurprising. Vitamin B1 is an example case where equivalence is clearly shown (and no significant difference is found): the Test values fall inside the range of the Reference values. The other boxplots (Niacin, Phytic Acid, Potassium. Vitamin E) show cases in between, where the ELSD is close to -1 (two slightly lower, two slightly higher). The boxplots for these four cases look similar. It can be observed that the two cases that lead to the outcome non-equivalence more likely than not (Phytic Acid, Potassium) have a relatively high residual variation (low ratio $\hat{\sigma}_R/\hat{\sigma}_E$) which tends to overwhelm the average T-R difference in the plots, but not so much in the statistical tests. For example, the raw data for Phytic Acid show quite some variation between the Reference sample means, but fitting the LMM attributes this variation to residual (sampling) error, whereas the estimated variance component for R was zero. Therefore, the boxplots may be helpful for interpretation, but do not tell the whole story.

Fig. 2 also shows the outcome of the EFSA and Perc procedures. For

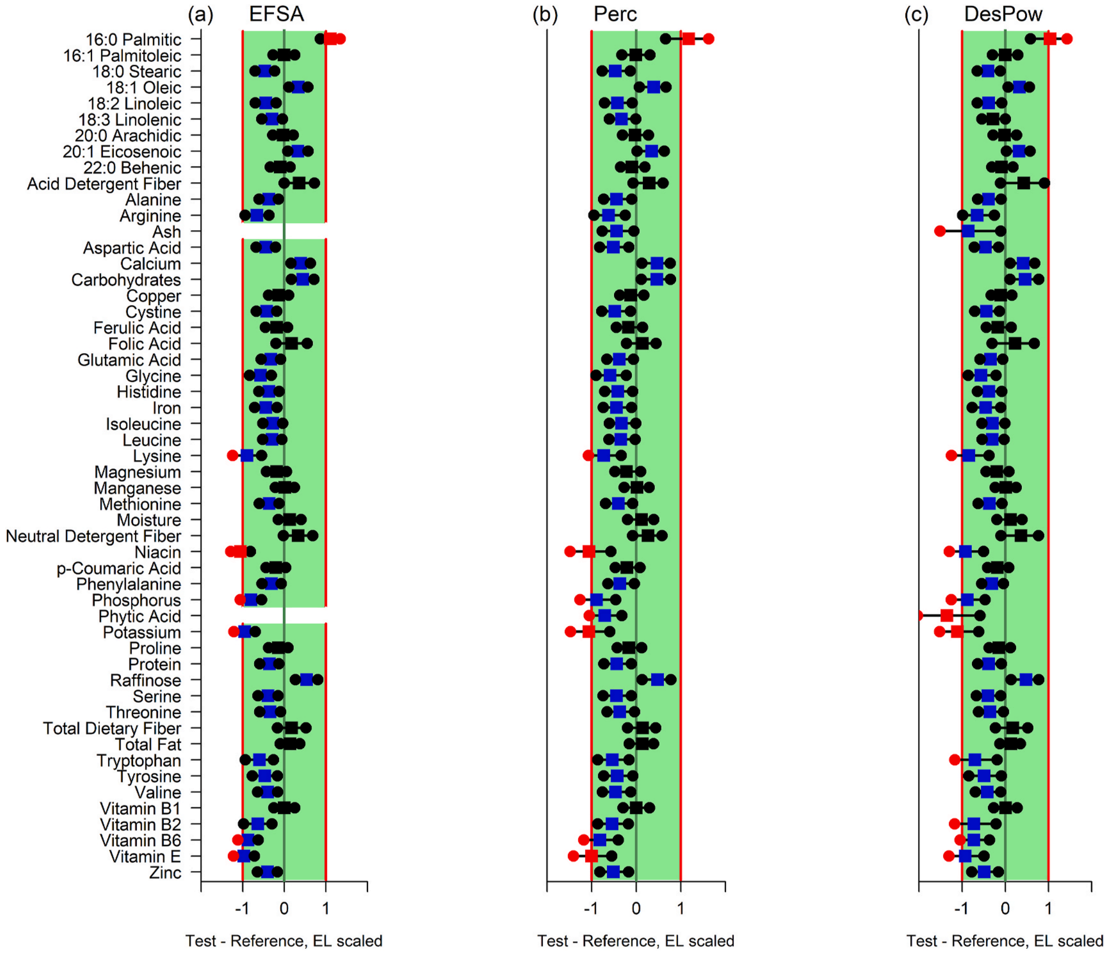

Fig. 2. Example of a) EFSA, b) Perc and c) DesPow applied to maize composition data (EFSA, 2011) for a comparison between Test and Reference genotypes. Significant Test-Reference differences (square symbols, with 95% confidence intervals as bars) are shown as blue or red squares. Differences for which equivalence could not be established are indicated with red dots, such cases are equivalent more likely than not when the point estimate is within the equivalence region (black or blue squares). Analytes in alphabetic order. Note that the EFSA test breaks down for Ash and Phytic Acid because of the estimated zero variance between References.

most analytes these methods reach the same equivalence conclusion as DesPow. Small differences are observed for Ash, Niacin, Phytic acid, Potassium, Tryptophan, Vitamin B2 and Vitamin E. Table 3 shows that Potassium is a borderline case: the equivalence limit of EFSA is slightly larger compared to Perc and DesPow resulting in a change in outcome from cat. III to cat. II. For Niacin the DesPow EL was slightly wider compared to that of EFSA and Perc. As a result, DesPow marks this analyte as “equivalent more likely than not” as opposed to “non-equivalent more likely than not”. Vitamin E was considered another borderline case with very similar equivalence limits. Those of Perc appear to be slightly smaller resulting in a “non-equivalent more likely than not” outcome. More notable differences can be observed for Ash and Phytic acid. For these cases the EFSA method is not applicable because the (estimated) variance between reference genotypes is extremely low (or zero). The Perc approach assumes that the difference between reference genotypes is the dominating source of variation in the data when setting the equivalence limit. Clearly, this assumption is unrealistic for Ash and Phytic acid and the Perc equivalence statements should be deemed unreliable. A similar difference between the methods can be observed for Tryptophan and Vitamin B2, for which a small $\hat{\sigma}_R/\hat{\sigma}_E$ ratio of about 0.5 was observed. Actually, the assumption $\sigma_R \gg \sigma_E$ from Perc is not met for most analytes: in this data set $\hat{\sigma}_R/\hat{\sigma}_E$ values from 0 to 3 were observed, see Table 3 and Supplementary Table S3 – S5.

Table 3 shows the equivalence limit found for each analyte by DesPow, Perc and EFSA. To allow for comparison between the methods, the limits and their confidence intervals are reported on the scale of the ratio between the Test and Reference means. These limits are plotted against each other in Fig. 4a. No clear pattern was observed between the EFSA and DesPow equivalence limits, except for the cases where $\hat{\sigma}_R/\hat{\sigma}_E$ was close to zero and no equivalence limit could be set by EFSA. In contrast, as shown in Fig. 4b, a clear pattern between the Perc and DesPow equivalence limits was observed because DesPow takes the (estimated) σ_R/σ_E ratio into account, whereas Perc effectively assumes $\sigma_R \gg \sigma_E$. The DesPow limit was smaller than the Perc limit for all analytes with small σ_R/σ_E estimates. The limits are similar when the estimated σ_R/σ_E is about one. For increasingly larger σ_R/σ_E estimates a wider equivalence limit was found by DesPow.

3.2. Simulation study

Fig. 5a presents the equivalence limits DesPow as a function of σ_R/σ_E, i.e. the ratio between the reference genotype and residual standard deviation (between-genotype to within-genotype variation), and n_R, the number of reference varieties. The solid lines show the DesPow limit that is obtained using the true simulation setting for σ_R/σ_E. They represent the optimal equivalence limit for the desired power criterion, given the experimental design and the values of the variance components set in the simulation. It can be observed that the DesPow ELs are strongly dependent on σ_R/σ_E as well as n_R. As expected, the equivalence limit increases with σ_R/σ_E, and remains at a lower value when there are more reference varieties. When n_R increases, a lower limit is sufficient to ensure that the power to show equivalence for cases where the Test is

Table 3

EFSA, Perc and DesPow upper equivalence limits calculated on the scale of the ratio of Test to Reference mean of maize composition data (EFSA, 2011). The DesPow point estimate of this ratio, $\exp(\widehat{\Delta}_{TR})$, is given in the column ratio. The EFSA and Perc estimates were almost the same, see Supplementary Tables S4 and S5. The column $\widehat{\sigma}_R/\widehat{\sigma}_E$ is based on the estimates from the LMM (1) of the reference and residual standard deviations. The estimates of the Perc and DesPow upper equivalence limits were obtained as the exponential of the median values from the distribution of $GPQ_{EL_{\Delta_{TR}}}$. In a similar fashion do the values between brackets follow from the 2.5 and 97.5 percentile points from this distribution.

Analyte	Ratio	$\widehat{\sigma}_R/\widehat{\sigma}_E$	EFSA	Perc	DesPow
16:0 Palmitic	1.39	2.01	1.34	1.32 (1.22–1.55)	1.37 (1.26–1.63)
16:1 Palmitoleic	1.00	1.63	1.27	1.26 (1.18–1.47)	1.29 (1.20–1.51)
18:0 Stearic	0.84	2.40	1.47	1.47 (1.32–1.83)	1.56 (1.38–1.99)
18:1 Oleic	1.15	2.98	1.52	1.44 (1.29–1.78)	1.54 (1.36–1.99)
18:2 Linoleic	0.89	1.68	1.30	1.31 (1.22–1.52)	1.34 (1.24–1.58)
18:3 Linolenic	0.92	1.65	1.31	1.27 (1.19–1.45)	1.31 (1.22–1.51)
20:0 Arachidic	0.99	1.75	1.27	1.25 (1.18–1.42)	1.29 (1.20–1.48)
20:1 Eicosenoic	1.11	1.85	1.38	1.35 (1.25–1.59)	1.39 (1.28–1.67)
22:0 Behenic	0.98	2.15	1.27	1.26 (1.18–1.44)	1.30 (1.21–1.50)
Acid Detergent Fiber	1.10	0.40	1.31	1.40 (1.31–1.59)	1.25 (1.20–1.39)
Alanine	0.91	2.24	1.29	1.24 (1.17–1.40)	1.28 (1.20–1.47)
Arginine	0.92	1.06	1.14	1.14 (1.11–1.22)	1.14 (1.10–1.22)
Ash	0.92	0.00		1.21 (1.17–1.27)	1.10 (1.08–1.14)
Aspartic Acid	0.91	2.19	1.23	1.19 (1.13–1.31)	1.22 (1.16–1.36)
Calcium	1.16	2.09	1.45	1.37 (1.25–1.63)	1.42 (1.29–1.73)
Carbohydrates	1.01	1.15	1.02	1.02 (1.01–1.03)	1.02 (1.01–1.03)
Copper	0.94	2.39	1.58	1.66 (1.44–2.21)	1.78 (1.51–2.48)
Cystine	0.93	1.64	1.19	1.17 (1.12–1.27)	1.19 (1.13–1.30)
Ferulic Acid	0.95	1.36	1.32	1.34 (1.25–1.57)	1.34 (1.25–1.56)
Folic Acid	1.09	0.42	1.54	1.72 (1.55–2.08)	1.45 (1.35–1.71)
Glutamic Acid	0.91	2.33	1.32	1.27 (1.19–1.45)	1.31 (1.22–1.52)
Glycine	0.93	1.38	1.13	1.12 (1.09–1.19)	1.13 (1.09–1.21)
Histidine	0.93	1.76	1.21	1.18 (1.13–1.30)	1.20 (1.15–1.33)
Iron	0.88	1.14	1.34	1.34 (1.25–1.56)	1.33 (1.24–1.53)
Isoleucine	0.93	1.89	1.30	1.25 (1.18–1.41)	1.29 (1.20–1.47)
Leucine	0.91	2.40	1.37	1.30 (1.21–1.51)	1.36 (1.25–1.62)
Lysine	0.93	0.72	1.09	1.11 (1.08–1.16)	1.09 (1.07–1.14)
Magnesium	0.97	1.64	1.21	1.18 (1.13–1.29)	1.19 (1.14–1.31)
Manganese	1.01	2.58	1.53	1.41 (1.28–1.72)	1.49 (1.33–1.85)
Methionine	0.91	1.91	1.30	1.26 (1.19–1.44)	1.29 (1.21–1.49)
Moisture	1.01	1.21	1.08	1.09 (1.06–1.13)	1.09 (1.06–1.13)
Neutral Detergent Fiber	1.07	0.56	1.23	1.29 (1.23–1.44)	1.21 (1.16–1.32)
Niacin	0.76	2.41	1.30	1.30 (1.20–1.50)	1.34 (1.24–1.59)
p-Coumaric Acid	0.89	2.30	1.74	1.70 (1.47–2.29)	1.82 (1.54–2.55)
Phenylalanine	0.92	2.49	1.32	1.26 (1.18–1.45)	1.32 (1.22–1.54)
Phosphorus	0.88	1.24	1.17	1.15 (1.11–1.24)	1.15 (1.11–1.23)
Phytic Acid	0.80	0.00		1.38 (1.3–1.510)	1.18 (1.14–1.23)
Potassium	0.88	1.14	1.14	1.13 (1.09–1.20)	1.12 (1.09–1.18)
Proline	0.96	2.08	1.31	1.26 (1.18–1.43)	1.30 (1.21–1.50)
Protein	0.92	2.32	1.24	1.19 (1.14–1.32)	1.23 (1.16–1.37)
Raffinose	1.31	1.30	1.66	1.75 (1.52–2.36)	1.76 (1.53–2.36)
Serine	0.91	1.85	1.25	1.21 (1.15–1.35)	1.24 (1.17–1.41)
Threonine	0.94	1.55	1.2	1.19 (1.13–1.31)	1.20 (1.14–1.32)
Total Dietary Fiber	1.04	0.71	1.25	1.32 (1.24–1.49)	1.26 (1.20–1.41)
Total Fat	1.04	2.36	1.26	1.27 (1.19–1.46)	1.31 (1.22–1.53)
Tryptophan	0.93	0.50	1.13	1.15 (1.12–1.21)	1.10 (1.08–1.16)
Tyrosine	0.88	0.82	1.32	1.35 (1.26–1.55)	1.30 (1.23–1.47)
Valine	0.92	1.78	1.23	1.20 (1.14–1.32)	1.22 (1.15–1.36)
Vitamin B1	1.00	1.86	1.24	1.27 (1.19–1.45)	1.30 (1.21–1.51)
Vitamin B2	0.90	0.55	1.17	1.20 (1.16–1.30)	1.15 (1.12–1.23)
Vitamin B6	0.82	2.15	1.26	1.28 (1.19–1.46)	1.32 (1.22–1.53)
Vitamin E	0.69	1.68	1.46	1.44 (1.31–1.76)	1.49 (1.34–1.86)
Zinc	0.90	1.57	1.31	1.24 (1.17–1.38)	1.25 (1.18–1.41)

'just another reference' remains at the desired level of $1-\beta$.

In practice, the "optimal" DesPow equivalence limit is unknown because σ_R/σ_E is unknown. DesPow uses the estimate $\widehat{\sigma}_R/\widehat{\sigma}_E$ when setting the equivalence limit. Since $\widehat{\sigma}_R/\widehat{\sigma}_E$ is a random quantity, the estimate for the equivalence limit thus obtained will be so as well. Even so, Fig. 5a shows that the DesPow estimates are close to their "optimal" counterparts. Their spread (shaded regions) tends to decrease as n_R increases. Irrespective of n_R, the median estimates (dashed lines) are close to the "true" values. When n_R is relatively small (e.g. $n_R \leq 13$) the spread of the DesPow equivalence limit estimates in Fig. 5a suggests that the uncertainty generated by using the estimate $\widehat{\sigma}_R/\widehat{\sigma}_E$ to set the equivalence limit should not be ignored. However, as shown in Fig. 5b the power of DesPow for showing equivalence when the Test is 'just another reference' was always close to the desired level of $1-\beta$.

The Perc equivalence limit is independent of σ_R/σ_E and n_R, and is indicated by the horizontal green dashed line in Fig. 5a. Its value is $[z^2_{0.975}+1]/2 = 2.42$ (Given our definition for θ, this limit is effectively the same as the equivalence limit for the DWE-C criterion defined by (Vahl and Kang, 2016)). In comparison to the DesPow EL, we observe that Perc uses a higher EL when σ_R/σ_E is low and a lower EL in the opposite case. In this example, the DesPow and Perc EL are similar when σ_R/σ_E is close to 1. This also matches the case study results in Table 3. For small σ_R/σ_E ratios, the DesPow EL is lower than the Perc EL because it accounts for the σ_R/σ_E ratio while Perc assumes $\sigma_R \gg \sigma_E$. For large σ_R/σ_E ratios, the DesPow EL is higher than the Perc EL because it accounts for the limited number of data to preserve the desired power. Note that

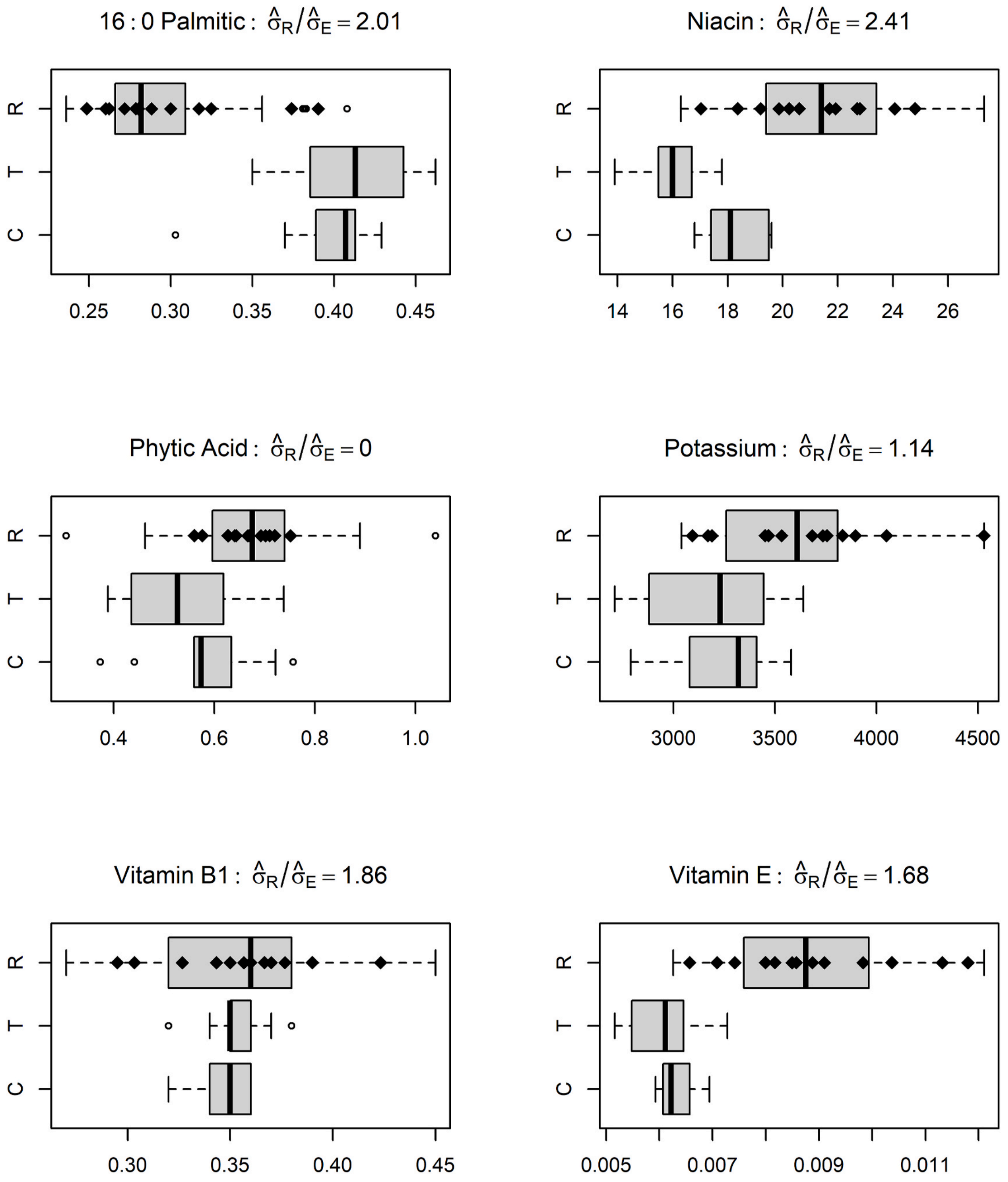

Fig. 3. Boxplots of the data (no log-transformation) for selected analytes. Filled diamonds correspond to the mean values of the Reference genotypes. Open circles correspond to potential outlier observations.

for $\sigma_R \gg \sigma_E$, with very large numbers of references and sites/blocks in an experiment, the DesPow equivalence limit would approach the Perc EL.

Fig. 6 shows the proportion of rejected null hypotheses (of no-equivalence) by Perc, and DesPow against the true difference between the test and reference means. The number of reference varieties was 13. Supplementary Fig. S1 compares the tests for the case of 50 reference varieties. In most cases in Fig. 6, the proportion of rejected hypotheses quickly rises to effectively 100% when the difference decreases to zero. Comparing the three graphs, with increasing values for $\sigma_R/\ \sigma_E$ the proportion of cases for which equivalence is shown remains high for larger deviations Δ_{TR} (note that the horizontal scales are not the same). The differences between the Perc and DesPow equivalence limits can also clearly be observed. When σ_R/σ_E is large (Fig. 6c), DesPow shows equivalence for more cases than Perc due to the desired power criterion. In contrast, when σ_R/σ_E is small (Fig. 6a) the null hypothesis of no equivalence is rejected often by Perc, even for Δ_{TR}-values that are not likely to occur when the Test "is just another reference" (horizontal bar of zero length at the bottom of Fig. 6). As mentioned earlier, in this case the assumptions of Perc when setting the equivalence limit are not met, because it is assumed that $\sigma_R \gg \sigma_E$. The Perc equivalence statements are deemed unreliable when σ_R/σ_E is small. DesPow improves upon Perc in this respect. In particular, we focus on the vertical dotted line in Fig. 6, which indicates Δ_{TR}-values at the optimal DesPow equivalence limit from Fig. 5a. For DesPow, the proportion of rejected null hypotheses at this point was close to or slightly deviating from 5% (α), with values up to 10% occurring for small ratios σ_R/σ_E and number of references n_R. In Fig. 6a, however, a value of roughly 70% was observed for Perc at this point.

Fig. 6 also compares DesPow to EFSA. In general, the curves for DesPow and EFSA are quite similar, although equivalence is shown more

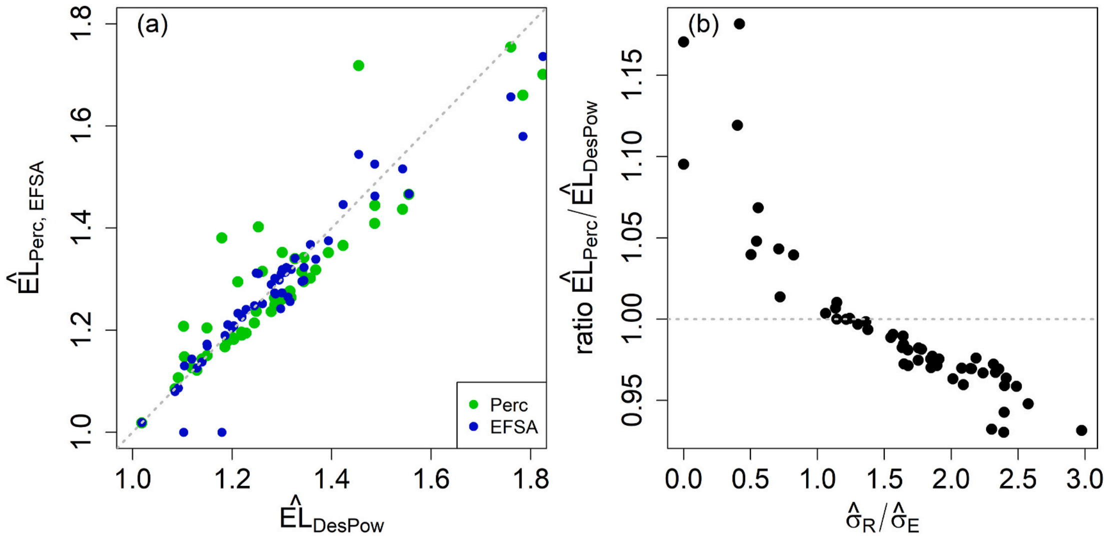

Fig. 4. Comparison of EFSA, Perc and DesPow upper equivalence limit point estimates (Table 3) of the maize composition data (EFSA, 2011). The grey dotted line indicates perfect agreement.

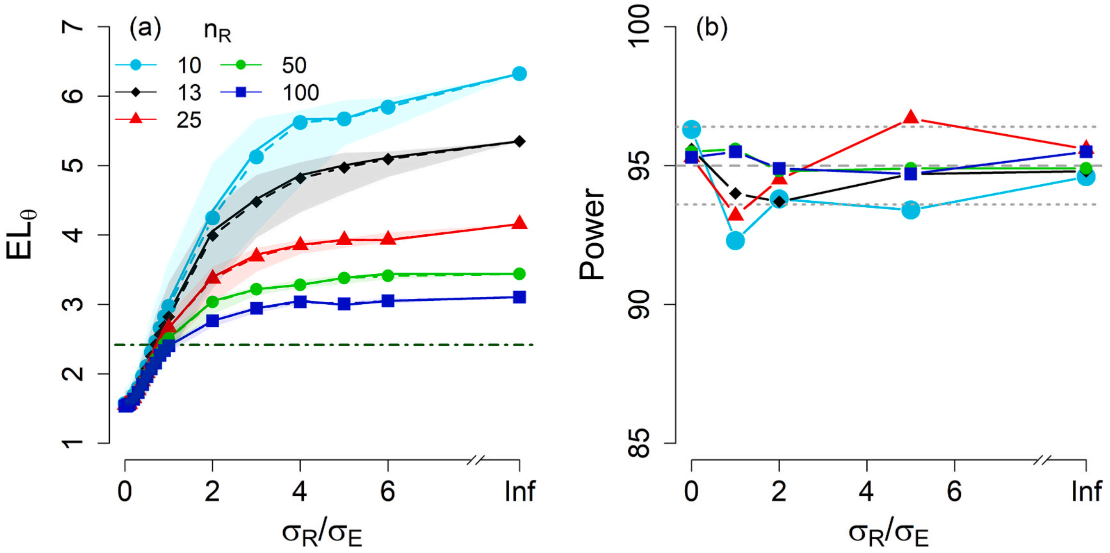

Fig. 5. *(a) Equivalence limits (EL) of DesPow and Perc against σ_R/σ_E, and (b) the power of DesPow against σ_R/σ_E when T is 'just another reference'. In (a), the Perc EL is given by the horizontal dashed line. The solid lines correspond to the optimal DesPow EL obtained using the true simulation setting for σ_R/σ_E. The median DesPow EL estimate is indicated by the other dashed lines. The shaded regions correspond to the 5th to 95th percentiles of these estimates. In (b), the horizontal dotted lines correspond to the 95% prediction interval of an optimal DesPow test with 95% desired power (which has a binomial (n,p) distribution with $n = 1000$ and $p = 0.95$).*

often by DesPow because it controls this power directly when specifying the equivalence limit. When $\sigma_R/\sigma_E = 0$ (panel a) EFSA has extremely low power to show equivalence: the method cannot set the equivalence limit when the estimate of the reference variance component is zero and hence the null hypothesis of no equivalence cannot be rejected. This was also observed for Ash and Phytic Acid in the case study. DesPow clearly improves upon EFSA in this respect.

4. Discussion

4.1. Desired power is a desired criterion

In this paper we introduced a new equivalence test to compare a test genotype to a collection of reference genotypes based on plant composition data. The fundamental principle behind the method is the wish to control the statistical power of the equivalence test, i.e. to know in advance that truly equivalent genotypes will pass the test with a pre-specified desired power $1-\beta$ (e.g. 95%). This is a reasonable approach in a practical biosafety policy to limit the frequency of more detailed investigations of test genotypes that are similar to the reference genotypes (i.e. test genotypes that are "just another reference").

4.2. A minimum number of reference varieties should be set by the regulator

A natural consequence of constructing equivalence tests using the desired power criterion is that the estimated equivalence limits become wider when the number of reference varieties n_R included in a study decreases. Therefore, it is of crucial importance that the regulator specifies a minimum value for n_R for the experimental design. Indeed, such a requirement is already part of the (EFSA, 2011) and (EC, 2013) regulations ($n_R \geq 6$). Higher numbers of reference genotypes seem common in practical datasets (e.g. $n_R = 13$ in the example case study); scanning EFSA opinions of the last two decades, we found 34 out of 38

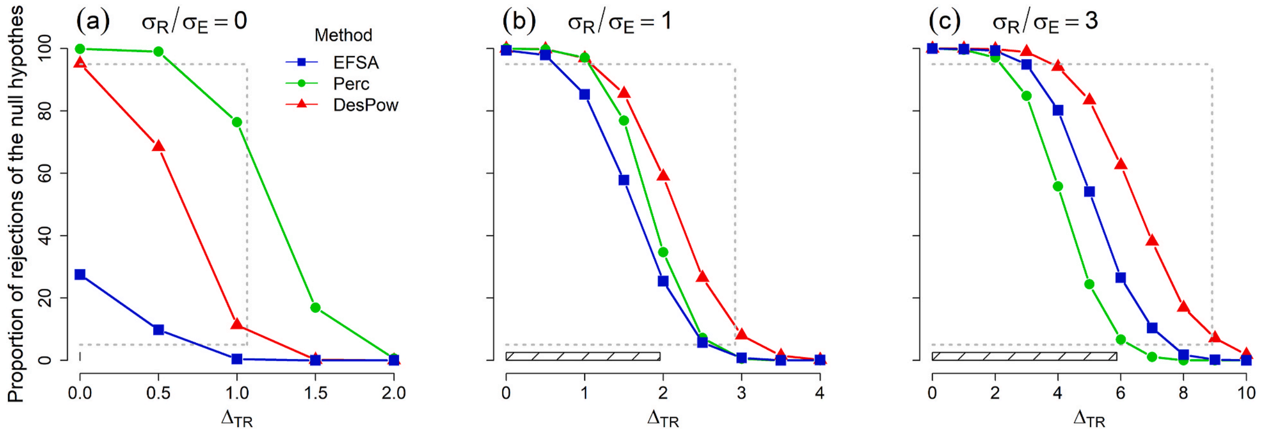

Fig. 6. *Proportion (as %) of rejections of the null hypothesis of no-equivalence by DesPow, Perc and EFSA as a function of the difference between the Test and Reference means (Δ_{TR}) for the case of 13 reference varieties. Three values for σ_R/σ_E are shown, namely (a) σ_R/σ_E=0, (b) σ_R/σ_E=1, and (c) σ_R/σ_E=3. In each panel $\sigma_E = 1$. The horizontal grey lines indicate the significance and desired power levels. The horizontal grey lines indicate the significance level (5%) and desired power level (95%). The vertical grey line indicates the value for Δ_{TR} at the optimal equivalence limit from Fig. 5a, i.e. $|\Delta_{TR}| = \sqrt{EL_\theta(2\sigma_R^2 + 2\sigma_E^2) - \sigma_R^2 - 2\sigma_E^2}$. The rectangle in the bottom left of each panel indicates the range of the 95% most probable values of $|\Delta_{TR}|$ when the Test is 'just another reference'. Similar results were obtained for $n_R = 10$.*

comparative assessments with $n_R \geq 10$ non-GM varieties, and 7 assessments with $n_R \geq 20$. Our simulation results (Fig. 5a) show that the equivalence limits needed to keep the desired power may become very high for small values of n_R. Therefore it may be good to look out for opportunities to enlarge the number of reference genotypes in the equivalence studies. The statistical uncertainty of the estimated equivalence limits was seen to be ignorable for larger numbers (e.g. $n_R > 13$ in our setup), Note that in our simulations we only considered one simplistic design and a few n_R values, so more research would be needed for a final advice on an appropriate regulatory minimum in practical designs. Such research would have to include statistical considerations relevant for practical, e.g. unbalanced, experimental designs in future assessments. It should balance costs and benefits of options for proportionate regulatory oversight of products from new genomic techniques. Moreover, future research should also address wider issues, such as societal values and the relation of risk assessment of genetically modified plants to normal testing procedures for commercial plant varieties.

4.3. Evaluating equivalence tests for practical, unbalanced, experimental design

The simulation study employed a balanced experimental design to allow for a systematic comparison of DesPow, Perc and EFSA. In real-world applications, however, often unbalanced designs are used. Therefore, the simulations were repeated for a partially balanced incomplete block design with 12 reference varieties, i.e. design 6 in (Kang and Vahl, 2014). Similar performance was observed as for the balanced case in Figs. 5 and 6. Although outside the scope of the present study, it may be of further interest to compare these methods for a range of practical (unbalanced) designs.

4.4. Unintended effects are normal, normal crop composition variation should be the basis for comparison

It has been argued that unintended effects are a normal phenomenon and should in fact be expected in plant breeding and that compositional safety should be considered in the context of normal crop composition (Herman and Price, 2013). This emphasises the importance for equivalence testing rather than difference testing as a primary instrument for safety assessment. Then a major issue is what to compare to. Some authors have argued that equivalence testing should only be performed to compare T and C to focus on changed traits (Jiang et al., 2019), whereas others stress the relevance of considering equivalence with respect to a collection of commercial R varieties (Kang and Vahl, 2014; Vahl and Kang, 2016). Recent criticisms on the EFSA method (Herman et al., 2019; Jiang et al., 2019) have been discussed elsewhere (van der Voet and Paoletti, 2019).

4.5. Focus on comparing the test genotype to a collection of references

This paper focuses on the comparison of T with a collection of R genotypes. Therefore, whenever we refer to significant differences, we mean the T-R difference (test vs. mean of the R genotypes), and the graphical representation is on a (equivalence limit scaled) T-R scale. It should be noted that this is not the same as in the EFSA procedure (EC, 2013; EFSA, 2011; van der Voet et al., 2011), where a T-R equivalence test was combined with a T-C difference test, and a rather complicated rescaling was used to express all results on the T-C scale. In our current opinion, we think that it is better to keep T-R assessments and T-C assessments apart, because they refer to different research questions. The T-R assessment is concerned about the comparison to a collection of commercial varieties, whereas the T-C assessment is just about direct trait effects of the genetic modification.

4.6. One-step vs. two-step approaches

A major criticism of the EFSA method has been that it is a two-step approach, where the uncertainty of estimating the equivalence limit is not included in the equivalence test itself. One-step methods where all uncertainty was included in the equivalence test were therefore proposed (Kang and Vahl, 2014; Vahl and Kang, 2016; van der Voet et al., 2017). The Perc method achieves this by assuming $\sigma_R \gg \sigma_E$ when setting the equivalence limit (Kang and Vahl, 2014; Vahl and Kang, 2016). However, this approach leads to equivalence limits which are too wide when $\sigma_R \ll \sigma_E$ due to the incorrect assumption. Note that it can be inferred from Fig. 5a that the Perc limits are wider than those of DesPow for small ratios σ_R/σ_E even when n_R is large. Typically, due to the desired power criterion, it is expected that the DesPow EL are larger than those of Perc for realistic values of n_R. In Fig. 5a, however, the Perc limits are larger for small σ_R/σ_E ratios because DesPow takes into account the ratio while Perc does not.

The method for animal studies, DesPow0 in Table 1, was set up for T-C equivalence and could assume $\sigma_R \ll \sigma_E$ for the safe case simulations as a limiting case (van der Voet et al., 2017). This latter assumption was not possible for the T-R equivalence criterion where σ_R is in both the numerator and the denominator of the criterion (equation (2)). In the proposed DesPow method we had to revert to a two-step approach by using an estimate of the variance component ratio. Nevertheless, our simulations (Figs. 5b and 6 and Supplementary Fig. S1) showed that the desired power and proportion of rejected null hypotheses at the equivalence limit were well controlled. Although DesPow0 was not found suitable for T-R equivalence tests on composition data, in light of the assumptions made by the methods there is an interesting argument to see DesPow as something in between DesPow0 (assumption $\sigma_R \ll \sigma_E$) and Perc (assumption $\sigma_R \gg \sigma_E$).

4.7. Limitations of the Perc method in relation to the proposed DesPow principle

The desired power principle was adopted because in practical risk assessment the number of data points will always be limited, and there is the pragmatic wish of having a sufficiently high probability that genotypes similar to the reference varieties would pass the equivalence test. This probability is explicitly controlled in the DesPow method, with the consequence that the equivalence limits vary with e.g. the number of reference varieties in the experiment (see Fig. 5a). As shown in Table 1, the Perc method is also a DWE-based T-R equivalence test employing a different strategy to set the equivalence limit. The equivalence limit, a function from a percentile point from a standard normal distribution, follows from the distribution of θ for the limiting case where $\sigma_R \gg \sigma_E$ assuming that the Test is 'just another reference'. Clearly, due to its construction the Perc method does not adapt the equivalence limit to different precision levels in relation to the number of reference varieties (as can be seen in Fig. 5a) and does therefore not control the desired power of the equivalence test. The power that is actually realized by Perc is strongly dependent on the true ratio of variance components and on the number of reference genotypes in the experimental design (see Fig. 6 and Supplementary Fig. S1).

4.8. Interpretation of equivalence limits on different scales

Comparing distributions rather than average values leads quite naturally to equivalence limits specified at a quadratic scale (θ in our notation). The proposed visualization using the ELSD scale has a better interpretability: it discriminates increases from decreases, and allows a direct interpretation in terms of equivalence and difference tests. However, on the ELSD scale the equivalence limits are $(-1,+1)$ by definition. To interpret equivalence limits at the original scale (log-transformed differences, or ratios), estimates can be made, but these will show uncertainty (see Table 3 and (van der Voet et al., 2019)).

4.9. Visualization with ELSD is appropriate for all discussed methods

We showed that results from different equivalence tests can be shown using the same visualization (Fig. 2). The ELSD visualization (see (van der Voet et al., 2019) for an extensive discussion) is simple to interpret in terms of difference and equivalence conclusions. This contrasts with more complex visualizations that have been proposed earlier (EFSA, 2011; Kang and Vahl, 2014; van der Voet et al., 2011).

5. Conclusion

In this paper we propose a new statistical test as an update to the EFSA equivalence test for GMO safety assessment based on plant composition data. In order to be suitable for practical safety assessment, the new DesPow method has a statistical power set to a desired value, e.g. 95% by construction. Contrary to the current EFSA method, DesPow can be applied to any analyte. We also propose an improved visualization of the equivalence test results as equivalence limit scaled differences. From simulations and a practical case study with the maize compositional analysis data available from EFSA, it is concluded that the proposed DesPow test has better statistical properties than the current EFSA equivalence test and another test proposed in the literature. Nevertheless, the results in a practical case study were broadly similar and the improvements were mainly relevant for analytes with hardly any variation between the reference genotypes.

CRediT authorship contribution statement

Jasper Engel: Conceptualization, Methodology, Software, Visualization, Formal analysis, Writing – original draft. **Hilko van der Voet:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledge funding from the Netherlands Organization for Scientific Research and the Ministry of Infrastructure and Water Management (NWO Proj. No. 15815) and from the Ministry of Agriculture, Nature and Food Quality (DDHT Proj. No. KB-38-001-003).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fct.2021.112517.

References

Chiu, S.-T., Chen, C., Chow, S.-C., Chi, E., 2013. Assessing biosimilarity using the method of generalized Pivotal quantities. Generics Biosimilars Initiatives to appear 2 (3), 130–135.
Codex, 2008. Guideline for the conduct of food safety assessment of foods derived from recombinant-DNA plants. CAC/GL 45-2003. revised 2008. . Retrieved from. http://www.fao.org/fao-who-codexalimentarius/thematic-areas/biotechnology/en/.
EC, 2013. Commission implementing regulation (EU) No 503/2013 of 3 april 2013 on applications for authorisation of genetically modified food and feed in accordance with regulation (EC) No 1829/2003 of the European parliament and of the council and amending commission regulations (EC) No 641/2004 and (EC) No 1981/2006. Off. J. European Union.
EFSA, 2010. Statistical considerations for the safety evaluation of GMOs. EFSA J. 8 (2), 1250.
EFSA, 2011. Guidance on the risk assessment of genetically modified microorganisms and their products intended for food and feed use. EFSA J. 9 (6), 2193.
EUGFL, 2002. Regulation (EC) No 178/2002 of the European Parliament and of the Council of 28 January 2002 laying down the general principles and requirements of food law, establishing the European Food Safety Authority and laying down procedures in matters of food safety. Retrieved from. http://data.europa.eu/eli/reg/2002/178/oj.
FDA, 2020. New plant variety regulatory information. Retrieved from. https://www.fda.gov/food/food-new-plant-varieties/new-plant-variety-regulatory-information.
FQPA, 1996. Summary of the food quality protection act. Retrieved from. https://www.epa.gov/laws-regulations/summary-food-quality-protection-act.
Herman, R.A., Huang, E., Fast, B.J., Walker, C., 2019. EFSA genetically engineered crop composition equivalence approach: performance and consistency. J. Agric. Food Chem. 67 (14), 4080–4088.
Herman, R.A., Price, W.D., 2013. Unintended compositional changes in genetically modified (GM) crops: 20 years of research. J. Agric. Food Chem. 61 (48), 11695–11701.
Jiang, C., Meng, C., Schapaugh, A., 2019. Comparative analysis of genetically-modified crops: Part 1. Conditional difference testing with a given genetic background. PloS One 14 (1), e0210747.
Kang, Q., Vahl, C., 2014. Statistical analysis in the safety evaluation of genetically-modified crops: equivalence tests. Crop Sci. 54 (5), 2183–2200.
Kang, Q., Vahl, C., 2016. Statistical procedures for testing hypotheses of equivalence in the safety evaluation of a genetically modified crop. J. Agric. Sci. 154 (8), 1392–1412.
Krishnamoorthy, K., Mathew, T., 2009. Statistical Tolerance Regions: Theory, Applications, and Computation, vol. 744. John Wiley & Sons.
McNally, R.J., Iyer, H., Mathew, T., 2003. Tests for individual and population bioequivalence based on generalized p-values. Stat. Med. 22 (1), 31–53.

Meeker, W.Q., Hahn, G.J., Escobar, L.A., 2017. Statistical Intervals: a Guide for Practitioners and Researchers, vol. 541. John Wiley & Sons.

Roy, A., Bose, A., 2009. Coverage of generalized confidence intervals. J. Multivariate Anal. 100 (7), 1384–1397.

Searle, S.R., Casella, G., McCulloch, C.E., 1992. Variance Components. John Wiley & Sons.

Vahl, C., Kang, Q., 2016. Equivalence criteria for the safety evaluation of a genetically modified crop: a statistical perspective. J. Agric. Sci. 154 (3), 383.

van der Voet, H., Goedhart, P.W., García-Ruiz, E., Escorial, C., Tulinská, J., 2019. Equivalence limit scaled differences for untargeted safety assessments: comparative analyses to guard against unintended effects on the environment or human health of genetically modified maize. Food Chem. Toxicol. 125, 540–548.

van der Voet, H., Goedhart, P.W., Schmidt, K., 2017. Equivalence testing using existing reference data: an example with genetically modified and conventional crops in animal feeding studies. Food Chem. Toxicol. 109, 472–485.

van der Voet, H., Paoletti, C., 2019. Equivalence testing approaches in genetically modified organism risk assessment. J. Agric. Food Chem. 67 (49), 13506–13508.

van der Voet, H., Perry, J.N., Amzal, B., Paoletti, C., 2011. A statistical assessment of differences and equivalences between genetically modified and reference plant varieties. BMC Biotechnol. 11 (1), 1–20.