SCIENTIFIC REPORT OF EFSA

Statistical considerations for the safety evaluation of GMOs:

response to the public consultation [1] [2] [3]

European Food Safety Authority (EFSA), Parma, Italy

Abstract

This document responds to the many institutions and individuals who provided comments on the field trial design and statistical approach proposed in the draft document of the WG Statistics of the EFSA GMO Panel on '*Statistical considerations for the safety evaluation of GMOs*', which was adopted by the GMO Panel on 2 July 2008 and published on the EFSA website for public consultation from 21 July 2008 until 21 September 2008. Following the public consultation, the original document has been revised taking into account all the relevant scientific comments, and a final opinion of the GMO Panel has been adopted on 21 April 2009 and published on EFSA website *http://www.efsa.europa.eu,* as a separate document.

Key words: GMO, equivalence limits, statistics, field trials, compositional analysis, mixed model, proof of hazard, proof of safety, confidence interval, difference test, equivalence test

[1] On a request of EFSA, Question No EFSA-Q-2008-04979 issued on 21 July 2008

[2] Correspondence: GMO@efsa.europa.eu

[3] This scientific report of EFSA is based on major contributions from: Marco Acutis, Hans Christer Andersson, Salvatore Arpaia, Gijs Kleter, Harry Kuiper, Jim McNicol, Joe Perry, Willem Seinen, Hilko van der Voet.

For citation purposes: European Food Safety Authority; Response to public consultation on Statistical considerations for the safety evaluation of GMOs, on request of EFSA. EFSA Journal 2009; 340. [37pp.]. Available online: www.efsa.europa.eu

TABLE OF CONTENTS

RESPONSE TO PULBIC CONSULTATION

In general the approach proposed in the draft document of the WG Statistics of the EFSA GMO Panel on '*Statistical considerations for the safety evaluation of GMOs*' was well accepted. There was clearly a need for clarifications to the methodology and calculations required for dossiers.

The main focus of the new proposals for both field trial design and statistical analysis was centered around the following issue: how potential differences in composition between GM plants and their conventional counterparts could be placed into the context of the natural variation due to biological and environmental factors, such as the variation due to the genetic backgrounds of commercial varieties with a history of safe use, whilst allowing for the usual uncertainties associated with the limited availability of data.

Many comments received during the public consultation requested clarification as to whether the proposed statistical approach related to animal feeding studies as well as to plant compositional data. The proposed approach does not relate to the experimental design of animal feeding trials with whole GMO foods/feed, for which EFSA has issued recent guidance separately. However, it might be used for the statistical analysis of data from such animal feeding trials with whole GMO foods/feed, particularly if commercial varieties have been included in those trials, where appropriate and on a case-by-case basis.

Some comments requested clarification whether agronomic/phenotypic data was included within the scope of the new proposed approach. The approach is designed to address those agronomic/phenotypic endpoints that can be measured and included within the same field trials as for compositional analysis.

In response to several requests for further clarification concerning what conclusions should be drawn in the event of endpoints showing difference or non-equivalence, note that:

(i) Neither difference nor non-equivalence by themselves necessarily imply lack of safety;

(ii) When any difference or lack of equivalence is found this should be evaluated in the context of a risk assessment process and interpreted within a risk assessment framework;

(iii) The function of the proposed statistical methodology is to produce results that may be interpreted by biologists, toxicologists or other safety experts and not, at the current state of development, to provide a decision-theoretic framework that allows direct inferences on safety;

(iv) The GMO Panel and its Statistics WG are aware that setting the size of the difference test at the 10% level will lead to a large proportion of tests being found to be significant by chance alone. *Per se* a large proportion of significant differences is not considered a sufficient reason for safety concern, unless the proportion would be larger than the proportion of significant results which can be expected for differences between two randomly chosen commercial varieties. Safety concerns may also be raised if the differences follow some systematic pattern such that endpoints of a certain type form a cluster that are significant;

(v) The opinion is quite specific that endpoints showing difference or non-equivalence must be further analysed to investigate possible site x treatment interactions. This approach is fully in line with the oft-stated philosophy of the GMO Panel that the requirements of toxicological testing shall be considered on a case-by-case basis. The original report has been amended to clarify some of these issues.

Some comments questioned why equivalence limits should be based on commercial varieties grown at the same sites when information on natural variation exists in sources such as the ILSI database. The GMO Panel and its statistics WG believe strongly that whilst it might be true that when a substantial database of commercial varieties has been established, future guidance might remove the necessity to include commercial varieties. However, that stage is nowhere near having been reached yet. It is essential that future implementations of suitable databases include detailed information on the particular variety concerned and a sufficient characterisation of the environments concerned to allow the elimination of major environmental differences in the comparison of GMO with commercial varieties.

Further questions asked why commercial varieties had to be integrated, as fully randomized and replicated treatments, into compositional field trials, and why the natural variability of varieties could not be estimated from unrandomized additional plots external to the field trial at a site, or even at other sites. The reason is that when commercial lines are included in the same experiment where the GMO is tested against the comparator(s) then data on commercial varieties are obtained in identical conditions to that of the GM and its comparator. This has the major advantage of eliminating uncontrollable confounding effects. The GMO Panel and its statistics WG affirm that randomization is a fundamental principle of good experimental design. Using information from unrandomized sources, as it was noted in some comments, would result in a biased estimate of the difference between the GMO and the commercial varieties. It would be completely unacceptable to place the entire basis for an equivalence test on an estimate of a difference that is biased.

In response to some comments received, the GMO Panel and its statistics WG have considered strategies to maximise the efficiency of trials on a given site, and have suggested designs by which the production of material for the comparative assessment of several different GM plants of the same crop species may be produced simultaneously at the same site and within the same field trial. This may be done by the placing of the different GM plants and their appropriate conventional counterpart(s) in the same randomized block, with some provisos. More details are given in the opinion adopted on 21 April 2009.

There was a plethora of requests for a clearer description of how the equivalence limits should be calculated. In response to these the GMO Panel and its statistics WG have done further work. The GMO Panel statistics WG has revised this section, and the opinion adopted by the GMO Panel on 21 April 2009 provides the needed clarification, also by supplying a worked example. The statistics WG recommendation for the estimation of equivalence limits is designed to effectively quantify the background variation between different varieties. Since each variety may be grown on a number of different sites (as long as these are appropriate), the estimated variation between varieties then takes automatically account of both genotypic variation and part of the full genotype x environment variation. The statistics WG has experimented with several formulae and has formulated guidance based on an estimate involving the standard error of the difference between the mean of the GM and of the commercial varieties. This formulation fulfils three important criteria:

(i) The width of the equivalence interval is positively related to the degree of the measured genotype x environment interaction;

(ii) GMO means fall within the equivalence limits with approximately the correct coverage (95%) under the null hypothesis that the GMO is exchangeable with any of the commercial varieties;

(iii) It is easily implemented via the statistical mixed model in common statistical software (as illustrated now in the example given in the opinion).

There were a large number of responses that questioned the basis for comparison of the equivalence test. The draft for public consultation had been misunderstood as giving the false impression that the equivalence test was based somehow on the mean of the comparator. This is not so, and the GMO Panel and statistics WG have now provided clarification in the opinion adopted on 21 April 2009. The methodology for the equivalence test compares the mean of the GM with the mean of the commercial varieties. In fact, the comparator mean does not influence the result of the test. Probably what gave rise to the misunderstanding was the form of the graph that is the basis of the required presentation of results. On this graph, the value of the GM and of the equivalence limits are all related to a baseline from which the mean comparator value has previously been subtracted. Indeed, simple mathematics will convince that the mean value of the comparator has no effect whatsoever on the test of equivalence, since the relevant difference tested is: [mean(GM) - mean(comparator)] – [mean(commercial varieties) – mean(comparator)]. Similarly, the equivalence limits themselves, when plotted on the graph, are related to a baseline from which the mean comparator value has previously been subtracted, and so are also recalculated to subtract this value from their raw values, but solely in order to achieve a consistent scale for the graph.

Some comments requested clarification that the proposed approach was not intending to prevent the widespread practice of the use of more than one background germplasm for GMO and comparator(s) in order to better accommodate different environmental conditions. Clarification has now been provided.

There was some misapprehension in the comments that the choice of the minimum number of replicates at a site was driven by a technical desire to have a minimum number of degrees of freedom for error in an individual trial. In reality, the requirements are driven by the desire to have adequate replication and hence to ensure that each field trial at each site has sufficient power to give the Panel and the Member States confidence that unintended effects will be detected. Whilst the key statistical analysis for comparing GMO and comparator(s) is the analysis which averages across locations, a fundamental principle underpinning the previous Guidance was that replication should be sufficient at each site to allow an adequate stand-alone analysis at each of those sites. This is especially important in the frequent circumstances of treatment x site interactions when there are differences detected at some sites but not at others. That principle remains undiminished within the current opinion adopted on 21 April 2009. Additionally, whilst it may be the case that coefficients of variation at the plot level for compositional trials are typically relatively low, this is not the case for all endpoints.

The GMO Panel and its statistics WG note with regret that, despite frequent emphasis on the need for statistical power calculations to guide replication levels in the previous food-feed (2006) Guidance, dossiers have only rarely included such studies. The result is that there is little specific data available on which to base decisions concerning replication levels. Furthermore, in the interests of transparency it is important that a high level of confidence be given to the public on this issue. The GMO Panel and its statistics WG are of the opinion that the proposed guidance concerning replication at a site was satisfactory and required no amendment or clarification.

There were several comments on the need for data transformation. Data transformations may be used either to stabilise otherwise heterogeneous variance, or to change the scale on which effects are additive, or both. Statistically, the more important of these is additivity. The draft for public consultation issues no mandatory instructions to transform to logarithms, although the statistics WG expects that the great majority of endpoints will be. Indeed, in its research, it found that analysis with transformation generally provided equal or better results than analysis without. A logarithmic transformation has the advantage that differences may be expressed as percentage changes on a multiplicative scale. Here again, the GMO Panel and its statistics WG concluded that the proposed approach was satisfactory and required no amendment or clarification.

The final opinion adopted on 21 April 2009 has been amended with clarifications on what should be the form of analysis for trials involving more than one comparator, and on how factors should be specified, as fixed or random, in the statistical mixed model.

ANNEX A: Comments received during the public consultation

CHAPTER TEXT	COMMENT
3.3.2 Use of concurrent data to estimate equivalence limits	3.3.2 Lns 824-825. The inclusion of commercial varieties may well have the advantage of "eliminating uncontrollable confounding effects", but the danger is that it may also eliminate the detection of any effects caused by the genetic modification that could be important to human and animal health. Lns. 864-867. The choice of varieties is left up to the applicant. It is their interests to make this as large a range as possible and simply not acceptable. Only the cultivar of the conventional counterpart can be used.
3.3.1 Which data can be used?	Lns 807-814. The inclusion of these wide ranges of different varieties and historical data on parameters is unacceptable. Comparisons must be limited to the control.
3.2.4 Multiple endpoints	continued from previous comment Lns 744-751. The assertion that there is a "discrepany between the usual assumption of GMO comparative assessment and the assumption that is used as a null hypothesis in the statistical test of difference" and the suggested allowance made for the "background variation that may exists between commercial varieties" is a direct contradiction of the concept of substantial equivalence where the GMO is regarded as the same as it's conventional counterpart except for the novel protein. To now broaden this comparison to include background variation of commercial varieties is not acceptable. The only comparison can be with the cultivar grown under the same conditions, i.e. the control. It is interesting to note that EFSA referencing it's assessment of MON863 as an example here. The assessment of MON863 is probably the most controversial assessment that EFSA has performed to date and has been highly criticised for it's approach using background variation in the population. An additional comment here is that, not only is it unacceptable to use background variation in commercial crop varieties, but also of control test subject (e.g. Norwegian Brown Rat). The assessment of MON863 was criticised also because of the inclusion of historical ranges in the animal parameters in the comparison of MON863 and its conventional counterpart (Séralini, G-E, Cellier, D. & Spiroux de Vendomois, J. 2007. Archives of Environmental Contamination and Toxicology DOI: 10.1007/s00244-006-0149-5.). Although not explicitly addressed in this document (possibly in 3.3.1 points 3 and 4?), inclusion of historical animal parameters as background variation data in the comparisons is unacceptable for the same reasons as the inclusion of commercial varieties stated above.

3.2.4 Multiple endpoints	This section is absolutely unacceptable. It reads like a manual to applicants on how to disregard significant differences they may find and stands to make a mockery of the concept of "substantial equivalence". Although, in our opinion, this concept is flawed because it cannot take into account the unexpected and unpredictable effects that GM crops are prone to, but to weaken the concept further by allowing applicants "get-out clauses", where significant differences do not have to be investigated further could be regarded as a serious failing of EFSA to protect consumers health. Indeed, the Royal Society of Canada recommended (2001: Elements of Precaution: Recommendations for the Regulation of Food Biotechnology in Canada) rigorous scientific assessment of such differences. Lns 716-718. The probability of obtaining significant differences by chance alone may be considerable, but this document does not consider the possibility of Type 2 errors, where the null hypothesis is not rejected, even though it is actually untrue. This needs to be considered, as it may also be considerable and needed to be taken into account by applicants. Lns 739-742. Making allowances for an increased number of significant differences that would be expected by chance alone is wholly unfounded. If endpoints are correlated, such endpoints should be clearly identified by independents experts. If significant differences are simply disregarded because the may be correlated, it is possible that an effect may go unnoticed. This is leaving subjective judgements in the hands of the applicant. Continued
2.6 Simultaneous assessment of multiple endpoints	lns 428-441 Whilst we welcome the fact that significant differences "should still be accompanied by an evaluation of the toxicological relevance" but it is not clear how this should be done. Relying on applicant's evaluation would not be acceptable here. Lns 446-473. A multivariate analysis framework for analysing multiple endpoints (especially is welcomed and we hope work in this area proves fruitful.
Summary	XXX welcomes the opportunity to comment on e Draft Report on statistical guidance. It is very pleasing to see this working group provide clarification on what is widely perceived as a grey area in the risk assessment. However, the suggest treatment of significant differences does nothing to increase confidence in the phrase often used by EFSA that such differences are "unlikely to be of biological relevance" for 2 reasons: 1) it leaves much of the evaluation of these differences up to the applicant and 2) it offers numerous opportunities to dismiss such differences, to the extent that any real effects are likely to go unnoticed. This could have implications for food/feed safety.

3.3.2 Use of concurrent data to estimate equivalence limits	**The use of concurrent data to estimate equivalence limits in order to reduce bias due to confounding is recommended. A mixed model approach is suggested for analysis. While this procedure has the advantage of transparency, it introduces new problems:** **- (manipulation) The approach stimulates the selection of extreme varieties to increase the variance estimates and thus to widen the equivalence margins.** **- (lack of robustness) While in balanced designs the estimation of equivalence margins and of genotype effects will be more or less independent, this is not the case if the design is unbalanced e.g. due to missing values: the EM algorithm may be unstable, in particular with small sample sizes. Convergence problems may result, estimates of genotype effects and variance estimates (and thus equivalence margins) will be correlated. Even with balanced designs, variance estimates on the CommercialVariety level will be rather imprecise, potentially resulting in poor estimates of equivalence margins.** **- (conceptual deficiency) With the recommended approach, equivalence margins are first estimated and then taken as fixed for the purpose of comparison to the genotype effect. The resulting confidence limits do not correctly address the random error of the comparisons since they ignore the randomness of the confidence limits as well as the mutual dependence of the estimates. In my eyes it would be more appropriate to incorporate the GM crop into the CommercialVariety random effect and to compare the genotype residuals directly to the level-2-residuals of the commercials after adjustment of both for control. Quantile estimates or ranks (GM crop residual as compared to the residuals of the commercials, with confidence limits) may be a better approach to find unusual deviations of the GM crop than the equivalence margin approach. Alternatively, if only a few commercials were available, GM crop and commercials may be taken as fixed and the GM crop contrast to control may be directly compared to each of the commercial varieties contrasts to control, possibly after further suitable adjustment.** **There will be even more problems that are linked to the challenge of model fitting. It may be more advisable not to recommend a specific modeling approach, but rather to define what an analysis should demonstrate. A mixed model fit to a certain data set may be added as an example in the appendix.**
3.3 Estimation of equivalence limits	**There seems to be a fundamental difference between safety concerns in drug trials and safety concerns in GMO trials. In drug trials, a safety concern relates to a potential harm to patients; equivalence margins are derived from discussion as to which extents of harm acceptable in relation to certain benefits. According to the arguments in chapter 2, in GMO trials safety concerns are not derived from potential harm to individuals, animals or environment but relate to effects that are out of range of ordinary biological variation. Obviously, the range of biological variation is felt to be identical to the region of harmlessness. This is not self-understood since this assumption may be wrong in both directions. The guideline should explicitly mention this assumption, give a rationale and discuss exceptions.**

3.2.3 Single endpoints in more complex experimental designs	**Lines 677-682: It should be mentioned that the formulae of 3.2.2 apply to more complex situations as well if s is replaced by the residual standard deviation of a model and the degrees of freedom are adapted to the number of model parameters.**
3.2.2 Single endpoints in simple two-group designs	**In lines 293-294, a switch from two-sided to one-sided confidence intervals with the same error level was recommended for power increase. In line 622-628, it is emphasized that the two-sided confidence interval of 90% should be used corresponding to a one-sided 95% interval when only one direction is of interest. These two notions are in conflict. The last notion is more reasonable since it unlinks power and direction. Moreover, the last notion is in line with the EMEA "Points to consider on switching between superiority and non-inferiority", chapter II.4. in this EMEA paper, the use two-sided 90% (95%) confidence limits and one-sided 95% (97.5%) confidence limits is recommended for bioequivalence (clinical equivalence) trials.**
3.2.1 Equivalence limits	**Page 14, line 553. It is not a question of logic whether a multiplication factor approach is preferable, but a question of scale type (ratio scale vs. other scales) and of the dependence of the standard error from level whether a multiplicative or an additive scale should be chosen for analysis.**
3.1 Introduction: choice of model and preprocessing of data	**Lines 513-520: standard normality tests are not appropriate in this context. Their power is high when sample size is high, but then they are not needed because of asymptotic theory (central limit theorem) , and their power is low in small sample sizes. Overall, size and power of tests/confidence intervals are not better with a normality-pre-test (two-step procedure) than without (one-step procedure).**
2.6 Simultaneous assessment of multiple endpoints	**The notion that a global assessment be the goal of a safety trial may be misunderstood. Safety concerns are specific concerns with respect to one endpoint or a few correlated endpoints, i.e. local differences are of interest. The aim is to receive a profile of a GMO substance relative to some standard or many standards, with sufficient precision. Such a profile would allow a scientific discussion of the special characteristics of the GMO that addresses chances and risks of the GMO in a comprehensive manner. While the plot on page 9 can be understood as such a profile, a global assessment may obscure important differences between GMO and iso-line. However, even if the perspective is local, when many endpoints are addressed, the probability of any type I error increases substantially. This alpha explosion can in general only be avoided by methods that reduce the power and thus result in larger confidence intervals, making superiority as well as non-inferiority assessments more difficult and thus increase the number of inconclusive results. While this is an honest result, it is unsatisfactory.** **A common solution is to ignore that problem. This means that the power is high, but type I error is only locally controlled. In this case, alpha explosion is almost certainly present. A rough calculation of the expected number of chance findings and a comparison to the number of significant results may help to give an impression of the number of real differences, but unfortunately does not allow to locate them. Thus, the method of multiple comparisons (lines 415-419) does not really help to solve the problem. Simultaneous confidence intervals reduce the alpha explosion but reduce the**

	power as well, so one problem is replaced by the other. **Experienced researchers try to avoid chance findings by searching for sensible patterns in the observed effects, i.e. by relying on correlations between endpoints. This can be formalized by applying PC (principal components) based methods that keep the local or multiple level alpha as suggested in No 2 (lines 420-422). This results in a dimension reduction and in a better balance of type-I error and power than with multiple comparisons. This method should be regarded as method of first choice for a comprehensive view on the safety issues. It can be supplemented with the more local multiple comparisons without multiplicity adjustments.** **Method No. 3 (a priori restriction, lines 425-426) is a dangerous concept since it introduces a parameter selection by the applicants that can be abused. In principle, safety trials should include all potential safety aspects and should not allow arbitrariness. Of note, with respect to drug safety, the EMEA states that multiplicity adjustments should not be performed. The 'EMEA Points to consider on multiplicity issues in clinical trials' state (in 2.4 Multiplicity in safety variables): 'In those cases where a large number of statistical test procedures is used to serve as a flagging device to signal a potential risk caused by the investigational drug it can generally stated that an adjustment for multiplicity is counterproductive for considerations of safety.'** **(The idea behind this statement is that otherwise any severe safety concern can be masked by adding irrelevant endpoints to render the analysis inconclusive after multiplicity adjustment.) According to the EMEA, the problem of alpha explosion in safety issues should not be solved statistically but by scientific reasoning: 'It is clear that in this situation there is no control over the type I error for a single hypothesis and the importance and plausibility of such results will depend on prior knowledge of the pharmacology of the drug."** **It may be a good compromise if a PC-based multivariate analysis is performed with strict type-I error control accompanied by a univariate calculation without any multiplicity adjustment.**
2.4 Types of possible outcomes between the GMO, its comparator and commercial varieties	**In drug trials, superiority analyses have to be performed first-line in the intention-to-treat (ITT) analysis set, while non-inferiority/equivalence analyses should be performed first-line in a per-protocol (PP) analysis set (cp. EMEA "Points to consider on switching between superiority and non-inferiority", IV.1.4 and iCH-E9 5.3). While these terms are not established in GMO trials, there should be recommendations with respect to analysis sets that regulate when observational units are to be excluded from analysis or have to be included.**

Summary	In summary, in my eyes, the pure existence of a guidance paper of that quality is an important milestone in the process of improving safety evaluations on GMOs. I would like to thank the authors and to congratulate them for their excellent work. Here are some ideas that may help to improve the paper further: While there is a special chapter on experimental design (4.1), there should be a section in chapter 2 that clarifies that randomized trials are the gold standard for safety evaluations and when observational trials are required or acceptable. Control of covariates will be advantageous in many situations for increase of power and bias correction. There should be a separate section with recommendations on adjustment. The paper will be a hard challenge for many researchers in the field. It will possibly better understood and implemented if a few examples were given in a appendix.
Table of contents	In summary, in my eyes, the pure existence of a guidance paper of that quality is an important milestone in the process of improving safety evaluations on GMOs. I would like to thank the authors and to congratulate them for their excellent work. Here are some ideas that may help to improve the paper further: While there is a special chapter on experimental design (4.1), there should be a section in chapter 2 that clarifies that randomized trials are the gold standard for safety evaluations and when observational trials are required or acceptable. Control of covariates will be advantageous in many situations for increase of power and bias correction. There should be a separate section with recommendations on adjustment. The paper will be a hard challenge for many researchers in the field. It will possibly better understood and implemented if a few examples were given in a appendix.
5 Conclusions and Recommendations	See comments made under ''Summary''

4.3 Choice of levels of replication	**As stated in the EFSA paper environmental variation will be added due to site-to-site variation and due to year-to-year variation. The approach of EFSA defines a minimum number of sites and allows to restrict experiments to one season. This seems problematic for several reasons. In practice this allowance led to a situation where the environmental variation which will be encountered in European agro-ecosystems has not been sufficiently tested. Applications usually lack the criteria employed to select the chosen field sites and to demonstrate their representativeness. The range of field sites tested should cover the full range of environmental variation which can be encountered and should also take the chosen endpoints into consideration. A restriction to the main growing regions, therefore, impairs the analysis substantially. We recommend including criteria for the selection of representative field sites into the document. One way to proceed would be to define the main different agronomical, geographical, and bio-geographical regions.**
3.3 Estimation of equivalence limits	**With regard of the use of concurrent data for the estimation of equivalence levels we want to stress the importance of carrying out field trials in multiple, representative environments. The objective here should be to cover the full background variance in relation to possible gene-environment interactions. The document so far lacks a respective guidance.** **In any case the document should be complemented by setting rules to prevent that the selection of varieties or the compilation of historical data will be used to increase variance estimates and thus widening the equivalence limits.** **Applications for the cultivation of GMPs in Europe should primarily include data from Europe. It should be stressed at this point that the market release of GMO according to Directive 2001/18/EC shall be carried out in a step-by-step manner. Prior to a market release sufficient data for the assessment should be collected during the test-phase (Part B releases) in representative European environments.**
2.6 Simultaneous assessment of multiple endpoints	**When carrying out multiple comparisons, an adjustment of alpha will lead to a decrease in statistical power with negative implications for the type II error. For this reason we have some reservations against multiple testing when looking at the biosafety of GMO.** **From the three options given in chapter 2.6 we consider analysing the data by PCA the most promising. This could be followed by 'local' multiple comparison of selected endpoints as opposed to a pure 'global' analysis.**
2.3 Decision analysis, tests and confidence intervals	**Lines 293 to 294: It is expressed here that the error level should be retained when switching from a two-sided to a one-sided test. This contradicts the notion in lines 622-628 where an adjustment of the error level is recommended. We suggest clarifying the advantages and disadvantages of both approaches.**

Summary	XXX welcomes the initiative of EFSA. EFSA produced a thorough document which can be regarded a milestone to guide the statistical analysis of data needed to establish the safety of GMO. As we commented on the new draft of the EFSA guidance for GMO, it should be kept in mind that both the assessment of effects on human and animal health and of effects on the environment partly rely on the same studies. This is especially true for data on the expression of the insert, for the compositional analysis, and the analysis of phenotypic equivalence. Mammalian toxicity studies will and must be used as a starting point to assess possible effects on wildlife including species of conservation concern. For this reason the guidance given in the actual statistical guidance document should also reflect specific requirements of the environmental risk assessment (e.g. need to define 'background variability'). As for the scope of the paper we caution against the use and interpretation of 'global equivalence' tests. We do not reject such tests per se but strongly feel that the scientific rationale behind this needs to be explained in more detail. Importantly, the use of global tests (multiple endpoints) will help to analyse interaction between variables. At the same time a global analysis will not be adequate to test for differences of variables of specific concern. This problem may be understood best by exploiting the experiences gained in other branches of risk assessment such as the assessment of medical or chemical substances. The paper would further improve by including practical examples and by stressing the rationale of the tests. In this regard the importance of the comparison between GMO and control (usually the iso-line) seems underrepresented. To our understanding testing is multi-layered: Whereas a comparison between GMO and isoline may indicate unexpected effects due to the genetic modification. The comparison with the expected background variability will add to the interpretation of observed differences. For the risk assessment it is important that both analyses are carried out separately.
APPENDIX. Proposed text for Chapter 7.2 of the EFSA Guidance Document	Lines 1236 – 1237: EFSA requests raw data and programming code in editable form. It would be appropriate if EFSA explains what is its objective with this information? Lines 1238 – 1240 and elsewhere: The emphasis is on using log transformed data for the responses. Lines 295-303 specify using the ratio of the test to control which becomes the difference of test and control under the log transformed data. The implication is that if the response is not log transformed and the raw data are used that the analysis should be on the ratio. However, lines 1241-1243 say a natural scale may be more suitable. If the response satisfies the ANOVA assumptions then we see no reason why the analysis cannot be conducted on the difference between test and control using the natural scale rather than the ratio. Another reason for using differences on the natural scale is for interpretability since it is easier to understand what the data tell you when differences using the natural scale are used. In addition, the guidelines only discuss transformation using a log scale for non-normally distributed data. What about the use of generalized linear models instead of applying a transformation?

4.3 Choice of levels of replication	**Line 1012: It does not necessarily follow that more residual degrees of freedom are required for highly variable response types (although more replicates might be prudent).** **Line 1017: It is recommended to include minimum 6 commercial varieties over all sites for the calculation of equivalent limits. However, if the distribution of commercial lines over the sites is not done homogenously this will lead to imbalance in the calculation of equivalence limits (interference of genotype and site factor).** **Lines 1050-1052: it is indicated that the site by (GMO vs Control) interaction is of interest, there should be some guidance on how to deal with a large interaction.**
4 Proposals concerning field trial design	**Lines 913 – 1066: It is appreciated that the document defines a minimum of eight sites for replication of the field trials while granted the flexibility in the number of years over which field trials are conducted, as our experience also suggested that within a single year, a range of likely receiving environments where the crop will be grown can be represented by the chosen sites if they cover a large geographic range.** **We agree that including commercial varieties concurrently in the same experiment has the major advantage of eliminating uncontrollable confounding effects and also providing additional degrees of freedom (df) for the comparisons between the GM crop and the control. However, since the total number of additional plots required for commercial varieties increases dramatically with the number of studies conducted concurrently at a trial site. For example, if a trial site has 10 studies, and each contains 3 replications of 3 commercial varieties, it ends up requiring a total of 3 x 3 x 10 = 90 plots for commercial varieties at that site! This additional resource requirement is difficult to achieve. Therefore, considering the practical situation, we'd like to propose an independent commercial study be conducted at each trial site in which we have at least one GM studies, and the commercial study will include a minimum of 3 commercial varieties with at least 2 replications of each variety. In trial sites where field homogeneity is of question, the commercial study will be replicated in multiple locations within sites.** **And we support the recommendation of using different commercial varieties in different sites so that these commercial varieties are representative of the sites at which they are grown.** **In recognizing the difficulty of providing characteristic-specific important difference between varieties required to be detected for a full power analysis, the document provides a specific recommendation concerning the minimum amounts of replication based on the number of residual df at each individual site in section 4.3. However, this recommendation is inconsistent with the recommendation in section 3.2.3 that a statistical analysis should be conducted for the complete data set over all sites rather than for each individual site as the power of a per-site evaluation is commonly too small. Therefore, the number of residual df for the across-site analysis provides a more reasonable criterion for the choice of the number of replications. For example, in an experiment with 5 varieties, 8 sites and 2 replications with a randomized block design at each site, there are 32 residual df, 4 df coming from each site. There are calculated as: total df (80 – 1 = 79) minus variety df (5 – 1 = 4) minus site df (8 – 1 = 7) minus block**

	(site) df ((2 – 1) x 8 = 8) minus variety x site df (4 x 7 = 28), i.e. 79 – 4 – 7 – 8 – 28 = 32.
3.3.3 Use of literature or databases to estimate equivalence limits	**Lines 877 – 880: This will be the case for all the next files submitted during the next three years. See above.**
3.3.2 Use of concurrent data to estimate equivalence limits	**Lines 860-863: the confidence interval is centered at (y2-y0)= the difference between the means for Commercial and comparator. In order to construct the equivalence limits in the chart, is this interval centered at 0 or at (y2-y0) ?**
3.3.2 Use of concurrent data to estimate equivalence limits	**Lines 833-858: The specific mixed model used in the analysis of variance is crucial for carrying out the tests of difference and tests of equivalence. The estimate of the variance among commercial lines (V) will depend on whether GxE interaction is included as a random effect, and on whether other interactions with site (e.g., site by GMO vs Control) are included. The error used for the test of difference (GMO vs control) depends on whether site and the interaction are viewed as random. Power will depend on the model. These issues are not dealt with clearly in this document.** **Lines 840-846: Section 3.3.2 delineates the model which should be used. How can all of the three factors given in lines 840-846 be included in the model? Since two of the factors have a number of missing values inclusion of all three factors will yield no data for the statistical analysis. Please give an example using data.** **Line 847: The description of the statistical model (mixed model with fixed effects and random effects) is not clear enough.** **On line 691, it is stated "A mixed effect model can be used for the analysis of the complete data set (all sites and years) where the factors site and possibly year are assumed to be random." This is compared to line 847, where the model is expanded to allow the calculation of equivalence limits from commercial lines included in the experiments and it is stated that: "The factors genotypegroup and testmaterial must be treated as fixed effects, whereas commercialvariety must be treated as a random effect." Other factors taken into account should be site, block (replicate??), year and interaction between genotype*site and site*year. More explanation should be provided when these factors have to be included as fixed or as random factors.** **Additionally, it is not clear how an ANOVA can be calculated with missing values for factors.** **Is there a test data set available to see how the statistical calculations need to be done?** **Lines 853-854: For multi-environment studies then regardless of whether environment is considered fixed or random, a term for the genotype by environment interaction must surely be fitted in order to obtain an appropriate error structure.** **Line 854: If interaction between factors is included what is the consequence when these interactions are significant?**

3.3.1 Which data can be used?	**Lines 807-808: The inclusion of commercial varieties in the experimental design is a new concept and up to now none of the field trials contain such varieties. Therefore, the alternative option using historical data should not be excluded in order to allow the application of this statistical approach on data to be provided in the existing or immediately upcoming application containing results obtained before the publication of this document.** **Lines 809-810: This option should not be excluded. As a matter of fact the files from different notifiers are likely to contain similar designs in similar locations. Would this option being accepted by EFSA? Industry could, therefore discuss internally on the opportunity of exchanging the information on the commercial varieties from their own proprietary field trials.**
3.2.4 Multiple endpoints	**Lines 735-737: This only holds if, in reality, no difference exists between GMO and its comparator in all p cases.** **Lines 778-780: This suggestion on FDR can it be a suggestion for further work? In which respect can industry co-operate in this work?**
3.2.3 Single endpoints in more complex experimental designs	**Lines 686-695: There seems to be some confusion here between (a) individual site vs. across site analyses, and (b) fixed vs. random effects. For an analysis across sites there is no reason to assume that a fixed effects approach will be less powerful than a random effects approach.** **Lines 691 and 847: The description of the statistical model (mixed model with fixed effects and random effects) is not clear enough.** **On line 691, it is stated “A mixed effect model can be used for the analysis of the complete data set (all sites and years) where the factors site and possibly year are assumed to be random.” This is compared to line 847, where the model is expanded to allow the calculation of equivalence limits from commercial lines included in the experiments and it is stated that: “The factors genotypegroup and testmaterial must be treated as fixed effects, whereas commercialvariety must be treated as a random effect.” Other factors taken into account should be site, block (replicate??), year and interaction between genotype*site and site*year. More explanation should be provided when these factors have to be included as fixed or as random factors.** **Additionally, it is not clear how an ANOVA can be calculated with missing values for factors.** **Is there a test data set available to see how the statistical calculations need to be done?** **Line 694: Replace "matter" with "manner"?**

3.2.2 Single endpoints in simple two-group designs	Line 622: The decision for the 90% confidence interval and the application of the 90% confidence interval in both statistical tests (equivalence test and t-test for mean differences) leads to an increase of the significance level from a = 0.05 to a = 0.1. Consequently, more statistical significant differences will be found by the t-test. However, in hazard approaches small significance level should be defined, to come to reliable and save results in the test. The significance level used should be a = 0.05 or even a = 0.01. Line 631: There is potential here for confusion as to whether the formula results in 90% or 95% confidence limits, depending on how the critical value is defined. Line 633: Log to base 10 (or indeed any other base) would be just as appropriate. Line 638: In a typical multi-site experiment, within group standard deviations are not derived for each group separately and then pooled because such an approach would not take proper account of site effects (nor block effects if blocks were present in the design). Instead, the pooled within group standard deviation needs to be derived from an analysis of variance in which all relevant sources of variation are taken into account. In addition, the text should read “$df1s1^2 + df0s0^2$” instead of “df1s1 + df0s0”. Lines 663-668: it is mentioned that there are R-library programs available for analyzing ratios. The R-library is a series of software programs generated by users and these programs are not validated. Because we do everything under GLP we would not use a series of non-validated software routines.
3.1 Introduction: choice of model and preprocessing of data	Lines 508-510: The statement that the assumption of constant CV typically breaks down at very low measurement values is questionable and certainly lacks generality; “may break down” would be a more appropriate choice or words. The meaning of the statement "log-transformed data may show too much variability" is unclear.

3 Statistical approaches	For more complex situations of simultaneous assessment of multiple endpoints, we are not sure how to approach the multivariate analysis which considers correlations among end points in the framework of mixed models. In addition, we have more characteristics than the number of observations in the data set which probably will prevent us from even conducting a conventional multivariate analysis of variance in the fixed model framework. We understand that all these issues will be addressed by future efforts of the EFSA WG on Statistics, and we are looking forward to further guidance. Meanwhile as we are still using the independent univariate evaluation of single endpoints, we do need to control the inflated type I error due to multiple comparisons on so many characteristics, especially under the framework of difference tests. We again recommend the use of false discovery rate (FDR) adjustment because it provides some degrees of control for type I error, and is a compromise to the family-wise error rate adjustment approaches that tend to cause loss of power. In addition, the beauty of FDR adjustment is that it can be applied across characteristics, and not just across multiple comparisons of means within an individual characteristic. We strongly require that the document reconsiders the necessity of multiplicity adjustment such as FDR approach. The recommendation in section 3.3.2 lines 840-848 will not work with organizations using The SAS System as observations with missing values for any of the terms in the model are deleted from the analysis. A better way would be to generate a variable like genotypegroup that has a value of 1 for the GMO, a value of 2 for the Comparator, and a value of 3 for the commercial varieties. Then the term variety nested within genotypegroup would provide the estimate of the commercial variety variance component, a contrast between 1 and 2 would provide the comparison between the mean of the GMO and the mean of the comparator, and a contrast between 2 and 3 would provide a comparison between the mean of the comparator and the mean of the commercial varieties.
APPENDIX. Proposed text for Chapter 7.2 of the EFSA Guidance Document	Line 1260: One should also look over what kind of interactions may appear. For example there may be interactions between non-random factors, between random factors and between non-random and random factors. For the latter case it has to be decided if the interaction should be regarded as random or fixed. As noted on several places above, the proposed methods in the Appendix have not been meant to cover mixed linear models. (These are the comments of an external advisor of XXX and does not necessarily represent the official view of XXX)

3 Statistical approaches	**Another important point we want to make for proof of difference and proof of safety is that they are both based solely on means, and no attention is devoted to individual data points. For the proof of difference, if the number of replications is not sufficient, confidence intervals for some characteristics will likely be too wide to indicate any difference. In contrast, for the proof of equivalence, if the number of replications is more than enough, confidence intervals for some characteristics will likely be too narrow to indicate any non-equivalence. For example, outcome 2 on the chart of line 332, the proof of difference indicates a significant difference, and the proof of equivalence indicates equivalence. So potentially if one increases number of replications, the confidence intervals could be shortened so that they will be contained within the equivalence limits even though they do not contain the non-difference value. Therefore, we think it is more appropriate to compare means using the difference tests, and supplement the mean comparisons with additional comparisons based on individual data points. Comparing individual data points of the GMO with a prediction interval or a tolerance interval computed based on commercial varieties is more appropriate because these intervals are not impacted by number of replications, and they capture the background variation of commercial varieties in the same experiment on the basis of individual data points. Finding one or more of the individual data points out side of the tolerance limits may be the most important information discovered. The final step in the analysis should be to evaluate each individual data point in comparison to the tolerance limits. Most established limits found in the literature are for individual data points and not for means. Thus, using prediction intervals or tolerance intervals from the data set would be a simulation of using established limits from the literature.**
5 Conclusions and Recommendations	**Many of the recommendations are not suitable for mixed linear models. (This is the comment of an external advisor of XXX and does not necessarily represent the official view of XXX)**

3 Statistical approaches	Lines 478 – 911: A joint graphic presentation that allows the quick comparison of the GMO and its comparator for many characteristics, in the light of background variability sounds very appealing. It is relatively straight-forward to generate the confidence intervals for the ratio of the GMO to its control and compare the confidence intervals with the no-difference value, which provides the test of difference. Working on the ratio basis turns the results scale-free and allows a simultaneous view of multiple characteristics. However, some concerns are raised regarding the scale-free equivalence limits. Exactly what are the recommended equivalence limits to use for composition of raw agricultural commodities or results from animal studies? 80% and 125% of the control mean? In section 3.2.1, the document points out that it is difficult to state that such values prescribed as a standard for pharmaceutical applications would also be suitable for the safety evaluation of GMOs. Since some compositional characteristics are inherently more variable than others, it makes more sense to use characteristic-specific equivalence limits that reflect the inherent variance of each characteristic. The procedures of estimating the characteristic-specific equivalence limits using data on commercial varieties are laid out in section 3.3.2. However, we found the formula for calculating the equivalence limits on line 863 is problematic. The center of this interval should be at the no-difference value (1 for a ratio), and not at the difference between the mean of the commercial varieties and the comparator. Applying the constant value of 1.96 assumes infinite df for the commercial variety variance component, which is not true with the number of commercial varieties included in the concurrent experiment. If an experiment contains six commercial varieties, then the number of df associated with the estimate of the commercial variety variance component would be 5 or less. These are not sufficient numbers of degrees of freedom to base the equivalence limits using the equation in line 863. Applying 1.96 assumes V is a known value with no uncertainty, which is far from truth. Even if we may be able to derive characteristic-specific equivalence limits, how can we turn them into scale-free equivalence limits for graphic presentation of multiple characteristics? More guidance is needed on this issue.

4 Proposals concerning field trial design	It is unclear to us if the concept of test site means that they should be clearly different from each other with respect to the environment. The total number of varieties in all sites, the minimum number of varieties within each site and the number of replicates within each site should be more clearly stated. In Section 4 on several places mixed (linear) models are suggested whereas in Section 3 a large part of the presentation considers simpler linear models. It is necessary to synchronize the sections somewhat. In mixed linear models it is not clear what kind of residuals should be used. For different factors to study there may be different residuals. In general (unbalanced case) estimators of the mean are weighted with functions of estimated variance components and therefore the estimated mean is not independent of the residual. For a special case of mixed linear models, i.e. the Growth Curve model and its extensions, different types of residuals have been considered. For example see Seid Hamid & von Rosen (2006), ”Residuals in the extended growth curve model”. Scand. J. Statist. 33, pp. 121–138. (These are the comments of an external advisor of XXX and does not necessarily represent the official view of XXX)
2.6 Simultaneous assessment of multiple endpoints	Lines 393 – 476: It is agreed that further work is necessary for addressing the issue of multiple endpoints, which is always the case in the case of comparative assessment of GMOs. Line 448: The test in question takes no account of relevance, so the word "relevantly" should be deleted.

2.4 Types of possible outcomes between the GMO, its comparator and commercial varieties	Line 309: To make the visualization of the test outcomes (Figure1) clearer results should be grouped by compound category (i.e. proximates, fibres, minerals, amino acids etc). Line 332 (Figure 1): This graphic representation is very useful however it implies that the lower and upper equivalence limits are identical for all endpoints, which is not the case if they are determined from the results of the commercial varieties as proposed in 3.3.1. Or it means that those lines do not correspond to a specific value. In such a case the graph will give a relative indication and it would be useful to indicate for each endpoint the actual value of each limit and of the confidence interval. In addition, in order to generate the graphical result which contains confidence intervals for the classical hypothesis testing situation, confidence intervals for the equivalence testing situation, and the equivalence limits estimated from either the commercial data or historical data, the classical hypothesis test of no difference must use a significance level of 0.10. This is contrary to our usual significance level of 0.05. Using an alpha of 0.10 will lead to more statistical differences. The concern we have is due to the perception by some regulatory agencies that statistical significance equals biological significance. Furthermore, in several instances it is stated that all responses should be included on the same graph. With composition studies which can contain approximately 60 analytes a graph with all 60 analytes will be very difficult if not impossible to interpret. The document further suggests creating the figure using relative differences which would avoid the problem of different measurement units, but this may not always be possible. Lines 346-356: The guidelines say to do both the classical test and the equivalence test separately at 0.10 level of significance. However, there is no mention of the simultaneous significance test for both tests together. In addition, there is a sequential nature to the tests which is given in lines 346-356 which states that the equivalence test will only be done under certain situations. Lines 349-356: It is not obvious that there is a need for a test per se as conclusions stem directly from the confidence intervals.
3.3.2 Use of concurrent data to estimate equivalence limits	Lines 842, 850: It is not clear how the idea of using missing values affects the analysis (model description). As it stands now it seems that one introduces a heavily unbalanced model where factors are strongly related. This leads to difficult interpretations. We think that it should be stressed that when working with unbalanced mixed linear model one has to rely on asymptotic results. For example all confidence limits are approximate. (These are the comments of an external advisor of XXX and does not necessarily represent the official view of XXX)

2.2 Error types and statistical power	**Line 206: We agree with the statement that “equivalence testing contrasts with other biological experimentation” like, in agriculture variety testing or testing of plant protection products, where the objective is to identify a product with increased performance compared to a standard.** **Line 249-252: Another positive aspect in the equivalence approach is that the problem inherent with the type II error (stating that GM and control are not different when they are in fact different = the risk for the consumer) disappears, since the null and alternative hypothesis are defined just contrary to the ones defined in the difference approach.**
3.2.4 Multiple endpoints	**Line 765: The Intersection-Union principle is mentioned but in some way we do not understand why. There are many other ways of constructing tests such as step down procedures, Union-intersection tests, etc. The problem usually encountered is that it is difficult to obtain the distribution of the test-statistic. A better reference than Berger (1982), is probably Tamhane & Logan (2004) ”A superiority-equivalence approach to one-sided tests on multiple endpoints in clinical trials” Biometrika, 91, 715–727.** **Line 792: We think one should be careful when using PCA in the present application. Usually one has few replicates, missing values cause a problem and last but not least components corresponding to small eigenvalues (indicate stability) may be of interest whereas PCA (PCR) focuses on large eigenvalues. (These are the comments of an external advisor of XXX and does not necessarily represent the official view of XXX)**
3.2.2 Single endpoints in simple two-group designs	**Line 670: It is not appropriate to refer to the R-library in the guidelines. There is no one who is responsible for the programs and the quality of the programs are very mixed: some are excellent some are less good. There is also no guarantee that the programs will be available or function in future. (This is the comment of an external advisor of XXX and does not necessarily represent the official view of XXX)**

2.1 Introduction	**Lines 177, 180 and elsewhere: The term "proof" in this context is not ideal because it makes the conclusion appear certain when in reality it is not; "evidence" might be a better term.** **Lines 198-203: The issue of availability of equivalence limits is crucial. They determine the experimental design. Therefore it would be useful to have a kind of early endorsement of the designs used by the notifiers prior to starting the tests and not discussing a posteriori when the GMO panel evaluates the data in the final notification.** **Lines 200-203: say that the equivalence limits can be estimated using the commercial varieties within the study or by using other available information. Furthermore, section 3.3.1 mentions that historical data can be used. However, the experimental design section 4 says references must be included in the field designs. It is not clear if the use of historical data refers to studies which are not composition or agronomic/phenotypic studies. This needs clarification.**
3.2.1 Equivalence limits	**Most what is mentioned in Section 3.2.1 – 3.2.2 only applies to very simple models and one may question if this should be presented in the guidelines.** **Multiple tests can be used when several response variables exist. However adjustment for multiple tests may also be used when for example testing for equality of different varieties. (These are the comments of an external advisor of XXX and does not necessarily represent the official view of XXX)**
1.2 Limitations	**Lines 138- 143: Additional datasets have been made available to EFSA for the specific purpose of this activity. The current approach has only been tested on one dataset and EFSA suggests that further proposals may be made in a second report. Such a process is not workable for product developers that have to submit studies for regulatory approval. Therefore, as mentioned above, it is premature to require changes in the statistical approaches and study design at this moment in time.** **Lines 150-152: The mandate of the self-tasking working group should be made available together with this report.**
3.1 Introduction: choice of model and preprocessing of data	**Section 3.1: Once again: the logarithmic transformation in mixed linear models gives principle problems of interpreting the model.** **Line 522: It should be Shapiro-Wilk test.** **In Section 3.1 it would be natural to comment on the variance assumption. (These are the comments of an external advisor of XXX and does not necessarily represent the official view of XXX)**

1.1 Scope	**Lines 119-135: It seems that paragraph 3 of the scope, i.e. “Undertake a feasibility study regarding the applicability of proposed statistical tools using suitable data”, is not reflected in the document.** **Line 134: The document suggests that a “feasibility study using suitable data” is required regarding the applicability of the proposed statistical tools. Indeed, it is premature to suggest important changes to generally accepted approaches in absence of such a feasibility study. Moreover, new methods need to be validated and international harmonization should be a prerequisite. In addition, applicants should be given the necessary time to adapt to new data requirements which have a major impact on the planning and design of safety studies and therefore such requirements should only be applicable for new studies carried out after the date of adoption of new guidelines, without discrediting earlier studies.**
2.6 Simultaneous assessment of multiple endpoints	**Line 424: If we may assume a multivariate normal distribution then mixed linear models multivariate analysis can be used. However for discrete or discrete/continuous response it is more complicated.** **Line 478: The graphical presentation has to be adjusted in some way if mixed linear models analysis is to be used. (These are the comments of an external advisor of XXX and does not necessarily represent the official view of XXX)**
Summary	**Line 19: mentions animal feeding trials. However the bulk of the document refers to statistical methods/analysis for composition and agronomic/phenotypic studies. Do the statistical guidelines also apply to rat and broiler feeding studies?** **In addition, do these guidelines also apply for the analysis of agronomic data?** **Lines 19 and 25: We find there to be lack of clarity in the guidance. For example in line 19, it is stated that the guidance “…proposes some minimum requirements to be met in experimental design of field trials such as the inclusion of commercial varieties in the experiments” while in line 25, there is the suggestion that background variability “may also be estimated from databases or literature”.** **Line 23-27: It seems that the WG will continue its work.** **- to test its approach on different datasets** **- to provide guidance on the simultaneous assessment of multiple endpoints** **- to provide guidance on data analysis for use in the environmental risk assessment** **Will the test design change again in the future? Will the transition phase be expanded to take into account the change in proposed statistical test design?**

2.2 Error types and statistical power	Most things which is presented in Section 2.2 – 2.4 is only valid for simple linear models and today one does not know how most of the results can be generalized to the analysis of mixed linear models which is the natural model class to consider in field experiments. Multiple comparisons in mixed linear models can only be performed under asymptotic conditions and they are certainly not satisfied in most field experiments. I think this should be made clear. (These are the comments of an external advisor of XXX and does not necessarily represent the official view of XXX)
Summary	Open comments • This document tackles a difficult topic. In general the guidance provided seems practical and accurate, but there are a few areas where clarification is needed. Particularly difficult topics, such as the problem of multiple testing due to the large number of endpoints, are appropriately left for future study. In this latter effort, the biotech industry through XXX could take part. • A better explanation is required as to the rationale for deviation from current internationally accepted practices. The concept of equivalence in the comparative assessment is the approach used worldwide. In this equivalence approach a lot of effort is put on the calculation and implementation of equivalence limits. These limits are generated from real composition data obtained from commercial lines currently on the market with a history of safe use that were grown in the same experimental design as the GM and its comparator. In consequence, the comparison of the confidence interval of the ratio between the GM and its control to these limits is much more realistic than the comparison to the no-difference value. Since the analysis of difference has to cope with ambiguous result as mentioned in line 189 of this document, a clear explanation is needed why two different test approaches for the evaluation of two data sets. In addition, as this proposed approach combining in one analysis both evaluations is really new, it will not only require validation but also further discussions between the experts of the notifiers and the experts of the GMO panel in order to ensure that we are all on the same line. • The document states that sites should be considered as random. However, it is further stated that the site by substance interaction should be tested. If sites are random then sites by substance will also be random and there is no test for interaction. It makes no sense to use sites as fixed to test for interaction and then to rerun the model using sites as random because of the affect on the significance level. • There will be instances – more complicated experimental designs (e.g., split-plots, repeated measures) where it may be difficult to apply the guidelines. If a different statistical analysis were to be done, what notification would have to be done regarding a statistical analysis different from what is recommended by the guideline?

Summary	**The comments below refer specifically to the document "Statistical considerations for the safety evaluation of GMOs". However, because of the relevance, it is important that the comments on Section 7.1.2 of the EFSA "Updated Guidance document for the risk assessment of genetically modified plants and derived food and feed" are also taken into consideration during the review of the "Statistical considerations for the safety evaluation of GMOs" document. For convenience these are included in Annex 1 of this document.**
2.1 Introduction	**Line 183: Exact equivalence limits have to our knowledged not yet been established for mixed linear models. (These are the comments of an external advisor of XXX and does not necessarily represent the official view of XXX)**
Summary	**Indeed, any new approach should be thoroughly evaluated and validated with appropriate experimental designs and datasets before being proposed for implementation. In that respect, XXX sees an opportunity for further cooperation between statistical experts of the applicants and the experts of the EFSA GMO Panel for the validation of the proposed study design and statistical analysis. Evaluation and validation will also be a prerequisite for international harmonization of the proposed study design and statistical analysis. There is a need to discuss this guidance at a global level with major regulatory bodies outside the EU to ensure harmonization of principles and study requirements.** **XXX is of the opinion that at this stage it would be premature to request the immediate implementation of the proposed study design and statistical analysis in the absence of the results of further evaluation and validation experiments.** **In any case, applicants should be given the necessary time to adapt to new data requirements which have a major impact on the planning and design of safety studies. Therefore such new requirements should only be applicable after a transition period of at least 3 years. Any retro-active applicability of the new requirements to the dossiers that have already been submitted or that are currently in preparation - the field trials for submissions up to 2011 are currently ongoing - should be avoided. As a consequence, XXX strongly recommends that EFSA continues to accept the current approaches/practices for the statistical evaluation of safety data.**
1.1 Scope	**The writing of guidelines for GMO's is an ambitious project. The main problem with writing the guidelines is that in field experiments nowadays one often uses mixed linear models whereas much of the proposed methodology is only fully developed for linear models. This one can also see from the text where a lot of detailed discussions are presented for linear models. This means that it is difficult to present guidelines which many statisticians would agree about.**

Summary	**XXX has integrated all the input from its agrifood member companies (XXX) into this one document and ensured a common position.** **XXX comments to the EFSA draft report Statistical considerations for the safety evaluation of GMOs (EFSA-Q-2006-080)** **General remarks** **XXX welcomes EFSA's efforts in providing guidance on the design of experiments and statistical approaches for the safety evaluation of GM products. However, XXX remarks that the described approach is completely new, both with regard to the proposed study design and statistical analysis for data interpretation. As stated in the guidance document, the proposed approach has only been tested on one dataset and EFSA therefore suggests that further testing on additional datasets is needed to confirm the applicability of the proposed statistical tools.**
1 Background	**Line 98: In toxicity studies one often uses dose-response modelling. For example when applying the Benchmark method (for an overview see Sand et al. (2008) ”The current state of knowledge on the use of the benchmark dose concept in risk assessment”. Journal of applied toxicology, 28, pp. 405-421). This method can be mentioned in the document. (This is the comment of an external advisor of XXX and does not necessarily represent the official view of XXX)**
Summary	**Line 14: The term compositional data is in the statistical literature something which is not considered in the present document (see Aitchison, J. The statistical analysis of compositional data. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986).** **On many places in the document one suggests to take the logarithm. In a linear model this is usually very natural. However in a mixed linear model, for example in a model with random blocks it is not natural to look at log-normally distributed random effects. In many biological material instead of taken the logarithm one assumes that the coefficient of variation is constant.** **It would be of interest to see explicitly what kind of endpoints should be analyzed. Moreover the term endpoint is often used in medicine and related areas whereas in the agricultural sciences it is less known. Why not just use response variable which indeed also is used in this document. (These are the comments of an external advisor of XXX and does not necessarily represent the official view of XXX)**

2.1 Introduction	r. 161 it is stated: "In the comparative safety assessment a GMO will be compared to an appropriate comparator1 or control organism/material. The comparison will be made by measuring a number of specific agronomical, phonotypical, and compositional characteristics of the GM plant derived foods/feed and its non-GM counterpart". A comparator is mentioned in the footnote to be: "A comparator is the non-genetically modified isogenic variety In the case of vegetatively propagated crops, and a non-GM line of comparable genetic background in the case of sexually reproducing crops (EFSA 2006)." XXX would like to request EFSA to be more explicit on this matter. What comparators are to be used in case of stacks? What comparators to use for stacks when comparing gene expression, agronomical characteristics, phenotypical traits or when performing a comparative analysis? Can EFSA give more guidance in this respect?
1.2 Limitations	EFSA states in r. 140-143: "The Working Group emphasizes that the current report represents a first analysis of approaches and limitations, and only after testing on several datasets it may be possible to make more specific proposals in a second report". However, in the APPENDIX (text to be included in the updated EFSA guidance, r. 1155 and further) the approach taken in this first analysis is presented as being requirements for experiments. We consider this as a contradiction from EFSA and we would like clarification on this issue.

3.2.2 Single endpoints in simple two-group designs	Line 650
	if y1 = y0 then t = -¿u/S.E. < 0 So p = Pr[tdf < t] < 0.50
	if y1 - y0 = ¿u then t = 0 and p = 0.50
	if y1 - y0 > ¿u then t > 0 and 0.50 < p < 1.00
	So decisions based on p-values are as follow (if we choose a = 0.05): If p > 0.95 we accept the null hypothesis = a difference between the GMO and its comparator of a certain minimum size exists at a significance level of 5%
	
	Line 651
	if y1 = y0 then t = -¿u/S.E. < 0 So p = Pr[tdf > t] lies between 1.00 and 0.50
	if y1 - y0 = ¿u then t = 0 and p = 0.50
	if y1 - y0 > ¿u then t > 0 and p < 0.50
	So decisions based on p-values are as follow (if we choose a = 0.05): If p < 0.05 we accept the null hypothesis = a difference between the GMO and its comparator of a certain minimum size exists at a significance level of 5%
	
	Do I see this correct?
	Anyhow, I believe it would be interesting to add some explanation.

5.1 Recommendations	Line 1110 : point 8 is too vague, what do you mean by « experts », « little practical relevance » and « external data » ?

4.2 Power of field trials	**Lines 1057-1058 : if spread over multiple years, we suggest to ask for a minimum number of replicates within each year.**
4.1 Experimental design	**Line 920-931: Information about DUS testing Remark: In the document it is noticed that DUS requirements comply with the general formats and crop-specific guidelines published by UPOV. This is partly correct. In the EU each country has a separate system for DUS testing, based on the Council Directive 2002/53/EC on the common catalogue of varieties of agricultural plant species (Official Journal of the European Communities (20.07.2002- L 193/1 –L 193/11). Most of the national DUS testing systems are based on the UPOV guidelines but this is not obligatory. Furthermore for a lot of crops there exist already CPVO - guidelines for DUS testing and in the framework of the EU these guidelines have to be used by the different countries in DUS-testing. Line 932- 939: Information about VCU testing Remark: In the text it is written that "In addition, from an agronomic perspective, the crop should have "value for cultivation and use" (VCU), i.e. it should have an advantage in terms of yield and/or quality over currently used varieties". Furthermore there is a reference to the CPVO website. This information given here is not really correct. The "VCU" value is described in the Council Directive 2002/53/EC on the common catalogue of varieties of agricultural plant species under Article 5 – 4. " The value of a variety for Cultivation and Use (VCU) shall be regarded as satisfactory if, compared to other varieties accepted in the catalogue of the Member states in question, its qualities, taken as a whole, offer, at least as far as production in any given region is concerned, a clear improvement either for cultivation or as regards the uses which can be made of the crops or the products derived therefrom. Where other superior characteristics are present, individual inferior characteristics may be disregarded". So this information from this directive is different from those in the Guidance document. The text in the Council directive gives also each EU-country the possibility to organise the VCU-tests in a somewhat different way and each country has also different rules for registering of new varieties. The VCU value is also only obligatory for "Agricultural species". In the directive the list of crops is given in annex. Furthermore in this directive a lot of obligations about testing GMO-varieties in the framework of registration on the common catalogue are given. So when you mention VCU- value in the document it is necessary to refer to the Council Directive 2002/53/EC and to give some additional information. Reference: Council Directive 2002/53/EC – 13th of June 2002 - on the common catalogue of varieties of agricultural plant species – Official Journal of the European Communities, L 193/1 – L 193/11.**
3.3.1 Which data can be used?	**Line 803-819: Which data can be used? … which we list in order of preference.. Remark: It is noticed that in general the first type of information is preferred. Should this not have to be changed in : "In general the first type of information is necessary". I think only in that way you will have accurate information about the new GMO variety, in comparison with other already existing varieties.**

3.2.4 Multiple endpoints	**Lines 758-764: See comment under title 2.6**
3.2.1 Equivalence limits	**Line 565-578 : From this section, it is difficult to understand if the regulation in the pharmaceutical industry is too conservative or too stringent to be translated in GMO regulation.**
3.1 Introduction: choice of model and preprocessing of data	**Line 501- 512: Logarithmic transformation of data Remark: A lot of logarithmic transformations can be used. Should it not be specified which kind of transformation is the most appropriate, depending on the type of evaluated characteristic and the distribution of data? Line 480 : We suggest to add subtitles within section 3.1. : 3.1.1 Data transformation, 3.1.2. Heteroscedasticity …**
2.6 Simultaneous assessment of multiple endpoints	**Lines 428-433: This paragraph is not clear. Confidence intervals for all the different endpoints have been determined. These are then plotted in a single graph. Some of them will lead to the conclusion of equivalence between GM and comparator, others will receive the label probable equivalence or probable non-equivalence and still others non-equivalence. But now, how can one decide by visual inspection that the GMO and its control can be termed equivalent? This problem returns in lines 758-764. Lines 439-440 : the number …. than expected. How can we calculate the « expected » ? Lines 451-452 : What do you mean by « problematic ». What is the « take home message » ? Line 476 : what do you mean by « an analysis of the frequency of significant results in the set of investigated endpoints ».**
2.5 More complex situations	**Lines 399-400 : the discussion of lognormal distribution is confusing. Sometimes, it seems that a log transformation is THE solution and, in other places, the discussion is about normality in general. Transformations such as arcsin are also useful in some cases.**
2.1 Introduction	**Lines 165-170: The issue of comparator may benefit from a better treatment.**
1 Background	**Lines 93-95: Is the use of « background variation » equivalent to the concept of « residual variation » ?**

Summary	General remarks XXX would like to remind here the remarks that have already been addressed to EFSA in a letter of XXX of 14 July 2008: - In several dossiers, XXX noticed that some animal tests were not designed according to scientific standards and should have included more animals per treatment to increase the power of the statistical analysis or the sensitivity of the trial. In consequence, it was not possible to draw any scientific conclusions from those trials. XXX is aware that in most cases, these trials were not required according to the EFSA guidance. But XXX is of the view that all scientific experiments presented in the dossiers (including supplementary studies not formally required on the basis of the EFSA guidance) should comply with standard of good design and quality for appropriate statistical analysis of the data, and should be fully considered in the context of the overall risk assessment. - XXX would like to request EFSA to be more strict with the applicants as regards the scientific quality of the dossiers. XXX would like also to discuss with EFSA how to deal with poorly designed experiments in cases where no formal requirements for such experiments are triggered. Additional comments: The guidelines should be clearer on the information that should be compulsory in the dossier : information on the a priori power, details of the design, discussion on normality …. The guidelines doesn't discuss the issue of the « agricultural system ». When studying herbicide resistant plants, what is the relevant basis of comparison : a system with another GM plant and herbicide or a system with a conventional variety without herbicide ?
APPENDIX. Proposed text for Chapter 7.2 of the EFSA Guidance Document	Line 1287: We suggest, with regard to the graphic representation of the results of the statistical analysis, using a classical box-whisker plot. The central squares would only have to be expanded by the information about medians and percentiles.
APPENDIX. Proposed text for Chapter 7.2 of the EFSA Guidance Document	Line 1328 to 1330: The document ''recommends'' further statistical assessment in case of significant difference and/or lack of equivalence: This should not be a recommendation but a requirement. Statistical analysis, however necessary, can only be the first step in further investigations. In most cases it will be necessary to obtain additional experimental data. A comparison to literature data, which will only widen the variability, and render the experimental set-up described in the document useless, should explicitly be discouraged in the document.

APPENDIX. Proposed text for Chapter 7.2 of the EFSA Guidance Document	**Line 1309-1311: Significant different results obtained by the experimental set-up described in this document should always lead to further investigation and cannot be rendered irrelevant by a discussion or comparison to literature data. It remains completely unclear what should be considered ''biologically relevant''.**
APPENDIX. Proposed text for Chapter 7.2 of the EFSA Guidance Document	**Line 1234 onwards: The presentation of the results of field trial data on plant composition should also include a site-specific analysis, and differences observed at particular sites should be addressed and discussed with respect to their environmental relevance.**
APPENDIX. Proposed text for Chapter 7.2 of the EFSA Guidance Document	**Lines 1224-1232: The whole paragraph concerning the comparative approach of GM plants containing stacked events is imprecise and not comprehensible. Therefore, we would like to ask for clarification.**
APPENDIX. Proposed text for Chapter 7.2 of the EFSA Guidance Document	**Line 1216 onwards: It is unclear, whether the recommendation that trials may be conducted in a single year or spread over multiple years refers to trials at a single site or to the entire set of trials. The definition of a ''site'' chosen for field trials is unclear.** **The applicant should be required to provide data on the exact geographical position (e.g. geographical coordinates) and a justification/prove that the chosen site is representative for a particular growing area of a particular crop.** **It should be required that field trials should take place in representative European environments if cultivation is included in the scope of the notification. Further guidance is needed with respect to the selection of such representative environments.**
APPENDIX. Proposed text for Chapter 7.2 of the EFSA Guidance Document	**Lines 1178-1196: The commercial varieties should correspond to varieties actually used in practice/available for commercial growing in the respective region where the field trials are located.** **The recommendation of only one season for field trials if a sufficient number of sites has been chosen does not take into consideration that this one season may not be representative due to extreme weather or other environmental conditions at a specific site. Therefore, field trials should last at least two seasons.** **It is also unclear, in which cases field trials for more than one growing season are requested. The criteria that this is only the case if ''the choice of sites is over a very restricted geographic range'' should be further specified. In general, all field trials at a specific site should be conducted for more than one season only (see above).**

3.3.1 Which data can be used?	**Line 811 and 813: The use of "historical data" should be avoided as in addition to the comment on Line 809, which also holds true for "historical data", the use of historical data may lead to masking significant differences, because this approach may lead to much wider equivalence limits.**
3.3.1 Which data can be used?	**Line 809: The reasoning behind using "other experiments" as a comparator is not clear. If a field trial is designed, it should not be a problem for the applicant to include a reasonable number of commercial varieties in order to obtain the relevant data. However, if data from "other experiments" are used, it has to be ensured that the data which are to be compared are obtained in experiments which are designed and carried out in the same way, otherwise it is nothing more than a comparison of apples and oranges. This has to be stated clearly in the document.**
2.4 Types of possible outcomes between the GMO, its comparator and commercial varieties	**Lines 389-391: According to the Draft Report, "risk assessors should specify if further evaluation of data is needed and, if so, what it should be". From our point of view, decisions on the necessity of further assessment should not depend on the assessor''s decision alone, but should be based on clear conditions specified in this document with regard to the obtained results of the statistical analysis.**
1.1 Scope	**Comprehensive guidance for statistic approaches generally needs to include information on mainly three parts, i.e. protocols for data collection, methods of statistical analysis and, finally, the interpretation of the results. All three parts are regarded equally important, and therefore, should be taken into account in the Draft Report. In particular, more guidance on the collection of data is strongly recommended.** **Relating to the collection of data is important that transparency is established by using standardised operating procedures, particularly with respect to possible small sample sizes (see line 452). This problem has clearly to be addressed in this Draft Report.** **Furthermore, guidance on the interpretation of the results of the statistical analysis is missing, and thus needs to be supplemented, taking account of all influencing factors including e.g. the chosen test design for data collection and the applied statistical method.**

2.1 Introduction	Comprehensive guidance for statistic approaches generally needs to include information on mainly three parts, i.e. protocols for data collection, methods of statistical analysis and, finally, the interpretation of the results. All three parts are regarded equally important, and therefore, should be taken into account in the Draft Report. In particular, more guidance on the collection of data is strongly recommended. Relating to the collection of data is important that transparency is established by using standardised operating procedures, particularly with respect to possible small sample sizes (see line 452). This problem has clearly to be addressed in this Draft Report. Furthermore, guidance on the interpretation of the results of the statistical analysis is missing, and thus needs to be supplemented, taking account of all influencing factors including e.g. the chosen test design for data collection and the applied statistical method.
2.6 Simultaneous assessment of multiple endpoints	Comprehensive guidance for statistic approaches generally needs to include information on mainly three parts, i.e. protocols for data collection, methods of statistical analysis and, finally, the interpretation of the results. All three parts are regarded equally important, and therefore, should be taken into account in the Draft Report. In particular, more guidance on the collection of data is strongly recommended. Relating to the collection of data is important that transparency is established by using standardised operating procedures, particularly with respect to possible small sample sizes (see line 452). This problem has clearly to be addressed in this Draft Report. Furthermore, guidance on the interpretation of the results of the statistical analysis is missing, and thus needs to be supplemented, taking account of all influencing factors including e.g. the chosen test design for data collection and the applied statistical method.
1 Background	This Draft Report, in principle, has been compiled to give additional guidance for the analysis of both: data from field trials and animal feeding studies (see lines 18-19). However, chapter 4 and 5 in the Draft Report are dealing with experimental design of field trials and provide advice on that matter only. Therefore, it is highly recommended that additional chapters regarding the experimental design and specific recommendations for animal feeding studies are compiled as well.
4 Proposals concerning field trial design	This Draft Report, in principle, has been compiled to give additional guidance for the analysis of both: data from field trials and animal feeding studies (see lines 18-19). However, chapter 4 and 5 in the Draft Report are dealing with experimental design of field trials and provide advice on that matter only. Therefore, it is highly recommended that additional chapters regarding the experimental design and specific recommendations for animal feeding studies are compiled as well.

5 Conclusions and Recommendations	This Draft Report, in principle, has been compiled to give additional guidance for the analysis of both: data from field trials and animal feeding studies (see lines 18-19). However, chapter 4 and 5 in the Draft Report are dealing with experimental design of field trials and provide advice on that matter only. Therefore, it is highly recommended that additional chapters regarding the experimental design and specific recommendations for animal feeding studies are compiled as well.
Summary	This Draft Report, in principle, has been compiled to give additional guidance for the analysis of both: data from field trials and animal feeding studies (see lines 18-19). However, chapter 4 and 5 in the Draft Report are dealing with experimental design of field trials and provide advice on that matter only. Therefore, it is highly recommended that additional chapters regarding the experimental design and specific recommendations for animal feeding studies are compiled as well.
References	test: references
APPENDIX. Proposed text for Chapter 7.2 of the EFSA Guidance Document	test: appendix
2.4 Types of possible outcomes between the GMO, its comparator and commercial varieties	test: 2.4 types of possible...
1.2 Limitations	test: 1.2 Limitations
3.2.4 Multiple endpoints	test: 3.2.4. multiple endpoints