TWF/45/15
ORIGINAL: English
DATE: May 21, 2014

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS

Geneva

TECHNICAL WORKING PARTY FOR FRUIT CROPS

Forty-Fifth Session
Marrakesh, Morocco, May 26 to 30, 2014

REVISION OF DOCUMENT TGP/8: PART I: DUS TRIAL DESIGN AND DATA ANALYSIS, NEW SECTION: MINIMIZING THE VARIATION DUE TO DIFFERENT OBSERVERS

Document prepared by the Office of the Union

Disclaimer: this document does not represent UPOV policies or guidance

1. The purpose of this document is to present a draft of a new section for document TGP/8 Part I: DUS Trial and Design and Data Analysis, on "Minimizing the Variation due to Different Observers".

2. The following abbreviations are used in this document:

CAJ:	Administrative and Legal Committee
TC:	Technical Committee
TC-EDC:	Enlarged Editorial Committee
TWA:	Technical Working Party for Agricultural Crops
TWV:	Technical Working Party for Vegetables
TWC:	Technical Working Party on Automation and Computer Programs
TWF:	Technical Working Party for Fruit Crops
TWO:	Technical Working Party for Ornamental Plants and Forest Trees
TWPs:	Technical Working Parties

3. The structure of this document is as follows:

BACKGROUND

4. Document TGP/8/1 Draft 7 PART I, paragraph 2.9.1: "Control of variation due to different observers", considered by the Technical Working Parties, at their sessions in 2007, states:

> *[If this section is required, TWPs are invited to contribute guidance on the control of variation due to different observers when statistical analysis is not used to determine distinctness and to consider it in relation to paragraph 2.7.2.9.]*

5. The TWC at its twenty-sixth session, held in Jeju, Republic of Korea, from September 2 to September 5, 2008 agreed that Mr. Gerie van der Heijden (Netherlands) would consult his Naktuinbouw colleagues in the Netherlands to see if they could contribute a draft for this section.

6. The TWV at its forty-second session, held in Cracow, Poland, from June 23 to 27, 2008, noted that it had encouraged the development of that section and agreed that it should provide suitable text for aspects which were not adequately covered in document TWC/25/12.

COMMENTS BY THE TECHNICAL WORKING PARTIES IN 2013

7. The TWO, TWF, TWV, TWC and TWA considered documents TWO/46/14, TWF/44/14, TWV/47/14, TWC/31/14 and TWA/42/14, respectively (see document TWO/46/29 “Report”, paragraphs 30 to 32, document TWF/44/31 “Report”, paragraphs 33 to 35, document TWV/47/34 “Report”, paragraphs 33 to 35, document TWC/31/32 “Report”, paragraphs 30 to 32, and document TWA/42/31 “Report”, paragraphs 33 to 35).

8. The TWO proposed that experts from Australia, Germany, the Netherlands and the United Kingdom help to develop further guidance on the proposed text to be included in TGP/8 part I: DUS Trial and Design and Data Analysis, New Section: Minimizing the Variation due to Different Observers, in a future revision of document TGP/8, with regard to guidance on PQ and QN/MG characteristics (see document TWO/46/29 “Report”, paragraph 31).

9. The TWO noted, however, the importance of the Test Guidelines in providing clear guidance for DUS examiners and to ensure consistency of observations (see document TWO/46/29 “Report”, paragraph 32).

10. The TWF agreed that the variation due to different observers was not relevant in fruit DUS testing as observations were usually made by a single observer, and therefore the TWF considered it unnecessary to provide experts to develop further guidance on the proposed text to be included in TGP/8 part I: DUS Trial and Design and Data Analysis, New Section: Minimizing the Variation due to Different Observers, in a future revision of document TGP/8 (document TWF/44/31 “Report”, paragraph 34).

11. The TWF noted, however, the importance of the quality of the Test Guidelines in providing clear guidance for DUS examiners and in ensuring the consistency of observations. In that regard, the TWF recalled the work done previously on the consistency of variety descriptions in strawberry and apple (see document TWF/35/4). The TWF proposed that the expert from New Zealand report at the forty-fifth session, on the work done on the “Publication of harmonized variety description for apple for an agreed set of varieties”, in order to consider if it could be relevant to further develop the study (document TWF/44/31 “Report”, paragraph 35).

12. The TWV proposed that experts from the European Union, France and Netherlands help the drafter to develop further guidance on the proposed text to be included in TGP/8 part I: DUS Trial and Design and Data Analysis, New Section: Minimizing the Variation due to Different Observers, in a future revision of document TGP/8 (document TWV/47/34 “Report”, paragraph 34).

13. The TWV and the TWA noted that the expert from the Netherlands would draft, in conjunction with other experts, a proposed text with regard to further guidance on PQ and QN/MG characteristics, to be circulated to the groups of experts of the other interested TWPs (document TWV/47/34 “Report”, paragraph 35 and document TWA/42/31 “Report”, paragraph 34).

14. The TWC noted that the drafter from the Netherlands was no longer participating in the TWC meetings and that it was not possible to indicate another expert(s) from the TWC to continue the work. However, the TWC noted that the TWO and TWV had proposed experts to help to develop further guidance, on the proposed text to be included in TGP/8 Part I: DUS Trial and Design and Data Analysis, New Section: Minimizing the Variation due to Different Observers, in a future revision of document TGP/8, with regard to guidance on PQ and QN/MG characteristics (document TWC/31/32 “Report”, paragraph 31).

15. The TWC noted that the TWF had proposed that an expert from New Zealand would report at its forty-fifth session, on the work done on the “Publication of harmonized variety description for apple for an agreed set of varieties”, in order to consider if it could be relevant to further develop the study (document TWC/31/32 “Report”, paragraph 32).

16. The TWA proposed that the TWA experts from Australia and the Netherlands assist the drafter to develop further guidance on the proposed text to be included in TGP/8 part I: DUS Trial and Design and Data Analysis, New Section: Minimizing the Variation due to Different Observers, in a future revision of document TGP/8 (document TWA/42/31 "Report", paragraph 35).

DEVELOPMENTS IN 2014

Technical Committee

17. The TC at its fiftieth session, held in Geneva from April 7 to 9, 2014, considered document TC/50/21 "Revision of document TGP/8: Part I: DUS Trial Design and Data Analysis, New Section: Minimizing the Variation due to Different Observers".

18. The TC noted that the TWF had requested an expert from New Zealand to report, at its session in 2014, on the previous work done on harmonized variety description for apple for an agreed set of varieties, as set out in paragraph 11 of this document (see document TC/50/36 "Report on the Conclusions", paragraph 46). The expert of New Zealand will present a document at the forty-fifth session of the TWF in 2014, in conjunction with document TWF/45/28 "Harmonized example varieties for Apple: historical data and possible new developments".

19. The TC invited the expert from Australia, with the assistance of experts from the European Union, France, Germany, the Netherlands and the United Kingdom, to draft further guidance to be included in a future revision of document TGP/8 on minimizing the variation due to different observers, including guidance on PQ and QN/MG characteristics, for consideration by the TWPs at their sessions in 2014 (see document TC/50/36 "Report on the Conclusions", paragraph 47).

20. In response to the request of the TC, the drafter from Australia (Mr. Nik Hulse) with the assistance of experts from the European Union, France, Germany, the Netherlands and the United Kingdom, provided draft guidance for future revision of document TGP/8 on minimizing the variation due to different observers, including guidance on PQ and QN/MG characteristics, as presented in the Annex to this document.

21. The expert from Australia proposed that the following points should be incorporated into the document:

- As already indicated in the draft, QN/MG can be dealt with in a similar way as QN/MS. However, it is important to note that in the case of QN/MG, the possible effect of random within-plot variation should also be considered.

- Further consideration is required on how to incorporate guidance on PQ characteristics. It should be explained that TGP/14 is another useful tool in clarifying many PQ characteristics (e.g. shape). Differences between observers could possibly be tested by frequency of deviations.

- Currently the document focus is on variation between observers at the authority level. Consideration could be given to whether minimizing observer variation between authorities should be mentioned in this document or, perhaps, a separate future document. Noting that such a consideration would introduce a greater number of factors such as GxE variation, sampling methods and ring tests.

22. The TWF is invited to:

(a) note that the expert from New Zealand will present, at its forty-fifth session, a document on the previous work done on harmonized variety description for apple for an agreed set of varieties, in conjunction with document TWF/45/28; and

(b) consider draft guidance in the Annex to this document, for inclusion in a future revision of document TGP/8 on minimizing the variation due to different observers, including guidance on PQ and QN/MG characteristics, in conjunction with the points raised by the expert from Australia in paragraph 21 of this document.

[Annex follows]

ANNEX

TGP/8/1: PART I: NEW SECTION: MINIMIZING THE VARIATION DUE TO DIFFERENT OBSERVERS OF THE SAME TRIAL

1. Introduction

This document has been prepared with QN/MS, QN/VG and QN/VS characteristics in mind. It does not explicitly deal with PQ characteristics like color and shape. The described Kappa method in itself is largely applicable for these characteristics, e.g. the standard Kappa characteristic is developed for nominal data. However, the method has not been used on PQ characteristics to our knowledge and PQ characteristics may also require extra information on calibration. As an example, for color calibration, you also have to take into account the RHS Colour chart, the lighting conditions and so on. These aspects are not (yet) covered in this document.

1.1 Variation in measurements or observations can be caused by many different factors, like the type of crop, type of characteristic, year, location, trial design and management, method and observer. Especially for visually assessed characteristics (QN/VG or QN/VS) differences between observers can be the reason for large variation and potential bias in the observations. An observer might be less well trained, or have a different interpretation of the characteristic. So, if observer A ~~measures~~ assesses variety 1 and observer B variety 2, the difference ~~measures~~ observed might be caused by differences between observers A and B instead of differences between varieties 1 and 2. Clearly, our main interest lies with the differences between varieties and not with the differences between the observers. It is important to realize that the variation caused by different observers cannot be eliminated, but there are ways to control it.

2. Training and importance of clear explanations of characteristics and method of observation

2.~~2~~1 Training of new observers is essential for consistency and continuity of plant variety observations. Calibration manuals, supervision and guidance by experienced observers as well as the use of example varieties illustrating the range of expressions are useful ways to achieve this.

2.~~2~~1 UPOV test guidelines try to harmonize the variety description process and describe as clearly as possible the characteristics of a crop and the states of expression. This is the first step in controlling variation and bias. However, the way that a characteristic is observed or measured may vary per location or testing authority. Calibration manuals made by the local testing authority are very useful for the local implementation of the UPOV test guideline. Where needed these crop-specific manuals explain the characteristics to be observed in more detail, and specify when and how they should be observed. Furthermore they may contain pictures and drawings for each characteristic, often for every state of expression of a characteristic. ~~The calibration manual can be used by inexperienced observers but are also useful for more experienced or substitute observers, as a way to recalibrate themselves.~~

3. Testing the calibration

3.1 After training an observer, the next step could be to test the performance of the observers in a calibration experiment. This is especially useful for inexperienced observers who have to make visual observations (QN/VG and QN/VS characteristics). If making ~~VG~~ visual observations, they should preferably pass a calibration test prior to making observations in the trial. But also for experienced observers, it is useful to test themselves on a regular basis to verify if they still fulfill the calibration criteria.

3.2 A calibration experiment can be set up and analyzed in different ways. Generally it involves multiple observers, measuring the same set of material and assessing differences between the observers.

~~3.3 In general, inexperienced observers are less likely to be entrusted to make VG observations but might be entrusted tomake MG and MS observations.~~

4. Testing the calibration for QN/MG or QN/MS characteristics

4.1 For observations made by measurement tools, like rulers (often QN/MS characteristics), the measurement is often made on an interval or ratio scale. In this case, the approach of Bland and Altman (1986) can be used. This approach starts with a plot of the scores for apair of observers in a scatter plot, and compare it with the line of equality (where y=x). This helps the eye gauging the degree of agreement between measurements of the same object. In a next step, the difference per object is taken and a plot is constructed with on the y-axis the difference between the observers and on the x-axis either the index of the object, or the mean value of the object. By further drawing the horizontal lines y=0, y=mean(difference) and

the two lines y = mean(difference) ± 2 x standard deviation, the bias between the observers and any outliers can easily be spotted. Similarly we can also study the difference between the measurement of each observer and the average measurement over all observers. Test methods like the paired t-test can be applied to test for a significant deviation of the observer from another observer or from the mean of the other observers.

4.2 By taking two measurements by each observer of every object, we can look at the differences between these two measurements. If these differences are large in comparison to those for other observers, this observer might have a low repeatability. By counting for each observer the number of moderate and large outliers (e.g. larger than 2 times and 3 times the standard deviation respectively) we can construct a table of observer versus number of outliers, which can be used to decide if the observer fulfills quality assurance limits.

4.3 Other quality checks can be based on the repeatability and reproducibility tests for laboratories as described in ISO 5725-2. Free software is available on the ISTA website to obtain values and graphs according to this ISO standard.

4.4 In many cases of QN/MG or QN/MS, a good and clear instruction usually suffices and variation or bias in measurements between observers is often negligible. If there is reason for doubt, a calibration experiment as described above can help in providing insight in the situation.

5. Testing the calibration for QN/VS or QN/VG characteristics

5.1 For the analysis of ordinal data (QN/VS or QN/VG characteristics), the construction of contingency tables between each pair of observers for the different scores is instructive. A test for a structural difference (bias) between two observers can be obtained by using the Wilcoxon Matched-Pairs test (often called Wilcoxon Signed-Ranks test).

5.2 To measure the degree of agreement the Cohen's Kappa (κ) statistic (Cohen, 1960) is often used. The statistic tries to accounts for random agreement:κ = (P(agreement) – P(e)) / (1-P(e)), where P(agreement) is the fraction of objects which are in the same class for both observers (the main diagonal in the contingency table), and P(e) is the probability of random agreement, given the marginals (like in a Chi-square test). If the observers are in complete agreement the Kappa value κ = 1. If there is no agreement among the observers, other than what would be expected by chance (P(e)), then κ = 0.

5.3 The standard Cohen's Kappa statistic only considers perfect agreement versus non-agreement. If one wants to take the degree of disagreement into account (for example with ordinal characteristics), one can apply a linear or quadratic weighted Kappa (Cohen, 1968). If we want to have a single statistic for all observers simultaneously, a generalized Kappa coefficient can be calculated. Most statistical packages, including SPSS, Genstat and R (package Concord), provide tools to calculate the Kappa statistic.

5.4 As noted, a low κ-value indicates poor agreement and values close to 1 indicate excellent agreement. Often scores between 0.6-0.8 are considered to indicate substantial agreement, and above 0.8 to indicate almost perfect agreement. If needed, z-scores for kappa (assuming an approximately normal distribution) are available. The criteria for experienced DUS experts could be more stringent than for inexperienced staff.

6. Trial design

6.1 If we have multiple observers in a trial, the best approach is to have one person observe one or more complete replications. In that case, the correction for block effects also accounts for the bias between observers. If more than one observer per replication is needed, extra attention should be given to calibration and agreement. In some cases, the use of incomplete block designs (like alpha designs) might be helpful, and an observer can be assigned to the sub blocks. In this way we can correct for ~~the~~ systematic differences between observers.

7. Example of Cohen's Kappa

7.1 In this example, there are three observers and 30 objects (plots or varieties).The characteristic is observed on a scale of 1 to 6.The raw data and their tabulated scores are given in the following tables.

Variety	Observer 1	Observer 2	Observer 3
V1	1	1	1
V2	2	1	2
V3	2	2	2
V4	2	1	2
V5	2	1	2
V6	2	1	2
V7	2	2	2
V8	2	1	2
V9	2	1	2
V10	3	1	3
V11	3	1	3
V12	3	2	2
V13	4	5	4
V14	2	1	1
V15	2	1	2
V16	2	2	3
V17	5	4	5
V18	2	2	3
V19	1	1	1
V20	2	2	2
V21	2	1	2
V22	1	1	1
V23	6	3	6
V24	5	6	6
V25	2	1	2
V26	6	6	6
V27	2	6	2
V28	5	6	5
V29	6	6	5
V30	4	4	4

Scores for variety	1	2	3	4	5	6
V1	3	0	0	0	0	0
V2	1	2	0	0	0	0
V3	0	3	0	0	0	0
V4	1	2	0	0	0	0
V5	1	2	0	0	0	0
V6	1	2	0	0	0	0
V7	0	3	0	0	0	0
V8	1	2	0	0	0	0
V9	1	2	0	0	0	0
V10	1	0	2	0	0	0
V11	1	0	2	0	0	0
V12	0	2	1	0	0	0
V13	0	0	0	2	1	0
V14	2	1	0	0	0	0
V15	1	2	0	0	0	0
V16	0	2	1	0	0	0
V17	0	0	0	1	2	0
V18	0	2	1	0	0	0
V19	3	0	0	0	0	0
V20	0	3	0	0	0	0
V21	1	2	0	0	0	0
V22	3	0	0	0	0	0
V23	0	0	1	0	0	2
V24	0	0	0	0	1	2
V25	1	2	0	0	0	0
V26	0	0	0	0	0	3
V27	0	2	0	0	0	1
V28	0	0	0	0	2	1
V29	0	0	0	0	1	2
V30	0	0	0	3	0	0

The contingency table for observer 1 and 2 is:

O1\O2	1	2	3	4	5	6	Total
1	3	0	0	0	0	0	3
2	10	5	0	1	0	1	17
3	2	1	0	0	0	0	3
4	0	0	0	1	0	0	1
5	0	0	0	1	0	2	3
6	0	0	1	0	0	2	3
Total	15	6	1	3	0	5	30

The Kappa coefficient between observer 1 and 2, κ(O1,O2) is calculated as follows:

- κ(O1,O2) = (P(agreement between O1 and O2) – P(e)) / (1 – P(e)) where:
- P(agreement) = (3+5+0+1+0+2)/30 = 11/30 ≈ 0.3667 (diagonal elements)
- P(e) = (3/30).(15/30) + (17/30).(6/30) + (3/30).(1/30) + (1/30).(3/30) + (3/30).(0/30) + (3/30).(5/30) ≈ 0.1867. (pair-wise margins)
- So κ(O1,O2) ≈ (0.3667-0.1867) / (1-0.1867) ≈ 0.22

This is a low value, indicating very poor agreement between these two observers. There is reason for concern and action should be taken to improve the agreement. Similarly the values for the other pairs can be calculated: κ(O1,O3) ≈ 0.72, κ(O2,O3) ≈ 0.22.Observer 1 and 3 are in good agreement. Observer 2 is clearly different from 1 and 3 and the reasons for the deviation requires further investigation (e.g. consider need for additional training) ~~probably needs additional training~~.

8. References

Cohen, J. (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20: 37-46.

Cohen, J. (1968) Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4): 213-220.

Bland, J. M. Altman D. G. (1986) Statistical methods for assessing agreement between two methods of clinical measurement, Lancet: 307–310.

http://www.seedtest.org/en/stats-tool-box-_content---1--1143.html (ISO 5725-2 based software)

[End of Annex and of document]