S

TC/49/22
ORIGINAL: Inglés
FECHA: 18 de febrero de 2013

UNIÓN INTERNACIONAL PARA LA PROTECCIÓN DE LAS OBTENCIONES VEGETALES
Ginebra

COMITÉ TÉCNICO

Cuadragésima novena sesión
Ginebra, 18 a 20 de marzo de 2013

REVISIÓN DEL DOCUMENTO TGP/8: PARTE I: DISEÑO DE ENSAYOS DHE Y ANÁLISIS DE DATOS, NUEVA SECCIÓN: CONTROL DE LA VARIACIÓN RESULTANTE DE LA EJECUCIÓN DE LOS ENSAYOS POR DISTINTOS OBSERVADORES

Documento preparado por la Oficina de la Unión

1. El presente documento tiene por finalidad presentar un borrador de una nueva sección del documento TGP/8 Parte I: Diseño de ensayos DHE y análisis de datos sobre el "Control de la variación resultante de la ejecución de los ensayos por distintos observadores".

2. En el presente documento se utilizan las siguientes abreviaturas:

CAJ:	Comité Administrativo y Jurídico
TC:	Comité Técnico
TC-EDC:	Comité de Redacción Ampliado
TWA:	Grupo de Trabajo Técnico sobre Plantas Agrícolas
TWC:	Grupo de Trabajo Técnico sobre Automatización y Programas Informáticos
TWF:	Grupo de Trabajo Técnico sobre Plantas Frutales
TWO:	Grupo de Trabajo Técnico sobre Plantas Ornamentales y Cultivos Forestales
TWP:	Grupos de Trabajo Técnico
TWV:	Grupo de Trabajo Técnico sobre Hortalizas

3. La estructura del presente documento es la siguiente:

ANTECEDENTES

4. El Comité Técnico (TC), en su cuadragésima octava sesión, celebrada en Ginebra del 26 al 28 de marzo de 2012, examinó la revisión del documento TGP/8 “Diseño de ensayos y técnicas utilizados en el examen de la distinción, la homogeneidad y la estabilidad” sobre la base del documento TC/48/19 Rev.

5. El TC, en su cuadragésima octava sesión, acordó solicitar al redactor que elabore un nuevo borrador de la sección "Control de la variación resultante de la ejecución de los ensayos por distintos observadores", basándose en los comentarios formulados por los TWP en 2011, que figuran en el Anexo II del documento TC/48/19 Rev. (véase el párrafo 51 del documento TC/48/22 “Informe sobre las conclusiones”).

6. El TC tomó nota de que sería necesario elaborar nuevos borradores de las secciones pertinentes antes del 26 de abril de 2012 para que puedan incluirse en el proyecto que se someterá al examen de los TWP en sus reuniones de 2012 (véase el párrafo 49 del documento TC/48/22 “Informe sobre las conclusiones”).

COMENTARIOS DE LOS GRUPOS DE TRABAJO TÉCNICO EN 2012

7. En sus sesiones de 2012, el TWA, el TWV, el TWC, el TWF y el TWO examinaron los documentos TWA/41/24, TWV/46/24, TWC/30/24, TWF/43/24 y TWO/45/24 respectivamente, en los que figura el texto propuesto de una nueva sección del documento TGP/8 Parte I: “Diseño de ensayos DHE y análisis de datos" sobre el "Control de la variación resultante de la ejecución de los ensayos por distintos observadores” preparado por el Sr. Gerie van der Heijden (Países Bajos), que figura en el Anexo del presente documento (sólo en inglés[*]), y formularon los siguientes comentarios:

Observaciones generales	El TWV examinó el documento TWC/46/24 y destacó la importancia de la calibración efectuada por el observador (véase el párrafo 35 del documento TWV/46/41 "*Report*").	TWV
	El TWC examinó el documento TWC/30/24 y recomendó que debería someterse a consideración del TC para su inclusión en el TGP/8 tras la modificación de la última frase de la sección 6.1 de modo que el texto sea: "las diferencias sistemáticas" (véase el párrafo 23 del documento TWC/30/41 "*Report*").	TWC

COMENTARIOS DEL COMITÉ DE REDACCIÓN AMPLIADO (TC-EDC) EN 2013

8. El TC-EDC, en su reunión celebrada en Ginebra los días 9 y 10 de enero de 2013, examinó el documento TC-EDC/Jan13/9: “*Revision of document TGP/8 part I: DUS Trial and Design and Data Analysis, New Section: Minimizing the Variation due to Different Observers*” (Revisión del documento TGP/8 Parte I: Diseño de ensayos DHE y análisis de datos, nueva sección: Control de la variación resultante de la ejecución de los ensayos por distintos observadores), que figura en el Anexo del presente documento, y formuló los siguientes comentarios:

Observaciones generales	El documento debe abarcar también los caracteres PQ (por ejemplo, color, forma, etc.)
Título	el título de la versión inglesa debe ser: REVISION OF DOCUMENT TGP/8 PART I: DUS TRIAL DESIGN ~~OF~~ AND DATA ANALYSIS, NEW SECTION: MINIMIZING THE VARIATION DUE TO DIFFERENT OBSERVERS
Párrafo 1 del Anexo	Se debe mencionar no sólo el carácter QN/MS sino también el carácter QN/MG
Párrafo 1.1 del Anexo	El texto debe ser: "... Así pues, si el observador evalúa ~~mide~~ la variedad 1 y el observador B evalúa la variedad 2, la diferencia observada ~~medida~~ podría deberse a las diferencias entre los observadores A y B, en lugar de tratarse de diferencias entre las variedades 1 y 2. Claramente, lo que más interesa son las diferencias entre las variedades y no entre los observadores…"
Párrafos 2.1 y 2.2 del Anexo	Modificar la numeración
2.1	Suprimir la última frase

[*] En su reunión del 9 y 10 de enero de 2013, el TC-EDC convino en que no era adecuado traducir el texto para la cuadragésima novena sesión del TC.

3.1	El texto debe ser: "El paso siguiente a la capacitación del observador podría ser probar el desempeño de éste en un experimento de calibración. Ello es útil en particular para los observadores inexpertos que tengan que realizar observaciones visuales (caracteres QN/VG y QN/VS). Si realizan observaciones visuales ~~VG~~, de preferencia, deberían superar un examen de calibración antes de realizar observaciones en el ensayo. Sin embargo, también para los observadores expertos es útil ponerse a prueba periódicamente para verificar que aún satisfacen los criterios de calibración".
3.3	Suprimir.
4.	El texto debe ser: "Prueba de calibración para los caracteres QN/MG o QN/MS"
4.1	Añadir una línea en blanco después del párrafo,

9. Se invita al TC a solicitar la elaboración de un nuevo borrador de la sección sobre el "Control de la variación resultante de la ejecución de los ensayos por distintos observadores" para someterlo a consideración de los TWP en sus sesiones de 2013, basándose en los comentarios formulados por los TWP y el TC-EDC.

[Sigue el Anexo]

ANEXO

(SÓLO EN INGLÉS)

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

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 variety 1 and observer B variety 2, the difference measured 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

2.2 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.

2.1 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 characteristics). If making VG 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 to make MG and MS observations.

4 Testing the calibration for 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 a pair 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 standard measurement methods as described in ISO 5725-2. Free software is available on the ISTA website to obtain values and graphs for seed laboratory tests according to this ISO standard.

4.4 In many cases of 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:$\kappa = (P(agreement) - P(e)) / (1-P(e))$, 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 $\kappa = 1$. If there is no agreement among the observers, other than what would be expected by chance (P(e)), then $\kappa = 0$.

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

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 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)

[Fin del Anexo y del documento]