



TC/49/22

ORIGINAL: Englisch

DATUM: 18. Februar 2013

INTERNATIONALER VERBAND ZUM SCHUTZ VON PFLANZENZÜCHTUNGEN

Genf

TECHNISCHER AUSSCHUSS

Neunundvierzigste Tagung Genf, 18. bis 20. März 2013

ÜBERARBEITUNG VON DOKUMENT TGP/8 TEIL I: DUS-PRÜFUNGSANLAGE UND DATENANALYSE,
NEUER ABSCHNITT: MINIMIERUNG DER VARIATION INFOLGE VERSCHIEDENER ERFASSER

Vom Verbandsbüro erstelltes Dokument

1. Zweck dieses Dokuments ist es, einen Entwurf für einen neuen Abschnitt zur Aufnahme in Dokument TGP/8 TEIL I: DUS-Prüfungsanlage und Datenanalyse, „Minimierung der Variation infolge verschiedener Verfasser“, vorzulegen.

2. Folgende Abkürzungen werden in diesem Dokument verwendet:

CAJ	Verwaltungs- und Rechtsausschuß
TC:	Technischer Ausschuß
TC-EDC:	Erweiterter Redaktionsausschuß
TWA:	Technische Arbeitsgruppe für landwirtschaftliche Arten
TWC:	Technische Arbeitsgruppe für Automatisierung und Computerprogramme
TWF:	Technische Arbeitsgruppe für Obstarten
TWO:	Technische Arbeitsgruppe für Zierpflanzen und forstliche Baumarten
TWV:	Technische Arbeitsgruppe für Gemüsearten
TWP:	Technische Arbeitsgruppen

3. Der Aufbau dieses Dokuments ist nachstehend zusammengefaßt:

HINTERGRUND	1
BEMERKUNGEN DER TECHNISCHEN ARBEITSGRUPPEN (TWP) IM JAHR 2012	2
BEMERKUNGEN DES ERWEITERTEN REDAKTIONSAUSSCHUSSES (TC-EDC) IM JAHR 2013	2
ANLAGE: VORGESCHLAGENER WORTLAUT ZUR AUFNAHME IN TGP/8 TEIL I: DUS-PRÜFUNGSANLAGE UND DATENANALYSE, NEUER ABSCHNITT: MINIMIERUNG DER VARIATION INFOLGE VERSCHIEDENER VERFASSEN (NUR AUF ENGLISCH)	

HINTERGRUND

4. Der Technische Ausschuß (TC) prüfte auf seiner achtundvierzigsten Tagung vom 26. bis 28. März 2012 in Genf die Überarbeitung von Dokument TGP/8 „Prüfungsanlage und Verfahren für die Prüfung der Unterscheidbarkeit, der Homogenität und der Beständigkeit“ auf der Grundlage von Dokument TC/48/19 Rev.

5. Der TC vereinbarte auf seiner achtundvierzigsten Tagung, den Verfasser zu ersuchen, auf der Grundlage der von den TWP im Jahr 2011 abgegebenen Kommentare, wie in Dokument TC/48/19 Rev.,

Anlage II dargelegt (vergleiche Dokument TC/48/22 „Bericht über die Entschließungen“, Absatz 51), einen neuen Entwurf des Abschnitts „Minimierung der Variation infolge verschiedener Verfasser“ auszuarbeiten.

6. Der TC nahm zur Kenntnis, daß neue Entwürfe maßgeblicher Abschnitte bis zum 26. April 2012 ausgearbeitet werden müssen, damit die Abschnitte in den von den TWP auf ihren Tagungen im Jahr 2012 zu prüfenden Entwurf aufgenommen werden können (vergleiche Dokument TC/48/22 „Bericht über die Entschließungen“, Absatz 49).

BEMERKUNGEN DER TECHNISCHEN ARBEITSGRUPPEN (TWP) IM JAHR 2012

7. Auf ihren Tagungen im Jahr 2012 prüften die TWA, TWV, TWC, TWF und die TWO die Dokumente TWA/41/24, TWV/46/24, TWC/30/24, TWF/43/24 und TWO/45/24 mit dem vorgeschlagenen Wortlaut für einen neuen Abschnitt zur Aufnahme in Dokument TGP/8 Teil I: „DUS-Prüfungsanlage und Datenanalyse, Minimierung der Variation infolge verschiedener Verfasser“, erstellt von Herrn Gerie van der Heijden (Niederlande), wie im Anhang zu diesem Dokument dargelegt (nur auf Englisch*), und machten folgende Kommentare:

Allgemein	Die TWV prüfte Dokument TWV/46/24 und betonte die Bedeutung der Kalibrierung des Erfassers (vergleiche Dokument TWV/46/41 „Report“, Absatz 35).	TWV
	Die TWC prüfte Dokument TWC/30/24 und empfahl, daß dieses nach der Änderung im letzten Satz von Abschnitt 6.1., wo es heißen sollte „systematische Unterschiede“, dem TC zur Prüfung für die Aufnahme in TGP/8 vorgelegt werden sollte (vergleiche Dokument TWC/30/41 „Report“, Absatz 23).	TWC

BEMERKUNGEN DES ERWEITERTEN REDAKTIONSAUSSCHUSSES (TC-EDC) IM JAHR 2013

8. Der TC-EDC prüfte auf seiner Tagung am 9. und 10. Januar 2013 in Genf Dokument TC-EDC/Jan 13/9: „Überarbeitung von Dokument TGP/8 Teil I: DUS-Prüfungsanlage und Datenanalyse, neuer Abschnitt: „Minimierung der Variation infolge verschiedener Erfasser“, wie im Anhang zu diesem Dokument dargelegt, und machten folgende Kommentare:

Allgemeine Anmerkung	Das Dokument sollte ebenfalls PQ-Merkmale abdecken (z.B. Farbe, Form usw.)
Überschrift	sollte lauten: ÜBERARBEITUNG VON DOKUMENT TGP/8 TEIL I: DUS-PRÜFUNGSANLAGE— DER UND DATENANALYSE, NEUER ABSCHNITT: MINIMIERUNG DER VARIATION INFOLGE VERSCHIEDENER ERFASSER
Anlage, Absatz 1	Es sollte nicht nur auf QN/MS sondern auch auf QN/MG verwiesen werden.
Anlage, Absatz 1.1	sollte lauten: „... Erfäßt mit also Erfasser A die Sorte 1 und Erfasser B die Sorte 2, so sind die erfaßten gemessenen Unterschiede eventuell auf Unterschiede zwischen Erfasser A und B statt auf Unterschiede zwischen Sorte 1 und 2 zurückzuführen. Unser Interesse gilt eindeutig den Unterschieden zwischen den Sorten und nicht den Unterschieden zwischen den Erfassern.“
Anlage, Absätze 2.1 und 2.2	Numerierung berichtigen
2.1	Letzten Satz streichen

* Der TC-EDC vereinbarte auf seiner Sitzung am 9. und 10. Januar 2013, daß es nicht zweckmäßig sei, den Text für die neunundvierzigste Tagung des TC zu übersetzen.

3.1	Sollte lauten „Im Anschluß an die Schulung eines Erfassers könnte der nächste Schritt darin bestehen, die Leistung des Erfassers anhand eines Kalibrierungsversuchs zu prüfen. Dies ist für unerfahrene Erfasser, die visuelle Erfassungen vornehmen müssen (QN/VG- und QN/VS-Merkmale), besonders nützlich. Bei visuellen VG Erfassungen sollten sie vorzugsweise einer Kalibrierungsprüfung unterzogen werden, bevor sie Erfassungen im Rahmen eines Anbauversuchs vornehmen. Aber auch für erfahrene Erfasser ist es nützlich, sich selbst regelmäßig zu testen, um sicher sein zu können, daß die Kalibrierungskriterien immer noch erfüllt werden.“
3.3	streichen
4.	sollte lauten „Prüfung der Kalibrierung für QN/MG- oder QN/MS-Merkmale“
4.1	Eine Leerzeile nach dem Absatz hinzufügen

9. Der TC wird ersucht, um die Erstellung eines Entwurfs für einen neuen Abschnitt über „Minimierung der Variation infolge verschiedener Erfasser“ zur Prüfung durch die TWP auf ihren Sitzungen im Jahr 2013, auf Grundlage der Kommentare durch die TWP und den TC-EDC, zu ersuchen.

[Anlage folgt]

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 OBSERVERS1. 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 $y=0$, $y=\text{mean}(\text{difference})$ and the two lines $y = \text{mean}(\text{difference}) \pm 2 \times \text{standard deviation}$, 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 account for random agreement: $\kappa = (P(\text{agreement}) - P(e)) / (1 - P(e))$, where $P(\text{agreement})$ 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, $\kappa(O1,O2)$ is calculated as follows:

- $\kappa(O1,O2) = (P(\text{agreement between } O1 \text{ and } O2) - P(e)) / (1 - P(e))$ where:
- $P(\text{agreement}) = (3+5+0+1+0+2)/30 = 11/30 \approx 0.3667$ (diagonal elements)
- $P(e) = (3/30).(15/30) + (17/30).(6/30) + (3/30).(1/30) + (1/30).(3/30) + (3/30).(0/30) + (3/30).(5/30) \approx 0.1867$. (pair-wise margins)
- So $\kappa(O1,O2) \approx (0.3667-0.1867) / (1-0.1867) \approx 0.22$

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: $\kappa(O1,O3) \approx 0.72$, $\kappa(O2,O3) \approx 0.22$. 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)

[Ende der Anlage und des Dokuments]