Organic Farming | 2021 | Volume 7 | Issue 1 | Pages 1–25
DOI: 10.12924/of2021.07010001
ISSN: 2297–6485

Organic Farming

Research Article

A Proposal for Improving Organic Group Certification. Quantification of Internal Control Systems' Performance and Sample Size Determination

Albrecht Benzing[1,*] and Hans-Peter Piepho[2]

[1] CERES GmbH, Bavaria, Germany
[2] Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Baden-Württemberg, Germany

* Corresponding author: albrechtbenzing@yahoo.com

Submitted: 16 March 2021 | In revised form: 23 June 2021 | Accepted: 1 July 2021 |
Published: 5 October 2021

Abstract: Organic certification, especially for smallholders, often uses group certification procedures. An internal control system (ICS) visits all farmers, and then the external certification body (CB) inspects a sample to assess the ICS' performance. Harmonised methods for measuring the ICS' reliability are missing so far. Here, we define criteria of "ICS performance", propose a new procedure for quantifying this performance and, based on this procedure, suggest that the sample size can be determined using classical statistical methods for survey sampling, instead of using the square root or a percentage of group size as in current practice.

Keywords: Internal control system; Organic group certification; Survey sampling; Systemic non-conformities; Witness audits

1. Introduction

1.1. Group Certification and One of Its Weaknesses

Group certification is used in different farm certification schemes (GLOBALG.A.P., Rainforest Alliance, Round Table on Sustainable Palm Oil, organic farming, etc.). The basic idea is to facilitate access to certification by building up an Internal Control System (ICS), the effectiveness of which is verified by an external inspection (also called "audit"). While under some programs (e.g. GLOBALG.A.P. and the National Organic Program of the USA, NOP [1], there is no restriction concerning size of the member farms, the EU regulation on organic farming restricts participation in group certification to small farms [2,3].

Research in relation to group certification so far has addressed its impact on market access and smallholder incomes [4–11], implementation of improved agricultural practices by the certified farmers [10,12], schooling [13], scalability [14], internal organisational problems of the groups and certification costs [7,15], environment and nature conservation [4,9], and adaptation to climate change [16], but not on the functioning of the ICS as such, their ability to ensure compliance with the standards, nor the way that certification bodies (CBs) deal with the ICS.

For a better understanding of the organic group certification process, Figure 1 describes the general workflow.

librello

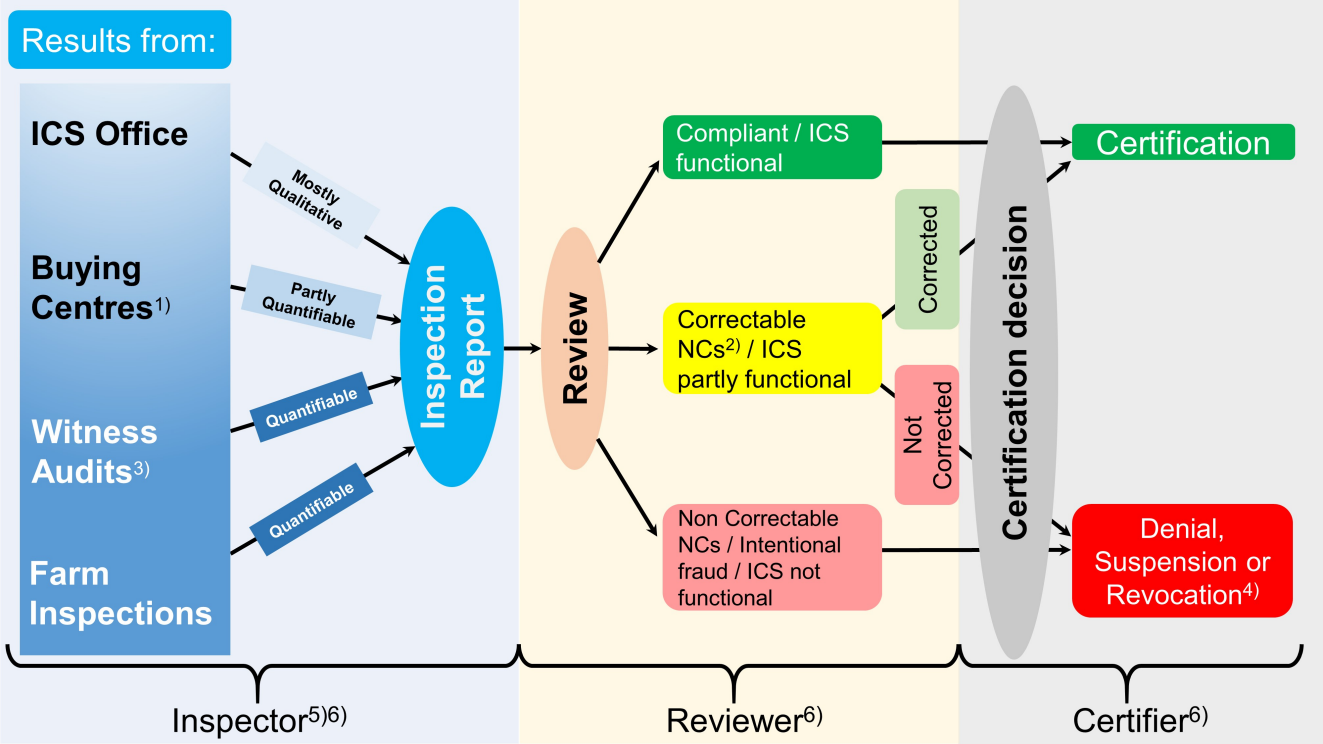

Figure 1. General workflow of an organic group certification process. On the left side, the four on-site inspection activities, for which both the need for, and possibility of, quantification increase from top to bottom. Our article deals mainly with farm inspections and witness audits.
1) Buying centres (also called collection points, wholesale points, buying points) are places, to which member farmers deliver their products. Sometimes the group contracts some of its members for this purpose, in other cases the group sets up its own structure. Some buying centres are permanent, others are active only during the harvest season. In other groups, the buying staff drives to the farmers for picking up the products.
2) NC: non-conformity.
3) Witness audits: the external inspector accompanies the internal inspector for observing her/his competence. See Section 8.
4) Denial: an initial application for certification is turned down; Suspension: existing certification is withdrawn temporarily; Revocation: existing certification is withdrawn terminally.
5) "Inspector" here refers to the external inspector who is an employee or contractor of the CB. Since the task is complex, group inspections are often performed by teams of several external inspectors.
6) Some certification programs require two, other three different persons to be involved in the certification process. The distribution of roles among these two or three persons depends on the certification program. All programs, however, require that the final certification decision is made by a person that is different from the inspector.

Table 1 summarizes the most important rules for an organic ICS and also explains at which level and through which methods an external inspector can verify compliance with each of these rules. Out of the eight rules in this table, (h) is the most important one, because an ICS cannot be considered functional if it does not identify the existing non-conformities (NCs) among its members, ensuring that these are either corrected or the non-compliant members are excluded. Also, for the CB the visit to a sample of farmers is the core part of the group inspection. The CB should not only assess compliance with basic organic farming rules like, e.g., having a proper crop rotation, protecting the soil from erosion, ensuring adequate storage conditions for organic products, using only allowed fertilizers, etc. at each farm in the sample, but also use these visits to the farmers for cross checking the accuracy of records kept at the group level, verify separation of certified from non-certified products on their way from farm to export, and find out if member farmers have received appropriate training and consultancy (Table 1).

However, little to no efforts have been made so far for a systematic assessment of the outcome of these external visits. A new EU regulation for the first time establishes official rules for group certification instead of unofficial guidelines [3]. But what exactly does it mean, when this new regulation says *"For the purpose of evaluating the set-up, functioning and maintaining of the ICS of a group of operators, the [...] control body, shall determine at least that the ICS manager*

takes appropriate measures in case of non-compliance, including their follow up, according to the ICS documented procedures that have been put in place" [3]? If in a sample of n farmers the CB finds one case where the ICS manager has not "taken appropriate measures": does that mean the ICS is not functional—which ultimately means the group cannot be certified (Figure 1)? Or is there a meaningful threshold, above which the CB should make that decision?

In a worldwide survey among organic CBs, including expert interviews, the lack of such thresholds was identified as one of the main weaknesses of the current situation of organic group certification (Textbox 1).

Table 1. Basic rules for the functioning of an ICS.

The ICS must:	**To be verified:**			
	At ICS office	**At farm**	**During witness audits**	**At buying centres**
a. Conduct at least one yearly inspection of 100% of the group members	Check availability of internal reports	Cross-check if farmer was visited		
b. Keep adequate records, including maps, of farm size, crops, buildings and production of each member	Check quality of records	Compare records to reality	Observe ability to correctly assess and record basic farm information	
c. Ensure that certified products are kept separate from non-certified products at any moment	General product flow, traceability check	How much did the farmer produce? How much did the farmer sell? Are the quantities plausible for the farm's size and production capacity?	Observe ability and thoroughness to check traceability and separation	Completeness and consistency of different records, traceability check, interviews with buying staff
d. Adequately train member farmers concerning rules and production techniques of organic farming	Training records	Cross-check participation in trainings; find out level of knowledge through farmer interviews		
e. Have a sufficient number of internal inspectors, who must be trained and supervised	Training records, monitoring records, interview with internal inspectors		Competence assessment during witness audits	
f. Prevent conflicts of interests among internal inspectors	Interviews, declarations	Farmer interviews	Interviews with internal inspectors	
g. Have a manual, which describes the functioning of the group, including a sanction catalogue	Review manual	Cross-check if manual matches reality	Cross-check if inspectors are familiar with manual	Cross-check if manual matches reality
h. Ensure that non-compliant farmers either implement corrective measures, or are excluded from the group	Review internal reports and records on how the ICS deals with non-conformities (NCs)	Compare the ICS' findings to the reality on the ground; especially, if the ICS has found the same NCs, which the internal inspector finds	Observe inspectors' ability to properly assess NCs	Cross-check if excluded members are no longer delivering to the group

Textbox 1:
One of the conclusions from a worldwide survey on organic group certification [17] (bold accentuation by the authors).

*Many experts mentioned the **lack of clarity** and [the] diversity of approaches when it comes to dealing with non-compliances found on farms, which may indicate a deficient ICS. There was a general concern that certifiers seem to be reluctant to sanction an entire group **when finding non-compliances on individual farms**, and have a tendency to put this down to problems with an individual farm, rather than a systematic ICS deficiency. [...] It is important to **improve guidance on dealing with weak ICS particularly in terms of: how to assess the percentage of farmers** (out of the visited sample) found to have major non-compliances **that are indicative of a systematic failure of the system**, and the sanctions and measures to be taken in case of a weak or failing ICS (e.g. follow up with an additional external inspection, suspension or withdrawal of certification).*

Non-organic group certification schemes are also vague in this regard. GLOBALG.A.P., e.g., differentiates between "structural" and "non-structural" NCs, but does not explain how often an NC must occur for categorising it as "structural" [18].

1.2. The External Sample Size

The size of the sample of farmers visited by the CB (the "external sample") has been subject of long standing discussions between the stakeholders involved. Currently, the most common approach is using the square root of the total number of group members, multiplied by a risk factor, which varies between 1.0 and 1.4. This is established in an unofficial guideline by the EU Commission [2]. Also GLOBALG.A.P. [18], Rainforest Alliance [19] and other programs use the square root as the basis for calculating the external sample size, although without applying risk factors.

The new Regulation (EU) 2018/848 on organic farming [20], which will come into force in January 2022, for the first time introduces official minimum requirements for the groups and their ICS [21] and for the procedures to be followed by CBs for this purpose [3]. Although clear evidence does not exist in this regard, according to the perception of regulatory authorities, fraud is more common under group certification than under individual farm certification [22]. To address the related risks, the EU Commission stipulates that (a) the maximum group size shall be limited to 2,000 members, and (b) organic CBs, instead of the square root shall inspect a minimum of 5% of the group members [3]. Figure 2 shows that for small groups the sample would be smaller with the 5% rule, while for large groups it would be much bigger.

This proposed change raises two concerns: (a) a fixed 5% sample disregards basic statistical principles of sample size determination and will lead to high standard errors for small groups, and (b) as long as the weaknesses in the system described above are not addressed, larger sample sizes (for big groups, see Figure 2) will only reproduce the existing problems at a larger scale.

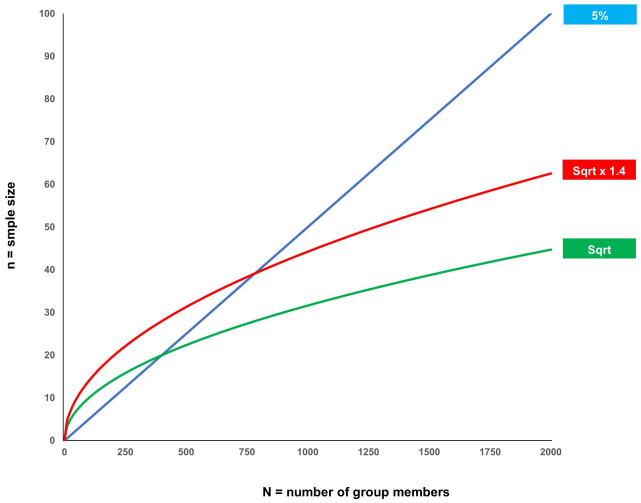

Figure 2. Sample size for group inspection, using $n = 0.05N$ compared to $n = \sqrt{N}$, and $n = 1.4\sqrt{N}$, for groups up to 2,000 members. For very small groups, [3] furthermore prescribes: If $N \leq 10 \rightarrow n = N$; if $N > 10 \rightarrow n \geq 10$. These special cases are not considered in the graph. $Sqrt$ = square root.

2. What is the Purpose of Sampling in Organic Group Certification?

As explained above, the performance assessment of an ICS takes place at four different levels: at the ICS office, in the buying centres, at the farms and during witness audits with the internal inspectors (Table 1 and Figure 1, also [23]). The results of the audits at the first two levels are mostly qualitative, but a meaningful assessment of the findings from the farm level requires some kind of quantification (Figure 1). Quantification of the results of the witness audits with internal inspectors may not be necessary in small groups with one or few inspectors, but becomes important in large groups with many internal inspectors (Section 8). A key underlying question is: What exactly is the goal of sampling a certain number of member farmers?

a. Is the goal to determine the exact percentage (incidence) of each kind of NC? Not really. Let us assume

we are dealing with a group, where many farmers use herbicides, which are prohibited in organic farming. Does it matter for the CB, if, say, 14%, 32% or 45% of the farmers use herbicides? The answer is "no", because in any of these cases, the conclusion would be the same: the ICS is not functional, and certification would have to be suspended, temporarily or terminally. Or let's imagine a group, where some farmers do not keep records of their daily field activities. Would it make a difference for the CB, if this problem were found among, say, 2%, 4% or 10% of farmers? No, because in any of these cases, the ICS would be requested to propose corrective actions, to ensure that farmers in the future keep their records. And in none of these cases would the group's certification be at risk.

b. Is the goal to find each and every NC that may exist in the group and has slipped through the ICS? Any type of sampling always involves the risk of a certain number of cases slipping through. This may not be acceptable when it comes, e.g., to high food safety risks, but it would not be appropriate for organic group inspections, because (i) compliance with organic production rules is not a food safety issue, (ii) the idea of "group" certification would become meaningless, since ultimately the sample size would have to be equal to the total number of farmers, and (iii) even with 100% external inspections, not all NCs existing at the time of the inspection will be detected, let alone those NCs, which may not be detectable on the day of the inspection.

c. Is the goal to ensure that non-compliant farmers identified during external inspections are excluded from the group? This is a common misunderstanding (see also Textbox 1), which completely misses the point of group certification. If the CB inspects, e.g., 10 out of 100 farmers, and finds in this sample two farmers using synthetic fertilizers, then we assume that in the entire group there are many more farmers with this problem, and excluding the two members would not solve the problem.

d. Is the goal to decertify groups, when the incidence of severe NCs exceeds a certain threshold? This is how, e.g., the Rainforest Alliance group certification works: *"if an irreversible non-compliant practice occurred on more than 5% (of the whole group, after extrapolation (...) and/or on at least 5 of the audited small farms this is considered to be a systemic issue (...) and therefore shall result in non-certification and/or cancellation"* [19]. There may be different opinions among CBs and regulatory authorities in this regard, but the authors believe that this approach does not sufficiently consider the efforts made by the ICS. Let's look again at the example above of a group with widespread herbicide use: When in a group of 100 farmers, the ICS has never detected any case of herbicide use, but then the CB in a sample of 10 farmers detects one case—this situation should be treated differently from the case where the ICS has already excluded 20 out of 100 farmers, but then the CB finds one more case.

e. The real goal of external inspections should be to determine (i) which existing NCs have been properly handled by the ICS and which not; (ii) among the latter, which are "systemic" and which are "isolated" cases; and (iii) which of the systemic cases put at risk the integrity of the products sold on the organic market, and the credibility of the certification system.

3. Judgement Sampling vs. Statistical Sampling

The U.S. Office of the Comptroller of the Currency [24] distinguishes between "judgement sampling" and "statistical sampling". The definition of judgement sampling is quoted in Textbox 2.

Textbox 2:
Definition of "judgement sampling" [24].

Judgement (i.e. nonstatistical) sampling includes gathering a selection of items for testing based on examiners' professional judgement, expertise, and knowledge to target known or probable areas of risk. [...] The key limitation with judgemental sampling is that the resulting conclusions cannot be extrapolated statistically to the population [...].

The current organic group certification procedures are mostly based on judgement sampling. The problem is, however, that the involved CBs do not always have the necessary level of "professional judgement" that would lead to satisfactory results (see [17]). A solution to the problem presented in Textbox 1 can only be found using "statistical sampling", which allows extrapolation of sample results to the entire group. Statistical sampling methods must select the sample randomly, not risk-based [24], otherwise the results would be biased. If a CB knows, e.g., that a specific problem is more frequent in one village belonging to a producer group, and therefore targets farmers from that village more than the rest of the group, the results from this inspection cannot be extrapolated to the entire group, because the problem would be over-estimated (Figure 3).

4. What Does "Systemic" Mean? What Does "Integrity" Mean?

For finding a solution to the problem described in Textbox 1, we must first define systemic NCs vs. isolated NCs and in which cases systemic NCs should lead to decertification. In this section, we propose a new procedure for quantifying these terms and for answering these questions, with the help of the variables defined in Table 2. Readers who are not so much interested in the statistical details, can jump directly to Table 3, from there to Figure 5, and then continue with the real life examples in section 5.

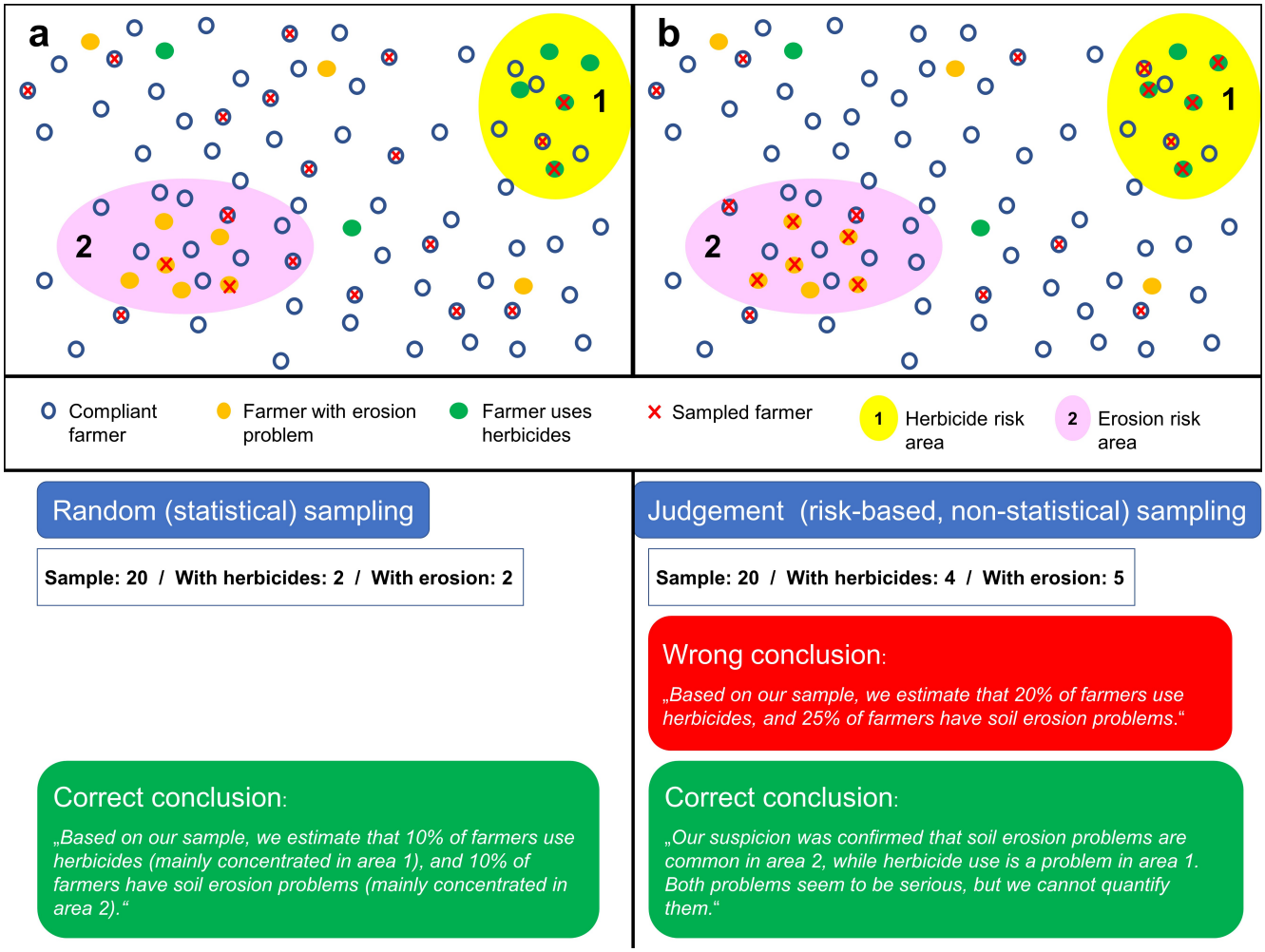

Figure 3. (a) Random (statistical) vs. (b) risk-based (non-statistical) sampling. In both cases, the group has 80 members, 9 of whom (11%) with erosion problems, and 7 (9%) with herbicide use. The sample is 20 in both cases. While (a) allows to estimate the probable dimension of the two problems, (b) does not. Therefore, the conclusion in the red box is wrong.

Table 2. Abbreviations and variables used in this article. For an illustration of π, refer to Figure 4, while Figure 5 illustrates some of the other variables.

Abbreviation	Variable	Definition
CB		Certification body (called "control body" in the EU Regulation on organic farming).
ICS		Internal control system.
NA		Not applicable.
NC		Non-conformity.
NC_1		A specific non-conformity occurring among the members of the group (see Table 4 and following for examples).
	HW	Half-width of 95% confidence interval (= standard error $\times$ 1.96).
	M	Number of farmers in the entire group with NC_1.
	M_a	Number of farmers in the entire group identified by the ICS for NC_1.
	M_b	Number of farmers in the entire group with NC_1 identified but not corrected by the ICS. Can be estimated from the sample by $m_b \times (N/n)$.
	M_c	Number of farmers in the entire group with NC_1 found by the CB, which were missed by the ICS. The CB estimates this variable from the sample using $m_c \times (N/n)$.
	m_b	Number of farmers with NC_1 found by the CB, which had previously been detected by the ICS, but not yet corrected at the time of the external inspection.
	m_c	Number of farmers with NC_1 found by the CB, which had not been detected by the ICS.
	m	$m_b + m_c$: These two cases are treated equally; number of farmers in sample taken by CB with NC_1.
	N	Size of population (number of all members of the group).
	n	Size of sample inspected by the CB
	π	$\frac{M}{N}$: Incidence of NC_1 in the entire group.
	π_a	$\frac{M_a}{N}$: Incidence of NC_1 in the entire group that are *detected and corrected* by the ICS.
	π_b	$\frac{M_b}{N}$: Incidence of NC_1 in the entire group, which were previously detected by the ICS, but not yet corrected at the time of the external inspection. Can be estimated from the sample by m_b/n.
	π_c	$\frac{M_c}{N}$: Incidence of NC_1 in the entire group, which were not detected by the ICS.
	π_e	$\pi_b + \pi_c$: Incidence of NC1 in the group, which either went undetected or were detected but not corrected by the ICS. This parameter is estimated by extrapolation from the sample by m/n.
	$\widehat{\pi}_{e(L)}$	Lower limit of the confidence interval for π_e; this can be obtained by an asymptotic method for large samples or by the exact Clopper-Pearson interval for small samples and populations.
	$\widehat{\pi}_{e(U)}$	Upper limit of the confidence interval for π_e; this can be obtained by an asymptotic method for large samples or by the exact Clopper-Pearson interval for small samples and populations.
	δ	$\pi_e - \pi_a$: For the sake of valuing the effort made by the ICS, π_a is deducted from π_e. Refer to section 3(b) in the text for more details. Small and negative values of this criterion are desirable. Values above a threshold δ_0 are considered as a systemic failure of the ICS.
	$\widehat{\delta}_L$	Lower limit of the confidence interval for δ : $\widehat{\delta}_L = \widehat{\pi}_{e(L)} - \pi_a$.
	$\widehat{\delta}_U$	Upper limit of the confidence interval for δ : $\widehat{\delta}_U = \widehat{\pi}_{e(U)} - \pi_a$.
	δ_0	Threshold, above which is considered "systemic".
	r	Repetition: number of subsequent external inspections, during which NC_1 is found at a systemic level. Normally, such inspections take place yearly, but they can also be more frequent.
	r_0	Threshold, above which the repetition of a systemic NC leads to decertification.
	s	Severity category of an NC (see Table 3).
	σ^2	Variance of a character trait within a group.
	σ_p^2	Pooled variance across groups.

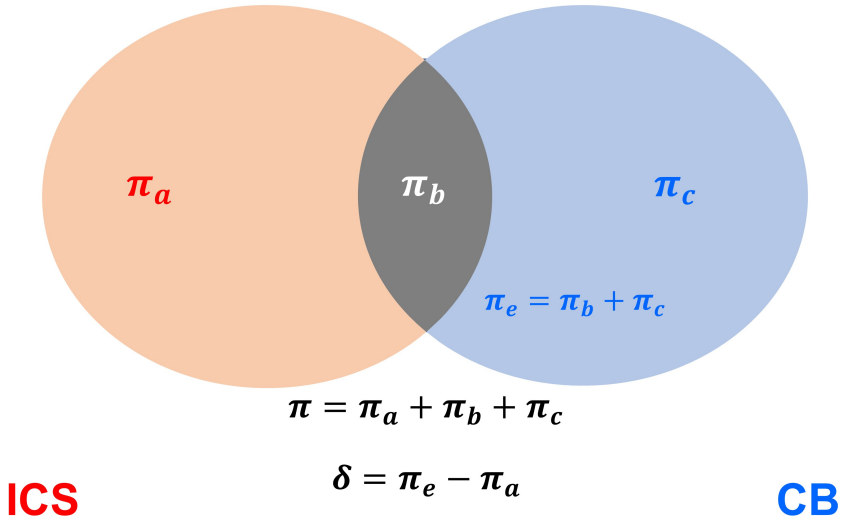

Figure 4. Venn diagram illustrating the incidence π of a specific NC in a producer group, its components π_a, π_b, π_c and π_e and the definition of δ. Refer to Table 2 for further details. ICS = Internal Control System, CB = Certification Body.

a. We count M and compute π_a (see Table 2 and Figure 4).

b. As explained in Section 2(d), one approach for assessing the performance of the ICS would be to simply define a threshold, above which a group should be decertified. This would mean using an estimate of π_e (Table 2) for this purpose, i.e. $\widehat{\pi}_e = \frac{m}{n}$. For the reasons explained in Section 2(d) (we want to value the efforts made by the ICS, which have already detected certain cases), we suggest to use the difference between the incidence of a specific NC *identified* by the CB in the sample (extrapolated to the entire group), and the incidence *identified and corrected* by the ICS in the entire group. This better values the efforts made by the ICS (an approach, which may not be shared by all CBs and regulatory authorities):

$$\widehat{\delta} = \widehat{\pi}_e - \pi_a \tag{1}$$

c. Next, to reflect that an estimate is used, we compute the lower and upper limits of a 95% confidence interval for π_e (Table 2) using standard procedures as described by Agresti ([25], pp. 15,18-21) and also described in detail below (see Equations 4 to 7). The lower and upper limits for π_e are denoted as $\widehat{\pi}_{e(L)}$ and $\widehat{\pi}_{e(U)}$, respectively. Those on δ are denoted as $\widehat{\delta}_L$ and $\widehat{\delta}_U$, respectively.

d. We define a threshold above which the incidence of an NC is considered systemic. Since this threshold should be different, depending on the type of NC, we group the existing NCs in five categories s, from 1 (least severe) to 5 (most severe). Refer to Table 3 for examples. These severity categories are associated with an acceptable threshold δ_0, above which δ is considered "systemic" (third column in Table 3).

$$\begin{aligned} &\text{If } \widehat{\delta}_L > \delta_0 \rightarrow \text{ NC is systemic,} \\ &\text{If } \widehat{\delta}_U < \delta_0 \rightarrow \text{ NC is not systemic.} \end{aligned} \tag{2}$$

The more severe the category, the lower the acceptance threshold. If neither of the two conditions hold, the sample size was too small to reach a definitive assessment. This is likely to happen only when $\widehat{\delta}$ is close to the threshold δ_0. Note that this step amounts to a significance test at the 5% level to decide if δ is significantly smaller or larger than the threshold δ_0.

e. As a second condition for considering an NC as "non-systemic", we introduce the requirement that π_e must be below 0.3 - regardless of $\widehat{\delta}$. The rationale is as follows: if the ICS makes serious efforts for handling NCs, but in spite of these efforts the CB still finds many undetected or uncorrected cases, there is a systemic problem.

$$\begin{aligned} &\text{If } \widehat{\delta}_L > \delta_0 \text{ or } \widehat{\pi}_{e(L)} \geq 0.3 \rightarrow \text{ NC is systemic,} \\ &\text{If } \widehat{\delta}_U < 0.3 \rightarrow \text{ NC is not systemic.} \end{aligned} \tag{3}$$

The assessment is inconclusive otherwise. If this happens for NCs with $s \leq 4$, we suggest the CB decides from case to case, if the NC is considered systemic or not. For NCs with s = 5, the sample should be increased until getting a clear picture.

Finally, we suggest how often a systemic NC can be repeated (r, see Table 3), before it seriously affects the integrity of the system and should therefore lead to (temporary or final) decertification. We call this threshold "repetition tolerance" r_0. r_0 is also related to s (Table 3, column 4). For NCs with s = 5, we have defined r_0 = 1, meaning there is no tolerance for systemic NCs of this category.

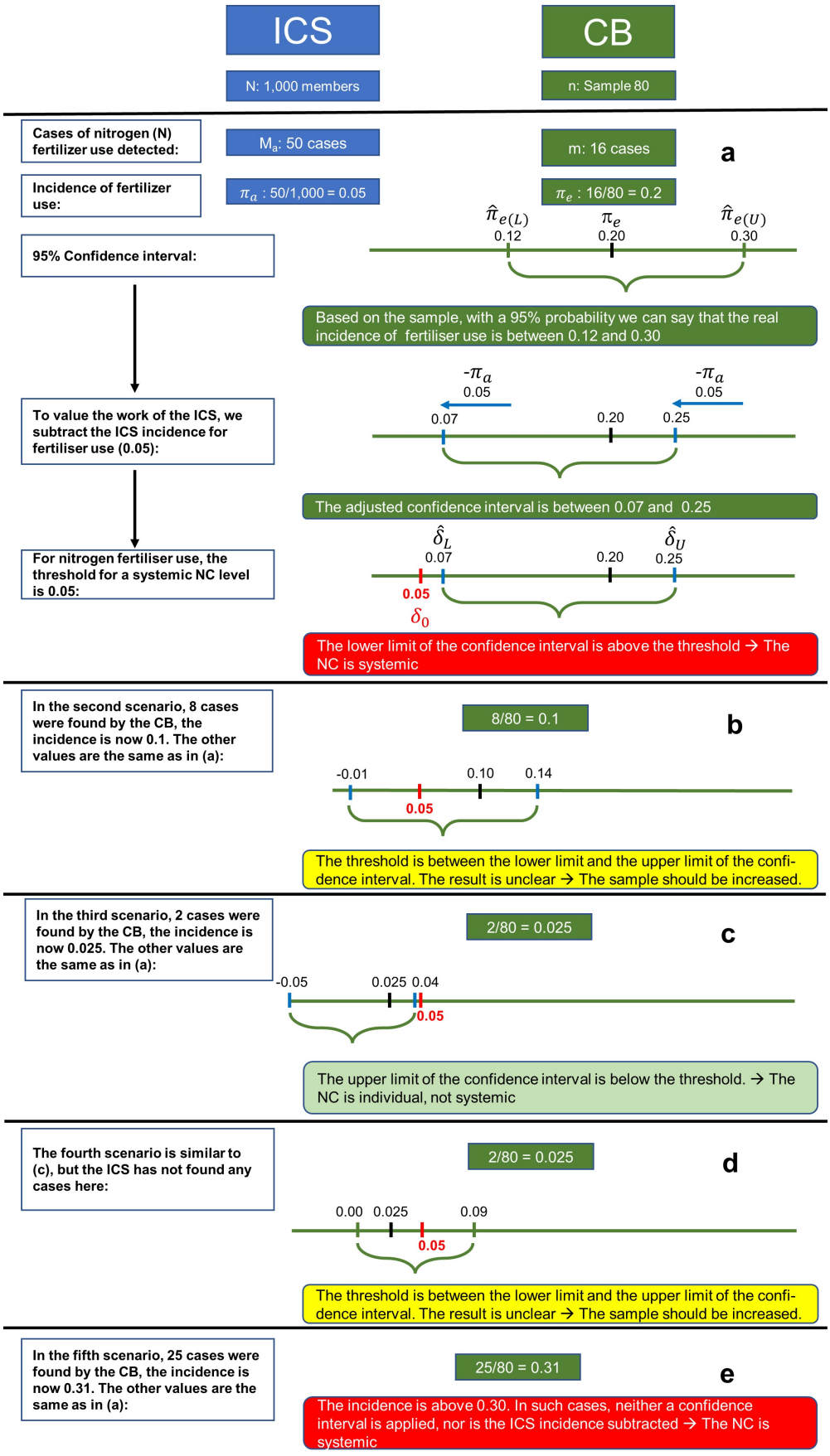

Figure 5. Five scenarios illustrating the procedure described in Section 4 The incidence of a specific NC (in this example synthetic nitrogen fertiliser use) found by the CB, the meaning of the confidence interval, and how the outcome can be affected by the performance of the ICS.

Table 3. Severity classes of NCs, examples, the corresponding thresholds δ_0 and repetition tolerances r_0.

Severity class s	Examples	Threshold δ_0	Repetition tolerance r_0
1	Minor inconsistencies in basic farm information (size, number of fields, accuracy of farm map, yield estimates), not involving risks of overdelivery. Use of non-organic but untreated green manure seeds. Farm records existing, but incomplete.	0.25	5
2	Soil erosion risk, no visible signs of erosion. Inorganic litter on organic fields. No farm records, but sales receipts available. Use of undeclared (but compliant) fertilizer or pesticide.	0.20	4
3	Measures to maintain soil fertility not adequately implemented. No farm records and sales receipts on farmer level. Undeclared parallel production. Use of conventional untreated seeds of certified crop without prior authorization of CB. Insufficient crop rotation.	0.15	3
4	Soil erosion visible. Incorrect figures concerning size of fields, yields, etc. (with probable implications for overdelivery).	0.10	2
5	Agrochemical use. Buying records show higher quantity than delivered by farmer.	0.05	1

5. Two Real Life Examples

For exemplifying the proposed method, we have selected two cases of group certification from the CERES database: a positive case of a group with a functioning ICS and only minor deficiencies, and a negative case, which lost its certification. If the method suggested had been applied, these results would have been confirmed—but based on a more transparent and reliable procedure.

The **first case study** refers to a cocoa farmers group with 1,079 members. Since this was the first inspection to this group, the risk factor had been calculated as 1.2, based on theoretical assumptions, leading to the sample size: $n = \sqrt{1,079} \times 1.2 \approx 40$.

Three NCs were found, two of which were systemic, but none of these with serious implications for integrity (Table 4).

The rather small NCs could be easily corrected, and the group was certified.

The **second case study** is for a group of 1,413 coffee farmers, spread over a large area, with highly heterogeneous geographical conditions. A risk factor of 1.4 had been determined, leading to the following sample size: $n = \sqrt{1,413} \times 1.4 \approx 54$.

During the four previous years, only minor NCs had been detected. During inspection planning in 2016, the CB found that the samples in previous years had not been random, because they had only covered a relatively small part of the region. This was corrected by randomly including farmers from all parishes in the new sample. Furthermore, the CB had learned that agrochemical use among coffee smallholders in the entire region had increased substantially. Therefore, coffee leaf samples were taken from 16 out of the externally inspected 54 farmers and tested for pesticide residues.

As a result of this change in inspection procedures, in addition to several other (systemic and non-systemic) NCs, on 10 farms the inspectors found synthetic pesticides and/or fertilizers. In 6 out of 16 leaf samples, residues of synthetic fungicides were found at levels, which could only be explained by application by the organic farmers (Table 5).

None of these NCs had been detected by the ICS, therefore the group's organic certificate had been withdrawn immediately. If the method proposed here had been used, the result would have been the same. These severe problems in the group, however, were detected not because the sample *size* was increased as compared to previous years, but because (a) the sample was chosen *randomly*, and (b) because the inspection procedure was improved by testing leaf samples, which had not been done in previous years.

Table 4. Incidence of non-conformities (NCs) found during inspections of a cocoa smallholder group, and determination of the systemic / non-systemic condition. $N = 1079$; $n = 40$. It was certified after correcting the indicated NCs. For illustration purposes, the table is made up similar to an MS Excel worksheet. The numbers in first column to the left stand for the horizontal rows. Please refer to the Excel template in the supplementary materials.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	**NCs found during the inspection**	**Cases detected by ICS**	**Incidence per ICS (M/N)**	**Additional cases detected by CB**	**Incidence of additional cases per CB (m/n)**	**Lower confidence limit**	**Upper confidence limit**	**Difference lower limit**	**Difference upper limit**	**Severity**	**Threshold for systemic condition**	**Systemic?**	**Repetition**[1)]	**Repetition tolerance**	**Decertification**
2[2)]	NC	M	π_a	m	π_e	$\pi_{e(L)}$	$\pi_{e(U)}$	$\widehat{\delta}_L$	$\widehat{\delta}_U$	s	δ_0	$\widehat{\delta}_L > \delta_0 \bigvee \widehat{\pi}_e \geq 0.3$	r	r_0	
3[3)]	Dropdown[4)]	Entry	=B4/B7[5)]	Entry	=D4/D7	=(F.INV)[6)]	=(F.INV)[6)]	=F4-C4	=G4-C4	=(VLOOKUP)[7)]	-J4∗0.05+0.3	=(IF...)[8)]	Entry	=(IF...)[8)]	=(IF...)[8)]
4	NC_1: Incorrect yield estimate[9)]	**0**	**0.00**	**38**	**0.95**	**0.83**	**0.99**	**0.83**	**0.99**	**1**	**0.25**	**Yes**	**1**	**5**	**No**
5	NC_2: Incorrect farm size	0	0.00	18	0.45	0.29	0.61	0.29	0.61	2	0.2	Yes	1	4	No
6	NC_3: Incorrect number of cocoa plots	0	0.00	2	0.05	0.006	0.17	0.006	0.17	2	0.2	No	1	NA	No
7	**N:**	**1079**	**n:**	**40**											

1) Repetition: in how many external inspection has this NC been found at a systemic level. In the present case, this is 1, because it was the first inspection. For NC_3 the value is 0, because this NC is not systemic.

2) Refer to Table 2 for an explanation of this row.

3) Many CBs use MS Excel or similar tools for such procedures. Row 3 shows how this can be done, for the example of NC_1 (row 4). Only the blue columns would require entries, the rest would be computed through formulas.

4) The common NCs in grower groups can be listed in a dropdown menu.

5) Cell B4 divided by cell B7 (both in yellow).

6) Here we calculate the lower Clopper-Pearson confidence limit ([25], p. 18); in Excel syntax we use the function F.INV. Refer to the template in supplementary materials.

7) VLOOKUP is a formula linked to the type of NC (column A).

8) Nested IF-THEN-ELSE formulas are used here.

9) In this case, incorrect yield estimates were assigned a “severity” of 1 only, because the group had intentionally used a very conservative estimate for kg cocoa beans per hectare.

Table 5. Incidence of NCs found during inspections of a coffee smallholder group, and determination of their systemic / non-systemic condition. $N = 1413$, $n = 54$. The group lost its certification because of these results. For further details regarding the different columns, refer to header and footnotes in Table 4.

A NCs found during the inspection	B Cases detected by ICS	C Incidence per ICS (M/N)	D Additional cases detected by CB	E Incidence of additional cases per CB (m/n)	F Lower confidence limit	G Upper confidence limit	H Difference lower limit	I Difference upper limit	J Severity	K Threshold for systemic condition	L Systemic?	M Repetition	N Repetition tolerance	O Decertification
NC_1: Insufficient records kept on farm	848	0.60	11	0.20	0.11	0.34	-0.49	-0.26	2	0.20	No	0	-	No
NC_2: Farm description not accurate	0	0.00	21	0.39	0.26	0.53	0.26	0.53	1	0.25	Yes	2	5	No
NC_3: Use of synthetic fungicides[1)]	0	0.00	6	0.38	0.15	0.65	0.15	0.65	5	0.05	Yes	1	1	Yes
NC_4: No training received	565	0.40	20	0.37	0.24	0.51	-0.16	0.11	3	0.15	No	2	3	No
NC_5: Agro-chemicals found during inspection[2)]	0	0.00	10	0.18	0.09	0.31	0.09	0.31	5	0.05	Yes	1	1	Yes
NC_6: Water pollution[3)]	0	0.00	6	0.11	0.04	0.23	0.04	0.23	3	0.15	Watch![4)]	1	3	No
NC_7: Littering	0	0.00	2	0.03	0.005	0.13	0.005	0.13	2	0.20	No	2	4	No
N:	1413	n:	54											

1) Figures for NC_3 are based on leaf sample tests. n is therefore not 54, but only 16, because samples from 16 farmers were tested for pesticide residues.

2) Figures for NC_5 refer to different agrochemicals found on the farms during field visits.

3) “Water Pollution” is caused by pulping coffee cherries in nearby creeks, which leads to heavy organic contamination.

4) The threshold in column K is between the lower (H) and upper (I) limit here, therefore a clear statement concerning the systemic condition of this NC cannot be made, and the warning “Watch!” appears. Since it is not an issue of severity 5, it would be up to the CB to decide if the sample is increased for arriving at a clear decision, or the group is requested to correct the problem and the CB will follow up next year. In this case, this was no longer relevant, because the group lost its certification anyhow.

6. Sample Size Determination in Scientific Surveys

In a scientific survey with the goal explained above, neither a fixed percentage nor a square root of the total population size would be used as sample size. Instead, a specification would be made regarding the precision with which a population parameter is to be estimated, based on a random sample, and then the necessary sample size would be determined accordingly [26]. Again, readers who are not so much concerned about the mathematical details at this point, can go directly to Figure 9, from there to Textbox 3, then to Figure 11 and then continue with section 7 on stratification.

Assuming we deal with a very large population (as, e.g., in consumer studies or pre-election polls), an asymptotic interval with 95% coverage probability could be employed, based on the estimate $\widehat{\pi}_e = m/n$ and these equations for the lower (L) and upper (U) 95% confidence limit for π_e [25]:

$$\widehat{\pi}_{e(L)} = \widehat{\pi}_e - HW, \text{ and} \tag{4}$$

$$\widehat{\pi}_{e(U)} = \widehat{\pi}_e + HW \text{ , where} \tag{5}$$

$$HW = 1.96 \times \text{ s.e. } (\widehat{\pi}_e) \text{ with} \tag{6}$$

$$\text{s.e. } (\widehat{\pi}_e) = \sqrt{\frac{\widehat{\pi}_e(1-\widehat{\pi}_e)}{n}} \tag{7}$$

is the half width of the confidence interval. Further, we may compute lower and upper 95% confidence limits for δ as $\widehat{\delta}_L = \widehat{\pi}_{e(L)} - \pi_a$ and $\widehat{\delta}_U = \widehat{\pi}_{e(U)} - \pi_a$, respectively.

It is important to point out that the half-width (HW) of the interval is inversely proportional to the square root of the sample size n (see Equations 6 and 7). Thus, the larger the sample size, the smaller the HW. This relation can be used to determine sample size, if we can make a specification of the desired HW.

Thus, the sample size to achieve a desired HW can be computed as

$$n = \frac{1.96^2 \pi_e(1-\pi_e)}{HW^2} \tag{8}$$

Note that population size N does not enter this equation. The sample size remains the same, regardless if the population is, e.g. 10^4 or 10^8, so long as the population size is large relative to sample size. What matters, are the variables π_e and HW. If, e.g., our rough guess in such a large group was that there is a proportion of up to $\pi_e = 0.10$ or $\pi_e = 0.20$ of undetected non-compliant members remaining for a specific problem (NC), then the sample size plotted against HW would take the form of Figure 6.

To shed further light on the equation for determining the sample size, we may give a second interpretation. If the sample size is chosen to equal n , then the probability is 5% that the estimate of π_e deviates from the true value by more than HW [26].

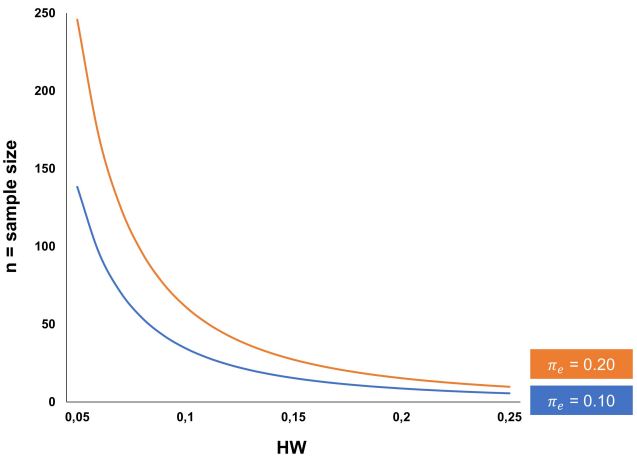

Figure 6. Sample size for half-widths (HW) ranging from 0.05 through 0.20, and two expected incidences of a given NC not detected (or not corrected) by the ICS (π_e), using Equation 8. As explained in the text, this method for determining sample size does not depend on the size of the total population (N)—provided the population is large enough.

So far, we have assumed that the population size is very large. In smaller populations, as in the case of group certification, the exact Clopper-Pearson interval should be used [25], which takes the population size into account. There are no exact equations to determine sample size for this procedure, which yields asymmetrical intervals. As an approximation, we may employ the fact that in finite populations the standard error ($s.e.$) takes the form described by Thompson [26]:

$$s.e.(\widehat{\pi}_e) = \sqrt{\frac{\pi_e(1-\pi_e)}{n}\left(\frac{N-n}{N-1}\right)} \tag{9}$$

As opposed to Equation 8, population size N does enter here. The factor $\frac{N-n}{N-1}$ relates to the *finite population correction*. From this, assuming approximate normality of the estimator of π_e, the sample size may be computed according to [26]:

$$n = \frac{N\pi_e(1-\pi_e)}{(N-1)\frac{HW^2}{1.96^2} + \pi_e(1-\pi_e)} \tag{10}$$

Note that for large N, this equation approaches the simpler one in equation 8. Also note that, even though there is a dependence on N, the required sample size is not proportional to N. And only in very small populations is the finite population correction at all noticeable. In Figure 7 we have plotted n against π_e, for four different HWs, showing that n is inversely related to HW (the higher our expectations on precision, the larger the sample must be), while in relation to π_e, n is biggest for 0.5, and decreases both towards 0 and towards 1.

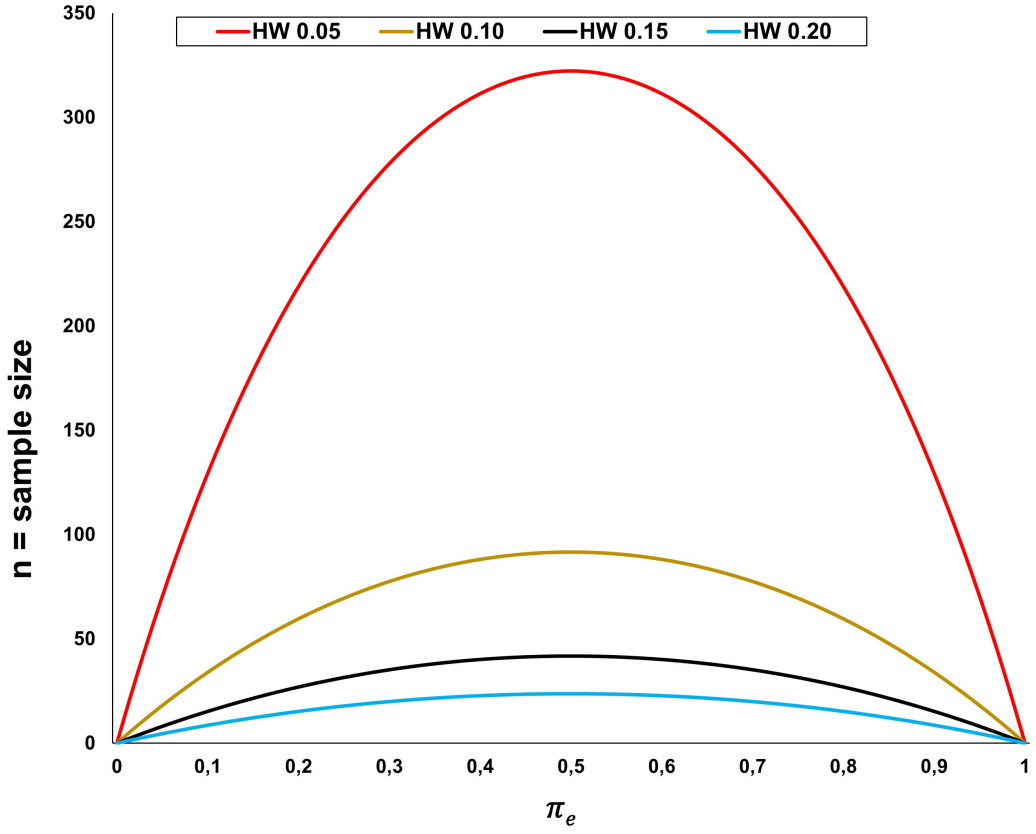

Figure 7. Sample size n for a population of 2,000, plotted against incidence π_e for four different HWs, using Equation 9.

In Figure 8 we see the impact of five different HWs and five different π_e on the required sample size for populations up to 1,000.

Two decisions remain to be made: (a) which is the highest π_e in the range from 0 to 0.5 that we must consider in an unknown group of farmers, and (b) which HW are we ready to accept? Statistics cannot answer these questions, which require normative or political answers.

Nevertheless, we can try an approximation:

a. In most cases, we do not know the incidence π_e, therefore it is reasonable to assume a value that is close to the worst-case scenario. The worst case is $\pi_e = 0.5$ (50% of the farmers have the NC we are dealing with)—for this scenario we need the largest sample for arriving at a correct decision (Figure 7). If we move too far away from this worst case, there is a risk of arriving at wrong conclusions.

b. Furthermore, we consider that the $s.e.$ should not be too far above 0.05, corresponding to a HW of 0.10.

c. Based on these two considerations, let us use $\pi_e = 0.50$ and $HW = 0.10$ as a starting point. The corresponding sample size is represented by the green line for HW 0.10 in Figure 8b. The sample sizes are substantially higher than the square root (e.g. $N = 100$: 48 vs. 10; $N = 500$: 78 vs. 23; $N = 2,000$: 84 vs. 45).

d. Then we looked for real life examples, where the CB CERES had used sample sizes, which were equal, higher, or at least close to these figures. Since CERES has also been using the square root multiplied by a risk factor, there are not many examples meeting these criteria. The examples we have found, are all from very large groups, because, as shown in Figure 8, above a certain population size, the sample sizes resulting from Equation 10 remain in the same range. We used the procedure explained in Section 4 for assessing the systemic condition of the NCs found during inspection of these example groups.

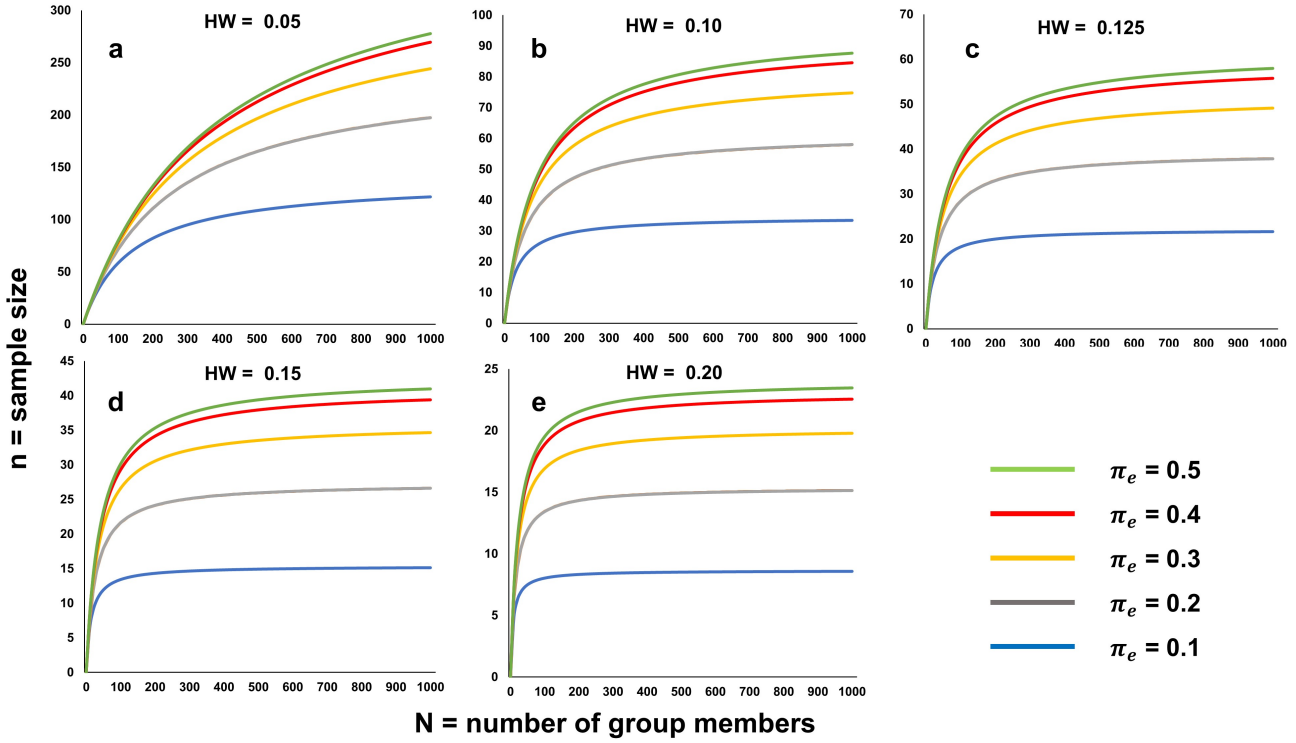

Figure 8. Sample size plotted against populations up to 1,000, for half-widths (HW) between 0.05 and 0.20 and π_e from 0.1 to 0.5, using Equation 10. Please note that the vertical scales are different for each HW. We omit displaying the results for larger groups, because for all HWs, the lines turn almost horizontal above 1,000.

e. As a result, we selected nine groups, all of them from Africa, because this is where the largest producer groups exist [17], with between 3,554 and 78,496 members each. At this point, we do not want to enter into the debate, if such large groups are certifiable or not—the groups were solely selected for the reasons explained in (d). Adding up all the NCs resulting from the nine inspections to these groups, in total CERES had identified 57 NCs, out of which (using the procedure explained in section 4), 29 were systemic, 19 were non-systemic, while 9 remained unclear (Method (I) in Figure 9).

f. Then we calculated for each of the nine groups different sample sizes, using Equation 10, with π_e ranging from 0.50 to 0.10, and HW from 0.10 to 0.20. The frequency of each NC was calculated proportionally to the sample size: When a specific NC had occurred 22 times in the original sample of 75 farmers, we assumed that in the same group, it would be detected 14 times in a sample of 58 farmers. From these proportional frequencies, we assessed the systemic condition of each NC, using the same procedure explained in Section 4. The results are shown in Figure 9 (Methods II to XIX).

g. To summarize what is represented in Figure 9:

- For achieving a result with only two “unclear” cases, we would have to use an unrealistically large sample size (Method II in Figure 9, with sample sizes between 2,594 and 8,577 farmers).
- As could be expected, the smaller the sample size, the higher the number of unclear (“watch”) cases (yellow in Figure 9).
- Because of the confidence interval, there is no NC, which would switch from “systemic” to “non-systemic” with decreasing sample size, or vice-versa. They switch from systemic to unclear, or from non-systemic to unclear (see also Figure 5d).
- If we use, e.g., a sample of 15 farmers per group (Method XIX), the interpretation of 38 out of 57 results would remain unclear. With all these unclear results, the sample size would have to be increased after the inspection—which is more complicated than planning for a bigger sample from the beginning.
- It becomes obvious from Figure 9 that the impact of a decreasing HW on sample size and on the number of unclear cases is much stronger than the impact of an increasing π_e. This is also confirmed through a regression analysis, where we get a steep and almost linear power function for unclear cases vs. HW, but a less steep power function for unclear cases vs. π_e. From $\pi_e = 0.35$ upwards, the results remain the same (Figure 10).

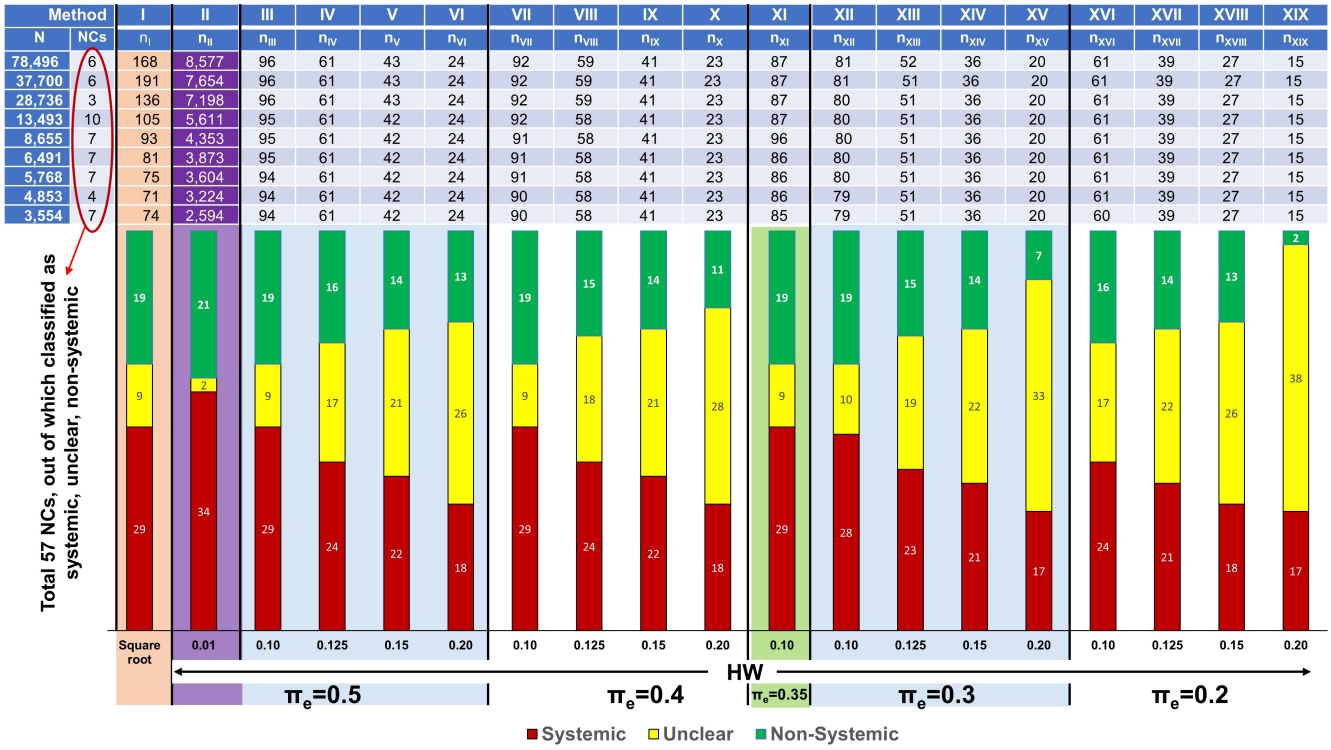

Figure 9. Evaluation of inspection results from nine large organic producer groups. The size of the groups (N) is indicated in the left column of the table on top of the graph. The second column of the table shows the number of NCs occurring in each group. For the nine groups together, a total of 57 NCs had been identified (red circle). The third column of the table (Method I) shows the real sample sizes, which were used by CERES, based on the square root approach (for extremely large groups, CERES has been using a risk factor < 1, therefore some of the samples are smaller than the square root). The sample sizes for Methods III to XIX were calculated using Equation 10, with different values for π_e (from 0.5 to 0.2) and HW (from 0.10 to 0.20). For demonstration purposes, also the sample for $\pi_e = 0.5$ and $HW = 0.01$ was calculated (Method II, in purple), resulting in extremely high sample sizes. As reflected in the table on top, the sample sizes vary substantially between methods, but very little between groups. The incidence of each NC was then calculated proportionally to the sample size. Then the classification of each NC was computed for each sample size, using the method described in Section 4. The red colour means the systemic condition of the NC was confirmed, the yellow colour means the systemic condition is unclear, because the threshold for qualifying an NC as systemic or not, lies between the lower and the upper limit of the confidence interval. The green colour means the NC is non-systemic. With decreasing sample size, the number of unclear cases increases. The only result with only two unclear cases was obtained with an unrealistically large sample (Method II), followed by Methods III, VII and XI, with nine unclear cases each.

We therefore suggest to use $\pi_e = 0.35$ and $HW = 0.1$. This is depicted as Method XI in Figure 9 and yields the following equation:

$$n = \frac{N0.35(1-0.035)}{(N-1)\frac{0.1^2}{1.96^2} + 0.35(1-0.35)} \approx \frac{0.2275N}{0.0026N + 0.23} \quad (11)$$

Another option would be to use a slightly larger HW, e.g. 0.125, being aware that many cases may remain in the "unclear" category, and especially when it comes to NCs of severity class 5, the sample size may have to be increased and the inspection extended, for getting a clearer picture. Figure 11a shows the sample size for Equation 11 ($HW = 0.1$ dotted black line) and $HW = 0.125$, dashed black line), as compared to square root and percentage approaches. In Figure 11b we have plotted HW against sample size, showing that for groups up to approximately 1,000 members, the method established by the European Commission accepts very large and questionable HWs, i.e. standard errors.

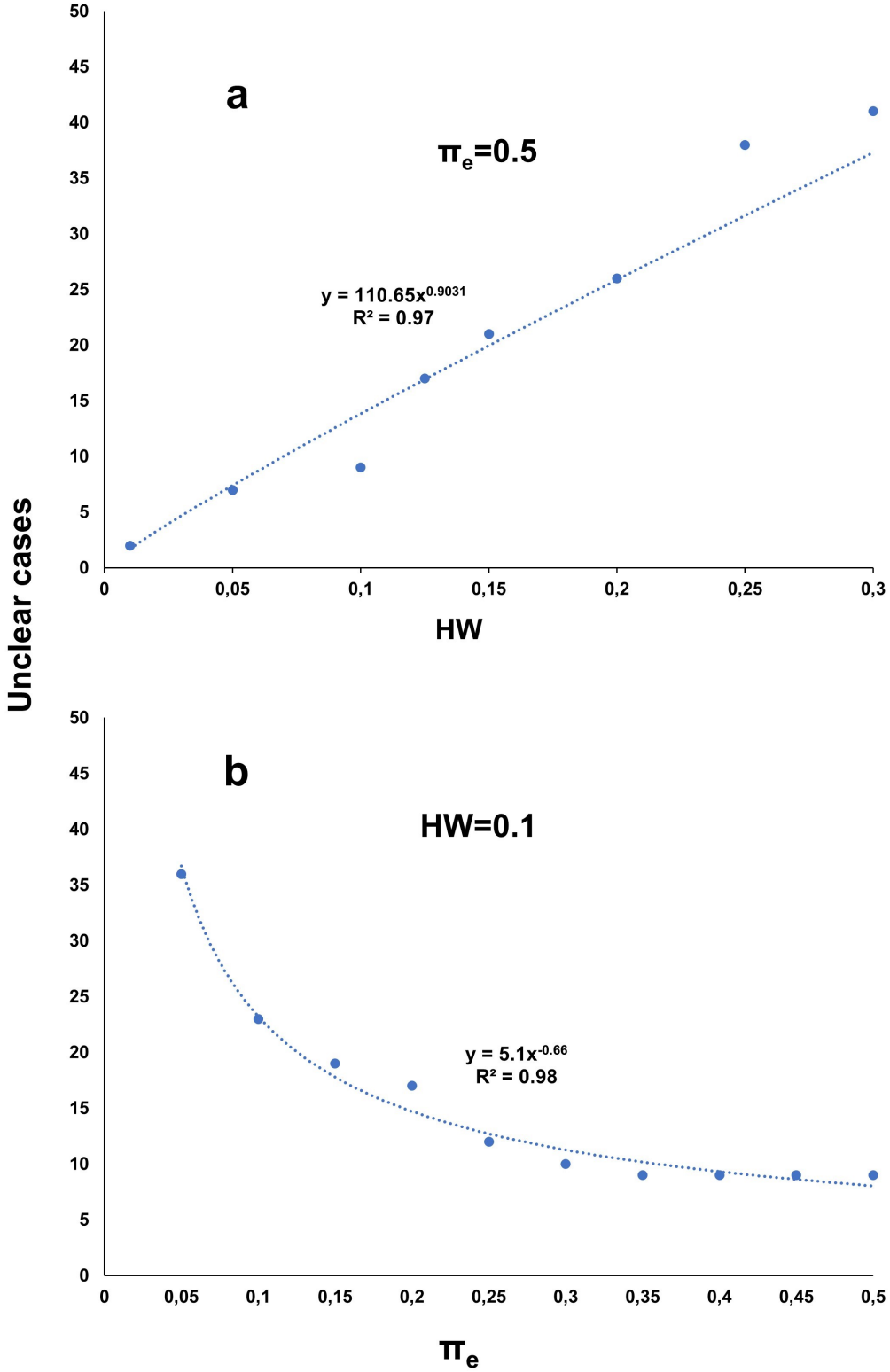

Figure 10. Regression function between (a) HW and the number of unclear cases, and (b) π_e and unclear cases, using the same data from Figure 9 (more scenarios were considered than shown in Figure 9). In (a) π_e is kept constant at 0.5, while in (b) the HW is constant at 0.1. For both HW and π_e, we have a very high coefficient of determination R^2, but for HW we have an almost linear correlation, while for π_e we have a power function with a less steep slope. In (b) from $\pi_e = 0.35$ to 0.5, the number of unclear cases remains constant.

Textbox 3:
Summarised and simplified explanation of Section 6: sample size determination using statistical standard methods.

We are looking at a binominal trait: the farmer either complies or doesn't comply with a certain requirement. For such traits, sample size in scientific surveys is determined by two variables:
a. *The probability of finding the trait, in our case the NC. We call this probability* π_e. *This variable is similar to what is commonly called "risk". But, as opposed to the common perception of "risk based sample size", the required sample size does not grow proportionally to* π_e. *It is highest for* $\pi_e = 0.5$ *(50% probability) and decreases both towards 0 and towards 1 (Figure 7). The problem is that normally we do not know* π_e *beforehand, because the number of non-compliant farmers is exactly what we want to find out. Therefore, we start from the worst-case scenario: 0.5. The real-life examples we checked, however, showed that for our purpose, we can go down to* $\pi_e = 0.35$ *without compromising the reliability of results.*
b. *The second variable is the standard error, which we are ready to accept. A common value used in many surveys for this purpose, is a standard error of 0.05. This means there is a 95% probability that the sample-based estimation for the entire group is correct. In our article we use the term "half-width"* (HW) *instead of standard error. A standard error of 0.05 corresponds to an approximate* HW *of 0.1.*

The combination of $\pi_e = 0.35$ *and* $HW = 0.1$ *yields the sample size represented by the black dotted line in Figure* 11*a.*

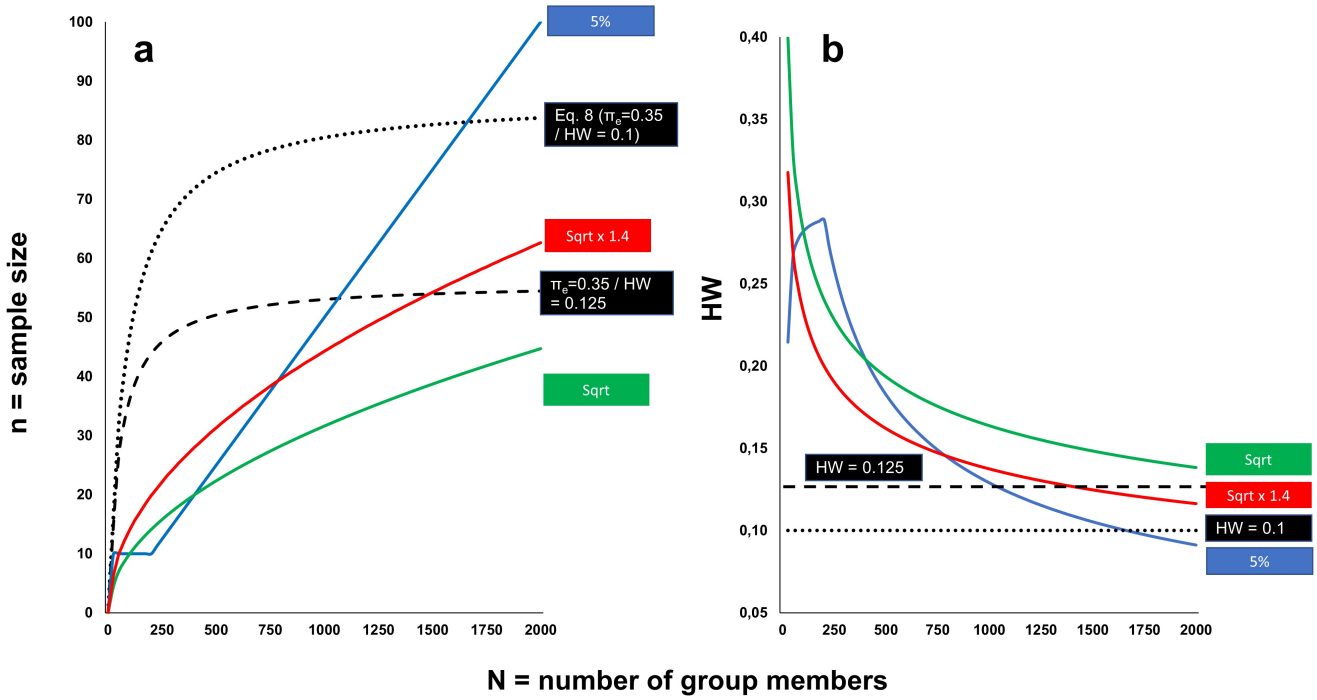

Figure 11. (a) Sample sizes plotted against group members, for four different procedures. The lines for 5%, square root, and square root multiplied by a risk factor 1.4 are the same as shown in Figure 2, but here presented in contrast to the sample size resulting from Equation 11 (black dotted line). The required sample for small groups is much bigger than with any of the other methods, while for a group of 2,000 members, it is slightly lower than the sample size required when using the 5% rule. (b) HW plotted against group members, for the same four methods. HW for Equation 11 is a horizontal line, because this is how it is defined. If we remember that HW = s.e. $\times$ 1.96 (Equation 9), this means that the accepted standard error is the same for all group sizes. If we look at the green curve for square root, we see that for a group of 20 farmers, HW is 0.41, for a group of 100 farmers, it is 0.29—meaning that we are ready to accept that 20 or 15% of NCs, respectively, slip through. The line for the 5% takes an irregular form in both (a) and (b), because according to [3] for groups with less than 200 farmers, the rules described in the caption to Figure 2 apply. Therefore, the HW reaches its highest point with 200 members, and then drops. This means that an NC in a 2,000 member group is three times more likely to be spotted than in a 200 member group.

7. Stratification

Even though most group certification rules include provisions for risk based sample selection (see Section 3), in real life these rules are mostly not followed, because the risks are generally unknown (with the exception of obvious risks, such as e.g. larger farms posing a higher risk than small ones, and farms on steep slopes being more prone to soil erosion than farms on flat land). Therefore, and because most group certification rules prescribe that members should be located in geographic proximity and have similar farming systems, the situation presented in Figure 3 is a rather exceptional one. If a CB faces such a situation, where a specific risk in one specific sub-group exists, which might slip through when applying random sampling, then the sampling method to the rest of the group is applied as described above, while for the "risky sub-group" one of the two following procedures is used:

a. If the risk situation is very clear, judgement sampling may lead to clear results, without the need for quantification. If, e.g., in a risk-based sample of 10 farmers there are three cases of insecticide use, while in the random sample from the rest of the group there are no similar problems, then the sub-group can be excluded, while the rest of the group remains certified.

b. The group can be stratified into two subgroups [26], and the sampling procedures described above are applied independently to each of the two subgroups. We should be aware, however, that a stratification, with a certification decision being taken separately for each sub-group, means that the overall sample size is increased substantially (often doubled) compared to simple random sampling.

8. Witness Audits: Sample Size and Quantification of Results

Witness audits with internal inspectors are an essential tool for assessing competence and compliance of an ICS [17,23]. Typically, such audits are combined with farm visits (see also Table 1 and Figure 1). For streamlining the assessment of the internal inspectors' performance, we suggest to use a scoring tool based on a weighted Likert scale [27]. To oblige users to make a clear decision between positive and negative scores, we recommend the use of a scale with four possible answers [28], as explained in Table 6.

The results are then summarized for all witnessed internal inspectors. If the total score for all witness audits is below a certain threshold (we suggest 70% of the maximum possible score), the ICS is considered to be not functional. If it is between 70 and 100%, corrective actions should be implemented (Table 7).

Table 6. Scoring tool using a Likert scale for witness audits with internal inspectors. For each criterion, the external inspector can make a choice: "Strongly agree / Agree / Disagree / Strongly disagree", corresponding to 3, 2, 1 and 0 marks, respectively. The results are weighted for calculating the sum, because not all criteria are equally important.

Subject: The internal inspector...	**Weight**	**Not applicable (NA) if:**
Brings all relevant records with her/him	1	
Acts in an impartial way	1	
Verifies things instead of simply interviewing the farmer	5	
Uses proper interview techniques	3	
Correctly assesses and records basic farm information	3	
Visits all relevant parts of the farm	3	
Correctly addresses any NCs observed on the farm	5	If the external inspector does not observe any NCs, this becomes NA
Writes a sufficiently detailed and accurate report	3	
Gives proper feedback to the farmer	2	
Spends enough time on the farm	2	
Follows up on implementation of previously agreed corrective actions	5	If no corrective actions had been agreed, this becomes NA
Total maximum score	**33 $\times$ 3**	**= 99**

Table 7. Summarising the scores from different witness audits for assessing the overall performance of internal inspectors. In these fictitious examples for two groups, six from a total of 10 internal inspectors have been witnessed. The maximum possible score (third column) differs from case to case, because not all questions are applicable to all farms (see Table 6, third column).

	Group 1				Group 2			
Internal inspector	**Score obtained**	**Maximum possible**	**% of maximum**	**Comment**	**Score obtained**	**Maximum possible**	**% of maximum**	**Comment**
N°1	99	99	100%	Excellent performance	45	99	45%	Unacceptable
N°2	81	99	82% Training needed	50	84	60% Needs training from scratch		
N°3	99	99	100%	Excellent performance	53	84	63%	Needs training from scratch
N°4	55	84	65%	Needs training from scratch	53	99	53%	Unacceptable
N°5	57	69	83%	Training needed	49	69	71%	Training needed
N°6	78	99	79%	Training needed	56	84	67%	Needs training from scratch
Total performance	**469**	**549**	**85%**	**In general good**	**306**	**519**	**59%**	**Very poor**

Small producer groups often have only one or two internal inspectors. In these cases the question of sampling does not come up. For groups with more internal inspectors, based on [29] we propose the following method for determining the sample of internal inspectors to be witnessed, out of a total of N internal inspectors (again: readers not interested in the statistical details, can jump to Figure 12):

$$n \geq \frac{1.96^2\sigma^2}{HW^2 + \frac{1,96^2\sigma^2}{N}} \tag{12}$$

While in Equation 10 we deal with a binominal distribution (farmers comply or don't comply with a specific requirement), here we are assuming an approximate normal distribution with unknown variance. Therefore, as opposed to Equation 10, the variance σ^2 of scores enters Equation 12 (in place of the variance $\pi_e(1-\pi_e)$ in Equation 10). Figure 12 shows the results of this equation, for $HW = 0.1$ and five variances.

From the CERES database, we evaluated the witness audit results from 18 producer groups from eight different countries, with a total of 72 internal inspectors. CERES has been working with a Likert scale with only three possible answers (Yes / Partly / No), but this should not substantially bias the variability of results, as compared to a scale with four answers. The within-group variance σ^2 for the performance of internal inspectors ranged from 0 to 0.34. For estimating the pooled variance σ_p^2 across k groups, we used [29]:

$$\widehat{\sigma}_p^2 = \frac{(n_1-1)\widehat{\sigma}_1^2 + (n_2-1)\widehat{\sigma}_2^2 + ... + (n_k-1)\widehat{\sigma}_k^2}{(n_1-1)+(n_2-1)+...+(n_k-1)} \tag{13}$$

which yielded $\widehat{\sigma_p^2} = 0.079$ for our case (orange line in Figure 12). To be on the safe side, we suggest to use a variance of 0.15 (black dotted line in Figure 12). Here, it is assumed that the underlying true variance is constant. If the performance of internal inspectors is more variable, larger samples must be used accordingly. According to our data, the variance tends to increase with lower score means. By way of analogy with the binomial distribution, and taking into account the fact that scores are integer values with a fixed lower and upper bound, it may be assumed that the variance drops to zero when the score mean μ attains the minimum or maximum value and follows a quadratic function of the mean in between. This model may be used to estimate a variance function for σ^2 which could then be used in Equation 12 with a prior estimate of the mean. Our estimate of the variance function based on the evaluation of the scores from 18 producer groups, is

$$\sigma^2 = 0.01192\mu(3-\mu) \tag{14}$$

In lack of such an estimate, the worst case scenario may be considered by plugging in the midpoint between the minimum and maximum score. Details are described in the Appendix. For the sake of simplicity we will assume a constant variance here.

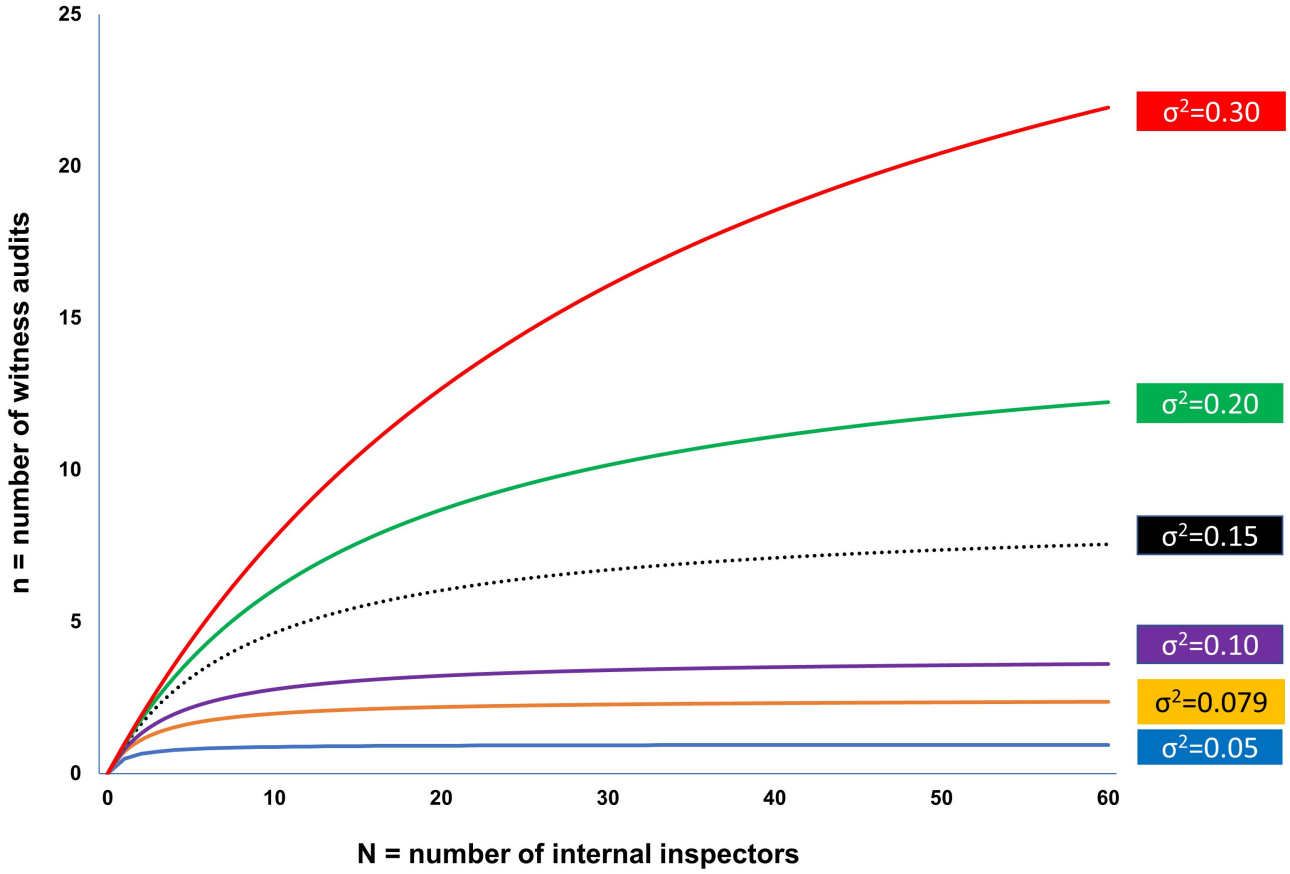

Figure 12. Sample size determination for witness audits with internal inspectors, based on Equation 12, for five different variances σ^2 concerning performance of the internal inspectors. Evaluation of 18 groups from the CERES database yielded a pooled variance σ_p^2 of 0.079 (orange line). The authors suggest assuming an average variance of 0.15 (black dotted line).

The suggested threshold of a minimum score of 70% is a political, normative proposal, and other choices are possible, of course. If the result is close to this threshold (see Table 7), the results should be assessed in combination with the results of the other inspection levels (Table 1, Figure 1). This can be done e.g. using the traffic-light system described in Table 8.

Table 8. Traffic light system for ICS performance from different inspection levels in a group certification scheme.

Performance	Farmer performance	Witness audit results	Buying system	ICS office	Conclusion
Good	No systemic NCs $> r_0$	$> 70\%$	No major inconsistencies	Good records	Certification (after corrective actions, if applicable)
Fair	Systemic NCs $> r_0$ of severity class 1-4	60 - 70%	Few inconsistencies	Some problems with farmer list, internal inspection reports, conflicts of interest, etc.	1 or 2 “Fair assessments: Certification granted, but follow-up inspection done for verifying implementation of corrective actions. More than 2 “Fair” assessments: Certification only after a follow-up inspection has confirmed implementation of corrective actions
Poor	Systemic NCs $> r_0$ of severity class 5	$< 60\%$	Major inconsistencies	Major problems with farmer list, internal reports, conflicts of interest, etc.	Case-to-case decision if: a) certification can be granted after a follow-up inspection has confirmed implementation of corrective actions, b) or certification must be denied, suspended or revoked

9. Conclusions

a. Experts agree that many CBs lack the ability of addressing NCs in producer groups at a systemic level. Our procedure for defining the systemic condition of NCs at farm level, depending on the incidence and severity of each NC, offers a tool for solving this problem. The method should be tested in practice, and the variables adjusted, as necessary.

b. Sample selection should be random, not risk oriented. If a combination of random and risk oriented sampling is used, then the group must be stratified, which leads to a larger sample size.

c. Neither a square root nor a 5% sampling rule are in line with the basic principles of sample size determination in scientific surveys. Especially for smaller groups, there is a high risk of cases slipping through with these methods. We suggest to use Equation 11 for sample size determination. If a larger HW (and thus smaller sample) is used, instead of 0.1 as in Equation 11, the CB must be ready to increase the sample if NCs of severity class 5 come up, for which it is not clear if they are systemic or not.

d. Similar to the quantification of farm inspection results, also results from witness audits with internal inspectors can be quantified and summarised in a meaningful way.

e. The combination of the results from farm inspections, witness audits, ICS office and buying system assessment, allows for differentiated certification decisions.

f. As a general rule, most important for assessing the functioning of an ICS are not large sample sizes, but personal integrity of inspectors, organisational integrity of CBs, inspector competence, inspection procedures (e.g. witness audits with internal inspectors, testing for residues where appropriate), asking the right questions to the right persons, cross-checking the right documents, and conducting inspections at the right time of the year.

References and Notes

[1] Certifying Operations with Multiple Production Units, Sites, and Facilities under the National Organic Program by the National Organic Standards Board (NOSB) to the National Organic Program (NOP). National Organic Standards Board (NOSB); 2008. Available from: https://www.ams.usda.gov/sites/default/files/media/NOP%20Final%20Rec%20Certifying%20Operations%20with%20Multiple%20Sites.pdf.

[2] Guidelines on Imports of Organic Products into the European Union. 2008 December 15; Available from: https://ec.europa.eu/info/sites/info/files/food-farming-fisheries/farming/documents/guidelines-imports-organic-products_en.pdf.

[3] Commission Delegated Regulation (EU) 2021/771 of 21 January 2021 supplementing Regulation (EU) 2018/848 of the European Parliament and of the Council by Laying Down Specific Criteria and Conditions for the Checks of Documentary Accounts in the Framework of Official Controls in Organic Production and the Official Controls of Groups of Operators. Official Journal of the European Union. 2021 May 11; Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32021R0771&qid=1622800043720.

[4] Tayleur C, Balmford A, Buchanan GM, Butchart SHm, Walker CC, Ducharme H, et al. Where are Commodity Crops Certified, and What Does it Mean for Conservation and Poverty Alleviation? Biological Conservation. 2018;217:36–46. doi:10.1016/j.biocon.2017.09.024.

[5] Kleemann L, Abdulai A, Buss M. Certification and Access to Export Markets: Adoption and Return on Investment of Organic-Certified Pineapple Farming in Ghana. World Development. 2014;64:79–92. doi:10.1016/j.worlddev.2014.05.005.

[6] Oelofse M, Høgh-Jensen H, Abreu LS, Almeida GF, Hui QY, Sultan T, et al. Certified Organic Agriculture in China and Brazil: Market Accessibility and Outcomes Following Adoption. Ecological Economics. 2010;69(9):1785–1793. doi:10.1016/j.ecolecon.2010.04.016.

[7] Handschuch C, Wollni M, Villalobos P. Adoption of Food Safety and Quality Standards among Chilean Raspberry Producers – Do Smallholders Benefit? Food Policy. 2013;40:64–73. doi:10.1016/j.foodpol.2013.02.002.

[8] Jouzi Z, Azadi H, Taheri F, Zarafshani K, Gebrehiwot K, Passel SV, et al. Organic Farming and Small-Scale Farmers: Main Opportunities and Challenges. Ecological Economics. 2017;132:144–154. doi:10.1016/j.ecolecon.2016.10.016.

[9] Rueda X, Lambin EF. Linking Globalization to Local Land Uses: How Eco-Consumers and Gourmands are Changing the Colombian Coffee Landscapes. World Development. 2013;41:286–301. doi:10.1016/j.worlddev.2012.05.018.

[10] Bolwig S, Gibbon P, Jones S. The Economics of Smallholder Organic Contract Farming in Tropical Africa. World Development. 2009;37(6):1094–1104. doi:10.1016/j.worlddev.2008.09.012.

[11] Subervie J, Vagneron I. A Drop of Water in the Indian Ocean? The Impact of GlobalGap Certification on Lychee Farmers in Madagascar. World Development. 2013;50:57–73. doi:10.1016/j.worlddev.2013.05.002.

[12] Innocenti ED, Oosterveer P. Opportunities and Bottlenecks for Upstream Learning within RSPO Certified Palm Oil Value Chains: A Comparative Analysis between Indonesia and Thailand. Journal of Rural Studies. 2020;78:426–437. doi:10.1016/j.jrurstud.2020.07.004.

[13] Akoyi KT, Mitiku F, Maertens M. Private Sustainability Standards and Child Schooling in the African Coffee Sector. Journal of Cleaner Production. 2020;264:121713. doi:10.1016/j.jclepro.2020.121713.

[14] Hajjar R, Newton P, Adshead D, Bogaerts M, Maguire-Rajpaul Va, Pinto Lfg, et al. Scaling up Sustainability in Commodity Agriculture: Transferability of Governance Mechanisms across the Coffee and Cattle Sectors in Brazil. Journal of Cleaner Production. 2019;206:124–132. doi:10.1016/j.jclepro.2018.09.102.

[15] Blanc J, Kledal PR. The Brazilian Organic Food Sector: Prospects and Constraints of Facilitating the Inclusion of Smallholders. Journal of Rural Studies. 2012;28(1):142–154. doi:10.1016/j.jrurstud.2011.10.005.

[16] Verburg R, Rahn E, Verweij P, Kuijk MV, Ghazoul J. An Innovation Perspective to Climate Change Adaptation in Coffee Systems. Environmental Science & Policy. 2019;97:16–24. doi:10.1016/j.envsci.2019.03.017.

[17] Meinshausen F, Richter T, Blockeel J, Huber B. Group Certification. Internal Control Systems in Organic Agriculture: Significance, Opportunities and Challenges; 2019. Available from: https://orgprints.org/id/eprint/35159/7/fibl-2019-ics.pdf.

[18] General Regulations, Part I – General Requirements. English Version 5.2. GLOBALG.A.P; 2019. Available from: https://www.globalgap.org/.content/.galleries/documents/190201_GG_GR_Part-I_V5_2_en.pdf.

[19] Certification Program. 2020 Certification and Auditing Rules. Rainforest Allience; 2021. Available from: https://www.rainforest-alliance.org/business/wp-content/uploads/2020/06/2020-Rainforest-Alliance-Certification-and-Auditing-Rules.pdf.

[20] Regulation (EU) 2018/848 of the European Parliament and of the Council of 30 May 2018 on organic production and labelling of organic products and repealing Council Regulation (EC) No 834/2007. 2021 January 21; Available from: https://eur-lex.europa.eu/eli/reg/2018/848/oj/eng.

[21] Commission Delegated Regulation (EU) 2021/715 of 20 January amending Regulation (EU) 2018/848 of the European Parlia-

ment and of the Council as regards the requirements for groups of operators amending Regulation (EU) 2018/848 of the European Parliament and of the Council as regards the requirements for groups of operators. 2018 November 14; Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32021R0715&qid=1622799975710.
[22] EU Regulatory changes and its effect on International Trade. Presentation during BIOFACH / VIVANESS 2021 eSPECIAL. European Commission; 2021.
[23] Smallholder Group Certification Training Curriculum on the Evaluation of Internal Control Systems A Training Course for Organic Inspectors and Certification Personnel. The International Federation of Organic Agriculture Movements (IFOAM); 2004. Available from: https://archive.ifoam.bio/sites/default/files/ics_manual_inspector_en.pdf.
[24] Sampling Methodologies. Office of the Comptroller of the Currency (OCC); 2020. Available from: https://www.occ.gov/publications-and-resources/publications/comptrollers-handbook/files/sampling-methodologies/pub-ch-sampling-methodologies.pdf.
[25] Agresti A. Categorical Data Analysis. John Wiley & Sons; 2013.
[26] Thompson SK. Sampling. Wiley; 2002.
[27] Likert R. A Technique for the Measurement of Attitudes. vol. 140. Archives of Psychology; 1932.
[28] Allen E, Seaman C. Likert Scales and Data Analyses. Quality Progress. 2007 July; Available from: http://rube.asq.org/quality-progress/2007/07/statistics/likert-scales-and-data-analyses.html.
[29] Guidance on sampling methods for audit authorities. Programming periods 2007-2013 and 2014-2020. European Commission; 2017. Available from: https://ec.europa.eu/regional_policy/sources/docgener/informat/2014/guidance_sampling_method_en.pdf.
[30] Searle SR, Casella G, MacCulloch CE. Variance Components. Wiley; 1992.
[31] Wedderburn RWM. Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. Biometrika. 1974;61(3):439–447. doi:10.2307/2334725.

Appendix

Based on the minimum and maximum possible mean scores (0 and 3, respectively), we may assume this variance function:

$$\sigma^2 = \phi\mu(3-\mu) \tag{15}$$

where σ^2 is the variance and μ the mean. This can be estimated by linear regression. The intercept is zero, and there is a single regression coefficient ϕ for the predictor variable $\times = \mu(3-\mu)$. Assuming approximate normality of the score means, we have for the sample variance $\widehat{\sigma}^2$ [30]:

$$var(\hat{\sigma}^2) = \frac{2\sigma^4}{n-1} \tag{16}$$

This function (Equation 16) for the variance estimate can be used in a quasi-likelihood approach [31] for fitting Equation 12. Here, we used the GENMOD procedure in SAS.

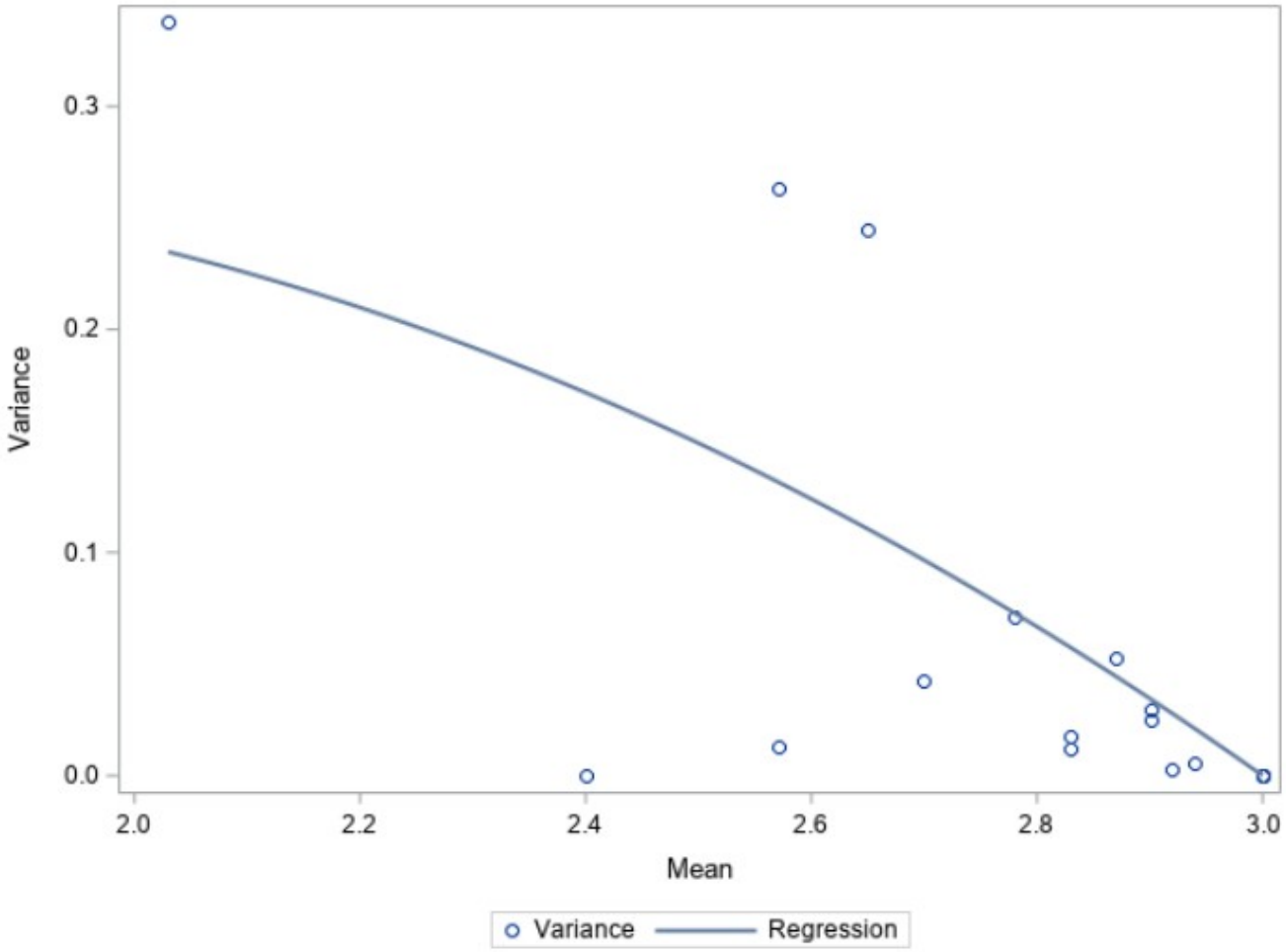

Figure A1. Variance plotted against mean for the scores given for internal inspector performance.

The estimated variance function is:

$$\sigma^2 = 0.1192\mu(3-\mu) \tag{17}$$

Using this function, the variance can be computed for an a priori estimate of μ, and this variance can then be used in an equation for determining sample size, such as Equation 9 in the main text. If a prior value is not available, one may plug in the worst-case value $\mu = 1.5$.

```
    data v;
input
Mean Variance n; x=Mean*(3-Mean);
datalines;
 3.00 0.000 3
 3.00 0.000 2
 3.00 0.000 3
 3.00 0.000 3
 2.94 0.006 2
 2.92 0.003 4
 2.90 0.030 3
 2.90 0.025 3
 2.87 0.053 3
 2.83 0.018 9
 2.83 0.012 4
 2.78 0.071 7
 2.70 0.043 7
 2.65 0.245 2
 2.57 0.263 3
 2.57 0.013 3
 2.40 0.000 3
 2.03 0.338 8
;
proc glimmix data=v;
_var_=2*_mu_*_mu_/(n-1);
model variance=x/noint solution;
output out=v predicted=p;
run;
proc sgplot data=v;
scatter y=variance x=mean;
reg y=p x=mean/degree=2;
run;
```

Table A1. Parameter estimates.

Effect	Estimate	Standard Error	DF	t Value	$Pr > \|t\|$
x	0.1192	0.02264	17	5.27	<0.0001
Scale	0.005708	0.001958			