Article

"Measure of agreement between experts on apple damage assessment"

C. Vincent et J. Hanley

Phytoprotection, vol. 78, n° 1, 1997, p. 11-16.

Measure of agreement between experts on apple damage assessment

Charles Vincent[1] and James Hanley[2]

Received 1995-12-29; accepted 1997-02-03

PHYTOPROTECTION 78 : 11-16.

Although damage evaluation is an important and frequent exercise in economic entomology, there are no quantitative studies on inter-rater agreement of experts. In this experiment conducted during the 50th New York, New England and Canadian Pest Management Conference, four teams of experts independently estimated the damage on 200 apples at harvest. The participants identified 22 types of damage caused by insects, 8 by diseases, and 8 related to other causes. For each type of damage an average measure of agreement was calculated. The lowest average agreements were found in plum curculio (*Conotrachelus nenuphar*) [Coleoptera : Curculionidae] damage (71.8%), tarnished plant bug (*Lygus lineolaris*) [Hemiptera : Miridae] damage (83.2%), and by early lepidoptera damage (87.1%). The usefulness of inter-rater agreement experiments is discussed in the context of many situations pertaining to crop protection.

[Mesure du degré de concordance entre experts pour l'évaluation de dommages sur des pommes]

Quoique l'évaluation des dommages soit un exercice important et fréquent en entomologie appliquée, il n'y a pas d'études publiées concernant le degré de concordance des évaluations de dommages par les experts. Au cours de cette étude, effectuée lors de la 50e Conférence en lutte intégrée des vergers de l'État de New York, de la Nouvelle-Angleterre et du Canada, quatre équipes d'experts ont évalué, de façon indépendante, les dommages causés sur 200 pommes. Les participants ont identifié 22 types de dommages causés par les insectes, 8 par des maladies et 8 reliés à d'autres causes. Nous avons calculé un degré de concordance pour chaque type de dommage. Les degrés de concordance les plus bas concernaient les dommages du charançon de la prune (*Conotrachelus nenuphar*) [Coleoptera : Curculionidae] (71,8 %), de la punaise terne (*Lygus lineolaris*) [Hemiptera : Miridae] (83,2 %) et les dommages causés par les larves de lépidoptères en début de saison (87,1 %). On discute de l'utilité de l'usage du degré de concordance dans le contexte de plusieurs situations de lutte intégrée.

1. Horticultural Research and Development Centre, Agriculture and Agri-Food Canada, 430 Gouin Blvd., Saint-Jean-sur-Richelieu, Quebec, Canada J3B 3E6. Contribution no. 335/97.03.04R
2. Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Ave. W., Montreal, Quebec, Canada H3A 1A2

INTRODUCTION

When a crop is attacked by several pest species, it is often desirable to know the proportion of damage attributable to a given species. Damage assessment allows the Integrated Pest Management (IPM) practitioner to evaluate the overall success of a program and, more importantly, to determine where the program or component of a program (*e.g.*, a biocontrol agent or a pesticide) failed. Damage assessment, either quantitative or qualitative, is therefore a very important part of IPM because, as a result, the performance of programs, management tactics and persons is evaluated.

Damage assessment is frequently done by scouts directly in the field or in packing facilities. Identification of the cause of damage often relies on a set of visual characters that appear on agricultural products. Scouts may also rely on information acquired through sampling and monitoring in the fields as indirect evidence suggesting the cause of damage. An unavoidable difficulty is that, on some agricultural products such as apples, damage caused by either insects, diseases or other factors may appear simultaneously and may even physically overlap. Occasionally positive identification of the damage can be achieved, *i.e.* when insects are found closely associated with the damage, or when fungi are isolated, cultured and identified.

Studies on the agreement between expert verdicts have been done in several circumstances in medical sciences (Feinstein 1985a) to measure, for example, the agreement between assessments of pneumoconiosis determined by radiography (Liddell 1963) or, the extent of dental cavities in clinical trials (Fleiss *et al.* 1979). Likewise, Caro *et al.* (1979) discussed the problem of inter-observer reliability in relation to animal behavioral studies.

How does this approach translate in crop protection? Intuitively it is known that agreement between damage evaluations of experts, scouts or growers is not perfect. But the question is : to what extent do these evaluations agree (or disagree)? In this study our objective was to assess the degree of agreement among experts on apple damage assessment. We first present the results of an experiment done to measure the agreement of apple damage identification between four teams of experts. Apples lend themselves easily to that kind of study because they are well defined discrete units and several types of damage may appear simultaneously on one fruit. We then briefly discuss the usefulness of this exercise in a variety of situations.

MATERIALS AND METHODS

The experiment was done during the 50th New York, New England and Canadian Pest Management Conference held at Stowe (Vermont), U.S.A. After explanation of the objectives and the rules of the experiment, four groups (variable number of persons from 2 to 5) hereafter denoted A, B, C and D were formed spontaneously (*i.e.* without constraint on expertise, location of work or number of persons per group). The persons were apple pest management specialists, mostly working for University or State Extension Services in Eastern North America. Damaged apples that were presented to the groups represented a typical sample of what can be observed in unsprayed apple orchards of New England. Each group assessed the damage on 200 apples that were collected from unsprayed apple orchards. Most fruit had more than one type of damage. Each fruit was examined individually and then replaced to its assigned position carved into a wooden tray. Consultation among the members of a team was allowed, but not between teams. Damage was recorded in a free format, *i.e.* with the groups' own codes and sequence of entry. Data were subsequently decoded and tabulated for each damage type to allow intergroup comparisons.

To calculate agreement, a two-way table was computed for each type of damage for each of the six possible team combinations (AB, AC, AD, BC, BD, CD) with procedure TABLES of Systat Software (version 3.2) for the Macintosh computer (Wilkinson 1987). An example of computation of agreement between group A and B for damage caused by the

plum curculio (*Conotrachelus nenuphar* Herbst) [Coleoptera : Curculionidae] is shown in Figure 1. The results presented are, for each type of damage, the frequency of damage as estimated by a group, and the average agreement for the six group combinations. We chose the method because : 1) the appropriate method, the Kappa statistic, is overly conservative (Maclure and Willet 1987); 2) the Kappa statistic is very sensitive to zero values present in datasets, which was unavoidable in our case; 3) interpretation of results is straightforward and thus easier than that of the Kappa statistic, especially when more than two raters are involved; and 4) it is recommended by Feinstein (1985b).

RESULTS AND DISCUSSION

The four groups of experts distinguished 22 types of damage caused by insects, 8 caused by diseases, and 8 caused by other factors (Table 1). There was 95.3% agreement on whether an apple was damaged or not. Among the damage caused by insects, the percent agreement was > 95% in 14 cases out of 22. The types of insect damage with the lowest average agreement were : the plum curculio (*C. nenuphar*) (71.8%), the tarnished plant bug (*Lygus lineolaris* Palisot de Beauvois) [Hemiptera : Miridae] (83.2%), and early lepidoptera damage (87.1%). Two diseases had an average agreement of less than 95%, namely fly speck (*Schizothyrium pomi* (Mont.: Fr.) von Arx) (91.8%), and apple scab (*Venturia inaequalis* (Cke) Wint.) (94.3%). Among other types of damage having less than 95% agreement are mechanical damage (91.3%), hail (93.0%), and unknown (90.2%).

The average percent agreements were lower for plum curculio, tarnished plant

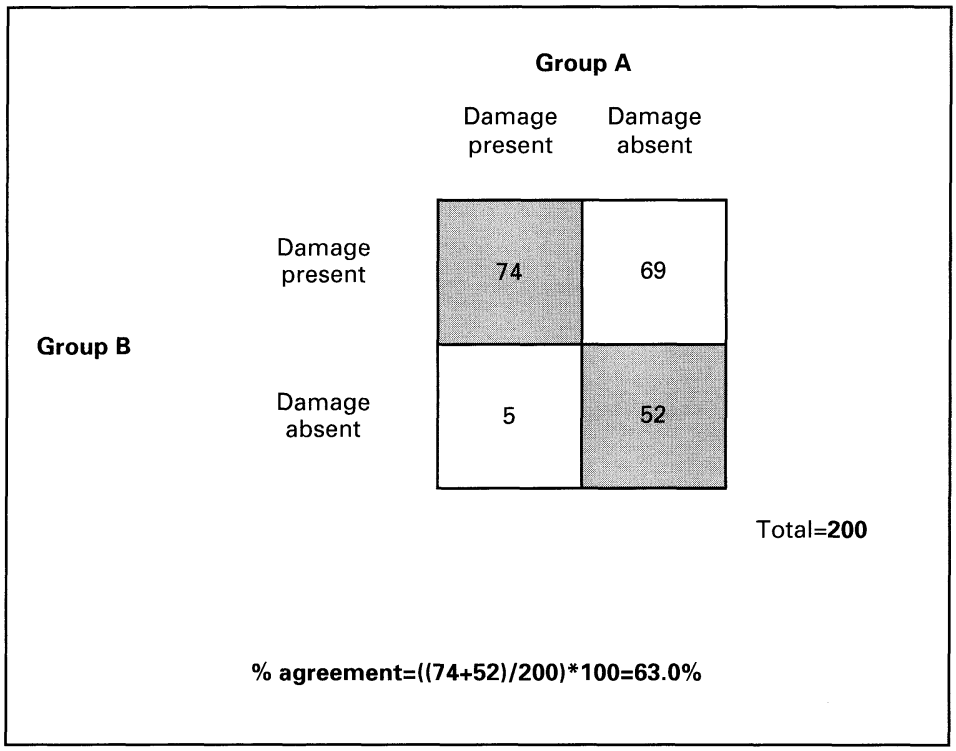

Figure 1. Example of computation of agreement between group A and B for plum curculio damage. Grey areas represent positive and negative agreements; white areas represent disagreements.

Table 1. Number of apples[a] considered to have the type of damage indicated and agreement of damage evaluation between four groups of experts

Type of damage	Group of experts				Average agreement (%)
	A (No. damaged)	B (No. damaged)	C (No. damaged)	D (No. damaged)	
No. of fruit damaged (n = 200)	197	184	197	194	95.3
Insects					
Aphid honeydew	2	0	0	0	99.5
Apple curculio	0	0	0	3	99.3
Apple maggot	1	6	18	0	94.3
Codling moth	3	4	0	9	95.8
Comstock mealybug	5	0	0	2	98.6
Early lepidoptera damage	1	52	7	32	87.1
European apple sawfly	7	8	6	10	97.1
Eye-spotted budmoth	0	0	0	1	99.8
Internal damage by lepidoptera	0	0	1	0	99.8
Late lepidoptera damage	0	0	27	0	93.3
Leafroller	21	1	2	15	91.9
Lesser appleworm	0	0	0	1	99.8
Oblique banded leafroller (overwintered generation)	0	19	0	0	95.3
Oblique banded leafroller (summer generation)	0	28	0	1	92.9
Plum curculio	143	79	105	95	71.8
Redbanded leafroller	0	6	0	9	97.3
San José scale	38	25	32	27	92.2
Sting	0	0	2	0	99.5
Sting bug	5	0	0	10	96.6
Tarnished plant bug	19	6	44	35	83.2
Tufted apple budmoth	0	0	0	1	99.8
White apple leafhopper	38	13	0	16	86.1
Diseases					
Apple scab	1	0	8	19	94.3
Bitter pit	3	0	5	3	97.9
Blackrot	1	0	0	2	99.3
Fly speck	1	0	0	32	91.8
Mildew	0	0	0	1	99.8
Russeting	1	0	0	0	99.8
Soothy blotch or soothy mold	0	0	0	2	99.5
Sterol inhibitor (Rubigan)	0	0	9	0	97.8
Other types of damage					
Bird	0	0	0	1	99.8
Boron deficiency	1	0	0	0	99.8
Calcium deficiency	0	0	0	4	99.0
Hail	0	0	0	28	93.0
Heat	0	0	1	0	99.8
Mechanical damage	3	0	24	15	91.3
Slug	0	0	1	1	99.5
Unknown	23	7	2	16	90.2

[a] n = 200 apples examined.

bug and early lepidoptera damages. Three factors may explain this result. First, the groups of experts had no information on the kind of pest problems experienced in the orchards. Second, the apples were examined only in surface : normally an IPM practitioner can use a knife to cut the fruit and get more information on the appearance of the damage. Third, certain types of damage, like those of the plum curculio, are typically variable in form.

For several types of damage, *e.g.*, eye-spotted budmoth (*Spilonota ocellana* D. & S.) [Lepidoptera : Tortricidae], the frequency of damage was very low and so the percent agreement was higher than 99%. In a field situation, the IPM practitioner faces various levels of pest damage and therefore has limited information on the levels of importance of minor pests.

It must be understood that our results have a relative value because we had no absolute reference for a given type of damage. To paraphrase Koran (1975) the findings of two physicians may agree (be reliable), and yet be wrong as compared to an independent standard of accuracy. The only way to achieve a definitive reference is to let an insect feed on a fruit covered with a sleeve cage and to wait for the damage to appear and develop. Again, intrinsic variations in damage appearance represent an unavoidable difficulty. The issue is not necessarily whether experts agree, but whether the experts are right. IPM practitioners must provide assessments that are both precise (*i.e.*, agreement or repeatability of assessments among experts) and accurate (*i.e.*, freedom from bias). In practice, however, growers can better tolerate lower levels of precision than accuracy.

There are limits to an exercise such as this one. Because the insect fauna infesting eastern North American orchards is different for other apple growing regions such as Washington State, Chile or France (regions where, for instance, the plum curculio is absent), the present exercise could be conducted with a different outcome. A similar exercise can be repeated in other crops where several types of damage appear simultaneously. It would allow one to measure the degree of agreement between IPM practitioners and, possibly, to clarify the reasons for disagreement. We have used a nominal (presence or absence of damage) scale to classify apples. An ordinal scale of damage assessment can also be used (Liddell 1963). Finally certain types of damage are common (and others uncommon) within a region. Because it is likely that IPM practitioners would face the situation of identifying common types of damages more often in their region, they may have a better agreement among themselves for common types of damage. The validity of this hypothesis could be checked by setting another series of experiments.

The objective measurement of agreement between IPM practitioners can be useful in a variety of situations. For instance research entomologists or phytopathologists can measure the limits of their expertise and thereby design experiments to reach a higher degree of agreement. IPM employers will surely be interested in objectively measuring the relative reliability of their employees to offer standardized advices to growers. IPM advisers can measure the agreement between growers, thus identifying their weaknesses in order to better target their extension courses. Growers confidence can be maintained (or restored) if they realize that there is an objective limit to the degree of agreement between two or more experts consulted independently. In the design of computerized expert systems to assist damage diagnosis, software developers can assign a probability of agreement for a given type of damage. Agreement experiments may help to settle lawsuits between two parties, especially when the objective identification of the damage is the focus of debate. Correct identification of damage is essential to assess the true performance of biocontrol agents or pesticides. Finally, the agreement on damage assessment of agricultural goods between the experts of two trading countries can be better understood and, in case of disagreement be settled more objectively. Considering agricultural phytoprotection worldwide, billions of dollars are at stake annually.

ACKNOWLEDGMENTS

We thank the anonymous volunteers of groups A, B, C, D, and Benoit Rancourt, Agriculture and Agri-food Canada, for assistance during the experiment and data compilation. The late Rick Weires, Hudson Valley Laboratory, Highland, NY and Ronald J. Prokopy, Department of Entomology, University of Massachusetts, Amherst, MA each provided 100 apples collected from NY and MA orchards, respectively. Pierre Joseph Charmillot, Station fédérale de recherches agronomiques de Changins, Switzerland, commented the manuscript.

REFERENCES

Caro, T.M., R. Roper, M. Young, and G.R. Dank. 1979. Inter-observer reliability. Behaviour 69 : 303-315.

Feinstein, A.R. 1985a. A bibliography of publications on observer variability. J. Chronic. Dis. 38 : 619-632.

Feinstein, A.R. 1985b. Clinical epidemiology, the architecture of clinical research. W. B. Saunders Co., Philadelphia. 812 pp.

Fleiss, J.L., S.L. Fischman, N.W. Chilton, and M. Park. 1979. Reliability of discrete measurements in caries trials. Caries Res. 13 : 13-31.

Koran, L.M. 1975. The reliability of clinical methods, data and judgments. N. Engl. J. Med. 293 : 642-646.

Liddell, F.D.K. 1963. An experiment in film reading. Br. J. Ind. Med. 20 : 300-312.

Maclure, M., and W.C. Willet. 1987. Misinterpretation and misuse of the Kappa statistic. Am. J. Epidemiol. 126 : 161-169.

Wilkinson, L. 1987. SYSTAT : the system for statistics (Macintosh computer, version 3.2), Evanston, Illinois.