Bot. Bull. Acad. Sin. (1995) **36:** 121–133

Sensory analysis in quality control—the agreement among raters

Hong-Pang Wu[1] and Lee-Shen Chen[2]

[1]*Institute of Botany, Academia Sinica, Taipei, Taiwan 115, Republic of China*

[2]*Department of Statistics, Ming Chuan College, Taipei, Taiwan, Republic of China*

(Received September 18, 1993; Accepted September 9, 1994)

Abstract. This paper considers cluster analysis, kappa methods, and log-linear association models, which are often used to study agreement in medical and psychology research but are rarely used in sensory evaluation, to evaluate the agreement of tea sensory data among panelists. We recorded sensory data six times during the period from 1989 to 1990. There were 8 to 10 well-trained panel members, giving 11 ordinal categories for each tea sample. We then combined the 11 categories into 5 and applied the three methods to measure the agreement among panelists, and we found that agreement among the panelists was not high enough, regardless of the proposed methods.

Keywords: Cluster analysis; Kappa coefficient (*k*); Log-linear association model; Panel; Sensory control.

Introduction

An experiment may produce quantitative or qualitative data. There are four measurement-scale types: interval, ratio, ordinal, and nominal. Interval and ratio are quantitative; they are described by a well known probability distribution theory, and there exist relatively simple methods to analyze and interpret the data. Ordinal and nominal are qualitative, and require transformation methods to convert data to quantitative form. These transformation methods are not always easy to use and interpret, especially with sensory data. The sensory data depends not only on the senses of taste, smell, and vision but also on the data scale and sensory evaluation method.

We consider two types of sensory evaluation method. The first evaluates characteristics, such as differences in color or smell, that can be detected by instruments or by the human sense organs. The other evaluates characteristics that can be detected only by human sensory organs, such as pungency, brothiness, and freshness. In this paper, we study evaluation by human sensory organs. Sensory evaluation is much faster and cheaper than machine evaluation, and humans can evaluate some sensory characters that machines can not measure.

The increasing standard of living in recent years has led to a market for higher quality food and has stimulated advances in crop breeding. We used tea as our experimental material. Its quality-factors, including its shape, color, odor, etc., were evaluated by panels of judges. From these evaluations, we can better understand the cultivated varieties of tea and their market potential. There are two main factors which affect the evaluation of quality—the agreement among panelists and the items selected for evaluation. In this paper, we analyze only agreement among panelists. Because this is the first paper to discuss the sensory evaluation of Paochung tea by a panel of judges, we do not use the transformation method, and weight each category in our analysis of the data. The samples were collected from 8 different counties in Taiwan, and were evaluated by 8 to 10 well-trained panelists. Based on the types of evaluation data, we apply the Kappa coefficient (*k*) (Cohen, 1960; Fleiss, 1971; Landis and Koch, 1977; Conger, 1980), the log-linear association agreement model (Agresti, 1988), and the cluster technique. We also compare the differences among these three methods, which are used in medical and psychological research, but are rarely used in sensory data analysis. We used the SAS statistical software package to compute the results.

[1]Corresponding author.

Materials and Methods

We used several kinds of tea, which were cultivated by the Taiwan Tea Experiment Station (TTES): TT 12, 13, 14, 15, 16, 17, 209, Chin-shin Oolong, and Chin-shin Dapong, which have been grown in 8 different counties of Taiwan since 1983. We harvested the tea after six years and then manufactured Paochung tea.

Method of Sensory Evaluation

The standard method for the sensory evaluation of tea is: steep 3 grams of tea in 150 ml boiling water (steep slender teas for 5 minutes, ball types for 6 minutes) and then decant. There were 8 to 10 well trained panelists participating in this evaluation. The panelists evaluated the overall quality of the teas according to appearance and brightness, color of liquid, and taste (e.g. odor, astringency, bitterness), and the quality was recorded by 11 categories, given the quality between 1 (lowest quality) and 11 (highest quality).

Methods of Statistical Analysis

Cluster analysis—In this paper, we use two-stage density linkage (Sarle, 1983) and density linkage (Wong and Lane, 1983) cluster techniques grouping the same agreement of the panelists. These analyses suggest there are no difference within a group, but many differences between groups.

Kappa coefficient method (k method)—The k coefficient is often used to study the agreement among judgments made by physicians or psychologists. This method is based on discrete data, such as ordinal or nominal scale data, on the opinions of each panelist about the same item. However, in literature this method has not been used in tea data with sensory evaluation. First, we consider the agreement of two raters, then extend our view to agreement among many raters. We also apply Cohen's (1960) k coefficient to determine the degrees of agreement between raters. We presume that each of a sample of n subjects (tea samples) is given independently to the same raters, with the ratings being on a discrete scale consisting of L categories (Table 1).

The overall proportion of observed agreement is,

$$P_0 = \sum_{i=1}^{L} P_{ii} ,$$

and the overall proportion of expected agreement is,

$$P_e = \sum_{i=1}^{L} P_{i.} P_{.i} ,$$

So the overall value of two raters' k (Cohen, 1960) is then

$$k = \frac{P_0 - P_e}{1 - P_e} \quad (1)$$

For testing, the ratings are independent. Fleiss, Cohen, and Everitt (1969) showed that the appropriate standard error of k is estimated by

$$STD(k) = \frac{1}{(1-P_e)\sqrt{n}} \sqrt{P_e + P_e^2 - \Sigma P_{i.} P_{.i} (P_{i.} + P_{.i})} \quad (2)$$

$$Z_{(k)} = \frac{k}{STD(k)} \quad (3)$$

(Z is a standard normal distribution)

Fleiss (1971) generalized two raters' k to measure the degrees of agreement among multiple raters, and demonstrated the statistic, k_d, that is used to measure multi-rater agreement by

$$k_d = \frac{P_0^* - P_e^*}{1 - P_e^*} \quad (4)$$

where P_0^* is the observed proportion of agreement given by

$$P_0^* = \frac{1}{NR(R-1)} \sum_{i=1}^{N} \sum_{k=1}^{L} X_{ik} (X_{ik} - 1) ,$$

N is all subjects (i.e. tea samples); R is all raters; X_{ik} is the number of raters who assigned the ith subjects to the kth category, where i = 1, ..., N, k = 1, ..., L

P_e^* is the expected proportion of agreement, which is given by

$$P_e^* = \sum_{k=1}^{L} P_k^2 ,$$

where $P_k = \frac{1}{NR} \sum_{i=1}^{N} X_{ik}$,

Also the appropriate standard error of k_d is estimated by

$$STD(k_d) = \frac{\sqrt{2}}{(1-P_e^*)\sqrt{NR(R-1)}} \times \sqrt{P_e^* - (2N-3)P_e^{*2} + 2(N-2)\sum_{k=1}^{L} P_k^3} , \quad (5)$$

and by central limit theorem $Z_{(k_d)}$ will be approximately distributed as a standard normal variate, where

Table 1. Contingency table.

Categories of panelist B	Categories of panelist A											
	1	2	.	.	.	K'	.	.	.	L	Total
1	P_{11}	P_{12}	.	.	.	$P_{1k'}$	.	.	.	P_{1L}	$P_{1.}$
2	P_{21}	P_{22}	.	.	.	$P_{2k'}$	.	.	.	P_{2L}	$P_{2.}$
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
K	P_{k1}	P_{k2}	.	.	.	$P_{kk'}$	.	.	.	P_{kL}	$P_{k.}$
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
L	P_{L1}	P_{L2}	.	.	.	$P_{Lk'}$	.	.	.	P_{LL}	P_{L}
Total	$P_{.1}$	$P_{.2}$	.	.	.	$P_{.k'}$	.	.	.	$P_{.L}$	*1*

$P_{kk'}$ is probability of panelist A at category K' and panelist B at category K.
$P_{.k'}$, $P_{k.}$ are marginal probability, where $p_{.k'} = \sum_i p_{ik'}$, $p_{k.} = \sum_j p_{kj}$

$$Z_{(k_d)} = \frac{k_d}{STD(k_d)} \quad (6)$$

Establishing the log-linear model—The k methods give only a general idea of the data. They do not tell us anything about the characteristics of the rating categories, nor how these may vary across raters. However the approached of log-linear model is potentially more informative. If two observers' opinions are independent, then the expected proportion will be,

$P_{ij} = P_{i.}P_{.j}$,

where $P_{i.}$ is the probability of classification in the *i*th category by panelist B, $P_{.j}$ is the probability of classification in the *j*th category by panelist A, and P_{ij} is the cell probability corresponding to X_{ij}, which is an observed count from the I × J contingency table with multinomial distribution. We use log to transfer the above equation as a linear model,

$\log P_{ij} = \log P_{i.} + \log P_{.j}$

The expected value of X_{ij} (viewed as a random variable) is $m_{ij} = NP_{ij}$, where N is the total sample size, and under the model of independence $m_{ij} = NP_{i.}P_{.j}$, which leads to

$$\log m_{ij} = \mu + \lambda_i^a + \lambda_j^b \quad (7)$$

where

m_{ij} : expected frequency of cases in cell (*i,j*)

μ : the common effect of total sample size

λ_i^a : A rater's category effects, $\sum_i \lambda_i^a = 0$

λ_j^b : B rater's category effects, $\sum_j \lambda_j^b = 0$

If two raters have similar opinions of the evaluated items, then following Tanner and Young (1985), the model (7) can be extended as

$$\log m_{ij} = \mu + \lambda_i^a + \lambda_j^b + \delta(i,j) \quad (8)$$

where

$$\delta(i,j) = \begin{cases} \delta_i , & i = j,\ i = 1,2,\ldots,L; \\ 0 , & \text{others.} \end{cases}$$

We call $\delta(i,j)$ a factor of agreement.

For ordinal data, Agresti (1988) added local log-odds ratio in (8) to get

$$\begin{aligned} \log m_{ij} &= \mu + \lambda_i^a + \lambda_j^b + \delta(i,j) + \beta\mu_i\mu_j \quad (9) \\ &= \mu + \lambda_i^a + \lambda_j^b + \delta(i,j) + \beta_{ij} \end{aligned}$$

where β is a noise factor caused by the structure of the contingency table, and μ_i and μ_j are scores associated with category i for rater A and category j for rater B. The noise factor can be interpreted as a tendency for raters to confuse categories having similar scores. Then, $\delta(i,j)$ would not be affected by the ranking order of the scale (which may reflect the actual agreement among the panelists). We may extend these methods to identify the agreement among 3 or more panelists.

Results

There were 487 tea samples, collected from eight different tea-growing areas during the years 1989–1990. We performed 6 evaluations during the experiment, and ten panelists (**a** – **j**) took part. All panelists attended the evaluations during 1990. Panelist **i** did not attend the first evaluation and panelist **j** did not attend the second and third evaluations, in 1989. The frequency distribution of panel evaluations (Table 2) shows that most of the panelists were distributed in the 4th, 5th, 6th, and 7th orders. Panelists **b** and **h** had a slightly lower rank than the others, but their ranks centered in the 4th, 5th, and 6th orders. We combined the data from ranks 1 to 5 (i.e. level 1 is a combination of the 4th order and below, level 2 is the 5th, level 3 is the 6th, level 4 is the 7th, and level 5 is a combination of the 8th and above), which is shown in Table 3.

Cluster analysis—We used the cluster procedure of the SAS statistical package to identify the agreement among

Table 2. Frequency table of each raters evaluation under 11 ranks.

Unit: Tea samples

Rank	Raters										
	A	B	C	D	E	F	G	H	I	J	Total
1	2	15	0	0	0	0	5	24	3	0	49
2	1	25	0	0	0	0	45	13	11	0	95
3	2	36	0	0	12	0	24	45	22	2	143
4	7	46	13	21	45	119	134	94	43	26	548
5	136	165	185	137	209	243	180	145	91	103	1594
6	274	102	218	211	162	91	88	99	119	104	1468
7	47	52	48	105	46	26	9	45	93	55	526
8	15	31	19	12	11	5	2	19	45	13	172
10	2	0	0	0	0	0	0	1	0	0	3
11	0	2	0	0	0	0	0	0	0	0	2
Total	487	487	487	487	487	487	487	487	433	308	4637

Table 3. Frequency table of each raters evaluation under 5 ranks.

Unit: Tea samples

Rank	Raters A	B	C	D	E	F	G	H	I	J	Total
1	12	122	13	21	57	119	208	176	79	28	835
2	136	165	185	137	209	243	180	145	91	103	1594
3	274	102	218	211	162	91	88	99	119	104	1468
4	47	52	48	105	46	26	9	45	93	55	526
5	18	46	23	13	13	8	2	22	51	18	214
Total	487	487	487	487	487	487	487	487	433	308	4637

Table 4. Cluster analysis of two identifiers at each evaluation time.

Unit: Panel

Group no.	1989 1	1989 2	1989 3	1990 1	1990 2	1990 3	1989	1990	Both
1	EFG	ABCDH	BDI	ABC	BCF	CDHIJ	ACD	ACDEIJ	ACD
2	ACDJ	EFG	EFH	DEIJ	ADEIJ	BEFG	EFGH	BFG	EFGH
3	BH	I	ACG	FGH	GH	A	B	H	B
1	none	ABCDHI	none	none	none	ACDHIJ	ACD	ACDEIJ	ACD
2	none	EFG	none	none	none	BEFG	BEFGH	BFGH	BEFGH

panelists. Under the defined two-group identifier of two-stage density linkage cluster techniques and the density linkage cluster techniques, we obtained similar results for agreement among panelists. Because the clusters in the density linkage model often merge before all the points in the tails have clustered, the two-stage density linkage cluster method assigns all points to the model clusters before the modal clusters are allowed to join. We use the two-stage density linkage cluster technique to determine the similarity of agreement among these panelists. The symbol 'X' in Figures 1 to 3 indicates the closeness of the relationship among panelists, i.e. more Xs indicates that panelists have high agreement, fewer Xs indicates that panelists have lower agreement. From the cluster fusion density plots (Figures 1 to 3) and the results (Table 4), we are aware that the panelists' evaluations do not have high correlation, and we reconsidered these panelists by examining the agreement between randomly selected pairs of panelists. We applied two methods, the kappa coefficient and the log-linear model, to refit the data and measure the closeness of the opinions among panelists.

Kappa coefficient (k)— High positive k value indicate close agreement among panelists, a value of zero means that the panel's opinions are unrelated, and negative values indicate the presence of random error. The k values for our panelists (Table 5, 6) are all below 0.50—this implies that the agreement between pairs of panelists is not high enough. We rearranged these panelists into several groups based on high k value. Figure 4 shows the relationship among panelists in each evaluated period; a straight line shows under $\alpha = 0.01$ significance, a dotted line shows under $\alpha = 0.05$ significance. We rearranged panelists into groups having high agreement in each evalu-

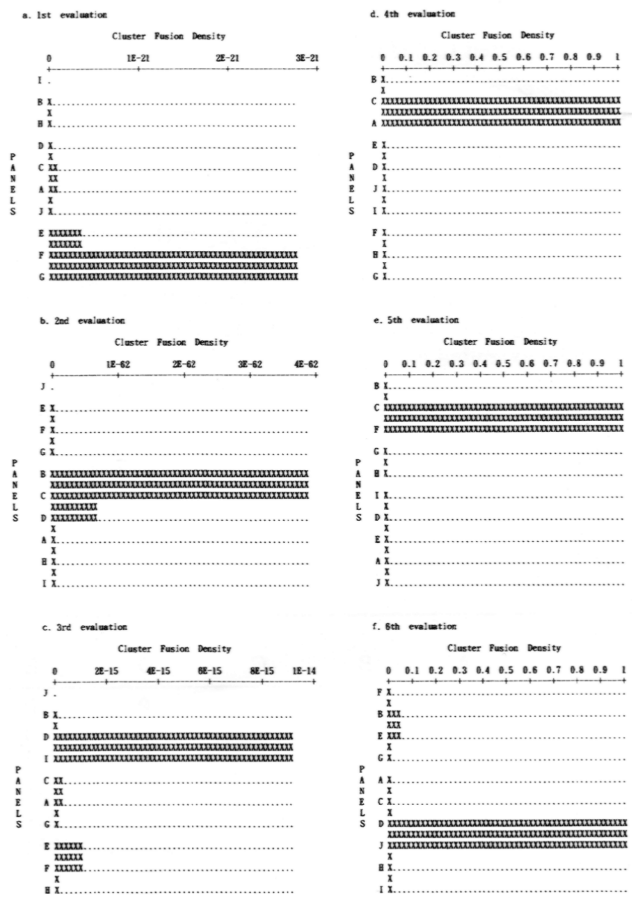

Figure 1. Cluster analysis of sensory evaluation at each evaluation time (under 2 identifiers).

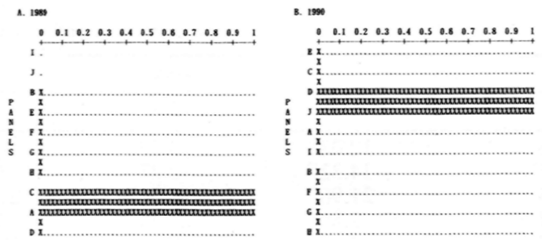

Figure 2. Cluster analysis of sensory evaluation in 1989 and 1990 (under 2 identifiers).

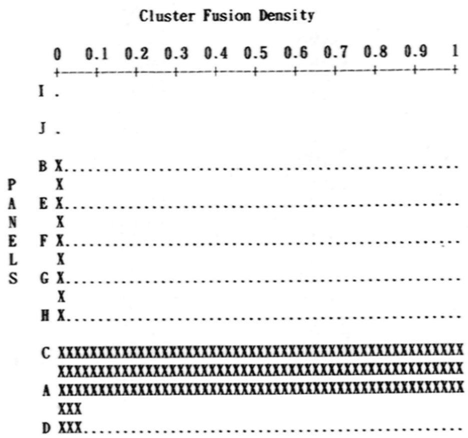

Figure 3. Cluster analysis of sensory evaluation in both years (under 2 identifiers).

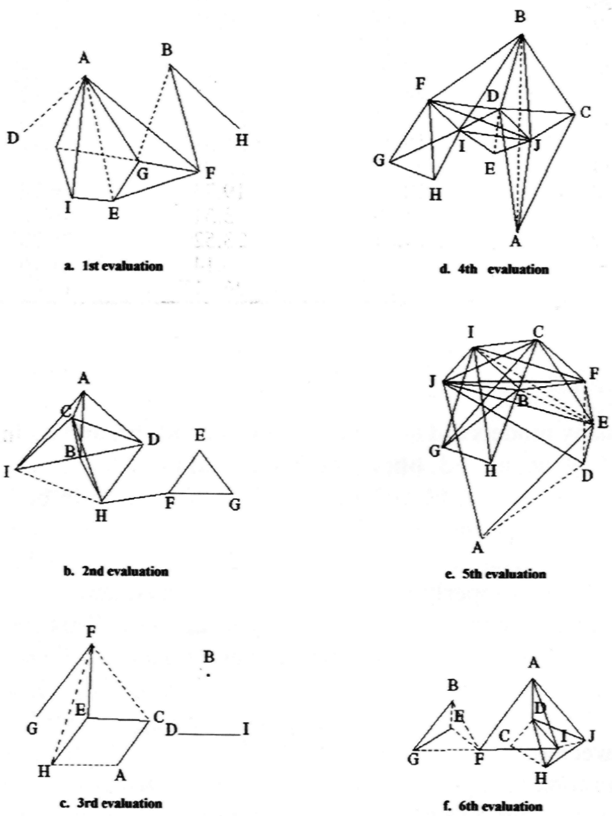

Figure 4. Kappa values among panels at each evaluation time (—, --- under 1%, 5% significant level, respectively).

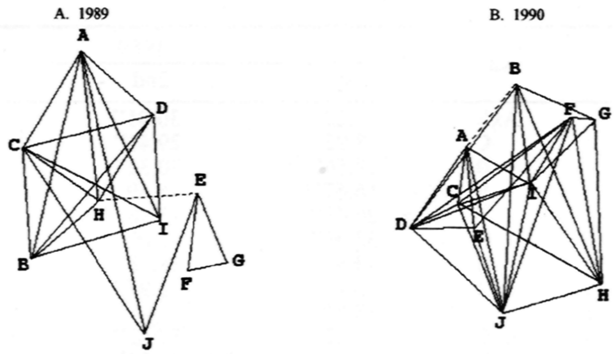

Figure 5. Kappa values among panels in 1989 and 1990 (—,--- under 1%, 5% significant level, respectively).

ation. We divided the panelist into two (**bh** and **acdefgj**) or three (**acd[j]**, **efg**, and **bh**; or **acde[j]**, **bgf**, and **h**) groups in the first evaluation period; into two groups (**abcdhi** and **efg**) in the second; into two (**bi**, **acdefgh**) or three (**b**, **di**, and **acefgh**) groups in the third; into two (**fgh** and **abcdeij**) or three groups (**acdj**, **ibe**, and **fgh**; or **abc**, **deij**, and **fgh**) in the fourth; into two (**ad** and **bcefghij**) or three groups (**ad**, **bcefij**, and **gh**; **ad**, **cbefg**, and **hij**; or **ad**, **cbef**, and **ghij**) in the fifth; into two (**bcfg** and **acdhij**) or three groups (**befg**, **aij**, and **cdh**; or **acd**, **befg**, and **hij**) in the sixth. This is a rough way to subgroup the panelists. Then, we used the multi-rater kappa method (k_d) to find further relationships among panelists having high two-rater k values.

Panelists **i** and **j** did not take part in 1989, so we did not calculate their k value for that year. The rest are described in Table 6 and Figure 5*A*, which show the k values for groups **efg** and **abcdh**. Table 6 and Figure 5*B* show that **b** and **f** have $\alpha = 0.05$ significant difference among other panelists. These values are wild, so we excluded **b** and **f** from our calculations and divided the panel into two groups, **acde** and **hg**. There were 10 panel members in 1990, and many ways to group them, e.g. **abcd**, **efij**, and **gh**; **fghij**, **ecb** and **ad**; **bghij**, **ade**, and **cf**, or **acdeij**, **bfg**, and **hi**. Because it is not easy to see the relationship among the combined panels from both years (Table 6, Figure 6), we divided them into two groups, **efgh** and **abcd**, for further discussion.

If we consider the k-values of three panelists under five categories, then there are 125 combinations, and we shall used one or two years data to analyze the degrees of agreement among panelists. Table 7 shows the top 7 k_d-values and the names of three panelists. That these values are less than 0.3 implies that the agreement among panelists is not high enough. Panel **efg** had the highest k_d value in 1989, panel **bgh** had the highest k_d value in 1990, and panel **acd** had the highest value among all three panelist groups in both years.

Establishing the model—We applied equation 9 as our model and assumed that the categories of μ_i and μ_j were 1, 2, 3, 4, and 5. We took δ as the common agreement parameter and β as the structure concordance

Table 5. Kappa value of any pairs of raters at each evaluation time.

Unit: %

	Kappa	1989 1st	1989 2nd	1989 3rd	1990 1st	1990 2nd	1990 3rd
	AxB	.22	30.01**	-5.95	12.08*	3.23	10.61
	C	38.92**	29.41**	27.97**	34.32**	2.23	3.51
	D	19.56*	30.32**	13.43	21.90**	16.54*	21.54**
	E	16.67*	-6.49	3.61	2.86	22.74**	-3.05
	F	26.41**	-3.98	16.75	2.63	7.83	27.70**
P	G	33.33**	-3.09	-2.90	-4.25	-6.73	10.35
a	H	-3.85	18.78**	24.26*	6.11	-2.33	11.89
i	I	—	8.97**	12.14	.89	3.12	41.73**
r	j	35.81**	—	—	14.38**	13.98**	22.50**
s	BxC	-1.32	37.62**	-2.43	20.34**	20.54**	-8.60
	D	2.95	38.03**	.11	25.05**	4.80	-5.71
	E	4.66	-2.69	-3.72	6.68	11.02*	26.24*
o	F	5.57**	-5.82	-8.65	15.47**	14.34**	21.26*
f	G	6.82*	-3.24	-3.03	7.37	23.12**	38.98**
	H	28.00**	31.98**	-9.04	7.94	17.78**	-3.42
	I	—	18.98**	5.56	30.53**	34.98**	1.46
	J	2.13	—	—	33.09**	26.58**	-13.39
R	CxD	12.90	24.46**	16.54	21.79**	6.01	17.25*
a	E	7.58	-4.50	21.48**	6.72	13.41**	-8.99
t	F	11.18	-2.17	25.51*	.83	27.71**	-3.18
e	G	15.50*	-.60	-2.43	-6.47	15.43**	-8.23
r	H	-2.32	17.05**	9.66	2.44	4.57	19.54*
s	I	—	20.99**	11.85	5.42	15.57**	9.38
	J	45.07**	—	—	20.36**	22.39**	11.71
	DxE	11.35	-2.69	7.61	13.07*	19.00**	-9.65
	F	6.94	-2.52	10.95	14.48**	11.48*	-.85
	G	8.89	.37	-7.22	2.09	-7.70	-1.55
	H	-3.77	36.37**	-4.67	6.47	-2.70	30.79**
	I	—	12.48**	38.97**	17.50**	1.12	29.31**
	J	14.49	—	—	35.91**	23.33**	44.16**
	ExF	31.65**	20.44**	33.01**	1.31	16.24**	22.95*
	G	27.65**	17.59**	9.74	-2.75	-2.27	30.76**
	H	.09	9.89	26.00**	-1.95	-2.31	-3.03
	I	—	-3.73	-1.49	18.23**	10.22*	-2.38
	J	19.21**	—	—	37.71**	31.92**	-4.35
	FxG	47.92**	41.94**	26.44**	15.70**	2.10	21.26*
	H	-.64	11.60**	26.26*	28.21**	-1.33	10.12
	I	—	-5.44	3.85	17.57**	22.82**	29.36**
	J	5.80	—	—	16.95**	35.34**	7.42
	GxH	7.22	8.67	19.34*	20.62**	33.06**	3.83
	I	—	-1.11	-5.82	12.82**	19.73**	6.78
	J	11.64	—	—	1.80	3.51	-6.76
	HxI	—	8.60*	-5.19	19.64**	23.52**	28.28**
	J	7.28	—	—	9.37	-.14	21.16**
	IxJ	—	—	—	31.29**	26.71**	17.85*

— : did not attained the evaluation.
* : under 5% significant level; ** : under 1% significant level.

caused by the structure of the contingency table. We assumed 0.01 for missing cells. Table 8 shows the common agreement between two panelists from each evaluation period. At a significance of $\alpha = 0.05$, panels **ag** and **fg** had common agreement in the 1st evaluation; panels **ce**, **dh**, and **fg** in the 2nd; panel **fg** in the 3rd; panels **bi** and **ej** in the 4th; and panels **fj** and **hi** in the 5th. None of panelists agreed in the 6th evaluation; i.e. all agreement between panelists were caused by structure concordance. Analyzed by year (Table 9), and adjusted by 0.1 for missing cells, panels **ae**, **be**, **bh**, **cj**, **dh**, **fg**, and **fh** had high agreement (δ) in 1989; panels **bi**, **ei**, **ej**, and **fj** in 1990; panels **bc** and **eh** in both years, with p-value below 0.05. The analysis showed that many models did not fit well—**ab**, **bc**, **bd**, **bf**, and **bg** in 1989, **ac**, **ag**, **bd**, **bh**, **bj**, **ce**, **eh**, and **hi** in 1990, and **ab**, **ac**, **ae**, **bd**, **bf**, **bh**, **bj**, **ce**, **dg**, **ei**, **fh**, **fi**, and **hi** in both years. This implies that the panels did not have a common structure coefficient (β) or that the grades μ_i and μ_j were not properly assigned. We modified equation 9 as β_j (no common β) and reanalyzed the data. This improved some panels' agreement, but the data still did not fit our model well. We recast the δ value as δ_i (μ_i, μ_j = 1,2,3,4,5; Table 10) and reanalyzed the agreement between panelists. In data from 1989, panels **a** and **b** show no common agreement among each grade. Because some of the panel agreements do not fit the model (9) well, we reranked the grades (μ_i) and reanalyzed the data. With μ_i

Table 6. Kappa value of two raters at various years.

Unit: %

Rater	Year 1989	1990	Both years
AxB	14.59**	7.52*	10.92**
C	29.24**	11.67**	19.19**
D	23.82**	21.18**	22.55**
E	-1.11	13.70**	7.30**
F	4.69	9.12	6.68**
G	4.18	-4.47	-0.22
H	14.00**	.38	6.64**
I	17.47**	7.24**	11.31**
J	35.81**	15.56**	18.57**
BxC	21.21**	13.86**	17.69**
D	24.07**	8.72**	16.29**
E	.48	12.37**	5.48*
F	-15.67	14.29**	-.78
G	-7.69	19.43**	6.30*
H	33.88**	11.79**	22.99**
I	18.27**	28.40**	26.24**
J	2.13	21.25**	18.07**
CxD	22.56**	14.59**	18.23**
E	4.24	4.72	2.87
F	4.04	10.90*	7.43**
G	2.65	4.28	3.67
H	10.86**	9.87**	10.71**
I	19.55**	13.60**	16.58**
J	45.07**	24.95**	29.06**
DxE	4.32	12.97**	8.48**
F	1.45	11.21**	6.32**
G	.99	-3.70	-1.22
H	19.51**	4.77	11.79**
I	18.29**	11.29**	14.23**
J	14.49	32.43**	29.47**
ExF	25.33**	11.13**	16.93**
G	19.10**	1.53	9.35**
H	9.76*	-7.33	.11
I	-2.63	10.42**	2.08
J	19.21**	28.46**	26.35**
FxG	43.73**	9.82**	24.41**
H	4.58	9.60**	5.53*
I	-3.14	22.06**	10.22**
J	5.80	23.99**	20.38**
GxH	6.06	23.88**	17.51**
I	-1.96	15.82**	8.77**
J	11.64	1.97	3.16
HxI	6.34	25.81**	20.06**
J	7.28	7.70**	7.52**
IxJ	—	28.13*	28.13**

— : missing.
* and ** : under 5% and 1% significant level, respectively.

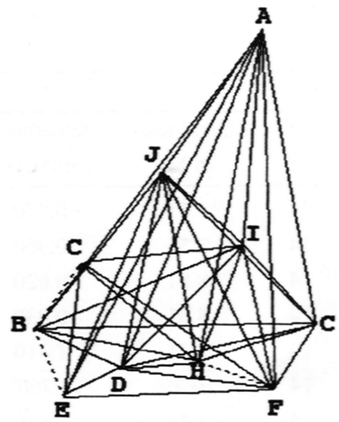

Figure 6. Kappa values among panels in both years (—, --- under 1%, 5% significant level, respectively).

Table 7. Kappa value of three raters.

1989 Raters	Kappa value	1990 Raters	Kappa value	Both years Raters	Kappa value
EFG	0.2890	BGH	0.1713	ACD	0.1944
BDH	0.2467	ADE	0.1524	BCD	0.1579
ACD	0.2462	ACD	0.1468	EFG	0.1519
BCD	0.2090	BFG	0.1236	BCH	0.1517
BCH	0.2057	BCF	0.1151	BGH	0.1512
ABC	0.1939	FGH	0.1120	BDH	0.1510
ABH	0.1936	BEF	0.1064	FGH	0.1477

= 9, 5, 3, 2, 1 or μ_i = 5, 1, 5, 4.5, 4, 1, models of panels **ei**, **bj**, and **ce** show a proper fit, and the agreement between panels **b** and **j** and the agreement between panels **e** and **i** have a significant difference with $\alpha = 0.05$ for the 1990 data. We rearanged the grade (μ_i = 10, 7, 4, 3, 1) and reanalyzed the models of panels. Panels **b** and **h** show high common agreement (δ) and have p-values below 0.01. Given grades μ_i = 1, 2, 3, 4, 5 and μ_j = 5.5, 5, 4, 3, 2, and with a different β_i, models of panels **bj**, **ce**, and **ei** fit well and panel **bj** and panel **ce** show high common agreement (δ) in both years' data.

We used different β_i and δ_j to reanalyze the agreement between panels **e** and **i**. From the analysis with μ_i, μ_j = 1, 2, 3, 4, 5, the agreement between panels **e** and panel **i** does not have significant difference. Because some of the models did not fit well, we rescaled the interval length among grades, which showed that **ei**, **bj**, and **ce** fit well with μ_i = 9, 5, 3, 2, 1 and 5.1, 5, 4.5, 4, 1; **bj** and **ei** had α = 0.05 significant difference, i.e. high agreement, in 1990. With the grades rearranged as μ_i= 10, 7, 4, 3, 1, panels **ab**, **ac**, **ae**, **bd**, **bf**, **bh**, **bj**, **ce**, **dg**, **ei**, **fh**, **fi**, and **hi**, **bh** had high common agreement (δ) and a p-value below 0.01; with β_i, we reassigned the grades μ_i = 1, 2, 3, 4, 5 and μ_j= 5.5, 5, 4, 3, 2. **bj**, **ce**, **ei** fit well, and **bj**, **ce** also had high common δ in both years. Because the grades of μ_i and μ_j for panel **ei** is not proper, we do different β_i's and δ_i's to reanalyze the data to get a better fit.

Discussion

Cluster analysis is affected by the assumption for the number of nearest groups, and the analysis will be changed by the nearest identifiers which you define, see Sarle (1983). Note that we have to decide how many groups

Table 8. Log-linear model in each evaluation time.

Unit: G^2-value

Panel	1989-1			1989-2			1989-3		
	Df	Goodness of Fit	Agreement parameters	Df	Goodness of Fit	Agreement parameters	Df	Goodness of Fit	Agreement parameters
AxB	14	9.15	0.070	14	14.16	0.010	14	1.15	1.020
C	14	3.55	0.060	14	9.96	0.260	14	7.64	1.160
D	14	8.81	0.020	14	6.60	0.330	14	2.63	0.830
E	14	5.31	0.030	14	2.46	0.610	14	3.94	2.260
F	14	2.31	1.410	14	12.04	0.130	14	6.45	0.040
G	14	10.79	5.760*	14	9.14	0.030	14	3.75	0.710
H	14	8.46	1.310	14	5.19	0.030	14	8.78	0.240
I	—	—	—	14	4.66	0.190	14	2.07	1.160
J	14	8.49	1.000	—	—	—	—	—	—
BxC	14	4.07	1.500	14	5.16	0.820	14	1.98	0.000
D	14	9.78	1.240	14	7.41	0.360	14	2.94	0.260
E	14	3.02	1.030	14	6.76	2.980	14	3.19	0.510
F	14	11.81	0.130	14	15.90	0.020	14	1.62	0.540
G	14	7.14	0.180	14	10.22	0.460	14	3.89	0.800
H	14	10.29	2.170	14	9.80	0.710	14	3.94	3.390
I	—	—	—	14	9.38	0.180	14	4.49	0.000
J	14	5.37	0.170	—	—	—	—	—	—
CxD	14	6.64	0.720	14	3.63	1.240	14	6.20	0.020
E	14	2.20	0.810	14	3.91	5.070*	14	3.70	1.070
F	14	3.64	0.040	14	3.92	0.180	14	1.80	0.470
G	14	6.38	0.170	14	7.83	0.120	14	8.96	0.000
H	14	9.06	2.010	14	2.69	1.610	14	5.61	0.080
I	—	—	—	14	3.77	1.610	14	1.88	0.120
J	14	4.56	1.800	—	—	—	—	—	—
DxE	14	8.48	2.350	14	8.06	3.750	14	1.83	0.020
F	14	6.83	0.810	14	5.47	0.110	14	5.54	0.280
G	14	6.52	1.500	14	5.36	0.600	14	3.53	0.000
H	14	10.82	1.120	14	10.46	4.670*	14	8.47	3.770
I	—	—	—	14	13.31	0.850	14	9.33	0.020
J	4	8.32	0.200	—	—	—	—	—	—
ExF	14	1.70	0.070	14	1.19	0.760	14	2.30	0.420
G	14	7.41	0.570	14	2.61	0.000	14	7.61	0.320
H	14	6.56	0.910	14	3.88	0.630	14	6.22	0.050
I	—	—	—	14	7.99	1.500	14	7.43	0.770
J	14	3.07	0.100	—	—	—	—	—	—
FxG	14	1.79	4.670*	14	8.08	3.900*	14	1.37	3.850*
H	14	6.51	0.900	14	11.64	1.670	14	10.05	0.220
I	—	—	—	14	5.14	2.340	14	7.93	0.410
J	14	5.18	0.100	—	—	—	—	—	—
GxH	14	6.81	0.590	14	9.91	0.080	14	8.18	1.690
I	—	—	—	14	11.41	0.010	14	4.72	0.010
J	14	8.62	0.020	—	—	—	—	—	—
HxI	—	—	—	14	14.27	0.520	14	7.13	0.530
J	14	8.29	0.320	—	—	—	—	—	—
IxJ	—	—	—	—	—	—	—	—	—

Table 8. Log-linear model in each evaluation time (Cont.)

Unit: G^2-value

Panel	1990-1 Df	Goodness of Fit	Agreement parameters	1990-2 Df	Goodness of Fit	Agreement parameters	1990-3 Df	Goodness of Fit	Agreement parameters
AxB	14	11.59	0.010	14	11.36	0.100	14	3.90	0.060
C	14	1.92	2.100	14	1.52	0.170	14	3.09	0.050
D	14	14.00	0.530	14	2.50	0.590	14	8.20	0.250
E	14	10.10	0.710	14	4.52	0.440	14	3.18	3.210
F	14	6.20	0.010	14	3.54	0.470	14	9.86	1.100
G	14	17.28	0.880	14	5.35	0.960	14	18.63	0.120
H	14	6.33	0.020	14	4.36	0.030	14	6.40	0.490
I	14	10.30	1.160	14	8.94	0.020	14	8.42	1.830
J	14	18.21	0.240	14	2.54	0.290	14	3.10	0.000
BxC	14	7.29	1.770	14	2.73	1.540	14	2.82	0.010
D	14	18.92	1.320	14	5.72	0.200	14	7.81	0.000
E	14	21.96	0.900	14	12.27	0.110	14	6.19	0.570
F	14	18.89	0.010	14	9.80	1.690	14	1.74	0.070
G	14	10.93	0.940	14	9.17	0.930	14	3.17	1.270
H	14	7.45	0.550	14	11.95	0.670	14	9.97	0.960
I	14	13.02	5.970*	14	19.14	1.660	14	3.94	0.470
J	14	22.28	3.710	14	16.67	2.490	14	2.41	0.270
CxD	14	5.04	0.000	14	4.94	0.030	14	5.61	0.380
E	14	11.18	1.790	14	2.32	0.330	14	8.75	0.100
F	14	12.01	0.000	14	2.95	1.400	14	5.05	1.470
G	14	8.41	0.080	14	3.97	2.500	14	5.03	2.020
H	14	7.19	0.050	14	7.96	0.050	14	5.93	0.300
I	14	6.39	0.840	14	8.01	1.500	14	5.76	0.160
J	14	13.96	0.050	14	7.20	1.190	14	3.28	0.070
DxE	14	5.56	0.730	14	7.82	0.180	14	8.74	2.030
F	14	4.95	0.080	14	5.10	0.040	14	8.40	2.130
G	14	9.68	0.030	14	2.58	0.150	14	7.88	0.040
H	14	12.62	3.120	14	4.77	0.010	14	2.48	0.030
I	14	10.84	0.260	14	13.86	3.430	14	13.33	0.010
J	14	12.49	0.310	14	8.17	0.980	14	2.16	0.070
ExF	14	6.80	0.380	14	11.41	0.080	14	7.64	0.030
G	14	16.98	0.190	14	9.46	0.020	14	2.62	3.130
H	14	11.07	2.360	14	4.79	0.520	14	4.08	1.110
I	14	13.91	0.000	14	7.85	0.420	14	4.67	2.260
J	14	9.40	5.190*	14	14.88	0.390	14	13.51	0.480
FxG	14	13.53	0.010	14	9.16	1.490	14	6.19	0.020
H	14	9.29	0.810	14	10.30	3.690	14	13.05	0.740
I	14	3.82	0.180	14	7.45	0.950	14	8.63	1.270
J	14	4.48	2.390	14	10.42	5.250*	14	2.56	0.430
GxH	14	11.55	0.140	14	5.36	0.380	14	16.25	0.130
I	14	5.19	2.500	14	9.85	1.490	14	12.78	0.120
J	14	4.78	0.480	14	9.26	1.200	14	4.95	0.650
HxI	14	13.83	0.670	14	3.76	7.530**	14	16.12	0.800
J	14	6.27	0.270	14	9.46	0.150	14	7.06	0.330
IxJ	14	10.53	0.010	14	15.60	0.030	14	4.99	0.460

—: missing.

*,**: under 5% and 1 % signifiacnt level, espectively.

Table 9. Table of log-linear model in each and both years.

Unit: G^2-value

Panel	1989			1990			Both Years		
	Df	Goodness of Fit	Agreement parameters	Df	Goodness of Fit	Agreement parameters	Df	Goodness of Fit	Agreement parameters
AxB	14	25.04*	3.550	14	18.42	0.000	14	24.46*	1.730
C	14	14.73	0.010	14	43.79**	0.370	14	41.21**	0.270
D	14	10.77	0.320	14	18.97	0.720	14	13.87	0.740
E	14	12.64	4.670*	14	17.93	0.010	14	24.75*	1.760
F	14	19.47	0.560	14	9.78	0.600	14	21.72	0.700
G	14	8.05	0.910	14	28.36*	5.940*	14	22.04	0.180
H	14	7.30	0.200	14	16.03	3.160	14	15.84	0.070
I	14	6.71	0.140	14	11.78	0.510	14	8.20	0.110
J	14	8.00	2.340	14	21.83	0.100	14	16.16	0.140
BxC	14	30.29**	12.100**	14	16.67	1.930	14	17.57	11.380**
D	14	46.03**	21.930**	14	24.81*	0.400	14	27.66*	7.860**
E	14	21.54	5.240*	14	23.27	0.320	14	9.81	3.170
F	14	54.80**	48.900**	14	21.18	1.720	14	32.24**	30.860**
G	14	26.40*	13.420**	14	16.60	2.490	14	21.46	1.450
H	14	22.85	17.820**	14	25.90*	0.000	14	31.74**	9.200**
I	14	19.17	0.330	14	16.64	3.910*	14	19.66	1.540
J	14	7.09	0.280	14	36.80**	1.580	14	24.35*	5.250*
CxD	14	13.20	0.580	14	17.88	1.400	14	17.26	2.350
E	14	14.06	2.620	14	51.02**	0.740	14	27.71*	6.980**
F	14	14.47	0.150	14	18.82	0.440	14	14.71	0.090
G	14	12.11	0.040	14	13.48	3.110	14	16.43	1.980
H	14	9.38	0.060	14	12.81	0.310	14	4.56	0.540
I	14	8.19	1.290	14	12.45	0.070	14	19.54	0.480
J	14	7.58	4.270*	14	13.25	0.060	14	10.38	1.020
DxE	14	11.44	0.230	14	9.80	0.010	14	11.70	1.760
F	14	20.50	0.000	14	7.57	0.050	14	15.70	0.050
G	14	19.33	0.730	14	15.49	0.710	14	25.60*	0.670
H	14	19.88	8.590**	14	7.46	1.760	14	13.14	2.240
I	14	12.70	0.070	14	23.27	2.450	14	14.81	2.210
J	14	9.75	0.010	14	11.96	0.440	14	15.96	0.280
ExF	14	9.94	0.220	14	10.78	1.080	14	14.31	2.050
G	14	13.75	1.800	14	14.39	0.060	14	13.73	3.270
H	14	10.55	0.410	14	28.31*	16.980**	14	18.16	4.760*
I	14	12.32	2.510	14	20.07	5.600*	14	30.50**	12.200**
J	14	6.71	0.220	14	12.10	5.350*	14	8.69	3.570
FxG	14	4.54	17.530**	14	17.30	1.070	14	14.19	2.880
H	14	19.22	8.790**	14	18.32	1.230	14	24.26*	14.510**
I	14	12.37	1.150	14	10.64	1.960	14	24.92*	0.130
J	14	8.11	0.110	14	8.21	6.470*	14	10.47	3.100
GxH	14	12.16	3.670	14	19.27	1.350	14	12.28	0.000
I	14	10.35	0.090	14	19.77	2.980	14	16.31	2.160
J	14	8.94	0.080	14	17.15	0.480	14	14.15	0.000
HxI	14	18.08	2.470	14	35.20**	6.520*	14	36.70**	0.980
J	14	7.55	0.080	14	8.32	1.130	14	10.06	0.570
IxJ	—	—	—	14	20.10	0.030	14	20.10	0.030

—: missing.

*,**: under 5% and 1% significant level, respectively.

Table 10. Table of log-linear model after adjusted in each and both years.

Unit: G^2-value

Panel	1989			1990			Both Years		
	Df	Goodness of Fit	Agreement parameters	Df	Goodness of Fit	Agreement parameters	Df	Goodness of Fit	Agreement parameters
AxB	14	17.61	==	14	18.42	0.000	11	18.26	1.570
C	14	14.73	0.010	11	10.04	1.330	11	9.13	0.200
D	14	10.77	0.320	14	18.97	0.720	14	13.87	0.740
E	14	12.64	4.670*	14	17.93	0.010	11	11.88	3.740
F	14	19.47	0.560	14	9.78	0.600	14	21.72	0.700
G	14	8.05	0.910	11	16.37	0.790	14	22.04	0.180
H	14	7.30	0.200	14	16.03	3.160	14	15.84	0.070
I	14	6.71	0.140	14	11.78	0.510	14	8.20	0.110
J	14	8.00	2.340	14	21.83	0.100	14	16.16	0.140
BxC	11	10.00	1.040	14	16.67	1.930	14	17.57	11.380**
D	11	14.98	4.000*	11	19.10	0.020	11	17.30	4.320*
E	11	9.19	0.880	14	23.27	0.320	14	9.8	3.170
F	11	9.78	8.670**	14	21.18	1.720	11	18.16	10.620**
G	11	4.67	0.890	14	16.60	2.490	14	21.46	1.450
H	14	22.85	17.820**	11	19.26	1.010	14	21.59	6.690**
I	14	19.17	0.330	11	11.77	2.020	14	19.66	1.540
J	14	7.09	0.280	11	18.21	5.490*	11	19.37	6.000*
CxD	14	13.20	0.580	14	17.88	1.400	14	17.26	2.350
E	14	14.06	2.620	11	24.00	0.740	11	17.19	11.020**
F	14	14.47	0.150	14	18.82	0.440	14	14.71	0.090
G	14	12.11	0.040	14	13.48	3.110	14	16.43	1.980
H	14	9.38	0.060	14	12.81	0.310	14	4.56	0.540
I	14	8.19	1.290	14	12.45	0.070	14	19.54	0.480
J	14	7.58	4.270*	14	13.25	0.060	14	10.38	1.020
DxE	14	11.44	0.230	14	9.80	0.010	14	11.70	1.760
F	14	20.50	0.000	14	7.57	0.050	14	15.70	0.050
G	14	19.33	0.730	14	15.49	0.710	11	18.06	1.460
H	14	19.88	8.590	14	7.46	1.760	14	13.14	2.240
I	14	12.70	0.070	14	23.27	2.450	14	14.81	2.210
J	14	9.75	0.010	14	11.96	0.440	14	15.96	0.280
ExF	14	9.94	0.220	14	10.78	1.080	14	14.31	2.050
G	14	13.75	1.800	14	14.39	0.060	14	13.73	3.270
H	14	10.55*	0.410	11	12.52	13.500**	14	18.16	4.760*
I	14	12.32	2.510	14	20.07	5.600*	8	14.95	==
J	14	6.71	0.220	14	12.10	5.350*	14	8.69	3.570
FxG	14	4.54	17.530**	14	17.30	1.070	14	14.19	2.880
H	14	19.22	8.790**	14	18.32	1.230	11	15.81	8.740**
I	14	12.37	1.150	14	10.64	1.960	11	9.38	0.100
J	14	8.11	0.110	14	8.21	6.470*	14	10.47	3.100
GxH	14	12.16	3.670	14	19.27	1.350	14	12.28	0.000
I	14	10.35	0.090	14	19.77	2.980	14	16.31	2.160
J	14	8.94	0.080	14	17.15	0.480	14	14.15	0.000
HxI	14	18.08	2.470	11	12.31	6.590*	11	18.90	0.930
J	14	7.55	0.080	14	8.32	1.130	14	10.06	0.570
IxJ	—	—	—	14	20.10	0.030	14	20.10	0.030

—: mising; ==: no unique parameter.

*,**: under 5% and 1% significant level, respectively.

needed before to execute the analysis, and have to try different identifiers to know the best relationship among panelists (Table 4). Two-stage density linkage cluster analysis gives us an easy and fast way to find the relationships among panel members. The cluster fusion density (Figure 1) provides more detail about the agreement among panelists, but it can not show the latent differences among the panels. We apply the *k* coefficient and model method to reanalyze the agreement between two panelists. Based on two panelists high *k*-value, we choose panel **efg** and **acd** with high agreement. From the future multi-rater k_d analysis, panel **acd** do show the highest value of all these panelists comparison from combined two years evaluation data. The k_d value is not over 0.5, which implies that the degrees of agreement is not high enough, (Landis and Koch, 1977). This phenomenon could be caused by the different taste habit among panelists, the unknown influential factors of panelists, the different levels of testing order, the choosing items, or other factors we have not found yet.

From the analysis model (9), we see that most of the panelists with high *k*-values are caused by the structure factors, i.e. $\beta\mu_i\mu_j$ in the model (9). Comparing with the kappa method, the modeling method shows the different results of panelists' agreement. Table 9 and 10 show that some of models are fitted not very well, because of many sparse data appear in the cells.

As we compare the above three methods, cluster analysis gives us a rough view of panelists' opinion; the *k* method provides by summary information about agreement among panelists, but does not consider the factors influenced the different panel rating; however, 'model establishment' informs us that the influential factors of agreement are a latent rating category value (μ_i), and the different structure (β_j) and agreement (δ_i) parameters under different grades. However, in the future we may try different grades to choose the possible high agreement panels in the modeling method.

Acknowledgment. This work was supported partly by grant NSC 81-0211-B-001-017 from the National Science Council of the Republic of China. Also, we like to thank for Mr. Yang, Sheng-Shium in TTES gave us a lot of help during the experimental period.

Literature Cited

Agresti, A. 1988. A model for agreement between ratings on an ordinal scale. Biometrics **44:** 539–548.

Cohan, J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20:** 37–46.

Conger, A. J. 1980. Integration and generalization of kappas of multiple raters. Psychol. Bull. **88:** 322–328.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. Psychol. Bull. **76:** 378–382.

Fleiss, J. L., J. Cohen, and B. S. Everitt. 1969. Large sample standard errors of kappa and weighted kappa. Psychol. Bull. **72:** 323–327.

Landis, J. R. and G. G. Koch 1977. The measurement of observer agreement for categorical data. Biometrics **33:** 159–174.

Sarle, W. S. 1983. Cubic clustering criterion. SAS Technical Report A-108, Cary, NC. SAS Institute Inc.

SAS Institute Inc. 1988. SAS/STAT 6.03 User's Guide, Release 6.03.

SAS Institute Inc. 1988. SAS/IML 6.03 User's Guide, Release 6.03.

Tanner, M. A. and M. A. Young. 1985. Modeling agreement among raters. J. Amer. Statist. Assoc. **80:** 175–180.

Wong, M. A. and T. Lane. 1983. A *k*th nearest neighbor clustering procedure. J. Roy. Statist. Soc., Series B. **45:** 362–368.

品質的官能檢定研究— 品評人員的同意程度探討

鄔宏潘[1] 陳立信[2]

[1]中央研究院植物研究所
[2]銘傳管理學院統計學系

本文乃就食品品質經人爲的官能檢定後，各品評員間對品評事項評分結果之一致性進行統計方法之研究。在醫學及心理學上經常應用群集分析，*k* 係數及對數線性相關模式進行評分員看法一致性之分析。對於品評員官能反應變化的一致性則鮮少探討。今將應用以上三種統計方法針對茶葉官能品評結果進行分析比較。本研究茶葉評分等第，由 1 品質最差依序漸進至等第 11 品質最佳。由 8 至 10 受過相當訓練之品評員參與，並依評鑑結果，重新劃分成五等級，按五等第進行相關研究。經由以上分析，可知目前茶葉品評員間之看法一致性均不高。經由群集分析，將看法相近之品評員初步區分成若干群集，同群間表示看法相近，不同群間表示看法較不一致，但當群集間群集密度不高時則不易明顯劃分；由 *k* 係數分析後，可知各品評員對共同事項看法一致性的變化關係，但未考慮各分級及其他事項之分布影響；應用模式可以估計各等第的給定值及在每個等級中一致性的變化。

關鍵詞: 品質官能檢定; 品評員; 群集分析; kappa(*k*) 係數; 對數線性相關模式。