12th IMEKO TC1 & TC7 Joint Symposium on
Man Science & Measurement
September, 3 – 5, 2008, Annecy, France

CONTROL OF VARIABILITY FOR MAN MEASUREMENT

Guerra Anne-Sophie [1], *Maurice Pillet* [1], *Jean-Luc Maire* [1]

[1] Laboratoire de Systèmes et Matériaux pour la Mécatronique (SYMME)
Université de Savoie – Polytech'Savoie - BP 806 - 74016 Annecy cedex , France
anne-sophie.guerra@univ-savoie.fr – maurice.pillet@univ-savoie.fr – jean-luc.maire@univ-savoie.fr

Abstract: In a visual control, the measurement system is human. Each appraiser must control the part and judge the conformity. Consequently, the subjectivity of measure is very present and the variability therefore high.
The subjectivity of this inspection is generated by two aspects: the exploration i.e. the capacity to perceive the defect by the appraiser and the evaluation i.e. the capacity to evaluate correctly the defect. So, after to realize the exploration and after to have made the evaluation, the appraiser could with more of facilities to give a reponse and so a judgement on the acceptability of the defect. The inspection must be separated into two operations: the exploration and the evaluation.

Thus, in this text, we propose a method to measure the variability of inspections and to know the causes: the R²E² test.

Keywords: variability, measure, subjective evaluation.

1. CONTEXT OF THE STUDY

The works which will be described in the paper are realized within the framework of collaboration between the laboratory SYMME of the University of Savoy and the famous watch-making factory situated in Switzerland. The objective of this collaboration is to formalize a process of visual inspection of these products.
One of the peculiarities of this process indeed is to have to lean on knowledge very subjective with a treatment which requires a very specific expertise. In fact, the visual inspection is composed of lot of step that the appraisers realize without Without being conscious of it and without formalized process...
This visual inspection of products involves a lot of problems. The most mattering of them is the one of the important variation of the results of inspection, leading decisions on the conformity of the product sometimes totally opposed. Without process, the variation of the results given by appraisers is very important and shows a veritable problem to guarantee a product conform at the end of the process of fabrication.
The principal difficulty is that Human is the only measuring instrument which can be used for this type of inspection. The various sensations which perceive this human for the controlled product are at the origin of the big variability regularly observed on the made measures. Furthermore, the defects of aspects are, among all the defects to be detected during a inspection, the least easy to be identified, to be described and to be interpreted to judge the conformity of product.
A part of our works permit to formalize the expertise associated with a visual control by proposing of new approaches that must contribute to describe better the defects, to define better the level of refusal of a product and to make reliability better of the results of the visual inspections of products. These approaches were developed on the various methods recommended by the sensory analysis. We define a formalized demarche to describe each step of process of inspection in order to improve the results of appraisers. This demarche is composed in 3 steps:

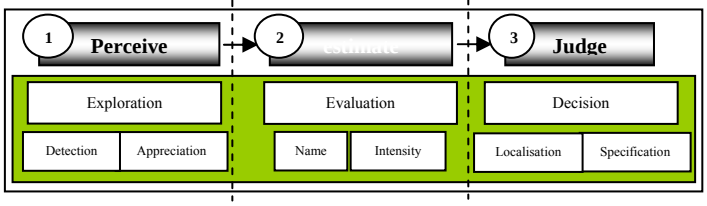

Figure 1 : 3 steps = 2 expertises for a inspection + 1 decision

To realize a inspection, the appraiser must:

- perceive the defect
- estimate the defect
- judge the defect

These 3 steps are very important and the respect of order is capital. Each step has got a particularity and an expertise. The expertise of exploration is trainning the detection and the appreciation of defect. So, it is the first step of control and without it, the rest of control is not possible.
The expertsie of evaluation depend on the first step because the appraiser gives the name and the intensity of defect. These both elements are chosen in the proposed lists and obtained by using the concept of sensory analysis.
The last step is the final decision: *Accepted* or *Refused* the defect. This decision could be given by the localisation of defect and the specification of product. These two elements are too in the proposed lists.
The new concept of control is exposed very rapidely in the article but [1] explain more in detail the demarche to build

the concept of control and the concept itself thank to the sensory analysis.

2. STUDY OF VARIABILITY OF RESULTS

As we indicated it in the context of our study, the visual inspection leads an important variability in the acceptability or not of the product. This variability can be taken into account by a R&R test very widely developed in the literature [2]. Our target is not to enter in the description of R&R test like it definie the literature, but to use this method to measure the variability of control and to look for the cause of this last one.

2.1. Application of R&R test

According to the recommendations of [3], we thus asked to 4 persons to estimate the acceptability of 30 products twice - being separated from week. So, each appraise controled product after product and in the first time given his judgement on the acceptability of parts. The reason of this judgement is too demanded but in the R&R test, it doesn't take in consideration. We obtained the results proposed in the table 1.

Part No	Just Value	Appraiser 1		Appraiser 2		Appraiser 3		Appraiser 4	
		1	2	1	2	1	2	1	2
1	C	C	C	NC	NC	NC	NC	NC	NC
2	C	C	NC	NC	C	C	C	NC	NC
3	C	C	NC	C	NC	C	NC	C	C
4	C	C	NC	C	C	C	C	C	NC
5	C	C	NC	NC	NC	NC	C	NC	NC
6	C	C	NC	C	C	NC	C	C	C
7	C	C	C	C	NC	C	NC	C	NC
8	C	C	C	NC	NC	C	C	NC	NC
9	C	C	NC	C	C	NC	C	C	C
10	C	C	C	NC	C	C	C	C	C
11	C	C	C	C	C	NC	C	C	C
12	NC	NC	NC	C	C	NC	C	NC	C
13	C	C	C	C	C	C	C	C	NC
14	C	C	C	NC	NC	C	C	NC	NC
15	C	NC	NC	C	C	C	C	C	C
16	C	C	C	NC	NC	C	C	C	NC
17	C	NC	NC	NC	C	C	C	NC	NC
18	C	C	C	C	C	C	C	NC	NC
19	C	NC	NC	NC	NC	C	C	NC	NC
20	C	NC	NC	C	NC	NC	NC	NC	NC
21	C	C	NC	NC	NC	NC	C	NC	NC
22	C	C	C	C	C	C	C	NC	NC
23	C	NC	NC	C	C	C	C	NC	NC
24	C	C	C	C	C	NC	C	C	C
25	C	C	C	NC	NC	NC	C	NC	NC
26	C	C	NC	C	C	NC	C	NC	NC
27	NC	C	C	NC	NC	NC	C	C	C
28	C	C	NC	NC	NC	C	C	C	NC
29	C	NC	NC	C	C	C	C	NC	NC
30	C	NC	NC	C	C	C	NC	NC	NC

* C: Conform – NC: Non-Conform

Table 1. Results of R&R test with 4 appraisers

The table 1 shows two essential points during a R&R test:
- The repeatability is "the variation in measurements obtained with one measurement instrument when used several times by an appraiser while measuring the identical characteristic on the same part" [2].
- The reproducibility is "the variation in the average of the measurements made by different appraisers using the same measuring instrument when measuring the identical characteristic on the same part" [2].

The table 1 gives a just value. This just value is a value of referance given by the experts of the demarche of inspection. These experts are chosed in using the recommendations of sensory analysis and so in agreement with the definition of [4] : *"a person who, through knowledge or experience, has to give an opinion in the fields about which helshe is consulted"*.

This just value is for us the value which must be comparing at other. Each apparaise give this one value and in the same conditions of inspection.

We see thus the problem building with these results: there is a lack of repeatability and reproducibility in the judgment of 4 judges. For example, the appraiser 1 give the product 12, 15, 17, 19, 20, 29 and 30 as *Non Conform* for the first time and for the second time, the products 2, 3, 4, 5, 6, 9, 12, 15, 17, 19, 20, 21, 23, 26, 28, 29 and 30 claimed as *Non Conform.* So, the first time, 7 parts are *Non Conform* and the second time, 17 are *Non Conform*. The problem of repeatability is very important beetwen these two reponses. And in comparing with the results of experts, we can too conclue at a problem of repoductibility between the results of the appraiser 1 and the just value. In fact, the just value claims like *Non Conform* the products 12 and 27 so only two parts are *Non conform*. We note the appraiser 1 don't give the product 27 like *Non Conform* whereas this product was the list of *Non Conform* of the just value.

In looking at the results of other appraisers, the problems of repeatability and reproductibility are too important.

As conclure on this R&R test, we can say that the variability of the answers is thus very important.

2.2. Limits of the R&R test of context

Results obtained and presented on the table 1, we show that the inspection made by the persons is very subjective. Indeed, the measure made by the persons establish on one hand on their perception of the defect and on the other hand on their personal feeling in front of defects. Indeed, it is necessary to note two important steps to realize a visual inspection:

- To perceive the defect
- To estimate the defect.

These two steps can be seen as two expertises, indeed they do not go the one without the other one to control. We shall thus call these two expertises: the expertise of exploration and the expertise of evaluation.

So, and it is the essential point in this notion of limits of the R&R test, from the obtained results on the table 1 how to know the reason of this variability. The variability can come

either from the expertise asked by the perception of the defect, or by that asked by its evaluation.

This R&R test is founded on the ability for the appraisers to estimate the conformity of the product. Sensory analysis [5] [6] propose an other way. It is possible to estimate the descriptor of the defect and the intensity of the perception in a numeric scale. The figure 1 of this article shows this all new demarch in the concept of visual inspection. Without enter in the details of this demarch of sensory analysis, we can say that this method permits to separate the two notions of inspection: the exploration and the evaluation.

Expertise of exploration

The expertise of exploration is the capacity of appraisers to detect the defect on the product but too the capacity to detect all defects on the product. So, the first limit of the R&R test is this expertise is acquired. In fact, the controlled products considered with the perceived defects and at the question: *"all the defects, are they perceived?"* The R&R test supposes *"Yes"*.

Expertise of evaluation

The expertise of evaluation is the capacity of appraisers to give the name and to give one intensity. So, the second limit of the R&R test is to take in consideration the final result. When a R&R test realised, the single data are *Conform* or *Non Conform* i.e. the step 3 of process of inspection: the final judgement.

Thus, Traditional R&R test are not applicable for estimate the capability of an appraiser in a context of sensory analysis. It takes in these data just the final result on the acceptability or not of the product but doesn't identify the root cause: repeatability and/or reproductibity.

3. PROPOSITION OF NEW TEST: THE R²E² TEST

We thus showed previously certain limits of the R&R test because of our problem. So, it is necessary to create a new test based on the steps of the demarche of inspection what we developped. The variability perceived in the tests can come to the exploration or to the evaluation. Thus, we must have a global test separing these two steps.

The R²E² test is founded on the results of an inspection by different appraisers. Each appraiser inspect two times the parts, check if they find a defect, and evaluate the intensity of the defect in a numeric scale [1-6].

Results of the R²E² test are for example for three parts where there are 4 defects – the product 2 has got two defects.

Part No	Just Value	Appraiser 1		Appraiser 2	
		1	2	1	2
1	3	#	3	3	4
2	5	4	5	5	5
2	2	3	3	#	#
3	0	0	0	*	0
…	…	…	…	…	…

0 : Product without defects
** : Product without defects for the reference but with defect for the appraiser*
 # : Product with defects for the reference but without defect for the appraiser

Table 2. Extract of results of R²E² test with 2 appraisers

When # or ***** are present in the reponse of appraisers, it is a problem of exploration. In fact, the defect doesn't perceive or a defect was perceived but don't exist for the reference i.e. the just value.
When the intensity is different, it is a problem of evaluation. The appraisers perceived the same defect that the reference but give a different intensity

We suggest making in the first time a global measure based on the Kappa test of Cohen, and if this test of Kappa show variability on the results, a study and so a new test will be realised to know the causes of the variability of the measure. Thus, in a second time, a new test allowing separating these two expertises (exploration and evaluation) is possible. This new test is named: the R²E² test (Repeatability and Reproductibility of the Exploration and the Evaluation).

3.1. Objectives of the test

From a coefficient given by the test Kappa, it is going to be possible to give a value to the variability of the measures and to deduct if it is necessary to intervene or not to reduce this last one. In case, where it is necessary to complete the study by a finer study, we suggest using the R²E² test. This test is going to allow then to realize:
- In a global way, on the capability of the process of measure and the variability of the measure between the appraisers (test Kappa)
- In a separate way, (when an important variability is detected)
  - of the Repeatability of exploration, that is the capacity of every appraiser to perceive the same defect on a part during both inspections
  - of the Reproducibility of the exploration, that is the capacity of the appraisers to collect the same defect as references on a part
  - of the Repeatability of evaluation, that is the capacity of every appraiser to give the same value of intensity to a defect during both inspections
  - of the Reproducibility of the evaluation, that is the capacity of the appraisers to give the same intensity as the reference to a defect

- of the inertia of the variability of the measure around the reference value

3.2. Global measure

This evaluation is made by means of the Kappa test proposed by Cohen [7], and then extended by Fleiss [8] [9] [10]. This non-parametrical test allows the calculation of the agreement between two or several observers or techniques when judgments are terms.

The Kappa coefficient *(K)* is integrated into the calculation based on the following principle: this agreement results, at the same moment, from a constituent based on an unpredictable answer *(Pe)* and from a constituent based on a true answer *(Po)*.

$$K = \frac{P_0 - P_e}{1 - P_e} \qquad (1)$$

with *Po*: the all right observed proportion.

$$P_0 = \sum_{i=1}^{r} p_{ii} \qquad (2)$$

Pe: the all right unpredictable proportion or the concordance expected under the hypothesis of the independence of judgments.

$$P_e = \sum_{i=1}^{r} p_{i.} p_{.i} \qquad (3)$$

with p_{ii}: the proportion of the defect for the same intensity given by every judge..

with $p_{i.}$: the proportion of defect for every intensity of a given judge

3.3. Detailed measure

When the global evaluation reveals a variability of the control that is too big, a more detailed evaluation can be carried out. The R²E ² test described previously can be used for it.

Expertise of exploration

The reproducibility of an appraiser's exploration *i* is estimated by means of the ratio:

$$\eta_i = \frac{n_i}{N} \qquad (4)$$

with

n_i : total number of defects perceived by the appraiser *i* for all the parts of the sample used in the test

N : total number of defects perceived by the group of experts for all the parts of the sample used in the test

The repeatability of the exploration of an appraiser is estimated by means of the ratio ρ_i :

$$\rho_i = \frac{m_i}{N} \qquad (5)$$

with

m_i : total number of defects collected by the appraiser *i* in two successive controls

N_i : total number of defects collected by the appraiser *i* for all the parts of the sample used in the test

Expertise of evaluation

The reproducibility of the evaluation between the appraisers and the reference is estimated thanks to the calculation of the mean value $\bar{z}$ allowing the estimation of the bias of the evaluation:

$$\bar{z} = \frac{\sum z_i}{k} \qquad (6)$$

with

$$z_i = C_i - E_i \qquad (7)$$

$C_{i:}$ value of intensity

$E_{i:}$ value given by the reference

The repeatability of the evaluation is estimated by the calculation of the standard deviation σ allowing to estimate the variability around the bias:

$$\sigma = \sqrt{\frac{\sum \left(z_i - \bar{z}\right)^2}{n-1}} \qquad (8)$$

with *n*, the number of defects estimated by every appraiser.

The reproducibility of the evaluation is estimated by the Bias :

$$t = \frac{\bar{z}}{\sigma / \sqrt{n}} \qquad (9)$$

The values obtained for the repeatability and the reproducibility also allow one to have a global indicator of the expertise of evaluation, by the calculation of the *"inertia" I:*

$$I = \sqrt{\frac{\sum z_i^2}{n-1}} \qquad (10)$$

"Inertia" is the mean square deviation from the reference value. Inertia includes the bias between the reference and the appraiser and the repeatability of the appraiser.

3.4. Results of the R²E² Test

The results of the R²E² test are resuming in the table 3.

Global evaluation	
Kappa coefficient (eq 1)	>0.8 excellent >0.6 Good >0.4 Moderate
Expertise of Exploration	
Repeatability (eq 5)	$\rho_i > 0.8$
Reproducibility (eq 4)	$\eta_i > 0.8$
Expertise of Evaluation	
Global (R&R) : Inertia (eq 10)	I > 0.75
Repeatability (eq 8)	$\sigma > 0.75$
Reproducibility (eq 9)	$t \le t_{1-\alpha/2,n-1}$

Table 3. Acceptance limits for the R²E² Test

The acceptation limits are determined by the assumptions:

1. The bias must be not significant
2. The total error of repeatability (4s) must be lower than ± 1.5
3. The ratios of similarity of exploration must be greater than 0.8

4. APPLICATION OF THE R²E² TEST

We shall illustrate the use of this test, by an industrial example realising in a manufactury where there are problems of exploration and evaluation on the visual inspection. We trained the apparaisers to use the new method of inspection in respecting the tree step: exploration, evaluation and decision.

Each apparaiser controled the product proposed twice in the same condition. The apparaisers must note down all the defects of each product, give the name and the intensity of defect, give the localisation and be based on specifications.

4.1. Results of the global measure

The table 2 is an extract of results of this application. Yet, the just value being the reference of test must have the same weight as the other value of appraisers. So, we must mutiple the number of just value as the number of repetition for calculation.

After calculation in respecting the demarch definied, the global evaluation gives a Kappa coefficient (*K)* equal to 0.459. We can conclue taht the agreement between the appraisers and with the just value can be considered as moderated (table 3), and give our case as very insufficient considering the quality requirements of the manufactury. So, we can demand : *Can this result be explained by a problem of repeatability? Or reproducibility? In exploration? In evaluation?* To answer, we used the new test : the R²E² test.

4.2. Results of the detailed measure

Results of test are in the first time to use to calculate the expertise of exploration and in the second time to calculate the expertsie of evaluation. The goal in the calculation is to definie the causes of the variability of results.

Expertise of exploration

The table 4 shows the data used for this expertise, as well as the results obtained for the measure of the reproducibility and the repeatability of the exploration for every appraiser.

	Appraiser 1	Appraiser 2	Appraiser 3	Appraiser 4	Appraiser 5
ni	54	31	48	34	14
N	56	56	56	56	56
mi	26	7	22	13	2
Ni	28	24	26	21	12
ηi=ni/ N	0.964	0.554	0.857	0.607	0.250
ρi=mi/ Ni	0.929	0.292	0.846	0.619	0.167

Table 4. Results of the expertise of exploration using the R²E² Test

Appraiser 1 is thus judged as being capable in reproducibility and repeatability in the expertise of exploration ($\eta_i = 0.964 > 0.8$, $\rho_i = 0.929 > 0.8$). It is also the case of appraiser 3 ($\eta_i = 0.857 > 0.8$, $\rho_i = 0.846 > 0.8$). On the other hand, appraisers 2, 4 and 5 are considered as unfit for exploration, both in reproducibility and repeatability ($\eta_i < 0.8$, $\rho_i < 0.8$).

Expertise of evaluation

The table 8 shows the data used for this expertise, as well as the results obtained for the measure of the reproducibility and the repeatability of the evaluation for every appraiser.

	Appraiser 1	Appraiser 2	Appraiser 3	Appraiser 4	Appraiser 5
Reproducibility *(Bias)*	-0.056	0.065	-0.021	0.059	1.357
Repeatability *(Sigma)*	0.359	0.250	0.385	1.906	2.170
Evaluation *(Inertia)*	0.363	0.258	0.386	1.907	2.587
texp *(Variable of Student)*	-1.137	1.438	-0.375	0.180	2.340
Tth	2.006	2.042	2.012	2.035	2.160

Table 5. Results of the expertise of evaluation of the R²E² Test

According to the levels of validation given in the paragraph of proposition of new test, appraiser 1 is thus judged as being capable in reproducibility, in repeatability and in global evaluation in the expertise of evaluation. It is also the case of appraiser 2 and appraiser 3. On the other hand, appraisers 4 and 5 are considered as unfit for the exploration except for appraiser 4 who is considered as capable for the reproducibility of the evaluation.

4.3. Conclusion of the application

The Kappa coefficient (*K*) is based on the concordance of the appraisers between other and with the just value. When

it is insufficient because the value given by the calculation is smaller, the R^2E^2 test comes to complete it to give the reasons of this variability: problem of exploration, problem of evaluation, in the repeatability or in the reproductibility.

In our application, we obtained Kappa coefficient equal to 0.459, considered as moderated.
The detailed analysis of the R^2E^2 test is thus going to allow us to identify on the origin of the failure more precisely.

The table 6 shows the synthesis of the results obtained and developped in the paragraph 4.2 on the capacity of the appraisers in the exploration of the products and so in the detection of defects for each product.

Exploration expertise	Appraiser 1	Appraiser 2	Appraiser 3	Appraiser 4	Appraiser 5
Reproducibility ηi	Capable	Not capable	Capable	Not capable	Not capable
Repeatability ρi	Capable	Not capable	Capable	Not capable	Not capable

Table 6. Results of each appraiser for the exploration

The table 7 shows the synthesis of the results obtained on the capacity of the appraisers to estimate the defects giving their name and their intensity and compare to the name and the intensity giving by the just value.

Evaluation expertise	Appraiser 1	Appraiser 2	Appraiser 3	Appraiser 4	Appraiser 5
Reproducibility	Capable	Capable	Capable	Capable	Not capable
Repeatability	Capable	Capable	Capable	Not capable	Not capable
Inertia	Capable	Capable	Capable	Not capable	Not capable

Table 7. Results of each appraiser for the evaluation

The appraisers 1 and 3 are the only ones to be considered as capable. So, they are as the sensory analysis can it define: the experts in exploration and evaluation of visual inspection. In dimensional metrology, we could tell that they are measuring instruments capable in comparing with the just value so the etalon of reference. The comparaison is easy between the visual inspection and the metrology control and even if we don't develop here we can refer us to [11].

In the expertise of exploration, other appraisers are *"Not capable"*. They have bad results in repeatabiliy and in reproductibility. They must continue to be still trained to improve their performance and to become *"Capable"* in the new test.

The appraisers 1, 2, 3 and 4 are considered as reproductive in the expertise of evaluation in comparison to the just value. The appraisers 1, 2 and 3 are repeatable in the same expertise, but not the appraiser 4. The inertia giving a global note of the expertise of evaluation based on the repeatability and on the reproductibility, indicates that only appraisers 4 and 5 are *"Not capable"* and so, don't correctly estimate the products in giving the good name and the good intensity in fonction to just value. The appraisers 4 and 5 must continue to learn to use this new method that we developed to estimate the defects of products. [1]

5. CONCLUSION

We thus developed in this article the necessity in our case of study to separate the various steps of the inspection to a product to reduce the variability of the measuring instrument. Yet, this variability can have different causes and we must define it to realize to corrective actions. In our case, the measuring instrument is human and so generated an important part of subjectivity. We propose a new method to formalize the visual inspection and a new test to satisfy requirements of this new method. In fact, the actual tests are not adapted to our problem as we showed in this article.

So, to reduce the variability and facilitate our objectives, we developped a new test: the R^2E^2 test. In a first time, we are working to the expertise of exploration i.e. to perceive the defect and in the second time, we are working to the expertise of evaluation i.e. to estimate the defect. Each expertise has got the calculated coefficient and the limits of acceptability to definie if the measuring instrument is *"Capable"* or *"Not capable"*.

The training of each measuring instrument will be realized in fonction of the results of capability. The more important objective is to reduce a maximum the varability of results of each instrument.

REFERENCES

[1] Guerra A.-S., Pillet M., Maire J.L., (2006) Formalization of subjective knowledge by the sensory analysis, Global Manufacturing and Innovation (GMI 2006), Coimbatore, India, pp. 103-110, July 2006

[2] Automotive Industry Action Group (2002). "Measurement systems analysis reference manual". 3e ed, Grays Essex (UK): Carwin Ltd., 225p

[3] Windsor, S. E., (2003). "Attribute Gage R&R". Delta Sigma Solutions LLC, Six Sigma Forum Magazine, pp. 23-28

[4] ISO: International Vocabulary of Basic and General Terms in Metrology, International Organization for Standardization, Geneva, 1993

[5] Meilgaard, M.C., Civille, G.V., Carr, B.T. (1999) "Sensory evaluation techniques", 3rd ed. CRC. CRC Press, London.

[6] Mojet, J. and Köster, E.P. (2005). "Sensory memory and food texture". Food Quality and Preference, 16, pp.251-266

[7] Cohen J. (1960). "A coefficient of agreement for nominal scales". Educ. Psychol. Meas., 20, pp. 27-46.

[8] Fleiss J.L., Cohen J., and Everitt B.S. (1969), "Large sample standard errors of kappa and weighted kappa", Psychol. Bull., 72, 323-327.

[9] Fleiss J.L. (1978). "Inference about weighted Kappa in the non-null case", Appl. Psychol. Meas., 1, pp. 113-117. Fleiss J.L., Cuzick J. (1979) "The reliability of dichotomous judgments: Unequal numbers of judges per subject". Appl. Psychol. Meas., 3, pp. 537-542

[10] Fleiss J.L. (1981) "Statistical Methods for Rates and Proportions", John Wiley and Sons, New York.

[11] Guerra A.-S., Pillet M., Maire J.L., (2007), Métrologie sensorielle pour un Contrôle Visuel des produits, 13ème Congrès Métrologique, Lille, 18-21 Juin 2007