Monitoring panel performances with the Mixed Assessor Model. Meta-analysis of the SensoBase.

C. Peltier[1*], P.B. Brockhoff[2], M. Visalli[1], P. Schlich[1]

[1] INRA, CSGA (UMR 1324), Dijon, France
[2] DTU-Informatics, Technical University of Denmark,Copenhagen, Denmark
* caroline.peltier@dijon.inra.fr

Context of the presentation

CAP Table *(Schlich)* :
Control Assessors Performances Table

A synthetic table for panel and panellist performances in the classical model (SensoBase project)

M.A.M *(Brockhoff* [1] *)* :
New tests of performances taking scaling into account

MAM CAP Table
- Synthetic table for panel and panellist performances in the MAM
- Intuitive set of colors

[1] ***Brockhoff, P.B. (2012).*** **Accounting for Scaling Differences in Sensory Profile Data: Improved Mixed Model Analysis of Variance** (to be submitted)
[1] ***Brockhoff, P.B. (2011).*** **Mixed Assessor Model.** Pangborn Oral Presentation
[1] ***Brockhoff, P.B. and Skovgaard, I.M. (1994).*** **Modelling individual differences between assessors in sensory evaluations**, Food Quality and Preferences 5,215-224

Outline

- The Mixed Assessor Model (MAM)
- The MAM-CAP table
- Application of the MAM to the SensoBase

Classical model of ANOVA

$$Y_{ijk} = \mu + \alpha_i + \nu_j + c_{ij} + \varepsilon_{ijk}$$

α_i : panellist effect
ν_j : product effect
c_{ij} : interaction effect

I : number of panellists
J : number of products
K : number of replicates

$$SS_{total} = SS_{prod} + SS_{subject} + SS_{prod*subject} + SS_{error}$$

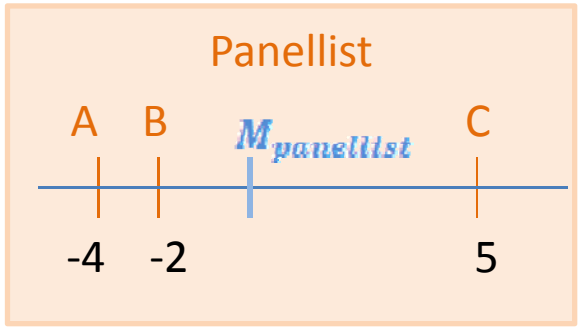
$$\sum_{i,j,k}(y_{ijk} - y_{...})^2 = J * K * \sum_{i}(y_{i..} - y_{...})^2 + I * K * \sum_{j}(y_{.j.} - y_{...})^2 + \sum_{i,j}((y_{ij.} - y_{i..}) - (y_{.j.} - y_{...}))^2 + \sum_{i,j,k}(y_{ijk} - y_{ij.})^2$$

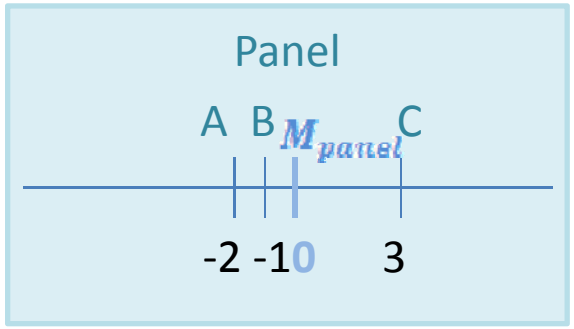
Interaction term: compares panellist means deviations and panel deviations

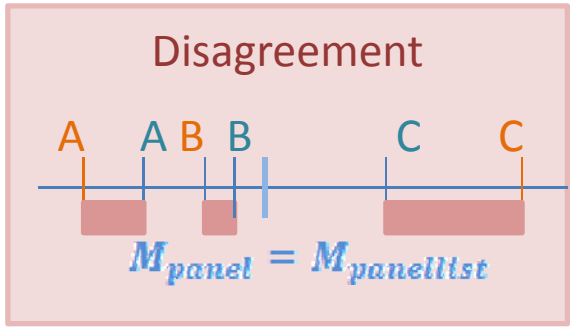
These deviations can be explained by:

- a disagreement with the panel
- a different use of scale = **scaling effect**

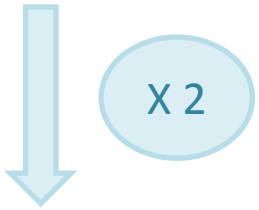
The scaling effect

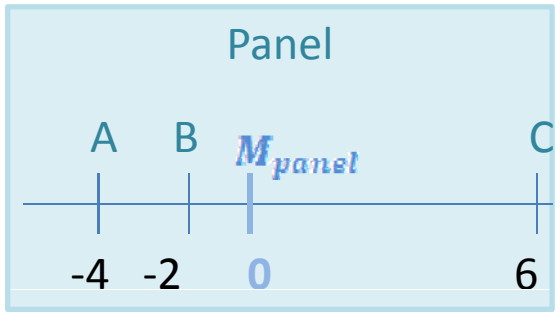






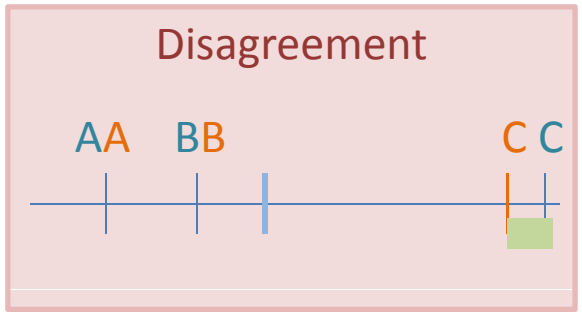
However the panellist seems to agree with the panel. He just uses a larger scale.
How to get rid of the scaling effect ?
We search the number by which we multiply the scores of panel to be as near as possible of the panellist scores.
2 is the coefficient of scaling.





Strong disagreement

Pure disagreement



A new interaction decomposition

- If the panellist i differently uses the scale, there exists β_i (scaling coefficient obtained by linear regression) and R_{ij} (pure disagreement) such as :

$$(y_{ij.} - y_{i..}) = \beta_i (y_{.j.} - y_{...}) + R_{ij}$$

- And so, the interaction $\sum_{i,j} ((y_{ij.} - y_{i..}) - (y_{.j.} - y_{...}))^2$ is decomposed such as :

$$SS_{inter} = SS_{scaling} + SS_{disag} \qquad SS_{disag} = K * \sum_{i,j} R_{ij}{}^2$$

The Mixed Assessor Model

$$Y_{ijk} = \mu + \alpha_i + \nu_j + \beta_i x_j + d_{ij} + \varepsilon_{ijk}$$

α_i : panellist effect

ν_j : product effect

x_j : $(y_{.j.} - y_{...})$

β_i : scaling coefficient

d_{ij} : pure disagreement effect

This new model implies new performances tests.

Panellist performances with the Mixed Assessor Model

Classical panellist performances

- Discrimination: to give different scores to different products
- Agreement: for a judge ⇒ to agree with the group;
  for a group ⇒ not to include too many judges in disagreement
- Repeatability: to give the same score to a same product presented twice
- Scaling: to spread the scores as the panel

Tests of performances with the usual model and the M.A.M.

	Panellist		Panel	
	Usual Model	**M.A.M**	**Usual Model**	**M.A.M**
Discrimination effect	$\frac{Product_{panellist}}{Error_{panellist}}$	$\frac{Product_{panellist}}{Error_{panellist}}$	$\frac{Prod_{panel}}{Inter_{panel}}$	$\frac{Prod_{panel}}{Disag_{panel}}$
Disag. effect	$\frac{Inter_{panellist}}{Error_{panellist}}$	$\frac{Disag_{panellist}}{Error_{panellist}}$	$\frac{Inter_{panel}}{Error_{panel}}$	$\frac{Disag_{panel}}{Error_{panel}}$
Scaling effect		$\frac{Scaling_{panellist}}{Disag_{panellist}}$		$\frac{Scaling_{panel}}{Disag_{panel}}$

We use the pure disagreement where we used to use the interaction.

MAM CAP Table

	Panel Performances					Panellist performances				
Attribute	Mean	F-Prod	F-Scal	F-Disa	RMSE	S2	S4	...	S1	S12
Sweet	4.89	37.80	1.44	0.55	1.91	|---|	> <	...	| |	| |
Sour	7.24	34.63	1.95	0.45	1.47	|--|	>---<	...	| |	| |
Chocolate	3.42	8.12	0.76	1.26	1.51	|---|	|---|	...	|-|	| |
...	...	...	...	...	...	...	...	...	...	...
Lemon	2.98	0.17	1.79	2.86	0.98	|---|	| |	...	|--|	<>
F-rank	-	-	-	-	-	2.1	2.7		4.2	5.7

Panel performances

F-Prod :	p<0.05	p>0.05
F-Scal :	p<0.05	p>0.05
F-Disag:	p<0.05	p>0.05

Panellist performances

Product discrimination :

--- p<0.01 -- p<0.05 - p<0.1

Scaling information :

| |: Average scale <>: Significant larger scale

>< : Significant smaller scale

Agreement information :

Yes	No	No test

MAM CAP Table

	Panel Performances					Panellist performances				
Attribute	Mean	F-Prod	F-Scal	F-Disa	RMSE	S2	S4	...	S1	S12
Sweet	4.89	37.80	1.44	0.55	1.91	|---|	><	...	| |	| |
Sour	7.24	34.63	1.95	0.45	1.47	|--|	>---<	...	| |	| |
Chocolate	3.42	8.12	0.76	1.26	1.51	|---|	|---|	...	|-|	| |
...	...	...	...	...	...	...	...	...	...	...
Lemon	2.98	0.17	1.79	2.86	0.98	|---|	| |	...	|--|	<>
F-rank	-	-	-	-	-	2.1	2.7		4.2	5.7

Discrimination

Scaling

Disagreement

Panel performances

F-Prod :	p<0.05	p>0.05
F-Scal :	p<0.05	p>0.05
F-Disag:	p<0.05	p>0.05

Panellist performances

Product discrimination :

--- p<0.01 -- p<0.05 - p<0.1

Scaling information :

| |: Average scale <>: Significant larger scale

>< : Significant smaller scale

Agreement information :

Yes | No | No test

Attributes sorted from the more discriminative to the less discriminative

MAM CAP Table

	Panel Performances					Panellist performances				
Attribute	Mean	F-Prod	F-Scal	F-Disa	RMSE	S2	S4	...	S1	S12
Sweet	4.89	37.80	1.44	0.55	1.91	|---|	><	...	| |	| |
Sour	7.24	34.63	1.95	0.45	1.47	|--|	>---<	...	| |	| |lr
Chocolate	3.42	8.12	0.76	1.26	1.51	|---|	|---|	...	|-|	| |
...	...	...	...	...	...			...	...	
Lemon	2.98	0.17	1.79	2.86	0.98	|---|	| |	...	|--|	<>
F-rank	-	-	-	-	-	2.1	2.7		4.2	5.7

Average rank of individual product F-ratio. The smaller is F-rank, the better is the panellist.

Panel performances

F-Prod:	p<0.05	p>0.05
F-Scal:	p<0.05	p>0.05
F-Disa:	p<0.05	p>0.05

Panellist performances

Product discrimination: Discrimination
--- p<0.01 -- p<0.05 - p<0.1

Scaling information: Scaling
| |: Average scale <>: Significant larger scale
>< : Significant smaller scale

Agreement information: Disagreement
Yes | No | No test

Repeatability: Repeatability
lr : less repeatable than the panel

Subjects sorted from the more discriminative to the less discriminative

Application of the MAM to the SENSOBASE

SensoBase contains data from :

- 1030 sensory descriptive studies
- 50 laboratories in 24 countries
- 4 734 panellists
- 6 234 products
- 17 275 descriptors
- 4 988 880 scores

Selection of the data sets with :

- At least 2 replicates
- At least 3 products
- Balanced

⇒ A study on 235 data sets.

A first study has been presented by Per B. Brockhoff in 2011.

- Are scaling differences among panellists a reality or not ?
- Does MAM provide us with better product discrimination and panellist agreement ?
- Is there an effect of age, gender and education level on performances ?

Are scaling differences among panellists a reality ?

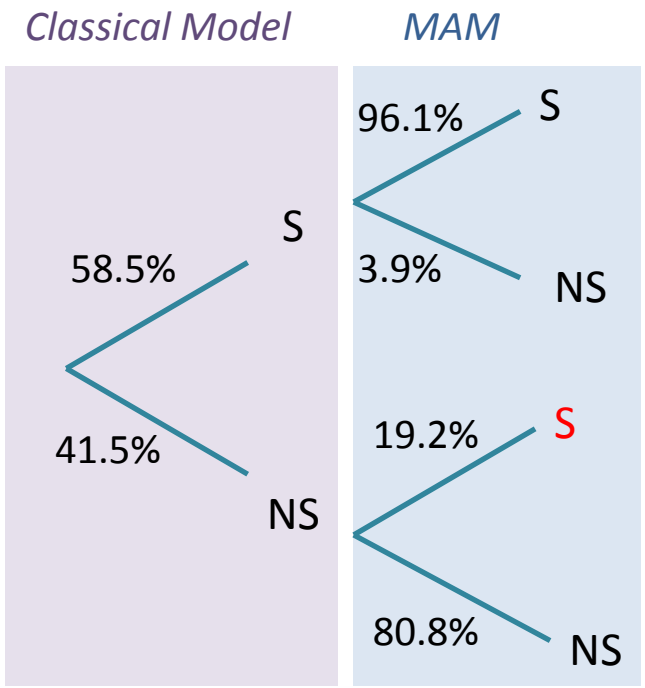
- A significant scaling heterogenity among panellists was found in **45%** of attributes

- A significant individual scaling effect was found in **23%** of the panellists by attribute pairs

- **92%** of the panellists had at least one significant scaling effect

Discrimination of panel for the classical model and the MAM

- With MAM, the product effect was significant (p=0.05) for **64.4 %** of the attributes, whereas it was only **58.5 %** with the classical model.



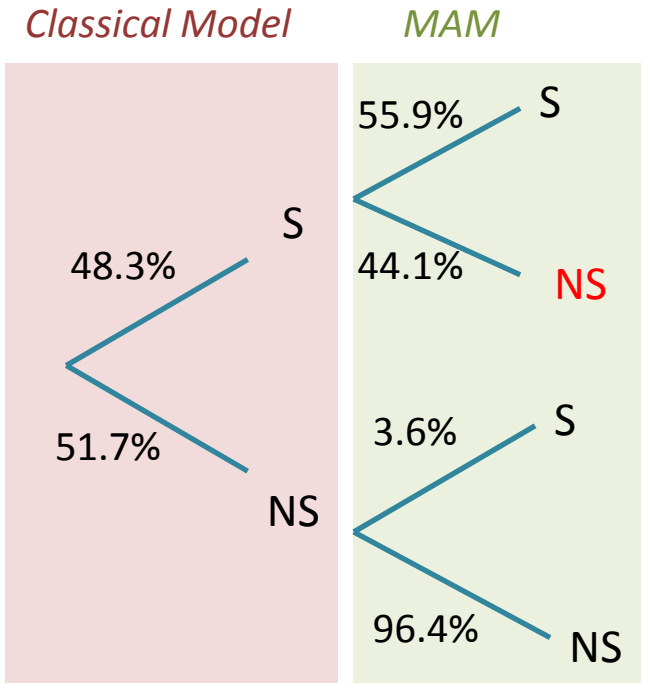
S : significant
NS : not significant

➡ 1/5 NS attributes becomes significant.

The Mixed Assessor Model moderatly improves the discrimination of the panel

Disagreement of panel for the classical model and the MAM

- With MAM, the disagreement effect was significant (p=0.05) for **28.9%** of the attributes, whereas it was **48.3%** with the classical model.



S : significant
NS : not significant

➡ 1/2 significant attributes for disagreement becomes non significant.

The Mixed Assessor Model dramatically improves the agreement of the panel

Percent of the significant attributes for the different effects according to the subject characteristics

Charact.	Level	N	Product	Scaling	Disag	Repet.
Gender	Male	9266	33.06	19.80	14.27	14.15
	Female	23436	29.78	20.69	14.01	13.38
Level of Education (12 datasets)	Secondary and Low	1199	37.95	41.00	20.60	13.43
	Sup	4955	44.04	44.00	23.23	19.60
Age	-30	4335	28.65 b	31.81 a	15.09 a	18.15 a
	30-60	45019	31.27 a	20.01 b	16.01 a	13.90 b
	+60	13010	27.86 c	18.38 c	11.19 b	9.84 c
Total		68998	32.00	23.14	16.13	13.94

P<0.05 P<0.1 P>0.1

- Men are more discriminative than women

Percent of the significant attributes for the different effects according to the subject characteristics

Charact.	Level	N	Product	Scaling	Disag	Repet.
Gender	Male	9266	33.06	19.80	14.27	14.15
	Female	23436	29.78	20.69	14.01	13.38
Level of Education (12 datasets)	Secondary and Low	1199	37.95	41.00	20.60	13.43
	Sup	4955	44.04	44.00	23.23	19.60
Age	-30	4335	28.65 b	31.81 a	15.09 a	18.15 a
	30-60	45019	31.27 a	20.01 b	16.01 a	13.90 b
	+60	13010	27.86 c	18.38 c	11.19 b	9.84 c
Total		68998	32.00	23.14	16.13	13.94

P<0.05 **P<0.1** P>0.1

- Men are more discriminative than women
- People with higher level of education are more discriminative but less repeatable than the others

Percent of the significant attributes for the different effects according to the subject characteristics

Charact.	Level	N	Product	Scaling	Disag	Repet.
Gender	Male	9266	33.06	19.80	14.27	14.15
	Female	23436	29.78	20.69	14.01	13.38
Level of Education (12 datasets)	Secondary and Low	1199	37.95	41.00	20.60	13.43
	Sup	4955	44.04	44.00	23.23	19.60
Age	-30	4335	28.65 b	31.81 a	15.09 a	18.15 a
	30-60	45019	31.27 a	20.01 b	16.01 a	13.90 b
	+60	13010	27.86 c	18.38 c	11.19 b	9.84 c
Total		68998	32.00	23.14	16.13	13.94

P<0.05 P<0.1 P>0.1

- Men are more discriminative than women
- People with higher level of education are more discriminative but less repeatable than the others
- The 30-60 years-old are the most discriminative but also the most in disagreement

Percent of the significant attributes for the different effects according to the subject characteristics

Charact.	Level	N	Product	Scaling	Disag	Repet.
Gender	Male	9266	33.06	19.80	14.27	14.15
	Female	23436	29.78	20.69	14.01	13.38
Level of Education (12 datasets)	Secondary and Low	1199	37.95	41.00	20.60	13.43
	Sup	4955	44.04	44.00	23.23	19.60
Age	-30	4335	28.65 b	31.81 a	15.09 a	18.15 a
	30-60	45019	31.27 a	20.01 b	16.01 a	13.90 b
	+60	13010	27.86 c	18.38 c	11.19 b	9.84 c
Total		68998	32.00	23.14	16.13	13.94

P<0.05 **P<0.1** P>0.1

- Men are more discriminative than women
- People with higher level of education are more discriminative but less repeatable than the others
- The 30-60 years-old are the most discriminative but also the most in disagreement
- The effect of scaling decreases with the age whereas the repeatability increases with the age

Conclusion

- The Mixed Assessor Model is a model taking the **scaling effect** into account
- The **MAM CAP table** sums up all the performances (panel and panellist) thanks to a intuitive set of colors
- With the SensoBase study, we have seen that :
    - The MAM improves the classical model (Panel performances : more discriminative, less disagreement)
    - Men, people with higher level of education and 30-60 years-old are more discriminative than the others
    - The effect of scaling decreases with the age whereas repeatability increases

Future prospects

- Extension of the M.A.M. to other cases (multi-dimensional, non balanced panels…)

Thanks for your attention