



HAL
open science

Apports de la méthode “ Best Worst ” à l’analyse interculturelle des critères des critères de choix des consommateurs : cas de l’huile d’olive

Sihem Dekhili, Eli Cohen, Lucie Sirieix

► To cite this version:

Sihem Dekhili, Eli Cohen, Lucie Sirieix. Apports de la méthode “ Best Worst ” à l’analyse interculturelle des critères des critères de choix des consommateurs : cas de l’huile d’olive. 2010. hal-02823522

HAL Id: hal-02823522

<https://hal.inrae.fr/hal-02823522>

Preprint submitted on 6 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

WORKING-PAPER – UMR MOISA

Apports de la méthode « Best Worst » à l'analyse interculturelle des critères de choix des consommateurs : cas de l'huile d'olive

Dekhili, S.
Cohen, E.
Sirieix, L.

WORKING PAPER MOISA 2010-6



WORKING-PAPER – UMR MOISA

Apports de la méthode « Best Worst » à l'analyse interculturelle des critères de choix des consommateurs : cas de l'huile d'olive

S. Dekhili¹, E. Cohen², L. Sirieix³

¹ Université de Strasbourg, Ecole de Management de Strasbourg, HuManiS, Strasbourg F-67000, France

² Ben Gurion University of The Negev, School of Business and Management, Department of Hotel and Tourism Management, Beer Sheva, I-84105, Israël

³ Montpellier SupAgro, UMR 1110 MOISA, Montpellier F-34000, France

Novembre 2007

Abstract

One frequent objective of market research, and particularly in cross cultural contexts, consists in comparing those attributes that determine the choice of different sub groups of consumers. Practitioners may resort to different types of importance scales, or use ranking methods, or even pair comparisons. This article first compares and shows the limits of these methods of evaluation that are widely discussed in the marketing literature. Then it describes the Best-Worst scaling method. An application of this method in the food domain (olive oil) in two different countries (France and Tunisia) shows that it allows easy product's attributes ranking. Finally, this article presents managerial issues, critics and relevant applications of the method.

Keywords

Choice criteria, Best Worst, Maxdiff, consumer behaviour, cross national research, product attributes.

Applying the “Best Worst” method to the cross comparison of olive oil choice criteria

Résumé

Un objectif courant des études interculturelles consiste à comparer l'importance des attributs déterminants du choix entre sous-groupes de consommateurs. Plusieurs types de mesures peuvent être mises en œuvre : échelles, classement ou comparaisons par paires... Après avoir comparé ces méthodes classiques d'évaluation et rappelé leurs limites, cet article présente la méthode Best-Worst, ses apports et les conditions de son utilisation. Une application dans le domaine alimentaire (huile d'olive), effectuée auprès de consommateurs de cultures différentes (Français et Tunisiens) montre que cette méthode permet de hiérarchiser facilement les attributs de choix d'un produit et de comparer la structure de préférence de deux groupes de population. Enfin, l'article évalue les résultats obtenus, les limites et les conditions d'utilisation de cette méthode.

Mots clefs

Choice criteria, Best Worst, Maxdiff, consumer behaviour, cross national research, product attributes.

JEL: D10, M31



INTRODUCTION

La mesure des critères de choix des consommateurs a toujours constitué un champ d'intérêt important pour les chercheurs et les praticiens, et notamment en marketing international. Elle permet aux décideurs de bien comprendre ce que les consommateurs espèrent, désirent, aiment et n'aiment pas, et d'adapter en conséquence leur offre de produits selon les marchés.

Dans la plupart des études qui portent sur les critères de choix, ou qui cherchent à hiérarchiser les attributs de produits qui sont pris en compte par les consommateurs, la collecte des données met en œuvre des échelles d'évaluation de type Likert ou des échelles sémantiques bipolaires, ou encore des méthodes de classement à modalités. Ces méthodes présentent cependant un certain nombre d'inconvénients. Par exemple, elles ne permettent pas toujours une discrimination nette entre les attributs. Ces derniers ne sont pas mesurés les uns relativement aux autres, ce qui rend difficiles les comparaisons entre les différentes mesures. Lorsqu'il s'agit par exemple de comparer les priorités de choix entre des groupes appariés, comme dans le cas des études interculturelles, la comparaison des échelles de mesure peut poser problème (Bartilowski et alii, 2006).

Pour dépasser certaines de ces limites, nous nous intéressons dans cet article à la méthode Best-Worst (ou encore méthode de la différence maximum, Maxdiff) développée notamment par Louviere et Woodworth (1990).

A titre d'illustration, nous présentons une étude cherchant à déterminer quels sont les critères de choix d'une huile d'olive des consommateurs Français et Tunisiens. Pour les opérateurs de la filière oléicole dans ces deux pays, cette approche comparative permettra de juger de la pertinence de la mise en valeur de l'origine des produits comme critère de qualité ; cette stratégie, qui a fait ses preuves dans d'autres secteurs comme le vin ou les fromages en particulier en France, reste peu familière en Tunisie.

Après avoir comparé les méthodes classiques d'évaluation et rappelé leurs limites, nous présentons la méthode Best-Worst et ses apports en l'illustrant par le cas de notre étude. Enfin, l'article évalue les résultats obtenus, les limites et les conditions d'utilisation de cette méthode.

COMPARAISONS ENTRE DIFFERENTES MESURES D'IMPORTANCE DES CRITERES DE CHOIX

Les méthodes d'enquêtes que nous comparons s'appuient explicitement ou non sur le modèle classique d'attributs de Rosenberg (1956) et Fishbein (1967). Il s'agit en général de mesurer l'importance des attributs qui influencent les choix à partir des déclarations d'individus repérées sur des échelles du type « pas important-très important » ou « d'accord-pas d'accord ». Ces échelles psychométriques ont l'avantage d'être faciles à administrer et sont devenues familières aux répondants. En associant l'importance de l'attribut avec sa performance, on est supposé pouvoir prédire les choix des consommateurs. Beaucoup d'auteurs ont souligné les limites de cette approche, et les critiques portent à la fois sur la qualité des données collectées mais aussi sur les capacités analytiques de ces échelles (Chandon et Bartikowski, 2004). Par exemple, en l'absence de choix alternatif, les répondants ont tendance à favoriser les réponses positives (tout est important), ou un regroupement vers le milieu de l'échelle, de sorte que les résultats sont parfois peu discriminants, ce qui est gênant lorsque l'objectif est de segmenter les marchés (Goodman et alii, 2005). Ce type d'échelle peut également être soumis à un biais culturel, certains répondants pouvant avoir une tendance particulière à « sur ou sous-noter ». Enfin, d'autres auteurs ont mis en évidence le manque de validité prédictive des modèles de Rosenberg et Fishbein (Wilkie et Pessemier, 1973) construits à partir de cette combinaison importance/performance.

On reproche aussi à ces échelles des défauts plus conceptuels. On peut s'interroger par exemple sur l'hypothèse d'équidistance entre les échelons d'une échelle d'intervalles

(Bartilowski et alii, 2006). Il est alors difficile d'interpréter l'ampleur des différences entre deux observations (par exemple, une moyenne de 4 signifie-t-elle que l'importance accordée à un attribut est deux fois supérieure à une moyenne de 2 ?).

Les chercheurs et les praticiens ont donc essayé de s'affranchir des limites de cette échelle largement popularisée à partir des travaux de Rensis Likert et de chercher d'autres voies et méthodes plus performantes. Nous allons d'abord examiner brièvement ces différentes méthodes selon le double point de vue de la collecte des données et des possibilités d'analyses, avant de présenter celle qui nous intéresse particulièrement ici, l'analyse « Best-Worst ». On peut les classer en deux groupes : dans un premier groupe, les méthodes qui cherchent à obtenir un score indépendant pour chacun des attributs, l'objectif étant que les mesures obtenues s'affranchissent des biais liés aux caractéristiques de chaque individu, tout en conservant des capacités discriminantes. Ces méthodes reposent finalement sur la notation. Un deuxième groupe rassemble des méthodes qui font plutôt appel à des tâches de tri, de classement ou de choix.

Méthodes utilisant la notation

Une première voie consiste à libérer le répondant des limites imposées par la taille de l'échelle de notation en lui proposant des notations libres ou « unbounded ratings » (Marder, 1997). Le répondant est ainsi libre d'attribuer des scores négatifs et de calibrer ses scores selon sa propre échelle de référence. Une manière courante de procéder est de proposer aux répondants un segment non gradué et borné par deux expressions opposées (pas d'accord du tout – tout à fait d'accord). Pour certains auteurs (Chandon et Bartikowski, 2004), le temps d'administration des questionnaires est deux fois plus long que pour l'échelle de Likert en 5 ou 7 points, surtout lorsque les répondants sont invités à mettre des évaluations négatives. Paradoxalement, et de l'aveu même de son auteur (Marder, 1997) la méthode n'améliore

sensiblement ni la discrimination entre attributs ni la validité prédictive. Elle impose évidemment une démarche d'harmonisation des échelles.

Une autre approche pour lever la contrainte des intervalles de l'échelle consiste à proposer au répondant d'évaluer les attributs par référence à la note donnée au premier attribut, en multipliant ou divisant cette note selon l'importance des différences perçues entre attributs « magnitude estimation » (Lodge, 1981). L'hypothèse sous-jacente est que l'individu procède de manière comparative et hiérarchique par rapport à ce qu'il connaît déjà. Cette approche séduisante fait cependant aussi l'hypothèse que les répondants sont capables de multiplier ou de diviser avec la même facilité, ce qui n'est pas démontré. Elle est également sensible aux effets d'ordre de présentation des attributs et à leur nombre, ce qui nécessite la mise en œuvre d'un plan expérimental équilibré. Le résultat est que la durée d'administration serait deux fois plus élevée que celle de l'échelle classique de Likert, mais en revanche, elle témoignerait d'une meilleure validité prédictive et de bonnes capacités de discrimination entre individus (Chandon et Bartilowski, 2004).

Méthodes s'appuyant sur un choix

Dans ces méthodes, il s'agit de placer les répondants en situation de comparaison ou de choix des attributs. La méthode des comparaisons par paires a été longtemps utilisée pour mesurer la dispersion des opinions en plaçant les répondants dans une situation de choix (Finn et Louviere, 1992). Toutefois, ces auteurs notent que ce type de méthode de classement par paires ne permet de prendre en compte qu'un nombre restreint des attributs à considérer, et reste réservée aux études de moins de dix stimuli.

La méthode de choix discrets, basée sur la mesure de l'utilité, permet elle aussi de dépasser cette contrainte du nombre des items (Finn et Louviere, 1992). Elle consiste à proposer à des répondants des tâches de choix successives selon un plan d'expérience mettant en

concurrence les modalités des attributs faisant l'objet de l'étude. Ces choix révèlent les préférences relatives d'un individu pour certains attributs. La méthode des choix discrets présente des qualités supérieures en termes de discrimination entre attributs et en termes de validité prédictive. Bien menée, elle permet de calculer des parts de marché théoriques. Mais elle présente deux limites importantes. La première est celle du coût élevé de sa mise en œuvre et de la nécessité de disposer de logiciels sophistiqués. La seconde est que cette méthode pose des difficultés d'interprétation des données, particulièrement lorsqu'il s'agit de comparaisons d'utilités issues d'expérimentations différentes (Goodman et alii, 2005). Enfin, si le nombre des attributs et de leurs modalités est élevé elle nécessite la mise en œuvre d'échantillons importants.

Enfin, une variante classique et souvent pratiquée de la méthode de choix consiste à faire classer les attributs par ordre d'importance. La méthode présente l'avantage d'une utilisation unique de chaque point de l'échelle, ce qui n'est pas le cas des échelles d'importance. Toutefois, cette méthode souffre des biais de l'effet d'ordre et n'est guère praticable dès lors que l'on dépasse un certain nombre d'attributs. Stephenson (1953) a proposé une amélioration qui consiste à placer chaque item dans un nombre pré-déterminé de catégories, allant du plus important au moins important, mais en imposant un maximum d'items dans chaque catégorie (tri forcé ou « Qsort »). A notre connaissance, cette piste n'a pas fait l'objet de travaux notables.

Présentation de la méthode « Best-Worst »

Cette méthode, appelée aussi « Max Diff », fait partie des méthodes dans lesquelles les répondants sont appelés à faire un choix. Le principe est simple : après avoir établi une liste d'attributs considérés comme pertinents par rapport à une situation d'achat ou de consommation d'un produit, il s'agit de présenter à des répondants des listes restreintes d'attributs parmi lesquels le répondant doit choisir celui qui est le plus important pour lui, et

celui qui est le moins important. Ces listes sont constituées et présentées selon un plan d'expérience combinatoire en blocs incomplet équilibrés (BIE) (Rao, 1960) permettant de s'assurer que chaque attribut est comparé le même nombre de fois à chacun des autres attributs. Le BEI est une extension du carré latin, dans lequel les n modalités à analyser sont disposées en n rangs et n colonnes, de manière à ce que chaque rangée et chaque colonne présentent chaque modalités dans une position différente. Chaque rangée peut être considérée comme un ensemble de choix (un « bloc »).

La première étape consiste à décider du nombre des critères de choix ou d'attributs qui seront comparés entre eux, puis à élaborer le plan incomplet en fonction du nombre de ces attributs et du nombre de modalités à faire figurer dans chaque tâche de choix. En pratique, il convient d'arbitrer entre le nombre des modalités et la liste des choix présentés au répondant. L'expérience suggère qu'il ne faut pas dépasser 20 tâches de choix, sous peine de démotiver les répondants, et qu'il faut pour chaque tâche limiter le nombre des modalités à 4 ou 6, ce dernier chiffre étant un maximum à ne pas dépasser, surtout si le libellé des modalités est long ou complexe (Remaud et Lockshin, 2009).

Par exemple si on a 13 attributs, on peut adopter un BEI de type (13, 4, 4, 1) : les 13 attributs (critères de choix supposés) permettent de construire 13 tableaux (blocs), comportant chacun 4 attributs. Chacun d'eux apparaît le même nombre de fois (4 fois). Et chaque paire d'attributs apparaît le même nombre de fois (1 fois).

Cohen (2003) a comparé les résultats issus de la méthode « Best-Worst » avec ceux obtenus avec des échelles de type Likert, d'une part, et avec des comparaisons par paires, d'autre part. En termes de durée d'administration, il constate qu'en moyenne les évaluations par les méthodes de comparaisons par paires et Best-Worst sont trois fois plus longues à accomplir que les évaluations par les échelles de type Likert. Il explique cette différence de durée par

une plus grande implication et réflexion dans le cas des méthodes de comparaisons par paires et Best-Worst.

En termes de capacité discriminante, il conclut au moyen d'un test d'égalité de moyennes des notes des attributs que la méthode Best-Worst est la plus performante des trois méthodes, tandis que l'échelle d'évaluation de type Likert est la moins discriminante. Dans le même sens, Goodman et al. (2005) mesurent l'évaluation de l'importance des attributs explicatifs du choix d'un vin, au moyen de l'échelle d'évaluation de type Likert et la méthode Best-Worst.

Cohen (2003) vérifie également que l'échelle d'évaluation de type Likert est la moins performante des trois méthodes pour discriminer entre les individus. La méthode des comparaisons par paires semble être légèrement plus discriminante que la méthode Best-Worst. Enfin il conclut que la méthode Best-Worst est la plus stable puisqu'elle montre la fiabilité test-retest la plus satisfaisante (97% contre 88% dans le cas de la méthode des comparaisons par paires, et 85% dans le cas des échelles de Likert).

Nous présentons une synthèse des évaluations de ces méthodes au tableau 1, en prenant en compte les critères liés à la qualité et la facilité d'obtention des données et les critères liés au pouvoir discriminant et prédictif des analyses.

Tableau 1 : Comparaison de méthodes de mesure d'importance des attributs

		Collecte des données			Analyse des données		
		Administration	Fiabilité des réponses	Durée administration	Discrimination entre attributs	Discrimination entre individus	Validité prédictive
Notation	Mesure directe des scores d'importance (Fishbein, 1967)	très facile	moyenne	courte	moyenne	bonne	médiocre
	Estimation comparative des grandeurs (magnitude estimation) (Lodge, 1981)	complexe	moyenne	moyenne	plus faible	bonne	bonne
	Echelle libre (Marder, 1997)	facile	moyenne	moyenne	plus faible	bonne	médiocre
	Somme constante des scores	facile	moyenne	moyenne	plus faible	bonne	bonne
Arbitrage	Tri forcé (Q sort) variante du classement (Stephenson, 1953)	complexe	bonne	moyenne	Plus forte	bonne	bonne
	Classement par paires	facile	bonne	courte	Très forte	bonne	
	Choix discrets (Finn et Louvière, 1992)	complexe	Très bonne	longue	Plus forte	Peu aisée	bonne
	Différence maximum ou choix des extrêmes « best worst » (Finn et Louvière, 1992)	facile	Très bonne	longue	Plus forte	bonne	bonne (la meilleure)

APPLICATION : UNE ETUDE INTERCULTURELLE SUR LES CRITERES DE CHOIX DE L'HUILE D'OLIVE

Les choix alimentaires sont des marqueurs des différences culturelles qui se reflètent dans les différences et la variété des dispositifs institutionnels ou privés qui encadrent l'offre et la demande de qualité alimentaire. Comme nous l'avons souligné, le choix d'étudier le secteur de l'huile d'olive est lié aux interrogations des responsables de la filière quant à l'opportunité et la manière de promouvoir des politiques d'appellations d'origine, en France comme en Tunisie. Notre objectif est donc de déterminer si les notions d'origine (région, pays) sont des critères de choix importants pour les consommateurs d'huile d'olive, au sein de deux populations différentes (Tunisie et France), tant sur le plan de la fréquence et des modes de d'achat de ce produit que sur la familiarité avec le concept d'appellation d'origine.

Un travail exploratoire préalable auprès de consommateurs français et tunisiens a permis de définir les critères de choix du produit (auteurs, 2006). Nous avons retenu 13 attributs ou critères de choix qui seront considérés dans la construction du plan d'expérience en blocs incomplets de quatre critères, chaque critère apparaissant quatre fois et étant comparé aux douze autres une fois et une seule (tableau 2).

Tableau 2 : Plan d'expérience équilibré en blocs incomplets pour les 13 critères de choix d'une huile d'olive

		13,4,4,1			
<i>n</i> =13		n°crit.	n°crit.	n°crit.	n°crit.
1	pays d'origine	1	2	4	10
2	région d'origine	2	3	5	11
3	prix	3	4	6	12
4	marque	4	5	7	13
5	variété d'olives	5	6	8	1
6	producteur/vendeur	6	7	9	2
7	goût	7	8	10	3
8	mention « vierge extra »	8	9	11	4
9	couleur	9	10	12	5
10	aspect	10	11	13	6
11	emballage	11	12	1	7
12	SOQet médailles	12	13	2	8

Les tâches de choix sont présentées sous forme de tableaux successifs. Les répondants doivent choisir dans chaque tableau le critère qu'ils jugent le plus important dans leur choix, et le critère le moins important. La figure 1 montre un exemple d'une tâche de choix.

Figure 1 : Exemple d'une tâche de choix

<i>Le moins important</i>	<i>Critère de choix</i>	<i>Le plus important</i>
<input type="checkbox"/>	1 A	<input type="checkbox"/>
×	2 B	<input type="checkbox"/>
<input type="checkbox"/>	3 C	<input type="checkbox"/>
<input type="checkbox"/>	4 D	×

128 consommateurs tunisois et 123 parisiens ont été interrogés au cours l'année 2006. Après dépouillement, 123 questionnaires français et 122 tunisiens se sont avérés utilisables pour l'analyse Best-Worst (tableau 3).

Tableau 3 : Caractéristiques socio-démographiques des répondants par pays

		Tunisie	France
Nombre de répondants		122	123
Sexe	Homme	93	41
	Femmes	29	82
Age	18 à 44 ans	67	49
	45 et plus	55	74
Niveau d'éducation	Secondaire	27	56
	Universitaire	95	67
Revenu	Faible et moyen*	59	85
	Haut*	63	38

* Faible et moyen revenu : Tunisie : <1000Dinars Tunisiens, France : 3693 Euros par mois

* Revenu élevé: Tunisie : ≥1000Dinars Tunisiens, France : 3693 Euros par mois

RESULTATS

Le niveau d'importance de chaque critère de choix est mesuré par la différence entre le nombre de fois où le critère a été choisi comme étant le plus important (Best) et le nombre de fois où il a été considéré comme étant le moins important (Worst). Marley et Louvière (2005) montrent que lorsque le plan expérimental est complet, cette différence est une représentation

dont l'échelle de valeur est à 95% aussi correcte que le modèle logit multinomial construit à partir des mêmes données.

Pour chaque attribut, on mesure la différence des cumuls de « best » et de « worst ». Dans notre situation, la différence maximum B-W que peut atteindre un attribut est + 4 ou -4. Le niveau d'importance dépend du nombre des répondants et de la fréquence d'apparition de chaque attribut dans les ensembles de choix. Il est donc pertinent de considérer un score standardisé afin de pouvoir comparer entre groupes de répondants qui diffèrent en termes d'effectifs.

$$\text{Score standardisé} = \frac{\text{score (Best-Worst)}}{4 n}$$

Où : n = le nombre des répondants

4 = la fréquence d'apparition de chaque attribut dans les ensembles de choix

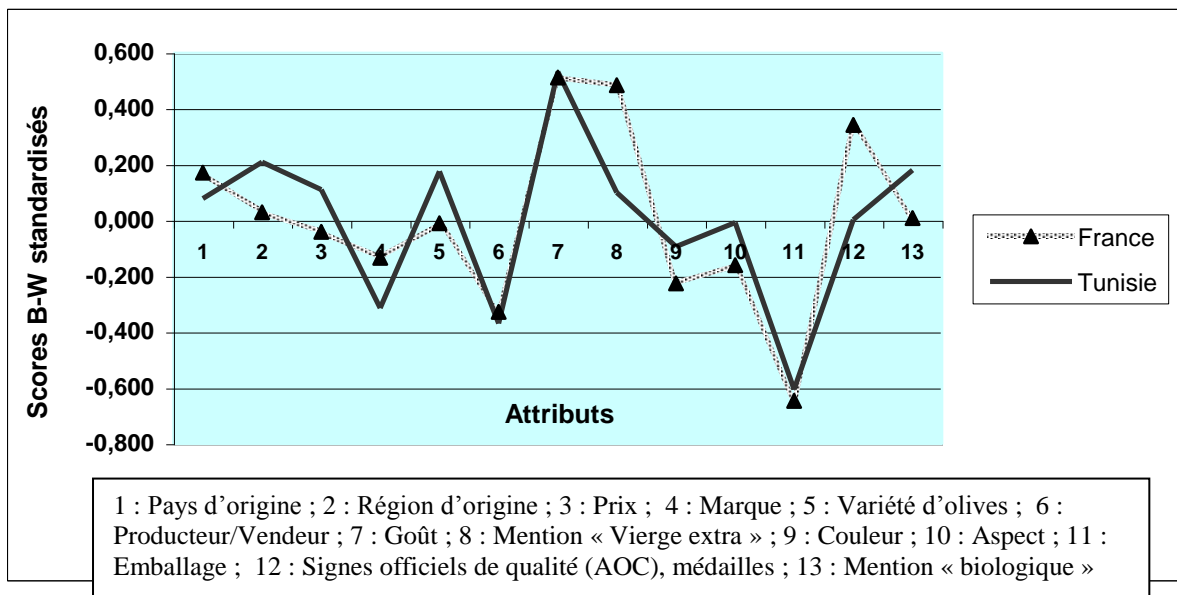
Les scores standardisés « B-W » figurent au tableau 4.

Tableau 4 : Importance accordée aux critères de choix (classée en ordre décroissant selon les score B-W France)

n°	Critères :	France n=123			Tunisie n=122			Sign. Difference B-W Valeurs de p
		Best	Worst	Score B-W Stand.*	Best	Worst	Score B-W Stand*	
7	goût	264	12	0.512	275	15	0.533	ns
8	« Vierge extra »	280	42	0.484	140	54	0.100	0.000
12	SOCs, médailles	228	60	0.341	172	70	0.002	0.000
1	pays d'origine	140	57	0.171	165	77	0.078	ns
2	région d'origine	119	105	0.028	152	98	0.209	0.002
13	mention « biologique »	140	136	0.008	118	80	0.180	0.019
5	variété d'olives	80	85	-0.010	154	105	0.176	0.000
3	prix	89	109	-0.041	137	136	0.111	0.025
4	marque	93	159	-0.134	90	94	-0.314	0.006
10	aspect	73	152	-0.161	54	100	-0.008	0.003
6	producteur/vendeur	50	211	-0.327	58	211	-0.369	ns
9	couleur	26	137	-0.226	51	231	-0.094	0.002
11	emballage	16	334	-0.646	20	315	-0.605	ns

Rappelons qu'un score négatif pour un attribut ne signifie pas que cet attribut influence de façon négative le choix du répondant, mais que cet attribut a été rarement cité comme étant le plus important, et/ou cité plus souvent comme étant moins important dans les différentes configurations de choix. Les scores standardisés de chaque attribut sont illustrés dans la figure 2.

Figure 2 : Les profils des attributs qui déterminent le choix de l'huile d'olive chez les consommateurs tunisiens et français



Les résultats sur les différences de score moyen (B-W) entre les deux échantillons peuvent se lire sur le graphique de la figure 2 en examinant celles de ces différences qui sont significatives (dernière colonne du tableau 4). Ainsi, le goût (7) est le critère déclaré le plus important dans le choix de l'huile d'olive, aussi bien chez les Tunisiens que les Français. C'est une réponse relativement attendue, s'agissant d'un produit alimentaire à forte vocation gustative. Le pays (1), la région (2) sont classés parmi les critères importants dans les deux pays, mais avec des différences, les Français accordant relativement plus d'importance au pays, les Tunisiens à la région.

Le critère « vierge extra » est particulièrement important pour les Français, ainsi que les SOQs (signes officiels de qualité), alors que la mention « bio » (13), la variété d'olives (5) et

le prix (3) présentent des scores moyens positifs chez les consommateurs tunisiens. Les deux attributs Producteur/Vendeur et Emballage (6, 11) sont les moins importants dans le choix de l'huile d'olive des deux groupes nationaux, avec des scores très comparables.

On peut se poser la question de l'importance relative des critères de choix. En d'autres termes, que peut-on dire de la hiérarchie entre « goût », « région d'origine » et « marque » ? Pour comparer l'importance relative des attributs entre eux sur une échelle positive, une méthode consiste à extraire la racine carrée du rapport entre B et W pour chaque attribut, en ajoutant arbitrairement une valeur positive à W, par exemple 0,5, afin d'éviter la division par zéro. La racine carrée de B/W est calée sur une échelle de sorte que l'attribut ayant le coefficient $\sqrt{[B/W]}$ le plus élevé devient 100 (dans notre cas, il s'agit du goût). Les coefficients obtenus expriment les probabilités de choix comparées de chaque attribut par rapport à l'attribut de référence (Marley et Louvière, 2005).

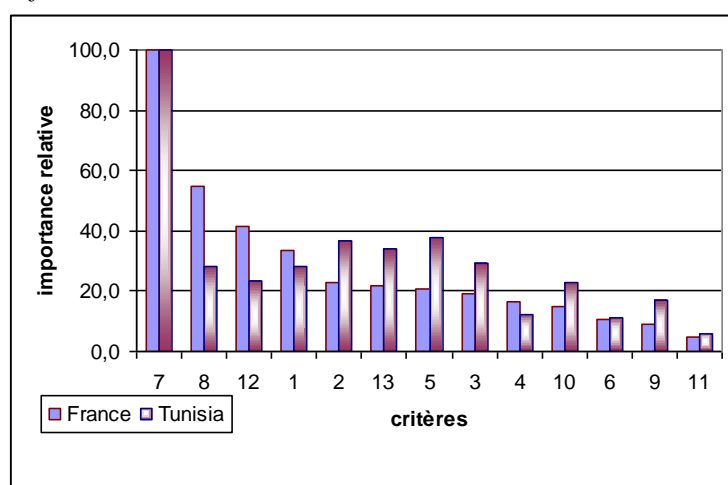
Le tableau 5 donne le résultat de ces calculs et permet de commenter la hiérarchie des attributs dans chaque pays, et de comparer cette hiérarchie entre les pays.

Tableau 5 : Hiérarchisation comparée des critères de choix selon les pays

n°	Critères :	France n=123		Tunisie n=122	
		$\sqrt{[B/W]}$	Importance Relative	$\sqrt{[B/W]}$	Importance Relative
7	goût	4.69	100.0	4.28	100.0
8	« Vierge extra »	2.58	55.0	1.21	28.3
12	SOQs, médailles	1.95	41.6	1.00	23.4
1	pays d'origine	1.57	33.5	1.21	28.4
2	région d'origine	1.06	22.7	1.57	36.6
13	mention « biologique »	1.01	21.6	1.46	34.2
5	variété d'olives	0.97	20.7	1.61	37.6
3	prix	0.90	19.3	1.25	29.1
4	marque	0.76	16.3	0.52	12.2
10	aspect	0.69	14.8	0.98	22.9
6	producteur/vendeur	0.49	10.4	0.47	11.0
9	couleur	0.44	9.3	0.73	17.2
11	emballage	0.22	4.7	0.25	5.9

On remarque que l'échelle d'importance relative des critères de choix autres que le goût est plus large pour les Français que pour les Tunisiens. Ainsi, pour les premiers, le critère le moins important, l'emballage, a une importance relative par rapport au goût de 4,7% et le critère le plus important, mention « vierge extra » a une importance de 55%. Pour les Tunisiens, les valeurs se situent entre 5,9% (emballage) et 37% (variété d'olives). Il semble donc y avoir une plus grande indécision en matière de critères choix de la part des Tunisiens. On peut représenter ces scores d'importance relative sur le graphique de la figure 3.

Figure 3 : Importance relative des critères de choix, exprimée par rapport au critère de référence



En étudiant les critères de façon plus détaillée, par rapport au goût qui constitue le critère le plus important (base 100), on remarque pour l'échantillon tunisien que la région d'origine, la mention « bio », et la variété d'olives ne représentent respectivement que 36.2%, 34.2% et 37.6%. Dit autrement, ces attributs ont environ trois fois moins de chances d'intervenir comme critère de choix principal que le goût. Quatre critères sont cinq ou six fois moins importants que le goût : la marque, l'influence du producteur/vendeur, la couleur, l'emballage. Pour l'échantillon français, le critère « Vierge extra » est relativement important avec un score de 49% par rapport au goût. Les signes de qualité (SOQs, médailles) se distinguent également par rapport à l'ensemble des autres critères de choix.

Il est également important de savoir si des critères socio-démographiques comme le sexe, l'âge, le niveau d'éducation et le revenu interviennent dans la hiérarchie des attributs. Pour répondre à cette question, nous avons adopté un modèle linéaire général au sein de chaque groupe de population, en considérant les segments suivants : hommes/femmes, les plus âgés/les plus jeunes, les consommateurs ayant un niveau universitaire d'éducation/ceux ayant un niveau secondaire, les plus aisés/les moins aisés. Les résultats de l'analyse ANOVA confirme que l'hétérogénéité des critères de choix est plus large dans le groupe français que dans le groupe tunisien. En ce qui concerne les variables de segmentation de ces groupes, l'âge semble être celle qui différencie le plus les critères de choix, à la fois à l'intérieur des groupes mais aussi entre les deux groupes.

Le tableau 6 montre les coefficients $\sqrt{[B/W]}$ pour les deux groupes France et Tunisie, en fonction de l'âge des répondants (au dessus et en dessous de 44 ans). De façon assez inattendue, on constate que pour les 7 critères de choix les plus importants, la hiérarchie selon les groupes d'âges est inversée dans les deux groupes nationaux: les critères les plus souvent cités par les plus de 44 ans en France le sont par les moins de 44 ans en Tunisie.

La variété d'olives, considérée comme plus importante pour les Tunisiens que pour les Français, l'est surtout pour la tranche des plus de 44 ans. La région et le bio sont les arguments les plus importants pour les tunisiens de moins de 44 ans, et sont également plus importants pour cette catégorie que pour les français de la même tranche d'âge pour qui ces attributs sont presque secondaires. Enfin, on remarque que si les Français privilégient les mentions « vierge extra » et les SOQs, c'est très nettement le fait du groupe le plus âgé, pour qui cette mention « vierge extra » est presque aussi importante que le goût.

Tableau 6 : Hiérarchie des critères de choix selon les tranches d'âge (France et Tunisie)

	France			Tunisie		
	<u>FR</u>	<u>FR</u>	<u>FR</u>	<u>TU</u>	<u>TU</u>	<u>TU</u>
	<u>Ens.</u>	<u>18-44</u>	<u>>44</u>	<u>Ens</u>	<u>18-44</u>	<u>>44</u>
	<u>123</u>	<u>49</u>	<u>74</u>	<u>122</u>	<u>67</u>	<u>55</u>
7 goût	100,0	100,0	100,0	100,0	100,0	100,0
8 « vierge extra »	55,0	30,9	88,2	28,3	32,2	23,5
12 SOQs, médailles	41,6	27,1	59,4	23,4	30,5	16,0
1 pays d'origine	33,5	26,1	40,1	28,4	32,1	23,6
5 région d'origine	22,7	16,7	28,1	36,6	40,7	31,1
3 mention « biologique »	21,6	16,5	25,9	34,2	48,1	23,6
2 variété d'olives	20,7	22,0	19,7	37,6	34,0	42,3
13 prix	19,3	19,2	19,2	29,1	30,3	26,8
10 marque	16,3	10,1	22,3	12,2	11,4	12,6
9 aspect	14,8	15,4	14,3	22,9	25,3	19,5
4 producteur/vendeur	10,4	8,1	12,3	11,0	13,3	8,3
6 couleur	9,3	10,5	8,3	17,2	17,5	16,2
11 emballage	4,7	4,6	4,7	5,9	6,4	5,1

INTERET ET LIMITES DE LA METHODE : DISCRIMINER ET HIERARCHISER

Certains contextes d'études nécessitent que l'on puisse à la fois discriminer entre les attributs des produits qui déterminent les choix, mais aussi quantifier les écarts d'importance entre les attributs, en s'affranchissant des biais méthodologiques liés aux échelles de mesure. Il faut donc disposer d'instruments qui favorisent ces deux objectifs, et le choix de la méthode Best-Worst paraît bien adapté. Le traitement de l'exemple que nous avons donné nous a permis de

comparer assez efficacement le rôle des attributs de choix d'une huile d'olive entre Tunisiens et Français, et nous a permis de vérifier que les attributs se hiérarchisent différemment selon les deux groupes nationaux, et à l'intérieur de ces groupes, selon les classes d'âge. La discrimination entre attributs est assez nette et le caractère significatif ou non des différences entre attributs a pu être mis en évidence à l'aide de tests statistiques simples. Par exemple, les consommateurs Français sont plus sensibles à l'appellation d'origine que les consommateurs en Tunisie, où la réputation des signes officiels de qualité est encore à construire. De même, les Français (surtout les plus âgés) ont adopté le signal « vierge extra » comme un critère de choix presque aussi important que le goût, et cet attribut est également valorisé par les jeunes Tunisiens. A l'opposé, la variété d'olives est plus importante pour les consommateurs Tunisiens que pour les consommateurs Français. Cependant, les différences ne sont pas seulement interculturelles : par exemple en France les consommateurs plus jeunes se montrent plus sensibles à la variété d'olives que les consommateurs plus âgés. La méthode Best-Worst se révèle donc être un outil intéressant pour segmenter de façon fine les marchés.

Le nombre de tâches de choix (treize) dans notre enquête était assez élevé, ce qui peut être un frein à l'application de cette méthode impliquante pour les répondants. Cependant la quasi totalité des questionnaires s'est avérée exploitable. Les échantillons de répondants mis en œuvre sont relativement restreints, de l'ordre de la centaine, et le coût de mise en œuvre est donc très comparable avec des études plus classiques.

Notre expérience montre que les répondants comprennent facilement la tâche demandée, l'enquêteur doit veiller cependant à ce que toutes les tâches de choix soient renseignées par le répondant, au risque de ne pouvoir utiliser le questionnaire, c'est donc un point important auquel devra être attentif le demandeur de l'étude. Contrairement aux échelles de mesure classiques, le tableau de données ne peut pas comporter de valeurs manquantes.

Si la méthode Best-Worst présente de nombreux avantages qui justifient son utilisation, elle ne résout cependant pas tous les problèmes.

Comme toute méthode d'analyse de données déclaratives, la méthode Best-Worst ne traite par nature que de l'importance déclarée, et non de l'importance réelle des critères proposés. Dans l'exemple donné par notre étude, il est peu réaliste de croire que la mention du producteur/vendeur, ou que l'emballage n'influencent pas le choix d'une huile d'olive par les consommateurs. Il est par ailleurs très « rationnel » pour un consommateur de produit alimentaire de désigner le « goût » comme attribut le plus important, et de ne pas se déclarer influencé par des indicateurs comme l'emballage ou la publicité. La méthode est donc sensible à l'hétérogénéité des attributs qui sont pris en considération dans l'analyse. La même prudence d'interprétation que pour d'autres études déclaratives doit donc prévaloir. Finalement, la méthode recèle les faiblesses inhérentes aux modèles d'attributs de type compensatoire qui mettent en comparaison des attributs appartenant à des catégories en réalité peu comparables entre elles. Les praticiens devront donc être particulièrement attentifs à cet inconvénient : dans notre étude, l'attribut « goût » aurait peut être dû être évité.

La méthode partage également avec toutes les autres méthodes d'analyse de données déclaratives effectuées dans un contexte inter-culturel la double difficulté du choix des items pertinents dans des cultures différentes et de la traduction du libellé des items. Dans notre étude, nous avons contourné la difficulté de traduction en sélectionnant des acheteurs tunisiens francophones. Mais ce n'est pas évidemment une situation généralisable. Par ailleurs l'huile d'olive est achetée en Tunisie pour ses propriétés sanitaires ou comme produit de beauté, ce qui n'est guère le cas en France. La liste proposée reflétait peut être mal cette finalité particulière du produit liée à la culture Tunisienne.

Finalement, on notera dans les remarques précédentes qu'il ne s'agit pas de limites propres à la méthode elle-même, mais à toute méthode qui s'appuie sur l'hypothèse d'un arbitrage entre attributs par le consommateur lors du choix.

La méthode Best-Worst souffre également de limites qui lui sont propres, concernant le travail du répondant et la sélection proposée des critères.

Une limite soulignée dans plusieurs études (notamment Marder, 1997) concerne le temps nécessaire aux répondants pour compléter le questionnaire. Notre expérience ne confirme pas cependant que ce temps nécessaire soit une limite réelle. En effet, les répondants ne semblent pas avoir été gênés ; il n'y a pratiquement aucun questionnaire non rempli. Cependant ce temps passé peut constituer une limite forte s'il correspond à une difficulté perçue de la tâche trop importante, avec le risque d'engendrer de la lassitude face au caractère répétitif des questions posées, voire une réaction de rejet de la part de répondants qui se demandent pourquoi la même question leur est posée plusieurs fois avec une sélection tournante parmi les mêmes attributs.

Il conviendra donc de prévenir les répondants que ces répétitions font partie de la méthode, et il est recommandé que les instructions soient rappelées à chaque répondant au début de la tâche. La présence de l'enquêteur lors de l'administration des questionnaires sera une précaution utile.

La méthode Best-Worst pose enfin le problème d'un choix forcé sur deux items (le plus et le moins important) dans le cas où le répondant ne perçoit aucune différence d'importance entre les critères proposés, ou entre plusieurs critères jugés importants, ou encore entre plusieurs critères jugés peu importants ou non pertinents. Ainsi, nous l'avons évoqué, le répondant sera sommé de choisir entre le goût et l'emballage, même si pour lui le choix ne se présente pas de cette manière là. Cette limite rejoint en partie celle qui est faite à propos des méthodes

d'analyse conjointe dans laquelle certains choix peuvent s'avérer aberrants. Elle est cependant compensée par le fait que le nombre de situations de choix qui est proposé offre une diversité dans laquelle la totalité des répondants doit pouvoir se retrouver. Contrairement à l'analyse conjointe, les attributs ne sont pas présentés dans une logique de construction d'offre.

CONCLUSIONS

A partir de la littérature, nous avons comparé dans cet article les performances de la méthode « Best-Worst », encore appelée « Maxdiff », à d'autres méthodes d'enquêtes, fondées soit sur des notations sur des échelles numériques, soit sur des comparaisons entre alternatives. Nous avons utilisé cette méthode pour déterminer si les notions d'origine (pays d'origine, région d'origine, variété d'olives) constituaient des critères importants du choix de l'huile d'olive chez les consommateurs en France et en Tunisie.

Nos résultats montrent qu'il existe des alternatives aux mesures classiques obtenues à partir d'échelles de Likert notamment pour améliorer le pouvoir discriminant des résultats et contourner les biais liés à l'appréciation différente des intervalles proposés par ces échelles. Compte tenu de la facilité de mise en œuvre, de sa capacité à hiérarchiser les attributs de choix des consommateurs, la méthode Best-Worst mérite tout particulièrement d'être prise en compte dans les études consommateurs par lesquelles les chercheurs et décideurs cherchent à comprendre les arbitrages que font les consommateurs entre les attributs d'un produit, et à segmenter les marchés en fonction des avantages recherchés.

Avec beaucoup de facilité, la méthode Best-Worst nous a permis de hiérarchiser les attributs de préférence de l'huile d'olive, et de comparer les résultats entre groupes différents. Les résultats issus de cette méthode sont par ailleurs cohérents avec des résultats antérieurs, obtenus par des méthodes qualitatives.

REFERENCES BIBLIOGRAPHIQUES

Bartikowski B., Chandon J-L., Gierl H. (2006). *Calibration internationale des échelles sémantiques*, Décisions Marketing, n°43-44, p : 207-219.

Chandon J-L., Bartikowski B. (2004). *Une échelle ordinale permettant de classer les répondants en "satisfait", "indifférent" et "insatisfait"*, Recherche et Applications en Marketing, n°.19, p : 39-51.

Chrzan K., Golovashkina N. (2006). *An empirical test of six stated importance measures*, International Journal of Market Research, v°48, n°6.

Cohen S. (2003). *Maximum Difference Scaling: Improved Measures of Importance and Preference for Segmentation*, Sawtooth Software Conference Proceeding, Sequim (WA).

Dekhili S., d'Hauteville F. (2006). *Les dimensions perçues de l'image de la région d'origine. Cas de l'huile d'olive*, 22ème Congrès de l'Association Française de Marketing (AFM), Nantes, France, 11-12 Mai, 19 p.

Finn A., Louviere J-J. (1992). *Determining the Appropriate Response to Evidence of Public Concern: The Case of Food Safety*, Journal of Public Policy and Marketing, n°11, p: 12-25.

Fishbein M. (1967). *A behaviour theory approach to the relations between beliefs about an object and the attitude towards the object*, Readings in attitude theory and measurement, New York: Wiley, p: 389-399.

Goodman S., Lockshin L., Cohen E. (2005). *Best-Worst Scaling: A simple Method to Determine Drinks and Wine Style Preference*, 2nd International Wine Marketing and Business Conference, Sonoma (USA).

Lodge M. (1981). *Magnitude scaling: quantitative measurement of opinions*, Sage University paper on quantitative applications in social sciences, Beverkly Hills, Sage Publications, p: 7-25.

- Louviere J-J., Woodworth G-G. (1990). *Best Worst Scaling : a model for largest Difference Judgements*, working paper, faculty of Business, University of Alberta.
- Marder E. (1997). *The laws of Choice : predicting consumer behavior*, N.Y, the Free Press.
- Marley A-A-J., Louvière J-J. (2005). *Some probabilistic Models of Best, Worst and Best Worst choices*, Journal of mathematical psychology, n°49, p: 464-480.
- Rao P-V. (1960). *The dual of balanced incomplete block design*, The Annals of Mathematical Statistics, p: 779-785.
- Remaud H., Lockshin L. (2009). *Building brand salience for commodity-based wine regions*, International Journal of Wine Business research, n°21.
- Stephenson W. (1953). *The study of behavior: the Q-Technique and its methodology*, Chicago, University of Chicago Press.
- Wilkie W-L., Pessemier E-A. (1973). *Issues in marketing's use of multi attribute attitude models*, Journal of Marketing research, n°10, p: 428-441.