Use of Check-all-that-apply (CATA) and Penalty Analysis for Product Development Guidance: A Case Study with Mexican-Style Sauces

by

Sheri Lynn Gordon

B.S., University of Cincinnati, 1993
M.S., University of Cincinnati College of Medicine, 1998

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Food Science
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2019

Approved by:

Major Professor
Edgar Chambers IV

Copyright

Abstract

In the industrial setting, one common inclusion in consumer testing is the measurement of consumer reaction to specific product attributes to provide product developers guidance for optimization. Just about right (JAR) scales are a popular choice for obtaining this re-formulation guidance when combined with overall liking to determine the potential impact of responses to overall acceptance. Since 2007, the practice of using Check-all-that-apply (CATA) in various consumer applications has steadily increased. CATA has been used successfully with consumers to understand the sensory characteristics of products. With the increasing popularity of the tool, researchers are studying new ways to analyze CATA data, including applying penalty analysis. However, scant literature exists to compare the outcomes from using JAR scales and CATA for guiding product development. In the present work, two approaches for eliciting hedonic and attribute responses from consumers were compared: one using Just About Right (JAR) scales for rating key attributes and the other Check-All-That-Apply to identify consumer likes and dislikes. Two separate sets of consumers evaluated the same 4 Mexican-style sauce products using one of the two methods in a central location test. Penalty analysis was used to help analyze the data and the interpretation of each data set and the subsequent guidance recommendations for product development were compared. There is a striking agreement between the penalties obtained from each method. Some advantages and disadvantages of both question types are discussed. Thus, intensity-rated CATA questions may have potential use as an alternative to JAR questions for providing formulation optimization guidance within the food manufacturing industry.

Table of Contents

List of Figures

List of Tables

Acknowledgements

The author wishes to thank the Campbell Soup Company for their support of this research, specifically for providing the sauce samples. We also wish to extend our thanks to Dr. Hal Macfie for providing expert guidance regarding data treatment.

Dedication

To my dear husband and best friend, Dan. Thank you for being my biggest cheerleader. To Emma and Sarah, who have survived on SpaghettiO's and pizza while I studied. And to Kathryn, you inspire me to be the best that I can be.

Chapter 1 - Introduction

Within the industrial setting, product design and development processes are vital activities. According to a Nielsen statistic, more than 85% of new products in the Consumer Packaged Goods, CPG, industry fail in the marketplace [1]. For some products, one factor involved is the development with less than optimal attributes. Consumer reaction, gathered during the course of product development, is critical in the identification of promising prototypes and direction for further improvement to increase product performance in-market [2]. Traditionally, liking scales such as the 9–point hedonic scale are used to gauge consumer acceptability [3]. In combination with descriptive analysis, such information can provide tremendous value to the developer [4]. Because of time and financial constraints, researchers have been looking for short cuts to traditional testing. One alternative to classical sensory profiling has been to have the consumer rate liking and provide information on the product's attributes at the same time. The Just About Right (JAR) scale in which consumers are asked to rate whether the attribute in question is too low, too high or just about right, has been one of the most popular techniques to acquire attribute perception directly from the consumer. Just About Right scales are bipolar in nature with a midpoint of "Just About Right" and measure the deviation from the consumer ideal levels for each attribute [5]. In the context of product development, it is important for the researcher to determine which attributes impact consumer liking the most when they are no longer "JAR". When combined with overall liking, the potential impact of responses that deviate from JAR can be analyzed through penalty analysis and those attributes with the highest negative effects on liking are determined. Those attributes are then adjusted by the product developers before re-testing with the target consumer. However, in some cases, JAR scales have been shown to provide incorrect direction to the developers [6].

There is a strong interest in the industrial sector to gain consumer information using new methods while decreasing cost and increasing speed to market. To adapt to the current business environment, there is an ongoing attempt to reduce development time, decrease expense and to enhance product quality through new methods to gain this critical information. Although CATA was being used in the industry many years prior, it didn't appear in the scientific literature until much later. The use of CATA, in which consumers are presented with a list of descriptors to check all that apply, has increased in various consumer applications including usage occasions, emotions, and sensory characteristics. CATA questions were used in consumer research to make it easier for consumers to respond providing terms rather than have consumers generate their own open-ended responses [7]. In one of the earliest published studies comparing CATA and Forced Choice questions, Rasinski et al. (1994) showed that the mean number of items marked "yes" in a Forced Choice questionnaire is higher than the mean number of items checked in the CATA questionnaire for the same three questions [8]. Research on the use of CATA in sensory research has demonstrated the success of this method in sensory characterization of products, providing similar data to other accepted methods [9-13]. However, several differences were identified between CATA and JAR in that CATA questions tended to identify fewer deviations from ideal when sample differences were small [14].

The aim of this paper was to assess CATA as an alternative to JAR questions by comparing sensory insights (optimization recommendations for product developers) from CATA-generated penalty analyses of the products and JAR-generated penalty analyses. It was hypothesized that CATA might be an alternate way of collecting attribute information from consumers which can then be used to advise product refinement activities.

Chapter 2 - Materials and Methods

2.1 Mexican-style Sauces

Mexican-style sauces were chosen for the product set tested because this category has a high use rate in the geographical regions where testing was conducted. To decrease variability among the studies, the same product lots were used for both methods. Currently marketed Mexican-style sauces were sourced commercially and the pilot plant-produced prototypes were supplied by the Campbell Soup Company. Prototype samples were formulated with different flavor profiles from the company's current product in an attempt to more closely match competitive product. Four jars of each Mexican-style sauce were combined into a 3-qt bowl labeled with a 3-digit blinding code and were held and served at a typical consumer consumption temperature. Sauces were held for a maximum of one hour, stirring regularly. Two ounces of sauce was served over three tortilla chips (Tostitos Scoops™, Frito Lay, Plano, TX) purchased locally. The chips and sauce were served in a 2-oz. Styrofoam bowl labeled only with a unique three-digit code.

2.2. Participants

Consumers were recruited in Dallas,TX (n=147) and Olathe, KS (n=141) by each test facility using a database of local consumers. Category users between the ages of 25-54 and responsible for most of their household shopping were recruited. Inclusion criteria also included other socio-demographic parameters. Participants had to pass a standard security screen of not working for a food or beverage-related company nor in marketing or sensory research. Informed consent was obtained for each participant and respondents were compensated for their participation. Studies included approximately 55% females and 46% males.

2.3 Experimental Designs

Each consumer evaluated all 4 samples which were served in sequential monadic fashion using a balanced serving order. In study 1, consumers in the Dallas, TX location evaluated the samples using liking and CATA. The order of CATA terms was randomized. The attributes selected as responses included those terms that differentiated the product set as determined by the research team and included those attributes considered to influence liking for the Mexican-style sauce category. CATA questions were divided into Likes and Dislikes both for Appearance and Flavor (See Table 1). Olathe, KS consumers in study 2 answered 5 JAR questions regarding flavor attributes in addition to the 9-pt hedonic scales used in study 1. A response to each JAR question was required by each consumer. The JAR questions were separated by modality and presented in a fixed order (See Table 1).

CATA Questionnaire N=147 Dallas, TX	
Overall Liking	
Appearance Appearance Likes - CATA 7 attributes Red Color Orange Color Light Color Dark Color Chunky Appearance Smooth Appearance Amount of Visible Onions Appearance Dislikes - CATA 10 attributes Too Much Red Color Too Much Orange Color Too Little Red Color Too Little Orange Color Too Light in Color Too Dark in Color Too Chunky Too Smooth Too Much Visible Onions Too Little Visible Onions	**Flavor** Flavor Likes – CATA 7 attributes Type of Pepper Flavor Type of Tomato Flavor Amount of Pepper Flavor Amount of Tomato Flavor Amount of Spiciness Amount of Tartness Amount of Saltiness Flavor Dislikes – CATA 10 attributes Too Strong Pepper Flavor Too Strong Tomato Flavor Too Weak Pepper Flavor Too Weak Tomato Flavor Too Spicy Not Spicy Enough Too Tart Not Tart Enough Too Salty Not Salty Enough.
JAR Questionnaire N=141 Olathe, KS	
Overall Liking	
Appearance Liking Appearance JAR - 5 attributes Color Too Light/Dark Color Red Not Enough/Too Much Color Orange Not Enough/Too Much Appearance Too Smooth/Too Chunky Visible Onion Too Little/Too Much	**Flavor Liking** Flavor JAR – 5 attributes Pepper Flavor Too Weak/Too Strong Tomato Flavor Too Weak/Too Strong Spiciness Not Enough/Too Much Tartness Not Enough/Too Much Saltiness Not Enough/Too Much

Table 2.1 Overall Questionnaire Design

2.4 Data Analysis

Overall Liking data was analyzed using XLSTAT (ANOVA, 90% confidence interval – the confidence level used by the company). The frequency of use for each CATA term was

calculated by counting the number of times a consumer chose that term. Data was collected electronically using Compusense at-hand (Compusense Inc., Guelph, Ontario) and analyzed using XLSTAT-MX 2015.2 (Addinsoft, New York, NY) to produce Penalty Charts for each data set [15-16].

Chapter 3 - Results

3.1 Hedonic Scores

Significant differences in Overall Liking scores (see Table 2) were found amongst the product set with market product 1 having the highest overall liking score in both studies. Regardless of question type, product liking followed the same ranking (market product 1> market product 2> prototype 1> prototype 2). Mean overall liking scores were significantly higher when presented with CATA/RATA questions than with JAR. It is hypothesized that JARs could cause consumers to think about sensory attributes not usually inherent to their decision process [16] and therefore change their hedonic perception [6,17] . There are examples in the literature on the effect of JAR on hedonic scores [17-19] and conversely those finding no effect [20-22]. Ares *et al.* (2017), when comparing 6 food studies, found that only 2 demonstrated decreases in overall liking when JARs were used with or without concurrent CATA [14]. Similar findings on CATA questions reveal that they are not as likely to cause hedonic bias [23-27] Multiple studies conducted by Jaeger *et al.* (2017) demonstrated lower hedonic scores when elicited with RATA, a "Rate All That Apply Scale" where participants select the attributes using a CATA method, then rate the intensity of each of the selected attributes. Scores in this study declined at a similar level for each question type when first position overall liking was compared to subsequent positions. There was also no difference in completion time between CATA and JARs [28].

Sample Code	Mean Overall Liking (CATA)	Mean Overall Liking (JAR)
158 (prototype 1)	6.0 b	5.8 b
323 (marketed product 1)	7.2 a	6.8 a
766 (prototype 2)	5.5 c	4.8 c
991 (marketed product 2)	6.4 b	6.4 a

Table 3.1 Overall Liking Scores
Statistical comparisons were conducted at the 90% confidence level, two-tailed

3.2. Penalty Analysis

Plaehn (2012) proposed a means of analyzing CATA data which was similar to that used for JAR responses to understand attribute importance on liking [29]. Penalty analysis was conducted on consumer responses to compare mean drop/increase in overall liking. JAR responses were collapsed into three categories and CATA data treated as a binary response. The criteria for CATA high penalties were above 20% and a less than a -1.5 point mean impact. Considering the CATA data in Figure 1, while there are no high priorities for reformulation, increasing the intensity of Pepper Flavor and Tomato Flavor could be considered by the developer for rework. Figure 2 identifies Pepper Flavor Too Weak and Too Tart as high penalties in prototype 2. The product developer might consider increasing Pepper Flavor and decreasing Tartness. Color Darkness was also identified as a moderate penalty for this prototype based on CATA penalty analysis. A moderate penalty is the percent respondents between 19% and 20% and Mean Drop between -1 and -1.5.

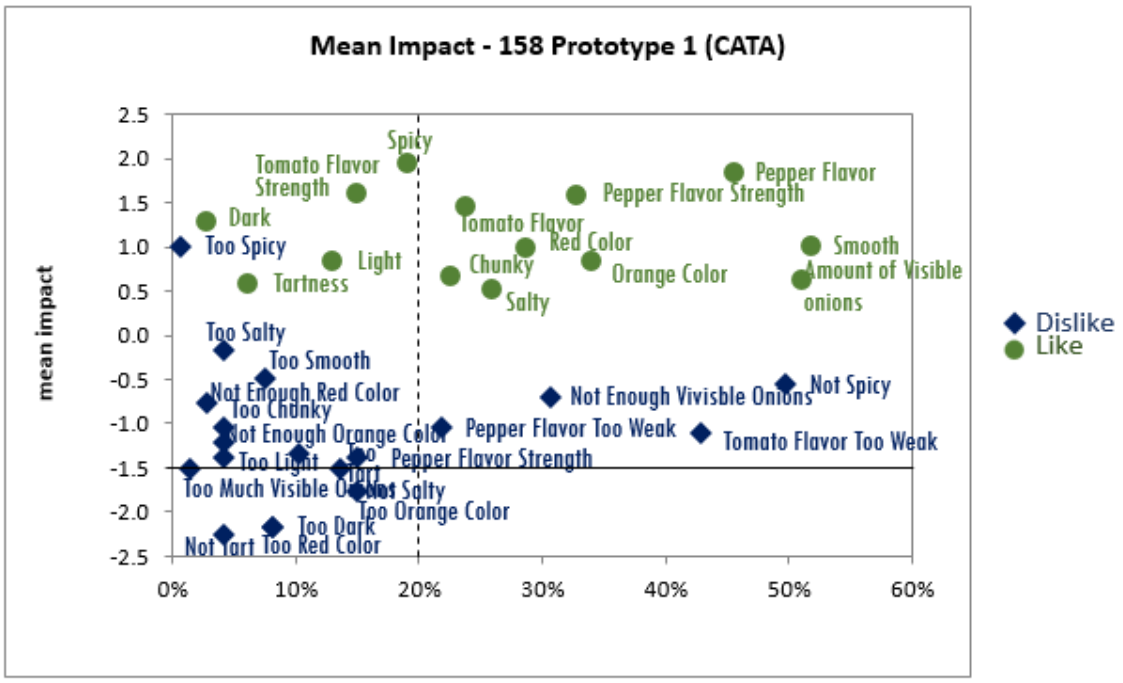

Figure 3.1 Mean overall liking impact for prototype 1 as a function of consumers that checked the attribute. Attributes labeled 'like' originated from the two CATA Appearance and Flavor like questions.

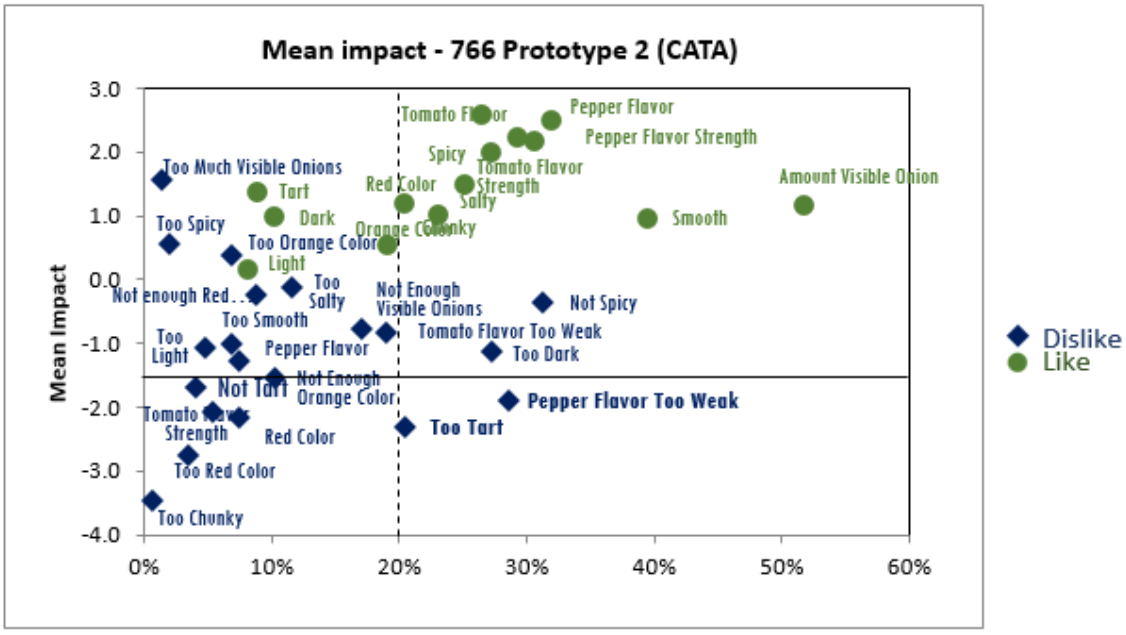

Figure 3.2 Mean overall liking impact for prototype 2 as a function of consumers that checked the attribute. Attributes labeled 'like' originated from the two CATA Appearance and Flavor like questions.

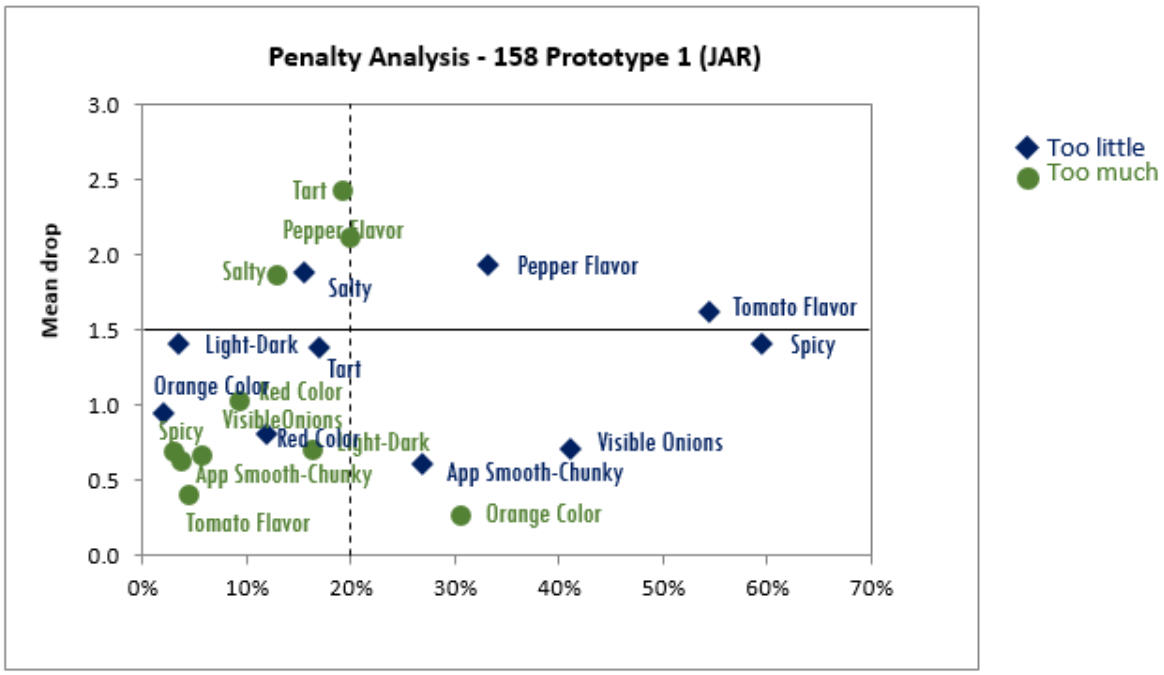

Figure 3.3 Penalty analysis for Prototype 1. The mean drop of interest for JARs is shown as a positive quadrant.

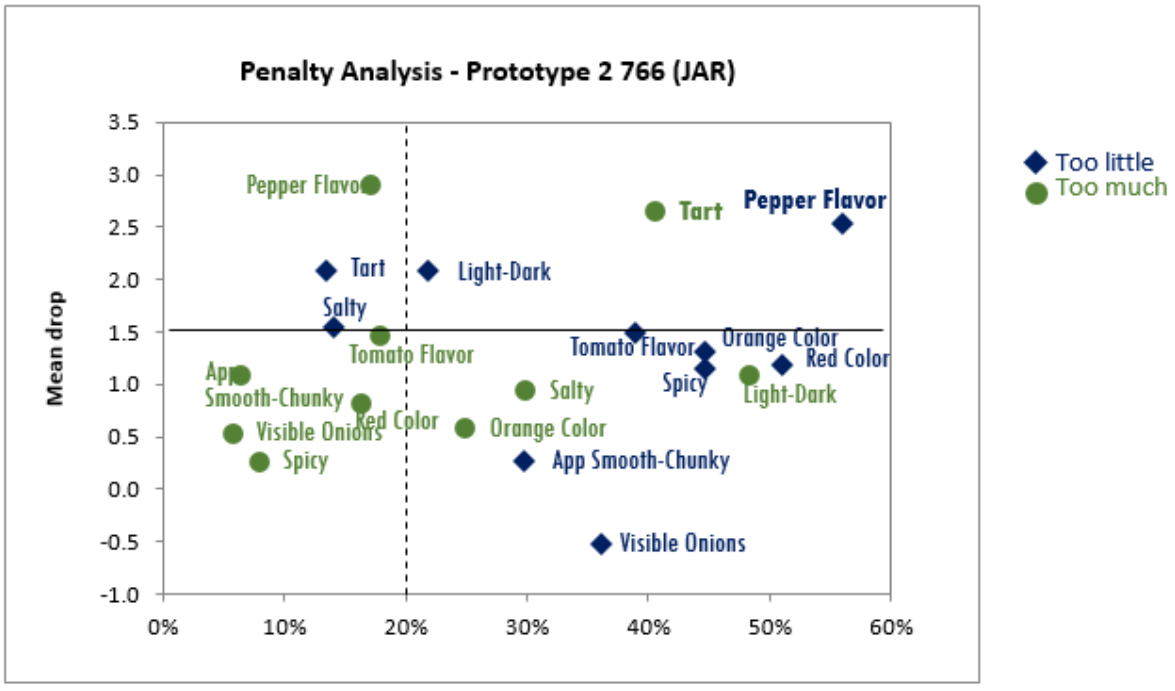

Figure 3.4 Penalty analysis for Prototype 2. The mean drop of interest for JARs is shown as a positive quadrant.

Recommendations for 158: Prototype 1	**CATA**	**JAR**
Increase Pepper Flavor	✓ moderate	✓ high
Increase Tomato Flavor	✓ moderate	✓ high
Increase Spiciness		✓ moderate
Decrease Tartness		✓ moderate

Recommendations for 766: Prototype 2	**CATA**	**JAR**
Increase Pepper Flavor	✓ high	✓ high
Decrease Tartness	✓ high	✓ high
Decrease Dark Color	✓ moderate	✓ high
Increase Tomato Flavor		✓ moderate
Increase Orange Color		✓ moderate
Increase Red Color		✓ moderate
Increase Spiciness		✓ moderate

Table 3.2 Comparison of CATA and JAR reformulation recommendations

3.3 Time to Complete Questions

The average time for consumers to complete each of the question types was not significantly different. The respondents did not spend significantly more time completing the CATA questions versus the JARs even though the eleven JAR questions each required a response. It appears that the participants were more thoughtful on the responses to the Like questions as the time to complete (27 sec) was longer than for Dislike questions (19 sec). There was no measure in this study of the panelists reading of all CATA responses in their entirety before moving on to the next question, however, previous work on visual attention to CATA questions showed that more attention was given to each of the terms when a ‘short’ CATA list was used (12 terms) vs longer lists (20 terms) [30], The findings of this study agreed in that longer time was spent on the CATA

Like options whereas the CATA Dislike options took shorter time to answer. Both lists were relatively short, only differing by three terms so it is unlikely that list length was a factor

3.4 Comparison of Analyses

Both JAR and CATA penalty analysis are available in XL-STAT-Sensory (Addinsoft, New York, NY). Once familiar with the CATA analysis procedure, there was no advantage found with one method of analysis over the other.

3.5 Term Usage in CATA

Consumers checked less than 3 boxes on average per CATA question. The consumers appear to be selecting only the most salient terms. Vidal *et al.* hypothesized that CATA questions may be slightly less sensitive because they demand less attention to samples as they allow more unstructured responses [31]. In the present study, this phenomenon could be due to the nature of the JAR questions where the consumer is required to provide an answer to each JAR before proceeding, thus ensuring measurement of all 5 attributes. CATA like and dislike questions required one response for each of the likes and dislikes. Additionally, JAR questions were formulated by the developer on what is believed to be salient features of the product. Often these questions are developed without prior knowledge of the specific attributes that strongly influence consumer liking. Therefore, the ballot may contain JAR attributes which do differ among products but may not be relevant to liking. Jaeger *et al.* (2015) observed the improved ability of the consumer to identify sensory differences among samples when using JAR questions [32].

Concurrent use of JAR and CATA may also increase discrimination among the products [24]. In a review of 6 studies, Ares *et al*. found that JAR and CATA questions provided similar insights regarding the most important deviations from ideal [14]. The comparison also concluded that CATA questions identified fewer deviations than JAR questions especially when there were small differences among the samples [14]. This does not present a problem to the product developer as there is little practical relevance for attributes without moderate to high penalties (lower mean drop scores and fewer participant responses).

Chapter 4 - Conclusions

The present study confirmed that the reformulation guidance was the same using both CATA and JAR questionnaires. Both CATA and JAR identified the priority reformulation effects, such as increasing Tomato and Pepper Flavor in Prototype 1. JAR additionally identified increasing Spiciness and decreasing Tartness as possible reformulation targets. For Prototype 2, increasing Pepper Flavor, decreasing Tartness and decreasing Darkness were common reformulation suggestions from both JAR and CATA, with JAR additionally identifying penalties such as increasing Tomato Flavor, Orange Color, Red Color and increasing Spiciness as reformulation guidance. The CATA questionnaire provided participants the opportunity to indicate the attributes that most strongly influenced their dislike. Alternatively, the use of the JAR questionnaire required participants to consider and rate every attribute that could be influencing liking. While the amount of time was the similar for both methods and liking ratings were the same for both methods, participants using the JAR method rated fewer attributes as being "just about right", finding more fault in the intensity of the attributes, versus the CATA method. The key findings, however, from major JAR penalties and CATA penalties, were similar regardless of method providing the same optimization information. Although main opportunities are identified by each method are similar, there are differences dependent on method which should be acknowledged.

A limitation of the research was the use of two separate consumer populations in different locations for testing. Cost constraints dictated the use of campus facilities for the second study. While the same batch of prototype product was used in each test, these products were not plant produced as were marketed products. Future research on CATA feasibility should include more

complex products to pinpoint optimal number of attributes for CATA and how this differs from JAR.

Chapter 5 - References

1 Nielsen. Looking to Achieve New Product Success? Listen to Your Consumers. Available online:https://www.nielsen.com/content/dam/nielsenglobal/dk/docs/Nielsen%20Global%20New%20Product%20Innovation%20Report%20June%202015.pdf (accessed on 5 September 2018).

2 Talavera, M.; Chambers E. IV. Using sensory science helps product succeed. British Food Journal 2017, 119, 2130-2144.

3 Lawless, H.T.; Heymann, H. Sensory Evaluation of Food: Principles and Practices, 2nd ed.; Springer: New York, NY, USA, 2010. ISBN 978-14-4196-487-8.

4 Muñoz, A.M.; Chambers, E. IV.; Hummer, S.; A Multifaceted Category Research Study: How to understand a product category and its consumer response. J. Sens. Stud. 1996, 11, 261-291.

5 Rothman, L.; Parker, M.J.; Just About Right Scales: Design, Usage, Benefits, and Risks. ASTM Manual MNL63, 2009. ASTM International, Conshohocken, PA.USA.

6 Epler, S.; Chambers, E. IV; Kemp, K. Hedonic scales are a better predictor than just about right scales of optimal sweetness in lemonade. J. Sens. Stud. 1998, 13, 191-198.

7 Adams, J.; Williams, A.; Lancaster, B.; Foley, M.; Advantages and uses of check-all-that-apply response compared to traditional scaling of attributes for salty snacks. In Proceedings of the 7th Rose Marie Pangborn Sensory Science Symposium (Pangborn 2007); Minneapolis, MN., USA, August 12-16, 2007.

8 Rasinski, K.A.; Mingay, D.; Bradburn, N.M.; Do respondents really "mark all that apply" on self-administered questions? The Public Opinion Quarterly 1994, 58 (3), 400-408.

9 Bruzzone, F.; Ares, G.; Giménez, A.; Consumers' texture perception of milk desserts II-comparison with trained assessors' data. J Texture Studies. 2012, 43, 214-226.

10 Fleming, EE.; Ziegler, GR.; Hayes, JE.; Check All That Apply (CATA) Sorting and Polarized Sensory Positioning (PSP) with Astringent Stimuli. Food Qual and Pref. 2015, 45, 41-49.

11 Ares, G.; Varela, P.; Rado, G.; Giménez, A.; Identifying Ideal Products Using Three Different Consumer Profiling Methodologies. Comparison with External Preference Mapping. Food Qual and Pref. 2011, 22, 581-591.

12 Cadena, R.S.; Caimi, D.; Jaunarena, I.; Lorenzo, I.; Vidal, L.; Ares, G.; Deliza, R.; Giménez, A.; Comparison of Rapid Sensory Characterization Methodologies for the Development of Functional Yogurts. Food Res Intl. 2014, 64, 446-455.

13 Dooley, L.; Lee, Y.S.; Meullenet, J.F.; The Application of Check-all-that-apply (CATA) Consumer Profiling to Preference Mapping of Vanilla Ice Cream and its Comparison to Classical External Preference Mapping. Food Quality and Preference 2010, 21, 394-401.

14 Ares, G.; de Andrade, J.C.; Antuñez, L.; Alcaire, F.; Swaney-Stueve, M.; Gordon, S.; Jaeger, S.R.; Hedonic product optimization: CATA questions as alternatives to JAR scales. Food Qual. Pref. 2017, 55 , 67-78.

32 Vidal, L.; Ares, G.; Hedderley, D.I.; Meyners, M., Jaeger, S.R.; Comparison of rate-all-that-apply (RATA) and check-all-that-apply (CATA) questions across seven consumer studies. Food Quality and Preference 2018, 67, 49-58.

15 Meyners, M.; Testing for differences between impact of attributes in penalty-lift analysis. Food Quality and Preference 2016, 47, 29-33.

16 Meyners, M.; Jaeger, S. R.; Ares, G.; On the Analysis of Rate-All-That-Apply (RATA) Data. Food Quality and Preference 2016, 49, 1-10.

17 Popper, R.; Rosenstock, W.; Schraidt, M.; Kroll, B.J.; The effect of attribute questions on overall liking ratings. Food Quality and Preference 2004, 15, 853-858.

18 Earthy, P.J.; Macfie, H.J.H.; Hedderley, D.; Effect of question order on sensory perception and preference in central location trials. Journal of Sensory Studies 1997, 12, 215-237.

19 Prescott, J.; Lee, S.M.; Kim, K-O.; Analytic approaches to evaluation modify hedonic responses. Food Quality and Preference 2011, 22, 391-393.

20 Mela, D. J.; A comparison of single and concurrent evaluation of sensory and hedonic attributes. Journal of Food Science 1989, 54, 1098–1100.

21 Vickers, Z. M.; Christensen, C. M.; Fahrenholtz, S. K.; Gengler, I. M.; Effect of the questionnaire design and the number of samples tasted on hedonic ratings. Journal of Sensory Studies 1993, 8, 189–200.

22 Gacula, M. Jr.; Mohan, P.; Faller, J.; Pollack, L.; Moskowitz, H.R.; Questionnaire Practice: What happens when the JAR scale is placed between two ” overall” acceptance scales? Journal of Sensory Studies 2008, 23, 136-147.

23 Ares, G.; Jaeger, S.R; Examination of sensory product characterization bias when check-all-that-apply (CATA) questions are used concurrently with hedonic assessments. Food Quality and Preference 2015, 40, 199-208.

24 Jaeger, S.R.; Ares, G; RATA questions are not likely to bias hedonic scores. Food Quality and Preference 2015, 44, 157-161.

25 Jaeger, S.R.; Ares, G.; Lack of evidence that concurrent sensory product characterization using CATA questions bias hedonic scores. Food Quality and Preference 2014, 35, 1-5.

26 Jaeger, S.R.; Giacolone, D.; Roigard, CM.; Pineau, B.; Vidal, L.; Gimenez, A. ; et al ; Investigation of bias of hedonic scores when co-eliciting product attribute information using CATA questions. Food Quality and Preference 2013, 30, 242-249.

27 Jaeger, S.R.; Hunter, D.C.; Kam, K.; Beresford, M.K.; Jin, D.; Paisley, A.; Chheang, S.L.; Roigard, C.; Ares, G.; The concurrent use of JAR and CATA questions in hedonic scaling is unlikely to cause hedonic bias but may increase product discrimination. Food Quality and Preference 2015, 44, 70-74.

28 Jaeger, S.R.; Kim, K-O.; Lee, S.M.; Hunter, D.C.; Kam, K.; Chheang, S.L.; Jin, D.; Lee, P-Y.; Xia, Y.; Ares, G.; Concurrent elicitation of hedonic and CATA/RATA responses with Chinese and Korean Consumers; Hedonic bias is unlikely to occur. Food Quality and Preference 2017, 56, 130-137.

29 Plaehn, D.; CATA penalty/reward. Food Quality and Preference 2012, 24, 141-152.

30 Ares, G.; Antuñez, L.; Giménez, A.; Jaeger, S.R.; List length has little impact on consumers visual attention to CATA questions. Food Quality and Preference 2015, 42,100-109.

31 Ares, G.; Alcaire, F.; Antunez, L.; Vidal, L.; Giménez, A.; Castura, J.C.; identification of drivers of (dis)liking based on dynamic sensory profiles: Comparison of Temporal Dominance of Sensations and Temporal Check-all-that-apply. Food research International 2017, 92, 79-87.