

Survey Design & Implementation

MSDS 402

Marketing Team:

Maria Aintablian

Michelle D'Agostino

Edward Chao

Yining Feng

Business Problem

The global grocery stores & supermarkets industry, excluding discount stores and warehouse clubs, has annual sales in excess of \$4 trillion and is expected to grow 6% per year driven by population growth and rising income of developing countries. Although it is a crowded industry, it is largely led by local players with major global retailers accounting for 30% of the market. At a national level, this industry is slightly more concentrated with 20 retailers sharing approximately 67% of the market.

The industry is changing with disruptions from online competitors and the rise of discounters. Although online grocery sales accounts for less than 6% of total sales today, this is quickly changing. "Online sales are expected to exceed 20% of the U.S. grocery market by 2025, growing to more than \$100 billion." (Aronson, 2018). With consumers developing a preference for online shopping, grocers are looking into increasing their online presence and adding social media to their marketing programs. Developing a digital presence in these early stages of migration will likely provide a competitive advantage. However, today little is understood about specific aspects of social media marketing and its impact on the likelihood of customers to shop online.

Our objective is to understand the role that social media marketing plays in shaping the decisions made by online grocery shoppers. Survey questions will be created to investigate how social media influenced shoppers to purchase groceries such as staple pantry items, specialist ingredients, beverages, personal care products, fresh produce, snacks, meats, fish, and deli products online.

Hypothesis Statements

The dependent variable is the influence of social media marketing on grocery consumers' decision to shop online, whereas independent variables entail factors such as grocery consumers' selection of various social media platforms, characteristics of social media platforms used by online grocery shoppers, their online shopping motives as well as ratings of their latest online shopping experience. Overall, we want to investigate the impact of all the independent variables on the dependent variable.

Null Hypothesis (H_{o1}): Following grocery brands on social media has no impact on whether a grocery consumer shops online.

Alternative Hypothesis (H_{a1}): Following grocery brands on social media has an impact on whether a grocery consumer shops online.

Null Hypothesis (H_{o2}): Social media marketing campaigns have no impact on whether a grocery consumer shops online.

Alternative Hypothesis (H_{a2}): Social media marketing campaigns have an impact on whether a grocery consumer shops online.

Null Hypothesis (H_{o3}): Product information/reviews available on social media platforms have no impact on whether a grocery consumer shops online.

Alternative Hypothesis (H_{a3}): Product information/reviews available on social media platforms have an impact on whether a grocery consumer shops online.

Null Hypothesis (H_{o4}): Coupons/promotions available on social media platforms have no impact on whether a grocery consumer shops online.

Alternative Hypothesis (H_{a4}): Coupons/promotions available on social media platforms have an impact on whether a grocery consumer shops online.

Null Hypothesis (H_{o5}): Social media contacts/network have no impact on whether a grocery consumer shops online.

Alternative Hypothesis (H_{a5}): Social media contacts/network have an impact on whether a grocery consumer shops online.

Null Hypothesis (H_{o6}): Celebrity endorsement on social media have no impact on whether a grocery consumer shops online.

Alternative Hypothesis (H_{a6}): Celebrity endorsement on social media have an impact on whether a grocery consumer shops online.

Null Hypothesis (H_{o7}): A specific social media platform or channel has no impact on whether a grocery consumer shops online.

Alternative Hypothesis (H_{a7}): A specific social media platform or channel has an impact on whether a grocery consumer shops online.

Sampling Plan

The target population for the survey are U.S. customers over 18 years old who have made an online purchase from a grocer within the last 12 months. Per a survey of online grocery shoppers published in August 2017 by Gallup, there were wide discrepancies among online grocery shoppers based on age, region, location, employment status and income. (Newport & Brennan, 2017) Thus, a stratified sampling method will be used to eliminate sampling bias and minimize sampling errors by dividing the target population into strata based on age, region, and location, the top three demographic characteristics noted by Gallup.

- The sampling pool will be first divided into regions (e.g. East, West, Midwest, South) to eliminate sampling errors on geographic regions.
- Then, these subpopulations will be further divided into four locations (e.g. City, Suburbs, Town and Rural area) to eliminate sampling errors on location.

- Lastly, the subpopulations will be divided into four subpopulations based on age (e.g. 18-29, 30-49, 50-64, 65+) to eliminate sampling errors on age.

Target survey participants will be sourced from a list broker such as ExactData or infoUSA. An email will be sent to target participants requesting them to complete our short online survey. We aim to obtain a minimum of 30 qualifying responses from each subpopulation or 1,920 total qualifying responses between December 1, 2019 and March 31, 2020. This will allow us to achieve a 95% confidence interval with a reasonable margin of error ($\leq \pm 2.5\%$) on an aggregated national level, along with a 95% confidence level (margin of error $\leq \pm 5\%$) for the regional strata, a 95% confidence level (margin of error $\leq \pm 5\%$) for the local area strata, as well as a 95% confidence level (margin of error $\leq \pm 5\%$) for the age group strata. Please see Appendix A for a matrix of the sampling subpopulations.

We expect a 30% response rate of all individuals contacted to participate in our survey which is based on the average online survey response rate (Kaplowitz, Hadlock, & Levine, 2004). To achieve 30 completed surveys per cell, we will solicit 110 people per cell or 7,040 in total. Although incentives would likely increase response rates, they will not be provided for the purpose of avoiding incentive bias.

As we are targeting customers that have made purchases online over the last 12 months, we assume that the customers have access to the internet and have an email address. The data collection process for this survey will use a mix-mode approach where an email will be sent to targeted participants that invites them to complete a web-based self-administered questionnaire. Once at the survey site, consumers will be asked to respond to 21 questions. To ensure a positive experience for survey participants and to ensure data quality, the survey structure will incorporate filtering and branching to the extent possible. For example, survey participants will be asked if they have already completed the survey. If they have, they will be advanced to a “thank you” screen. If they have not, additional screening questions will be presented to guide the participants through the remainder of the survey.

To minimize response bias, response options are grouped, or the participant is offered an alternative response option. For example, rather than asking a participant to provide their specific age, they are asked to identify their age group. And, to determine the geographic location of a participant, they are asked to provide either their zip code or their state in case they don’t feel comfortable in providing their zip code. The Survey Instrument that will be administered is included in Appendix B for reference.

Data Plan

Sampling Bias

We are using a convenience sampling methodology which impacts the randomness of our sample and its projectability to all online shoppers. By sourcing survey participants from a list broker such as ExactData or infoUSA, we may be introducing coverage bias as there is no way to determine if the third-party lists are equally comprehensive across all regions and demographic groups. To minimize such bias, we are using a stratified sampling methodology and randomly

selecting participants within the top three demographic characteristics (e.g. age group, region, and local area) which were noted by Gallup as having wide discrepancies among online grocery shoppers. We chose to use the stratified sampling methodology because such a sample will have no sample bias, as well as no sampling error on the variables used for stratifying. [Nice rewrite]

Survey Bias

Our research uses a mix-mode data collection approach where an email will be sent to targeted participants and they will be asked to complete a self-administered web-survey. Self-administered surveys have the following advantages and disadvantages:

Advantages:

1. Reduced cost as a result of not having any expenses related to hiring and training interviewers
2. Elimination of interviewer bias since the survey will be self-administered
3. Reduced response bias given the increased privacy with use of a self-administered survey

Disadvantages:

1. Lower response rates expected with self-administered surveys
2. Potential for question interpretation bias

While we have taken steps to reduce non-response bias, it is still expected to persist. To determine if non-response is an issue, we will compare response rates across all subgroups as surveys are completed. For example, we will compare the response rates of respondents aged 18-29 living in cities in the East with those aged 18-29 living in cities in the West. If we detect any issues, we will evaluate options and make necessary corrections such as increasing solicitations for the impacted cell to ensure we obtain a sufficient sample for that cell or clarifying the survey question.

Additionally, cognitive biases of survey respondents may lead to question interpretation bias as well as recall bias, thereby having an impact on accuracy. Question interpretation bias may arise from the way individual questions are designed, the way the questionnaire as a whole is designed, and how the questionnaire is administered or completed. To address interpretation bias, questions were simplified using easy to understand language and responses are mutually exclusive which also provide additional context. Clarifying instructions and/or examples are also provided where needed. And, key terms appear when respondents hover over certain words. In terms of recall, we are relying on survey participants to recall their online shopping experience and motives in the past 12 months. We cannot know definitively whether some survey respondents could be systematically misreporting their feedback, concern, and experience. Similarly, we cannot know definitively whether some survey respondents fail to provide their honest answers or feedback because they are rushing to get their survey done.

Cleaning and Coding

Before the online survey is made available and administered to participants, it will be tested to ensure that there are no programming errors. Assuming we have collected a sufficient number of responses by March 31, 2020, the data will be compiled into a single data source and processed as follows:

- Cleaning
 - Zip Codes and States will be reviewed for accuracy, when provided, and translated into a geographic region
- Coding – Non-numeric responses will be coded to numeric data as follows:
 - Non-responses will be coded as “9”
 - Responses based on an agreement scale will be standardized and coded as follows:
 - Strongly Agree = 5
 - Agree = 4
 - Neutral = 3
 - Disagree = 2
 - Strongly Disagree = 1
 - Yes / No responses will be standardized and coded as follows:
 - Yes = 2
 - No = 1
 - Geographic Regions will be coded as follows:
 - East = 1
 - West = 2
 - Mid-West = 3
 - South = 4
 - Locations will be coded as follows:
 - City = 1
 - Suburbs = 2
 - Town = 3
 - Rural Area = 4

In the event that we have missing values and data, we will use a regression imputation model. Regression imputation fits a statistical model on a variable with missing values. Predictions of this regression model are used to substitute the missing values in this variable. Relationships of X and Y (i.e. correlations, regression coefficients etc.) are preserved, since imputed values are based on regression models. This is a big advantage over simpler imputation methods such as mean imputation or zero substitution.

Analysis

After data has been cleansed, edited, and coded, we will begin analysis to determine if we should reject or accept our hypotheses. Simple linear regression will be used to qualify whether or not our dependent variable, social media, has any correlation to any of our independent variables (celebrity endorsements, reviews, promotions, etc.). In other words, we want to take a look at whether celebrity endorsements, recommendations, and/or promotions, etc. have any effect on shopping online. A t-test will be conducted to determine whether our sample falls within the 95%

confidence interval or outside of it, so we can ultimately decide to reject or accept our hypotheses.

Our analysis begins by obtaining a set of parameters that are specific to the stratified sampling method. To start, we would calculate the population parameter, represented by the following equation below, where x is the sample estimate of the mean of the population, N_h is the number of observations, h is the stratum of the population, and x_h is the mean score from the sample in h .

$$x = \sum (N_h / N) * x_h$$

Next, we use the following equation below to estimate the population variance, where s^2_h is the sample estimate of the variance of the population in h , x_{ih} is the i th value of h , and n_h is the number of sample observations from h .

$$s^2_h = \sum (x_{ih} - x_h)^2 / (n_h - 1)$$

Then, we compute the standard error:

$$SE = (1 / N) * \sqrt{\sum [N_h^2 * (1 - n_h / N_h) * s^2_h / n_h]}$$

Based on our confidence intervals of 95%, our significance level of committing a Type I error is equal to 0.05. And as we use two-tailed hypothesis, we will use the z-score with a cumulative probability equal to 0.975. To find the region of acceptance of our hypotheses, we need to find the upper and lower limits of the region by using the standard error and critical value together. If our estimate of the parameter does not lie within the region of acceptance, then we have no choice but to reject the null hypothesis. If, however, our estimate does fall within, then we do not reject the null hypothesis.

[My comment on draft #3 likely drove you to eliminate the discussion about clustering. I meant for you to add the discussion on analysis of hypothesis to your discussion on clustering. I think the clustering of results – if significant – could add greatly to grocers understanding of how to influence different online shopper segments. So, I would do both the clustering and the regression analyses.]

Conclusion [Great wrap-up. The inclusion of cost and timeline really adds to your proposal quality.]

Once the analysis has been completed, we will determine the best format to share the results (i.e. white paper, infographic, presentation, blog post, webinar, etc.) Within this report, we will highlight the current industry problem and survey results for if, and how, social media marketing shapes the decisions made by online grocery shoppers.

Conducting this research will provide us with the opportunity to influence social media marketing strategy for grocery companies, expand brand presence and reach within the grocery

industry, establish grocery company as an industry leader, and help differentiate them from other competitors, which will ultimately in turn increase online grocery sales.

The budget for purchasing consumer contacts and conducting this research will be approximately \$11.7K, with the following breakout:

- \$704 – this is the cost of purchasing 7,040 contacts from Info USA. There could potentially be up to a 50% discount on this price if we only purchase email lists
- \$10,940 – this is the cost of conducting our research via Survey Monkey. This will generate a total of 1,920 completed responses for a 21-question survey

Please reference the following high-level timeline of the next steps needed in order to implement and conduct the survey, should we decide to move forward with this research study.

To-Do Item	Nov	Dec	Jan	Feb	March	April	May	June
Research Approval								
Reeceive approval to move forward								
Planning & Design								
Recruit and train research team								
Finalize survey questions and implementation								
Data Collection								
Conduct survey								
Data Analysis								
Data clean up								
Conduct analysis								
Shareout of final report								

Bibliography:

- Aronson, G. (2018, January 29). FMI and Nielsen: 70% of Consumers Will be Grocery Shopping Online by 2024. Retrieved November 4, 2019, from <https://www.nielsen.com/us/en/press-releases/2018/fmi-and-nielsen-online-grocery-shopping-is-quickly-approaching-saturation/>.
- Kaplowitz, M. D., Hadlock, T. D., & Levine, R. (2004). A Comparison of Web and Mail Survey Response Rates. *Public Opinion Quarterly*, 68(1), 94–101. doi: 10.1093/poq/nfh006
- Newport, F., & Brennan, M. (2017, August 8). So Far, American Grocery Shoppers Buck Online Shopping Trend. Retrieved November 12, 2019, from <https://news.gallup.com/poll/215597/far-american-grocery-shoppers-buck-online-shopping-trend.aspx>.

Appendix A - Sampling Matrix

	East	East	East	East	East
	Cities	Suburbs	Towns	Rural	Total
18-29	30	30	30	30	120
30-49	30	30	30	30	120
50-64	30	30	30	30	120
65+	30	30	30	30	120
TOTAL	120	120	120	120	480
	West	West	West	West	West
	Cities	Suburbs	Towns	Rural	Total
18-29	30	30	30	30	120
30-49	30	30	30	30	120
50-64	30	30	30	30	120
65+	30	30	30	30	120
TOTAL	120	120	120	120	480
	Mid-West	Mid-West	Mid-West	Mid-West	Mid-West
	Cities	Suburbs	Towns	Rural	Total
18-29	30	30	30	30	120
30-49	30	30	30	30	120
50-64	30	30	30	30	120
65+	30	30	30	30	120
TOTAL	120	120	120	120	480
	South	South	South	South	South
	Cities	Suburbs	Towns	Rural	Total
18-29	30	30	30	30	120
30-49	30	30	30	30	120
50-64	30	30	30	30	120
65+	30	30	30	30	120
TOTAL	120	120	120	120	480
Grand Totals					
National Grand Total	1920	95% +/- 2.5	(you have a sufficiently large sample size to be more rigorous)		
Each Regional Grand Total	480	95% +/- 5	(you have sufficient sample size at each of these levels)		
Each Local Area Grand Total	480	95% +/- 5	(you have sufficient sample size at each of these levels)		
Each Age Category Grand Total	480	95% +/- 5	(you have sufficient sample size at each of these levels)		

Appendix B - Survey Instrument

Survey Instrument:

This survey should take less than five minutes to complete. The objective of this survey is to understand the role that social media played in shaping your decision to purchase groceries such as staple pantry items, specialist ingredients, beverages, personal care products, fresh produce, snacks, meats, fish, and deli products online. We appreciate your participation. If you have completed this survey previously, please indicate so in the first question or click cancel.

[Note: Programming instructions are provided in braces. Unless otherwise specified, present questions in order presented.]

1. Have you completed this survey within the last 12 months?
 - a. Yes **[Terminate from survey]**
 - b. No
2. Have you made a grocery purchase online within the last 12 months?
 - a. Yes
 - b. No **[Terminate from survey]**
3. How many times have you made a purchase in the last 12 months?
 - a. <6
 - b. 6 - 12
 - c. 13 - 24
 - d. 25+
 - e. [a Don't Remember option may be needed here too]
4. What is your age group?
 - a. 18-29
 - b. 30-49
 - c. 50-64
 - d. 65+
5. Please specify your geographic region. Zip code is preferred; however, you may provide your State of residence instead.
 - a. Zip code? _____
 - b. State: **[Provide list box of States.]**
 - c. National Region: [This should be programmed into the survey but not visible to the respondent. It is not an option but rather your own grouping based on their responses.]
[Program the survey tool to group respondents into regions based on the state where they live as follows:

East (E): Florida, Georgia, South Carolina, North Carolina, Virginia, Maryland, Delaware, New Jersey, New York, Connecticut, Rhode Island, Massachusetts, New Hampshire, Maine

West (W): Montana, Idaho, Wyoming, Colorado, New Mexico, Arizona, Utah, Nevada, California, Oregon, Washington, Alaska, and Hawaii.

Midwest (M): North Dakota, South Dakota, Nebraska, Minnesota, Iowa, Missouri, Wisconsin, Illinois, Kansas, Michigan, Indiana, and Ohio.

South (S): Virginia, Tennessee, Arkansas, Louisiana, North Carolina, South Carolina, Mississippi, Alabama, Georgia, Florida, Texas.]

6. Please state your geographic location: [You are going to have interpretation bias with this question. If you get a person's zip code, then you can group them appropriately using coding in the survey and can skip the question entirely for them. However, if you get a state only, then you need to provide definitions so each respondent interprets these geographic areas the same way. How does suburb differ from town? Every suburb is comprised of one or more towns. Work on this question.]
- City
 - Suburb
 - Town
 - Rural Area
7. Do you use social media? [It seems to me this question should move up in your order. If a respondent declares that he/she does not use social media, then shouldn't they be eliminated from your survey?]
- Yes
 - No
8. Do you follow grocery stores on social media platforms (i.e. Instagram, Facebook, Twitter, etc.)?
- Yes **[Go to Question 9]**
 - No **[Go to Question 19]**
9. Which of the following social media platform/s have you used to follow grocery brands? (Please check off all the item/s applied to you in the following table) The question and responses do not match. Based on your question, shouldn't the answer options be don't use, use some, use a lot to follow grocery brands.]

	No prior user experience	Have some user experience	Have extensive user experience
Facebook			
Instagram			
Twitter			

Pinterest			
LinkedIn			
YouTube			
Other (please specify)			

10. How likely are you to recommend the following social media platforms to another online grocery shopper who follows grocery brands on social media?

	Not likely	Somewhat likely	Very likely
Facebook			
Instagram			
Twitter			
Pinterest			
LinkedIn			
YouTube			
Other (please specify)			

11. Has social media ever led you to make a grocery purchase online?

- a. Yes **[Go to Question 12]**
- b. No **[Go to Question 19]**

12. Considering the times when social media had led you to make an online grocery purchase, what was the primary reason?

- a. Discount Coupon
- b. Other Promotional Coupon
- c. Celebrity Endorsement
- d. Endorsement from someone else
- e. Advertised Recipe
- f. Other

13. How strongly do you agree or disagree [because you lead with agree, your responses should lead with 'Strongly Agree', reverse the order of your response selections] that social media has influenced your decision to shop online?

- a. Strongly Disagree
- b. Disagree
- c. Neutral
- d. Agree
- e. Strongly Agree

14. How strongly do you agree or disagree [same comment on order of response options as noted above] that your contacts/network/friends on social media platforms had an influence on your online grocery purchasing decision(s)?

- a. Strongly Disagree
- b. Disagree
- c. Neutral
- d. Agree

e. Strongly Agree

15. How strongly do you agree or disagree [order] that social media advertisements had an influence on your online grocery purchasing decision(s)?

- a. Strongly Disagree
- b. Disagree
- c. Neutral
- d. Agree
- e. Strongly Agree

16. How strongly do you agree or disagree [order] that product information/reviews available on social media platforms had an influence on your online grocery purchasing decision(s)?

- a. Strongly Disagree
- b. Disagree
- c. Neutral
- d. Agree
- e. Strongly Agree

17. How strongly do you agree or disagree [order] that recipe information available on social media platforms had an influence on your online grocery purchasing decision(s)?

- a. Strongly Disagree
- b. Disagree
- c. Neutral
- d. Agree
- e. Strongly Agree

18. How strongly do you agree or disagree that coupons/promotions available on social media platforms had an influence on your online grocery purchasing decision(s)? **[After selection made, go to Question 21]**

- a. Strongly Disagree
- b. Disagree
- c. Neutral
- d. Agree
- e. Strongly Agree

19. As social media was NOT what led you to make an online purchase, what was the primary reason you purchased groceries online?

- a. Store reputation
- b. Delivery options
- c. Product selection
- d. Product availability
- e. Convenience

20. As you don't follow any grocery brand on social media, which of the following items might influence you to begin following a grocery brand on social media?
(Please check off all the item/s applied to you in the following table)

Coupons	
Rewards	
Recipes	
Consumer Reviews	
Nothing	
Others (please specify)	

21. How would you rate your latest online grocery shopping experience?
- Very dissatisfied [need consistency of language]
 - Dissatisfied
 - Neither satisfied nor dissatisfied
 - Satisfied
 - Very Satisfied