Comparing Mystery Shopping with Smartphone and Traditional Mystery Shopping

Catherine Gabrielle Santoso [a], Pei-Luen Patrick Rau [a] and Yubo Zhang [a]

[a] Department of Industrial Engineering
Tsinghua University
Beijing, 100084, China

ABSTRACT

This study aims to investigate different methodologies in performing mystery shopping programs, i.e. new (call, SMS, and notes) and traditional methodologies in the aspect of shoppers' workload, satisfaction, consistency, accuracy and timing. For that purpose, a single factor experiment with four levels was designed in this research. A smartphone and computer-mediated experiment was conducted with 40 Chinese students. The results showed that the traditional method was the most consistent method and there was no significant difference among the other three mystery shopping methods. However, the results also demonstrated that traditional method was the least accurate and there was no significant difference among the other three methods. The researcher also found that when conducting a mystery shopping program, the SMS method took the longest time, followed by the calling method. Finally, this research discusses the implications of these findings for developing mystery shopping guidelines.

Keywords: Mystery Shopping, Convenience Store, Smartphone, Workload, Satisfaction, Consistency, Accuracy, Timing

INTRODUCTION

Roger Mayland, VP of Martiz's Quality Controlled Services Division, defined mystery shopping as a "process for measuring service quality, with feedback, that is understandable to the front-line employees" (Erstad, 1998). Mystery shopping uses trained researchers to act as customers or potential customers of an organization with the intention of monitoring and assessing the quality of the customer service level, and the processes and procedures used in the delivery of the service (Calvert, 2005).

Mystery shopping is a research technique that has frequently been used by many retailers to gather observational data about a store and to collect data about customer-employee interaction (McDaniel & Gates, 2010). It is estimated that 70% of America's national retailers use this technique. However, in measuring customer experience, this research method has not been really widely used in retailing sector in China such as convenience store.

In a typical mystery shopping experience, a key skill of mystery shoppers is the keen memory because no notes can be taken during the shopping experience (Lusch, Dunne, & Carver, 2010). This is then coherent with Wilson's (1998) statement which stated that retention and recording of information was particularly important, as the shoppers could not complete an assessment form during the service encounter.

https://openaccess.cms-conferences.org/#/publications/book/978-1-4951-2109-8

Mystery shopping is a useful method only if it is conducted in a right manner. However, several issues may affect the reliability and validity of the results (Low, 2011). As mystery shopping relies on human judgment, the item of subjectivity must be reckoned and any bias must be reduced from the beginning. Moreover, many companies set a limit on frequency to conduct the program and number of audits due to high cost, which may affect the reliability of the results. Moreover, several factors such as timing, stores being studied, and mystery shoppers themselves might affect the validity of the results obtained. For example, shoppers might forget one or more things to be assessed, and stores condition is different during day and night.

A traditional mystery shopping involves visiting the business as a customer, then completing an evaluation form or narrative within 12 or 24 hours describing the customer service, cleanliness, quality, sales skills, and other aspects of the experience after the shopping is done (Stucker, 2004). For individuals who are not proficient at this method, there are difficulties they have to face in the process of mystery shopping. For example, they may leave out important details due to the big amount of information. These difficulties may influence the evaluation outcome of the method. However, the emergence of smartphone together with its multiple functions such as audio recording, note taking may assist mystery shoppers to maintain a high level of professionalism.

This study will compare two ways of mystery shopping: one with the assistance of smartphone and one with traditional methodology in terms of mystery shoppers' workload, satisfaction, accuracy, consistency and completion time. Afterwards, this study will try to provide guidelines for improving the evaluation outcome of mystery shopping.

CONCEPTUAL FRAMEWORK AND HYPOTHESES

Figure 1 shows the conceptual framework of this study. In traditional mystery shopping, shoppers' memory is the only tool which can be used to evaluate the service quality and customer experience. With the assistance of smartphones, new methods such as shorting messages, phone calls, taking notes and the function of audio recording can be used to improve mystery shopping. In summary, four methods of mystery shopping were compared in terms of shoppers' mental workload, satisfaction, consistency, accuracy and completion time.

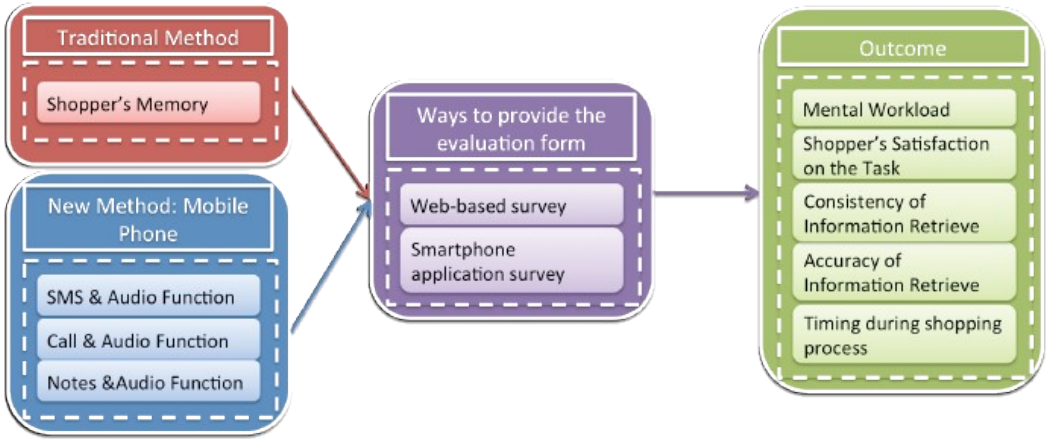

Figure 1. Conceptual framework

According to the framework, five hypotheses were come up with as follows.

By traditional method, mystery shoppers may forget some details during the process. By using smartphones especially the audio recorder on the smartphone, shoppers do not have to pay much attention to memorizing details. When shoppers use smartphone functions such as the audio recorder, they will have fewer items to remember and they are able to retrieve it again after the process; hence, it will reduce the amount of workload. Therefore, the first hypothesis is:

H1: When conducting mystery shopping, using smartphone application survey during the shopping process will reduce shoppers' workload compared with using traditional mystery shopping.

Since using mobile phone to conduct mystery shopping can reduce workload for shoppers, they can have a more relaxed status when they carry out the tasks. Besides, with the help of smartphone functions, shoppers can record more data and information, which facilitates further evaluation for the convenience stores. All these benefits can lead to higher satisfaction compared with traditional method. Hence, the second hypothesis is:

H2: When conducting mystery shopping, shoppers will have higher satisfaction by using new methods mystery shopping than by using the traditional mystery shopping.

In the traditional mystery shopping process, shoppers have to act as real customers and they cannot bring any tool to take notes. Thus, they have to read the evaluation guideline very carefully and try to remember every detail, otherwise they may have left out some aspects. Besides, every shopper has to complete a detailed form after the process. The details provided in the form should be as objective as possible. For a mystery shopper who is not proficient enough, it is difficult for them to observe and memorize every detail required by the guideline.

However, if the shoppers bring along smartphones, they are able to interact with the researcher who is outside the mystery shopping spot and directs the whole process in real time. They can keep contact via phone calls or short messages. Shoppers can write down important information in the smartphone before evaluation. Thus, the researcher does not need to give a tedious guideline to shoppers beforehand. Instead, the researcher just needs to give a brief guideline before the evaluation and interacts with shoppers during the evaluation. The interaction between the participant and the researcher can attempt to manage the retrospective problems of a complicated experience (Echeverri, 2005).

Traditional mystery shopping requires shoppers to fulfill the report within 12 to 24 hours after the evaluation. Since fresh mystery shoppers tend to forget more details than proficient shoppers, there might be bias or inconsistency in the report they submit. Calvert (2005) and Morrison et al. (2010) stated that mystery shopping could be coupled with other methods to improve reliability and validity.

Since traditional mystery shoppers have to memorize a lot of key points before the evaluation, the deep impression of those points may lead to a better consistency of those points they have memorized. However, it is easy for them to leave out details. Hence, the third and fourth hypotheses are:

H3: When conducting mystery shopping, using traditional method is better in terms of consistency of information retrieve than by using smartphone application survey.

H4: When conducting mystery shopping, using smartphone application survey during the shopping process is better in terms of accuracy of information retrieve than after the shopping process.

Although traditional mystery shopping has its weakness, there is one aspect that new methods may be inferior in: completion time. A study by Dawson and Hillier (1995) in the retail sector showed that the mystery shopping process should be at most half an hour and no longer. A mystery shopping that takes too much time may impact the business of the organization and risk the identification of the mystery shopper by the employee, thus jeopardizing the entire effort. The new method may consume more time than traditional method by which shoppers memorize every detail by their memory. They have to either ask the researcher about what to do next or wait for the answer via SMS or peek at their phone to see notes on smartphones. Therefore, the last hypothesis is:

H5: When conducting mystery shopping, using traditional mystery shopping will consume less time than using new methods to conduct mystery shopping.

METHODOLOGY

Independent Variable

The independent variable is different methods in mystery shopping, including calling and audio recording functions on smartphone, SMS and audio recording functions on smartphone, notes and audio recording functions on

smartphone, and participant's memory, i.e. no smartphone.

For calling and audio recording functions using smartphone, the mystery shoppers called the researcher once they arrived at a convenience store, thus real time interaction occurred during the whole process. The audio recorder functioned during the whole process. Mystery shoppers filled in the evaluation forms on the smartphone application during the process. During the interaction, the researcher guided and reminded shoppers and they were expected to ask questions secretly if there were points that they were unsure of.

For SMS and audio recording functions using smartphone, the mystery shoppers texted the researcher once they arrived, thus real time interaction occurred. The audio recorder functioned during the whole process. Mystery shoppers filled in the evaluation forms on the smartphone application during the process. During the interaction, the researcher guided and reminded shoppers via SMS and they were expected to text any questions for any points that they were unsure of. They then texted the researcher once they were done with the shopping.

For notes and audio recording functions using smartphone, the mystery shoppers were given a list of points written on the notes which showed on smartphone before the process. They were allowed to add or delete some points accordingly and they got the notes ready right before entering the store. No real time interaction occurred during the whole process. The audio recorder functioned during the whole process. During the process, they had to fill in the forms on the smartphone. They were expected to peek the key points on the notes to make sure they obtained the necessary data.

For the traditional method, without smartphones as tools, mystery shoppers were expected to memorize key points before the process and observe them during the process. After the evaluation, each shopper filled in an evaluation form right after the process.

All mystery shoppers had to fill in another evaluation form within 24 hours after the process.

Dependent Variables

There were five dependent variables measured in this study. First, in order to measure participants' mental workload, NASA-TLX was used. Second, to measure the participants' satisfaction about the tasks, a satisfaction questionnaire was used. It consisted of two sections. The first one had 8 questions measuring how participants felt and the second one had 6 questions measuring their opinions about the process. The third one is the consistency of information retrieving. It was measured by the comparison of two evaluation forms. The fourth one is the accuracy of information retrieving. It was measured by the number of key points which were listed in the guideline but forgotten to observe by mystery shoppers. The last one is the length of time from the moment when shoppers entered the spot to the moment when they stepped out of the spot.

Participants, Spot and Apparatus

In this study, 40 university students were recruited via social media as mystery shoppers. All of them were Chinese and were familiar with smartphone applications. They were divided into four groups. Each participant conducted one method of mystery shopping. A 7-Eleven convenience store was selected as the spot of mystery shopping.

During the process, each participant was equipped with two smartphones except those who conducted traditional mystery shopping. One was SAMSUNG smartphone which had Android OS and it was used to make a call or to send SMS as well as to fill in the evaluation form during the mystery shopping. The other smartphone was used to record the audio. They were both equipped with new SIM cards to prevent disturbance from others' call. Besides, for participants who were allocated to deploy the calling method, each one was equipped with an earphone. For participants who deployed the traditional method, each was provided with a laptop right after the mystery shopping in order to fill in the first evaluation form and each was equipped with a stopwatch to record the time.

Task and Procedure

Before conducting the evaluation, each participant was assigned individually to read a guideline of the key points and details that they needed to observe and ask in the store. Memorizing the detailed information was not required for participants who were involved in the three new methods. In addition, they had to fulfill a practice test on using the smartphone application prior to evaluation because they had to use the application during the process. Participants had to report every negative observation; for example, "the floor has debris all over the store". The result of the evaluation form would be sent directly to the researcher via text message once they clicked "submit".

During the whole process, the other smartphone was running to record conversation. For the traditional methodology, no smartphone was used. Thus, memorizing the detailed information was required.

The details in different groups were slightly different from one another. For the calling and audio recording method, each participant had to call the researcher once they arrived at the convenience store and the researcher would record the beginning time. During the shopping, real-time interaction via phone call was carried out to remind participant what to do. For SMS and audio recording method, all interaction between the participant and the researcher was via SMS. For notes and audio recording method, there was no direct interaction between the participant and the researcher. Each participant had to prepare the notes ready on the smartphone and recorded the beginning time all by himself. After the shopping process was done, he recorded the ending time again. For traditional method, each participant had to deploy the process all by his memory of the guideline and record the time by himself.

After the shopping process, they had to meet the researcher again in the nearest coffee shop to fill in post-evaluation questionnaires. Three questionnaires were given to them. The first was a web-based evaluation form, they were given a link with username and password, and they had to fill in the form anytime within 24 hours after the shopping. The result was submitted directly to the researcher's email immediately after they clicked the 'submit' button. Audio recording was uploaded together with the second evaluation form. For the traditional method, each participant also had to fill in the first evaluation form. Then the second and the third post-experiment questionnaires regarding participant's workload and satisfaction were provided and filled in.

RESULTS

Among the 40 participants, there were 15 male and 25 female. The participants' age ranged from 18 to 30 years old (Mean=23.1, SD=2.49). There was only 1 participant who had experienced mystery shopping before and there were 3 participants who had not used smartphone. Half of the participants majored in industrial engineering and the other half were from other departments. There was no significant difference among each group in the aspect of age, gender, education level, major and mystery shopping experience.

To test the hypotheses, this study compared the mean value of four methods in the aspect of workload, satisfaction, consistency, accuracy and timing.

K-S test was conducted to test if the sample was normally distributed. As a result, the accuracy was not normally distributed ($Z= 2.205$, $p<.001$) and the other four variables were normally distributed. Therefore, one-way ANOVA was conducted as the variance analysis on workload, satisfaction, consistency and timing while nonparametric test was conducted as the variance analysis on accuracy.

The result of variance analysis is revealed in Table 1. There was no significant difference on participants' workload ($F=.315$, $p=.814$) and satisfaction ($F=1.776$, $p=.169$). There were significant differences on participants' consistency, accuracy and completion time.

Pairwise test was conducted to confirm where the difference existed between groups in terms of consistency, accuracy and completion time. The result is in Table 2, Table 3 and Table 4.

Table 1: Variance analysis of dependent variables

Dependent variable	Calling		SMS		Notes		Traditional		$F(\chi^2)$	p
	Mean	SD	Mean	SD	Mean	SD	Mean	SD		
Workload	13.6	2.02	13.0	3.17	12.4	4.57	12.5	2.85	.315	.814
Satisfaction	5.3	.56	4.9	.44	5.0	.44	5.3	.59	1.776	.169
Consistency	.48	.22	.46	.10	.46	.19	.69	.08	4.955	.006
Timing	11.8	1.8	16.9	3.43	8.4	1.60	6.2	2.81	33.972	<.001

Accuracy*	1.00	.01	.98	.03	.99	.03	.89	.08	16.695	.001

Note: The variance analysis of accuracy was conducted via nonparametric test.

Table 2: Pairwise comparison of four methods on timing

Dependent variable		Mean difference	*p*
Calling	SMS	-5.1*	<.001
	Note	3.4*	.005
	Traditional	5.7*	<.001
SMS	Calling	5.1*	<.001
	Note	8.5*	<.001
	Traditional	10.7*	<.001
Note	Calling	-3.4*	.005
	SMS	-8.5*	<.001
	Traditional	2.3	.054
Traditional	Calling	-5.7*	<.001
	SMS	-10.7*	<.001
	Note	-2.3	.054

Note: The mean difference is significant at the level of p=.05

Table 3: Pairwise comparison of four methods on consistency

Dependent variable		Mean difference	*p*
Calling	SMS	.02	.777
	Note	.03	.724
	Traditional	-.21*	.006
SMS	Calling	-.02	.777
	Note	.01	.944
	Traditional	-.23*	.003
Note	Calling	-.03	.724
	SMS	-.01	.944
	Traditional	-.23*	.002
Traditional	Calling	.21*	.006
	SMS	.23*	.003
	Note	.23*	.002

Note: The mean difference is significant at the level of p=.05

https://openaccess.cms-conferences.org/#/publications/book/978-1-4951-2109-8

Table 4: Pairwise comparison of four methods on accuracy

Dependent variable		Mann-Whitney U	p
Calling	SMS	34.500	.121
	Note	44.500	.503
	Traditional	11.000*	.001
SMS	Calling	34.500	.121
	Note	41.000	.399
	Traditional	16.000*	.007
Note	Calling	44.500	.503
	SMS	41.000	.399
	Traditional	14.000*	.003
Traditional	Calling	11.000*	.001
	SMS	16.000*	.007
	Note	14.000*	.003

Note: The mean difference is significant at the level of p=.05

It can be seen that SMS method took the longest time compared with other three methods and traditional method took the shortest time. As for pairwise comparison of timing, all methods showed significant difference with each other except for notes and traditional method.

As for consistency, it can be seen that traditional method had the highest consistency among the four methods while there was no significant difference among the other three new methods about consistency.

As for accuracy, it can be seen that traditional method was the least accurate method among the four and there was no significant difference among the other three new methods about accuracy.

Then a summary review of the five hypotheses is:

Hypothesis 1 was to examine the influence of four methods on participants' workload. The one-way ANOVA result showed that there was no significant difference. Hence, hypothesis 1 was not supported.

Hypothesis 2 was to examine the influence of four methods on participants' satisfaction. The result also showed that there was no significant difference. Hence, hypothesis 2 was not supported either.

Hypothesis 3 was to examine the influence of four methods on participants' consistency of information retrieving, i.e. the similarity between the results of the two evaluation forms which were filled in when participants finished mystery shopping just in time and afterwards within 24 hours respectively. The result showed that traditional method had significantly higher consistency than new methods and there was no significant difference among the other three methods. Therefore, hypothesis 3 was supported.

Hypothesis 4 was to examine the influence of four methods on participants' accuracy of information retrieve, i.e. how many key points the participants did evaluate during the mystery shopping process. The result showed that traditional method had significantly lower accuracy than new methods and there was no significant difference among the other three methods. Therefore hypothesis 4 was supported.

Hypothesis 5 was to examine the influence of four methods on the length of time. The statistical result showed that SMS method took the longest time and every two methods had significant difference of timing except for notes method and traditional method. Hence, hypothesis 5 was partially supported.

DISCUSSION

Within the aforementioned conceptual framework, some interesting results were found.

This study did not find significant difference among four methods on shoppers' workload or satisfaction, which indicates that when related organizations conduct mystery shopping programs, shoppers' workload and satisfaction would not be important factors influencing the adoption of different methods.

As for consistency of information retrieving, traditional method had significant advantage. This is due to the fact that for traditional method, participants have to memorize everything that is stated in the guideline in detailed manner before the evaluation and have to memorize everything that they observe in the store after the evaluation. Whereas for the new methods, shoppers are not required to memorize the guideline and things that they observe inside the stores as they are allowed to bring smartphone inside the store and interact with the researcher. Therefore, participants who undergo traditional method are able to memorize the information retrieved than those who undergo new methods.

Although traditional method beats the other three in the aspect of consistency, it shows low accuracy according to the result. This shows that having interaction with the researcher who appoints specific task to the shoppers and having additional tools to enhance the mystery shopping have higher accuracy in terms of information obtained. This is because shoppers can interact with the research in real time and the researcher can direct shoppers on information that needs to be observed. For notes method, shoppers can write as much information as they want in the smartphone application for it serves as a reminder. Therefore, organizations should trade off between consistency and accuracy of information retrieving or deploy mixed methods combing the advantages of human memory and enhanced assisting tools.

Finally, as for the length of shopping time, the result shows that SMS and calling methods take longer time than the other two. This is due to the fact that for SMS methods shoppers need to write an SMS once they arrive in the assigned store, and have to wait for the researcher's reply and direction for any information that needs to be observed. Likewise, calling method takes more time than notes and traditional methods because they have to do multitasking jobs, i.e. listening, observing, and filling-in the evaluation form on Android smartphone at the same time. As for notes and traditional methods, no interaction between the shopper and director occurs in this methodology so the shopping process takes less time than the former two.

In summary, depending on the type of company, budget, period of time, expected results, organizations interested in conducting mystery shopping programs can choose suitable method for them or mix them together to achieve comprehensively good results.

CONCLUSION

The objective of this study is to investigate different methods in performing mystery shopping programs, i.e. new and traditional methods, focusing on China's convenience stores. On one hand, the results should help the mystery shopping company to learn and to apply new approaches for its clients to enhance clients' satisfaction and experience. On the other hand, it should be seen as a guideline for any organization interested in conducting mystery shopping program in a way that ensures that their product and service quality can be improved after they implement the mystery shopping program. It should also help the mystery shopping company to evaluate which new method is the best.

An important finding is that traditional methodology is regarded as the best method in terms of consistency; whereas new methods are better in terms of accuracy. For timing, methods that are meant to interact in real-time situation with the project managers, i.e. call and SMS method, take more time than methods that are not meant to interact with the project managers. Although not all hypotheses formulated have significance differences among the methodologies, the results also indicate that those hypotheses have positive implications. For example, for workload and satisfaction, there is no significant influence among all methods. This means that organizations are allowed to choose any method depending on their budget, preference, and decision to conduct mystery shopping programs

without having to worry about the effect of shoppers' workload and satisfaction.

In addition, it is also possible to combine the new and traditional methodology so that the information obtained will be more accurate and consistent. The most important prerequisite is to provide the right and proper training to mystery shoppers prior to conducting the program.

However, the convenience store is limited to 7-Eleven in Beijing and participants' culture, education background is also limited, which may cause bias to the results. Besides, since most participants have no prior mystery shopping experience, whether or not the results will be the same on professional mystery shoppers is worthy of further validation.

REFERENCES

Calvert, P. (2005). "*It's a mystery: mystery shopping in New Zealand's public libraries*", LIBRARY REVIEW Volume 54 No. 1. pp.24-35

Dawson, J., & Hillier, J. (1995). "*Competitor mystery shopping: methodological considerations and implications for the MRS Code of Conduct*", *JOURNAL OF THE MARKET RESEARCH SOCIETY* Volume 37 No. 4. pp. 417-427

Echeverri, P. (2005). "*Video-based methodology: capturing real-time perceptions of customer processes*", INTERNATIONAL JOURNAL OF SERVICE INDUSTRY MANAGEMENT Volume 16 No. 2. pp. 199-209

Erstad, M. (1998). "*Mystery shopping programmes and human resource management*", INTERNATIONAL JOURNAL OF CONTEMPORARY HOSPITALITY MANAGEMENT Volume 10 No. 1. pp. 34-38

Low, I. (2011). "Mystery Shopping in Singapore's Retail Sector: A Case Study", UNLV Theses/Dissertations/Professional Papers/Capstones. Paper 1065.

Lusch, R. F., Dunne, P. M., & Carver, J. R. (2010). "*Introduction to Retailing* (7th ed.)", Tsinghua University Press.

McDaniel, C., & Gates, R.H. (2010). "*Marketing Research with SPSS*", John Wiley & Sons, Incorporated.

Morrison, M., & Mundell, M. (2010). "*Connecting, Engaging and Exciting Shoppers*", in: Shopper Marketing: How to Increase Purchase Decisions at the Point Of Sale, Stahlberg, M., & Maila V. (Ed.). pp. 75-81

Stucker, C. (2004). "*The Mystery Shopper's Manual: How to Get Paid to Shop in Your Favorite Stores, Eat in Your Favorite Restaurants and More*", Special Interests Publishing.

Tucker, RB. (1991). "*Ten driving forces of dynamic change*", EXECUTIVE EXCELLENCE Volume 8 No. 3. pp. 16

Wilson, A.M. (1998). "*The role of mystery shopping in the measurement of service performance*", MANAGING SERVICE QUALITY Volume 8 No. 6. pp. 414-420