博士論文

A Study on Utilization of Customer Ratings in Services

(サービスにおける消費者評価情報の利用に関する研究)

高橋 裕紀

Abstract

Services have unique characteristics, and it is difficult for consumers to know their actual quality before consumption. Therefore, consumers and service providers should know customer satisfaction to understand a service's current quality. The widespread use of smartphones in the recent years has enabled consumers and mystery shoppers to write ratings of services effortlessly, and many people have benefited from referring to these ratings. Although writing and observing the cycles of service ratings are both expected to boost service markets, how these cycles affect service markets in reality remains unclear. Service rating cycles should be improved to suit service characteristics better. In this study, economic experiments and empirical analysis were conducted on the current trends in service ratings to examine utilization of service ratings. Two topics were focused on: customer ratings and mystery shopping. Economic experiments provide empirical evidence that consumers can use services better with rating systems considering the heterogeneity of consumer preferences or the dynamic consumption of services. Data analysis proposes the possibility of new methods of summarizing mystery shopping results with mystery shoppers' lifestyles. Additionally, this study provides empirical evidence of the direct feedback from mystery shopping reports to employees that can benefit service providers, by analyzing the effect of introducing a new app that enables employees to observe the results of mystery shopping directly. In summary, this research clarified the effect of consumer ratings through both consumer ratings and mystery shopping. It empirically clarified how consumers rate services and how they are affected in turn by these ratings through these two methods.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Services have four unique characteristics: intangibility, inseparability, perishability, and heterogeneity. These characteristics make it difficult to understand their quality before consumption (Zeithaml 1981). Therefore, both consumers and service providers must rely on customer satisfaction as a quality index. In addition, a wider variety of services are available owing to the customization trend and diversity in customer preferences (Gilmore 1997). It has become more critical for providers to predict customer preferences based on their reactions. Therefore, service providers should pay attention to customer reactions.

Owing to the widespread use of smartphones, consumers can easily rate services based on their satisfaction. These satisfaction values or service ratings are collected and posted on platforms such as Amazon, Trip Advisor, and Yelp to inform customers about the quality of services. As the number of posted ratings increases, the value of platforms also increases for customers. Eventually, the number of users of the platform increases, and the number of posts consequently increases. These platforms affect consumers, manufacturers, and service providers, and thus they must reconsider their marketing strategies.

In addition to these anonymous and less-censored reviews from mass customers, ratings by qualified professionals or market research companies are gaining attention. One such method of rating is through mystery shopping (MS). Mystery shopping is a type of market research wherein mystery shopping companies send mystery shoppers specific checklists to make the shoppers experience and examine service quality of their client companies. The mystery shopping market has become significant. The Mystery Shopping Professional Association estimated the turnover of global mystery shopping to be more than $2 billion in 2017 (MSPA 2018). Mystery shopping companies specialize in service evaluations with qualified checklists and training systems for mystery shoppers. Mystery shopping results are sent to client company managers and employees to improve their management and services.

Customer satisfaction value plays an increasingly important role in the service industries for consumers and market research companies. However, it is unclear whether increasing the collection and display of customer satisfaction values can benefit service markets and is optimal for the current service markets. Customer satisfaction may be affected by human psychological aspects such as satiation, expectations, and lifestyle. Customer ratings may be biased, and may confuse future users. In addition, a popular satisfaction display is the average value rating by consumers. However, more optimal rating systems or display styles may exist for consumers and employees.

1.2 Research Purpose

This study aims to examine two main satisfaction data collection methods, namely, customers' and mystery shoppers' ratings, through economic experiments and data analysis. Fig 1.1 shows the target area of this research in the service market. The image exhibits important players and service elements: receivers, providers, ratings, and services. Bold arrows indicate the effects of one player on the other. There are two cycles in this image: the smaller cycle shows the circulation of service receivers' ratings to other service receivers, and the larger cycle shows the circulation of service receivers' ratings to service providers, which affects the services they provide. Inclusion of mystery shoppers in the service receiver group in both cycles may seem confusing. Mystery shoppers were initially recruited from among ordinary consumers. They were trained to fill the service checklists and evaluate the service subjectively. Although mystery shopping companies check them, these subjective evaluations can have effects similar to those of customer ratings. Therefore, the mystery shoppers are grouped with consumers in Fig 1.1.

This study examines two main research questions: (1) Is current customer satisfaction display truly beneficial to consumers and employees? (2) Is there an optimal rating system for services? To answer these questions, sub-research questions are proposed in Chapter 2, and each of them is analyzed in Chapters 3, 4, 5, and 6.

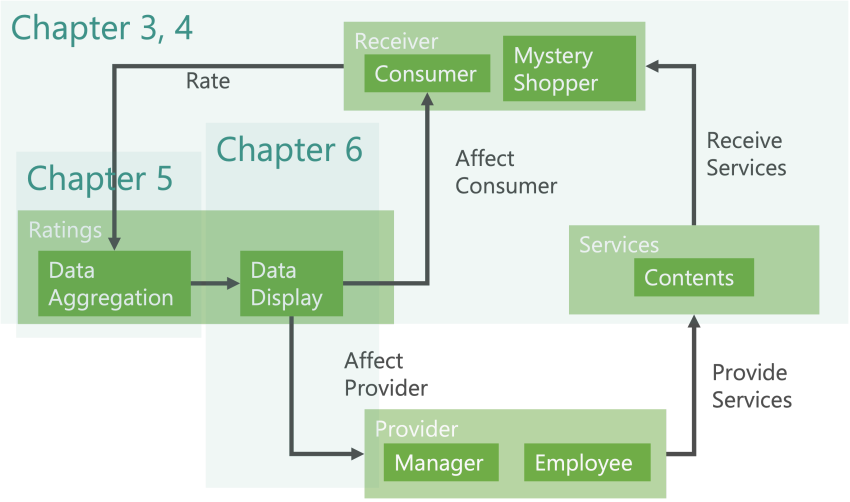


Fig. 1.1: Overall image of this study

Economic experiments and data analyses were conducted in this study. The economic experiments aim to examine the effect of a rating system on consumers. A rating system can control consumer preferences while engaging in irrational rating behavior. This clarifies the relationship between the actual utility of services and rating scores. By contrast, we did not conduct experiments in mystery shopping research because we had access to the actual data on mystery shopping, and conducting experiments in this research was not very advantageous. Examining other useful data collection methods for mystery shopping and the effect of mystery shopping on employees was more important for our research purpose. Real data analysis can provide more direct evidence than that from experimental methods.

The contributions of this study to the literature are as follows: First, this study clarified the effect of the display of customer ratings through a controlled experiment, which is scarcely conducted compared with data analysis research. Second, it used unique data from mystery shoppers to examine the more explicit role of mystery shopping in service industries.

1.3 Organization of This Paper

This paper consists of three parts. Part 1 discusses the effects of customer ratings and two experiments are presented in this section. In Chapter 3, three simple display methods of customer ratings are compared for heterogeneous goods. In Chapter 4, the rating system is extended to consider the dynamic consumption of services to explore a better fitted rating system for services. Part 2 focuses on mystery shopping. Chapter 5 examines the effect of lifestyle on the subjective evaluation of services. This study proposes an advanced mystery shopping research that considers the heterogeneity of consumers' preferences for services. In Chapter 6, the effect of displaying service receivers' satisfaction on service providers is examined by analyzing the effect of adopting a new device for mystery shopping. In Part 3, the results of the four studies are summarized to answer the two research questions, and this study is concluded in Chapter 7.

Chapter 2

Theoretical Background

2.1 Overview of Subjective Evaluation of Services

In our study, customer satisfaction is an important factor for service provider companies. The famous service model consisting of customer satisfaction is the service profit chain (SPC) (Heskett et al. 1994). Figure 2.1 shows a conceptual model of SPC. It models the dynamic flow of the effects from internal service quality to profit. Customer satisfaction plays a crucial role in the relationship between external service quality and customer loyalty. Several studies have provided empirical evidence on the importance of customer satisfaction by showing the relationship between other objective indices (Gupta and Zeithaml 2006, Andersen and Simester 2004, Mittal et al. 2005, Rust Anthony and Roland 1993). For example, Huang and Sudhir (2020) showed that customer satisfaction maintains customer revisit intentions and acquires new customers as higher customer satisfaction has an advertising effect on new customers. Many other studies have examined the relationship between customer satisfaction and sales (Kumar et al. 2013) and the lagged effect of customer satisfaction on sales (Bernhardt et al. 2000, Evanschitzky et al. 2012).

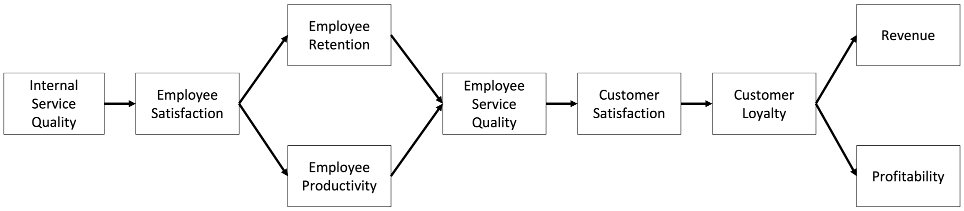


Fig. 2.1: Service Profit Chain (Heskett et al. 1994)

Market research companies and researchers have developed several methods for evaluating customer satisfaction, which can be divided into two types: asking customers about their satisfaction directly, and using market research professionals to observe service quality indirectly. An effective method in the former type is to conduct questionnaires with customers. The ASCI is a well-known questionnaire used in both practice and academics. It considers both customer expectations and perceived quality based on the disconfirmation theory (Fornell et al. 1996).

Owing to the widespread use of smartphones, new trends in customer satisfaction data collection have become popular: rating systems and mystery shopping. However, the effects of these two popular systems on service markets are still being studied. In this study, several points that were insufficiently studied in the past are clarified, and sub-research objectives are proposed based on Figure 1.1.

2.2 Consumer Reviews

Before rating systems spread through the Internet, the effect of word of mouth (WOM) between real consumers was studied (Eugene W. Anderson 1998). However, these studies used questionnaires to collect data on customer satisfaction and self-reported activity related to word of mouth. An important change in the adoption of a rating system is that an enormous number of consumers can now share a single WOM or review on a website. The reviews have become increasingly influential and have gained attention.

A considerable number of prior studies have analyzed the effect of reviews on purchase decisions. Many studies show that sales are correlated with review statistics in various fields (e.g., books, travel agencies, restaurants, video games, and retail) (Chevalier and Mayzlin 2006, Chintagunta et al. 2010, Ye et al. 2009, Zhu and Zhang 2010, Resnick et al. 2006, Wu et al. 2015). Some studies focused more deeply on consumers' purchase decisions, such as how they consider reviews with supplemental information (Zhao et al. 2013, Forman et al. 2008, Ameri et al. 2019). In addition to purchasing behavior, several studies have focused on how previous reviews affect consumers' reviewing behavior. Reviews can contain information that is unrelated to the goods themselves (Moe et al. 2011). Review behavior is affected by the negativity in prior reviews (Schlosser 2005) or expectations Ho et al. (2017). It is natural to question whether these effects assist consumers in or hinder them from buying better goods.

Several empirical studies have analyzed how well these rating values of reviews can be substituted for the actual quality of goods, and show gaps in (Gao et al. 2015, de Langhe et al. 2016, Lu and Rui 2018) and dependencies on elements not related to the quality of goods (Godes and Silva 2012). How these reviews reflect actual quality or other objective evaluations still needs to be investigated. Several studies have analyzed whether consumers can find which goods are better than others and whether the collected reviews are accurate. Acemoglu et al. (2017) consider Bayesian agents and theoretically show the conditions under which rating scores converge into true quality in both cases—that of only showing average ratings and the other showing all history of ratings. By contrast, Besbes and Scarsini (2018) not only considered Bayesian agents, but also focused on irrational agents. They showed that if a naive agent assumes that the rating is the actual quality of goods, then, in the case of only showing an average rating, the rating value will be biased and become overestimated. These theoretical studies show that the conversion of average scores depends on agents' irrationality. It is better to analyze how people interpret rating scores and rate services. From this perspective, an economic experiment is more useful for researchers to consider irrational agents in a controlled environment.

Research Question 1 *Under the circumstance that consumer preference is controlled, do consumers benefit from the reputation system?*

2.2.1 Time series ratings

In Chapter 4, time-series ratings for service evaluations were studied. A service has four unique characteristics: intangibility and separability characterize services as goods that consumers experience in dynamic processes. Optimal allocation of resources to experiential services is important for service providers. Several studies have focused on optimization of activity levels of services defined as an instant excitement from the service in each period. Baucells and Sarin (2010) constructed an experience utility model considering psychological aspects of human satisfaction: satiation and habituation. Their study analyzed the optimal amount of consumption in a time series to maximize total utility, considering satiation and habituation. Gupta et al. (2016) considered memory decay and acclimation—other human psychological aspects—and showed optimized allocation of service allocation. They showed that crescendo and U-shaped patterns are optimal under certain conditions.

However, it remains unclear how consumers can share services' information of activity level before consumption. In Chapter 4, time-series ratings are examined to see how this system can increase consumer benefits. The most popular rating format is posting a rating value for each product or service. By contrast, several websites adopt multidimensional rating systems (Chen et al. 2018). For example, Trip Advisor offers ratings for multiple attributes, such as location, cleanliness, service, and hotel value. We expect this time-series rating for services to assist consumers in increasing their utility.

Research Question 2 *Can a time-series rating system benefit consumers?*

This rating system may change the providers' decisions. Several studies have clarified how firms are affected in their decisions based on rating systems (Hong and Pavlou 2014, Sunder et al. 2019). Some service patterns may benefit from the system while some others may not. Our study also concentrated on this topic.

Research Question 3 *Whether service benefit or disadvantage under time-series rating system depends on its activity level patterns?*

2.3 Mystery Shopping

Mystery shopping has been used to evaluate service quality (Wiele et al. 2005, Heskett et al. 1994, Dutt et al. 2019). Mystery shopping providers send such shoppers to a client's company and allow them to evaluate services using a checklist.

This checklist is typically customized for each client. Clients use the mystery shopping results to improve their service offerings (Latham et al. 2012). Improvements in services offered by providing feedback to employees and managers are discussed in Chapter 6. As most mystery shoppers are ordinary people, mystery shopping companies work to improve the reliability of service evaluation done by them. For example, companies prepare instruction manuals to train mystery shoppers. In addition, companies evaluate their output and provide feedback to improve their service evaluation skills. Researchers have also analyzed the reliability of mystery shopping. Finn (2001) conducted an empirical study, demonstrating the psychological validity of mystery shopping. Several other studies have also demonstrated its validity (Finn and Kayandé 1999, Lowndes and Dawes 2001, Brito and Rambocas 2016).

Mystery shopping is used in a variety of service industries such as restaurants (Luria and Yagil 2008, Latham et al. 2012, Liu et al. 2014, Chen and Barrows 2015), hotels (Beck and Miao 2003, Yaoyuneyong et al. 2018), retail (Wilson 2002, Blessing and Natter 2019), B2B services (Mattsson 2012), banks (Tarantola et al. 2012), libraries (Zorica et al. 2014), and exhibitions (Peterman and Young 2015).

Mystery shopping has several advantages over other methods. First, as a previous study pointed out, mystery shopping is more cost-efficient than customer surveys (Finn 2001). Second, mystery shopping reports are helpful as complementary data to those obtained from traditional surveys (Cervellon et al. 2019, Eger and Mičík 2017, Barber and Tietje 2004, Mendes and Cardoso 2006). For example, Takenaka et al. (2020) used mystery shopping data along with employee satisfaction and year-on-year sales data to confirm the relationship between customer satisfaction, employee satisfaction, and firm profitability, known as the service profit chain (Heskett et al. 1994). Third, a client company can collect opinions from non-users through mystery shopping. Fourth, in most cases, mystery shoppers are registered based on their demographic traits. Therefore, mystery shopping companies can use their data to obtain more profound findings in market research. This cannot be done if a company uses only online reviews written by anonymous reviewers.

However, Blessing and Natter (2019) question the validity of mystery shopping as a proxy for customer reviews. They compared mystery shopping reports with customer surveys and store sales performances. They demonstrated that mystery shopping has less predictive performance because of its smaller sample size. By contrast, Wilson (2001) argues that because a customer could visit any store, service companies are recommended to listen to all their opinions. Although Wilson (2001) has a point, mystery shopping requires improvements to overcome its weaknesses.

One way to improve mystery shopping is to optimize by sending the mystery shoppers to each client. Porter and Heyman (2018) showed that although mystery shopping companies provide elaborate manuals to mystery shoppers, their reports are affected by mystery shoppers' heterogeneity or experience to some extent. Mystery shopping companies should understand each mystery shopper's tendency to evaluate services to match mystery shoppers and clients. Chapter 5 proposes the possibility of improving mystery shopping by understanding the characteristics of mystery shoppers' subjective evaluations of services given their lifestyles.

2.3.1 Lifestyle

One candidate variable that characterizes mystery shoppers is their lifestyle. Lifestyle has gained attention in many market fields (Massara et al. 2020, Zhang et al. 2020, Težak Damijanić 2019, Zhang et al. 2020). For example, Dahana et al. (2019) showed that lifestyle is related to customer lifetime value. In early research, several fields developed lifestyle concepts. Lifestyle was a way to segment markets that captured how people spend their leisure time, their interests, opinions, and demographics (Plummer 1974). Many researchers have developed lifestyle metrics in various fields Green et al. (2006), Wells (1974), Takenaka et al. (2011). In the recent times, lifestyle metrics have been used in various fields such as luxury fashion (Li et al. 2012), computer usage (Ye et al. 2011), and cinema (Palomba 2020).

Because mystery shopping is used in various service industries, we must use an index related to the universal features of services to classify mystery shoppers. Service evaluation is considered to be more related to psychological aspects than other demographic or behavioral aspects (Ishigaki et al. 2010). We consider lifestyle as the key factor in service evaluation because it captures a holistic picture of the general public's hobbies, interests, recreational and cultural activities, and work (Green et al. 2006), many of which are related to services.

Although many lifestyle metrics exist, we use (Takenaka et al. 2011)'s lifestyle metric mainly because of its relevance to Japanese service research. It consists of personality and consumption style. It was designed to characterize consumers with their lifestyle factor scores and demographic information to be adapted for use in many fields in Japan. The lifestyle factor score was calculated from a survey with approximately 20 question items related to consumers' consumption habits and personalities. The literature in Psychology contains research related to personality. The most famous categorization is the Big 5, which consists of five personality traits: extroversion, agreeableness, conscientiousness, neuroticism, and openness for experience/intelligence (Tupes and Christal 1992). This categorization was subsequently improved by Goldberg (1992). Using factor analysis, each person is given a five to seven factor score that demonstrates the strength of each lifestyle factor. It enables managers and researchers to comprehend each customer's lifestyle factor distribution and understand how a particular lifestyle group favors certain services. Several studies have adopted (Takenaka et al. 2011)'s lifestyle metric and examined its relationship with observational behavior (Ishigaki et al. 2010, Takenaka et al. 2011, 2013, 2016).

In Chapter 5, a lifestyle metric is adopted for mystery shoppers' service evaluation. To the best of our knowledge, no study has examined the relationship between mystery shoppers' lifestyles and their service evaluations using large-scale data. The research questions were as follows.

Research Question 4 *Can lifestyle factors improve the understanding of service evaluation by mystery shoppers?*

Research Question 5 *Will mystery shoppers with different lifestyles have different preferences for existing brands based on their service attributes?*

Testing this research question will contribute to research on both lifestyle and mystery shopping. Furthermore, if we prove Research Question 4, we can suppose that mystery shoppers with different lifestyles evaluate established brands differently because each brand has different strengths and weaknesses to differentiate their service from their opponents. Therefore, research question 5 is proposed. These research questions provide insights into the implications for managers of lifestyle in mystery shopping.

2.3.2 The feedback of customer satisfaction by introducing digital device

Mystery shoppers' customer satisfaction feedback for employees or managers can improve their service quality. Managers can revise their service operations based on this feedback. Several studies have shown that feedback can improve employee productivity (Jung et al. 2010, Huang et al. 2019). In the service profit chain, Heskett et al. (1994) called the relationship between employee satisfaction (ES) and customer satisfaction (CS) a satisfaction mirror, and Mortimer and Laurie (2016) examined the effect of CS on the ES.

Chapter 6 introduces a new app that allows employees to concisely view mystery shopping reports. The introduction of digital devices in the service industry to improve operations and management has recently attracted attention. However, whether these digital devices are truly beneficial considering their implementation costs remains a question. Some industries in the service industry have low profitability, which may limit the use of these devices. Some studies have empirically examined the effects of these devices on productivity (Tan and Netessine 2017, Pierce et al. 2015, Hitt and Tambe 2016, Tambe and Hitt 2012, Bavafa et al. 2018, Lu and Rui 2018, Staats et al. 2017). In Chapter 6, the effect of the adoption of the mystery shopping app on mystery shopping is examined to determine the effect of feedback from service receivers on service providers.

Research Question 6 *Can the adoption of a mystery shopping app improve service providers by encouraging employees to check mystery shopping feedback?*

2.4 Chapter Summary

This section reviews prior studies on both customer ratings and mystery shopping. Six sub-research questions are proposed, which are important for answering the two main research questions proposed in Chapter 1. The relationships are as follows:

(1). Is the current customer satisfaction display truly benefits consumers and employees?

- RQ1 Under the circumstance that consumer preference is controlled, do consumers benefit from the reputation system?
- RQ2 Can a time-series rating system benefit consumers?
- RQ6 Can the adoption of a mystery shopping app improve service providers by encouraging employees to check mystery shopping feedback?

(2). Is there a more optimal rating system for services?

- RQ3 Whether services benefit or disadvantage under time-series rating system depends on its activity level patterns?
- RQ4 Can lifestyle factors improve the understanding of service evaluation by mystery shoppers?
- RQ5 Will mystery shoppers with different lifestyles have different preferences for existing brands based on their service attributes?

Part I

Consumer Reviews

Chapter 3

Effects on Purchasing and Rating Behaviors through Economic Experiments

3.1 Introduction

Experience goods like services cannot be evaluated before consumption due to their nature. Such a characteristic has been driving consumers to rely on word-of-mouth since long ago, and more recently, an online review platform such as Yelp spurs consumers to check the reviews posted by those who had experienced the service. For instance, such a platform generally provides subjective evaluations rated on a five-star scale. Nowadays, it has become common for consumers to consult the information to decide whether to purchase goods/services.

However, it is not guaranteed that a review system always enables consumers to purchase their desirable goods/services. Intuitively, if a consumer had a sufficient number of reviews to let them be urged to purchase, they could have appropriately inferred the quality of goods/services. Nevertheless, the reviews consumers rely on might be biased by other prior reviews. Many studies clarify that there potentially exist several types of effects or biases caused by reviews (Magnani 2020). For example, it is empirically demonstrated that biases are arisen by factors such as the difference between positive and negative reviews (Schlosser 2005), social or demographic information of prior reviewers (Forman et al. 2008, Ameri et al. 2019), the user interface of web sites (Jiang and Guo 2015, Chen et al. 2018), etc.

Therefore, under a situation with the trade-off between accuracy and bias, sharing review information is one of the key concerns for consumers to reach better goods/services. In practice, each platform differs in how they share the rating information: for example, Amazon.com, Inc. uses a histogram for expressing customer's ratings, and in the meantime, Expedia, Inc. presents average scores, including sub item scores. Even differences in the user interface will distort what consumers see when browsing websites. So, when one has a little information only, it is further difficult to infer the quality of goods. On the other hand, even if consumers have too much information, there is a possibility that consumers may misunderstand the information. Also, it may cause an information cascade (Duan et al. 2008), where customers ignore their own experiences and believe in others' voices. This phenomenon may cause unhappy matching of customers and goods/services.

This research demonstrates how review scores make customers be able to purchase better goods and eventually what effects are engendered on the market efficiency. To this end, economic experiments are adopted because its controlled laboratory environment enables us to elucidate the mechanism that is focused on, especially by excluding unrelated factors. The causality effect of a review system on purchase behavior is investigated. Since human's rating decisions may include bounded rationality and/or irrationality, and in some cases, may include strategic behavior to manipulate scores or others' behavior, it motivates us to understand the rating and purchasing mechanism by observing in a controlled laboratory how actual subjects decide which signals to send, which is not generally considered by a theory that assumes rational agents. Human subjects purchase an imaginary product and rate it. Afterward, other subjects decide whether to purchase it or not after checking the ratings that the prior subjects did. This approach could bring about deep insights that are impossible to obtain simply by analyzing empirical data. Thanks to the characteristic of preference controllability in experiments, such an analysis can be realized, which prior studies have not fully explored.

This research consists as follows; in Section 3.2, an experimental design and actual procedures is explained. In Section 3.3, the results are reported and discussed in Section 3.4. Conclusion is summarized in Section 3.5.

3.2 Experiment Design

3.2.1 Model

Consider a model where N consumers sequentially decide on purchasing a good. More precisely, for the same kind of goods, the first consumer decides whether to purchase it or not; and after that, the second consumer does, and so forth. If the preceding consumer(s) did the review(s), the subsequent consumers could check the review(s) for consideration of purchase. Herein, the consumer i's utility function is defined as:

$$U_i = \begin{cases} q^b + q_i^h & \text{(if purchasing)} \\ c & \text{(otherwise)} \end{cases} \tag{3.1}$$

where q^b and q_i^h respectively signify the base quality and the heterogeneous quality for the good; and c stands for an initial endowment which is given to each consumer. This formula means that a consumer can purchase the good in exchange for the endowment. If not purchasing, the endowment becomes his/her utility.

The base quality q^b and heterogeneous quality q_i^h are modeled based on the prior researches of theoretical analysis (Besbes and Scarsini 2018, Acemoglu et al. 2017). q^b is the same among all consumers, whereas q_i^h is different for each consumer. Both are determined by the uniform distributions: $q^b \sim U[\underline{q}^b, \overline{q}^b]$ and $q_i^h \sim U[\underline{q}^h, \overline{q}^h]$. Therefore, once q^b is determined before the first consumer makes a decision, it never changes afterward. In the meantime, q_i^h is determined whenever each consumer decides.

If a consumer has made a purchase decision, he/she leaves a review for good. For simplicity, a consumer rates the good on a five-point scale. So, after a consumer confirms the utility value realized by purchase, he/she chooses a value from 1 to 5 as a rating. Now let us suppose that the consumer's index number i means the order of decisions. Then, letting consumer i's rating be $s_i \in \{1,2,3,4,5\}$, the three types of rating scores confronted with consumer i in purchase are defined as:

$$d_i = \begin{cases} s_{i-1} & \text{(recent rating)} \\ \frac{1}{i-1} \sum_{1 \le j \le i-1} s_j & \text{(average rating)} \\ (f(1), f(2), f(3), f(4), f(5)) & \text{(histogram)} \end{cases} \tag{3.2}$$

where $f(x)$ means the number of consumers who left the rating value of $x \in \{1,2,3,4,5\}$. In case of d_1, no rating score is available because no one have not purchased yet.

To sum up, consumer i decides to maximize their utility, consulting the rating score d_i for purchase. Because the heterogeneous quality q^h is independently determined, a consumer has to reason only with the displayed rating score whether the base quality q^b is high enough or not.

3.2.2 Treatments

Four treatments conducted in this study are summarized in Table 3.1. Treatment A is the control group, where consumers have to purchase a good without any rating score. Although any rating score will not be presented to the following consumers, each consumer must decide the rating value. So, in this treatment, their rating behavior will not affect other people's behavior. With that, how participants' behavior on both purchase and rating decisions alters when others see their reviews are examined. Also, whether people reflect their utility value to rating decisions without incentives to notify others is observed.

Treatment B is the group where the recent rating is used. Participants can only see the most recent rating score. It is not easy to find the platform that adopts this system in reality; however, similar situations may exist in reality. For example, it could be the case when consumers may only see several recent evaluations because of a website's cumbersome user interface.

In treatment C, participants can see the average rating that has been posted on the good. It is rounded off to one decimal place. On the contrary, participants can see the histogram of ratings in treatment D. Consumers can know the respective numbers of how many reviewers chose each scale level.

In summary, these four settings are compared to see how these differences affect people to buy better goods.

3.2.3 Arrangement for efficient data collection

In the model, a group in the experiment consists of N subjects. If each subject decides for the good of one kind, at most N reviewers leave his/her rating score for

Tab. 3.1: Treatments

Treatment	Explanation (Consumers can ...)	Example
A (control)	not see any reviews.	No reviews
B (recent rating)	only see the last review on the good	The last rating score was 3
C (average rating)	only see the average value of review score on the good.	The average rating score is 3.2
D (histogram)	see the number of reviews for each rating value on the good.	Score 5: 1 Score 4: 3 Score 3: 2 Score 2: 1 Score 1: 0

the good. In order to obtain the rating scores for a different good, another set of N participants is required. This setting is not an efficient way of data collection. In order to overcome this issue, the experiment is arranged as follows:

- Each subject encounters J kinds of goods for rating.
- Each subject repeatedly decides for T periods, so he/she makes T decisions in total.
- Each subject makes a decision only once for the same good. A subject never has an opportunity to purchase the same good that he/she has encountered before.
- Respective subjects decide for a different good in parallel.

Accordingly, under the arrangement above, based on the decision sequence shown in Table 3.2, which is an instance of $N = T = J = 8$, the review scores for N kinds of goods can be collected only with N subjects. In Table 3.2, each sell presents the consumer's index number i, a column stands for a kind of goods, and a row means a period.

In addition, this arrangement enables subjects to equally encounter situations where the different accumulated number of reviewers have rated. Any rating score is not available for the first in the decision sequence. At the last of the sequence, the subject can see the scores accumulated by all other subjects. That arrangement could avoid unbalanced opportunities to see the reviews are given to the subjects.

Tab. 3.2: Subject's decision sequence in experiments (The case of $N = T = J = 8$)

Period t	Good j							
	a	b	c	d	e	f	g	h
$t = 1$	1	2	3	4	5	6	7	8
$t = 2$	2	3	4	5	6	7	8	1
$t = 3$	3	4	5	6	7	8	1	2
$t = 4$	4	5	6	7	8	1	2	3
$t = 5$	5	6	7	8	1	2	3	4
$t = 6$	6	7	8	1	2	3	4	5
$t = 7$	7	8	1	2	3	4	5	6
$t = 8$	8	1	2	3	4	5	6	7

3.2.4 Monetary rewards

Participants are incentivized as follows. 1600 Japanese Yen (JPY) is guaranteed as a show-up fee. Besides, incentives are calculated based on the results of the experiment. After all experiments are finished, one utility value is selected randomly from each treatment, with four values. These values are added and divided by ten and paid as incentives. The reason why the way of paying only for randomly selected results is adopted is to reduce the wealth effect (Charness et al. 2016). By this incentivized scheme, subjects are expected to decide that maximize their utility at each period. The average payment was 3626 yen.

3.2.5 Setting up experiments

Parameters are set as $N = T = J = 16$. The within-subject design is employed, so all subjects join Treatments A to D. Thus, there are a total of 64 decision-makings for each subject in each group. In order to decrease the order effect, the order of treatments is changed in each group [1]. Review scores and qualities are reset in each treatment.

Then, the parameters are set as $\underline{q}^b = \underline{q}^h = 0$ and $\overline{q}^b = \overline{q}^h = 5000$. As the distribution of qualities, thereby, the base quality $q^b \sim U[0, 5000]$ and the heterogeneous quality $q^h \sim U[0, 5000]$ were used. Besides, c is set as $c = 5000$. Under these parameters, the following holds:

$$E[q^b + q^h] = c. \tag{3.3}$$

[1] We set the order of each treatment as follows: 1:ABCD, 2:DCBA, 3:CADB, 4:BDAC, 5:DCBA, 6:ABCD

This parameter setting means that when any rating score is not available for a subject, purchase option and no purchase option are equivalent to the expected payoff. So, especially if a subject is risk-neutral, he/she equally evaluate the two options. This setting is because decisions could be neutral in case of no rating score, which enables us to focus on the effect of seeing the rating score on decisions.

3.2.6 Experiment procedure

Undergraduate and graduate students of The University of Tokyo are recruited via SNS. A total of 96 subjects participated in this experiment. The experiment was conducted from October 13, 2020, to October 15, 2020, a total of six groups [2]. Each experiment consisted of 16 subjects, and it took approximately one and a half hours to finish all four treatments.

Due to Covid-19, these experiments were conducted online instead of lab experiments, where participants accessed the website set up by oTree (Chen et al. 2016) from home. Firstly, participants are asked to access the Zoom URL 25 minutes before the experiment for check-in. After that, 16 participants were selected for a group randomly and provided with them oTree URL to access the experiment web page. After the experiment starts, the experiment's instructions are read in about 20 minutes. When all treatments are finished, participants are asked to answer a questionnaire and inform them of the results of monetary rewards.

Participants are asked to turn on their videos on Zoom during the experiment. In an online experiment that participants join from home, it is not easy to control the environment precisely in the same way as an on-site lab experiment. However, to realize the equivalent environment to a laboratory, participants are asked to share their videos during the experiment for the experimenter to check whether they are seated in front of the computer and do not conduct any suspicious or cheating behavior during the experiment. However, there might be the effect of seeing other participants' faces thorough Zoom on the result of the experiment. In order to avoid this, participants are asked to minimize the Zoom window in order for them not to see the videos of other participants during the experiment[3].

[2] Ethical approval for this study was obtained from the Research Ethics Committee of School of Engineering, the University of Tokyo (KE20-37)

[3] In some experiments, several participants lost connection to Zoom temporally. All of them managed to reconnect soon after the connection failed, so it is supposed that there is no severe effect on the result of the experiment.

3.3 Findings

3.3.1 Descriptive statistics

Tab. 3.3: Descriptive statistics

		A	B	C	D	All
Base Quality[1]	Mean	2500	2500	2500	2500	2500
	SD	(1537)	(1537)	(1537)	(1537)	(1537)
Heterogeneous Quality[1]	Mean	2463	2533	2472	2501	2492
	SD	(1433)	(1437)	(1449)	(1454)	(1443)
Purchase Decision	Mean	0.56	0.37	0.46	0.50	0.47
(1:Purchase, 0:Not)	SD	(0.50)	(0.48)	(0.50)	(0.50)	(0.50)
Utility	Mean	4968	5239	5384	5433	5256
	SD	(1569)	(1305)	(1418)	(1517)	(1466)
Selected Rating Score	Mean	2.88	3.32	3.32	3.37	3.20
	SD	(1.48)	(1.49)	(1.49)	(1.46)	(1.49)
Displayed Rating Score	Mean		2.20	2.72	2.69	2.54
	SD		(1.41)	(1.19)	(1.21)	(1.29)
# of Decisions		1536	1536	1536	1536	6144
# of Decisions with Displayed Rating Score		0	1376	1367	1371	4114
# of Purchases/Rating Decisions		853	562	714	774	2,903

Note: [1] As a uniform distribution, a subject's base quality value is set randomly in the same group from $\{0, 333, ..., 4667, 5000\}$ without duplication, by which the range between 0 and 5000 is equally divided with the interval of 333. This is because the perfect equality is realized for base quality. In the meantime, regarding heterogeneous quality, an integer is generated from $U[0, 5000]$ at every decision stage. Subjects are only notified that base quality and heterogeneous quality are drawn from $U[0, 5000]$.

Table 3.3 shows descriptive statistics of this experiment. In each treatment, according to the arrangement in subsection 3.2.3, the total number of decisions is 1536 (= 16 [subjects] × 16 [periods] × 6 [groups]). It can be confirmed that base quality is uniformly distributed, and the mean presents the same in each treatment. Also, as for heterogeneous quality, it can be confirmed that the mean values are around 2500 in all treatments [4].

[4] The distribution of realized quality may affect decisions. If high/low heterogeneous quality are generated consecutively, participants may doubt the reliability of rating scores since it does not sufficiently reflect base quality. However, the probability of its occurrence will generally decrease as periods pass. The results considering these effects are shown in Appendix Section A.2

The purchase ratio through all experiments is 47%. Treatment A shows the highest purchase ratio, and B is the lowest. The ratio increases from B to D as information increases. A chi-square test is used to see the difference between treatments; except for the case between C and D, all showed significant difference with $p < 0.01$, whereas C and D showed $p < 0.05$ ($\chi = 4.537, \mathrm{df} = 1$). The average utility is increasing from treatment A to D, wherein although there is no significant difference only between C and D, all the other cases are significantly different with t-test ($p < 0.01$).

In the table, the selected rating score is defined as the rating score that the subject chose after they experienced goods. In contrast, the displayed rating score is defined as the prior rating scores presented for the subject at their decision stage. Thus, the displayed rating score does not include how subjects interpret their observed scores; it only represents the displayed number itself. In addition, the displayed rating score has a different form depending on treatments. Especially in treatment D, the average value of histograms is calculated for the displayed rating score. Compared with the selected rating scores, the displayed rating scores are low. It may be because the lower scored goods do not have enough chances to be bought and updated with new rating scores by participants. T-test shows that the selected rating scores are significantly different between treatment A and any of the others ($p < 0.01$), but not in the other combinations. Regarding the displayed rating score, the B-C and B-D cases demonstrate significant difference ($p < 0.01$), but not in the C-D case.

The number of decisions with displayed ratings shows the total number of decisions that participants can see prior to rating scores. It can be confirmed that there is no significant difference between B to D treatments. Since in treatment A rating scores are not given to participants, the number is 0.

3.3.2 Effects on purchase behavior

Whether participants could have appropriately purchased goods with higher base quality is examined for each treatment. Figure 3.1 shows the relationship between base quality and purchase ratio. In treatment A, irrespective of base quality values, the purchasing ratio is around 0.5. The reason is obvious: subjects have no clue about the purchase decision, so its probability becomes close to 0.5. On the other hand, treatments B, C, and D present a clear linear relationship. Furthermore, treatments C and D demonstrate a sharper slope of the linear relationship between the two variables compared with treatment B.

In order to ascertain the detailed difference among treatments, binary logit regression is conducted for each treatment and summarized the results in Table 3.4. From the table, it is clear that in treatments B, C, and D, *Base Quality* significantly

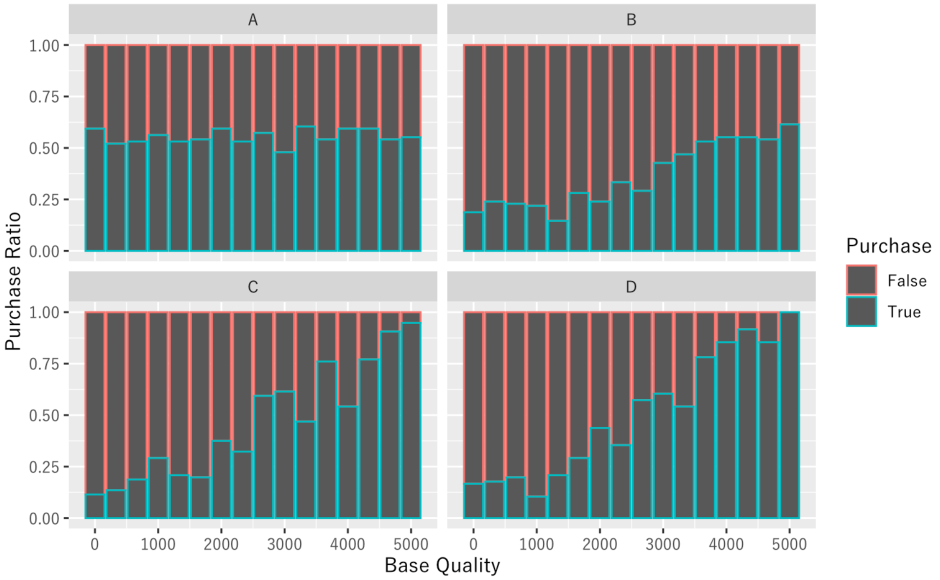


Fig. 3.1: The relationship between base quality and purchase ratio in each experiment

Tab. 3.4: The results of binary logistic regression on purchase decision

	Dependent Variable: Purchase Decision			
	Treatment A	Treatment B	Treatment C	Treatment D
	(1)	(2)	(3)	(4)
Base Quality	0.005	0.14^{***}	0.25^{***}	0.28^{***}
	(0.01)	(0.01)	(0.01)	(0.01)
Period	-0.03^{**}	-0.06^{***}	-0.05^{***}	−0.01
	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.48^{***}	0.30^{***}	0.41^{***}	0.50^{***}
	(0.03)	(0.03)	(0.03)	(0.03)
Group Dummies	Yes	Yes	Yes	Yes
Observations	1,536	1,536	1,536	1,536
Log Likelihood	−1,084.82	−954.53	−862.44	−802.24
Akaike Inf. Crit.	2,185.63	1,925.05	1,740.87	1,620.48

Note: *p<0.1; **p<0.05; ***p<0.01
Binary logit estimations. Standard errors in parentheses.

has a positive effect on *Purchase Decision* at the 0.01 level. Furthermore, since the coefficients of treatments C and D are higher than that of treatment B, it exhibits that participants could realize more correct purchases than treatment B. However, strictly speaking, the difference of coefficients cannot be statistically assured by the result of regression. Further comparison is conducted in Section 3.4. Meanwhile, although treatments C and D are similar in terms of the coefficient of *Base Quality*, a different tendency can be presented on *Period*. Treatment C shows a significant effect on *Purchase Decision* at the 0.01 level, whereas there is no significant effect in treatment D. This implies that the two styles of rating may differentiate purchase decisions.

Next, Figure 3.2 depicts how the average purchasing ratio changes over periods. As demonstrated in Table 3.4, treatments A, B, and C present a declining tendency as the period increases. On the other hand, treatment D does not show any clear increasing/decreasing inclination. However, it is noteworthy that the high ratio at the beginning is due to no rating information. As the period passes, the ratio decreases, meaning that rating scores make participants not purchase low-quality goods. However, it is surprising that treatment D does not exhibit such a trend. This result implicitly indicates a possibility that a histogram type of rating display (treatment D) might lead consumers to make the more appropriate purchase. More details will be discussed in section 3.4.

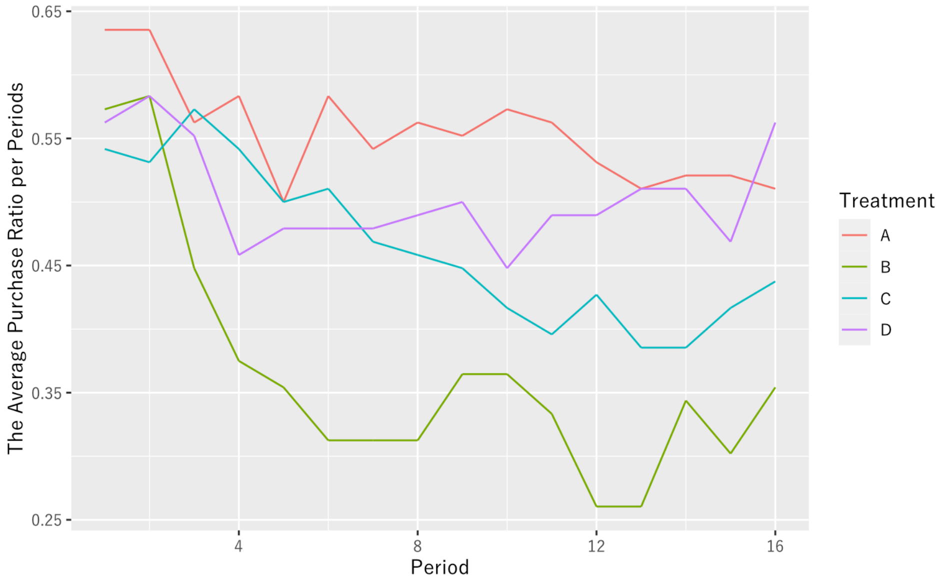


Fig. 3.2: The average purchase ratio per periods

3.3.3 Accuracy of displayed rating score

As seen in the previous subsection, except for treatment A, the result shows that the purchasing ratio increases in proportion to the magnitude of a base quality value. One of the possible reasons for that could be that the displayed rating score is accurate.

First, let us consider what the accurate score means in this context. As explained in section 3.2.1, in the model, the utility is defined as the sum of base quality plus heterogeneous quality. Because the heterogeneous quality for each subject is determined with the IID (Independent Identically Distributed) condition, whose mean value is 2500 in the experiments, rational agents shall choose to purchase goods if the base quality value is greater than or equal to 2500. Therefore, it could be regarded that if the rating score is accurate, it should be proportional to base quality values.

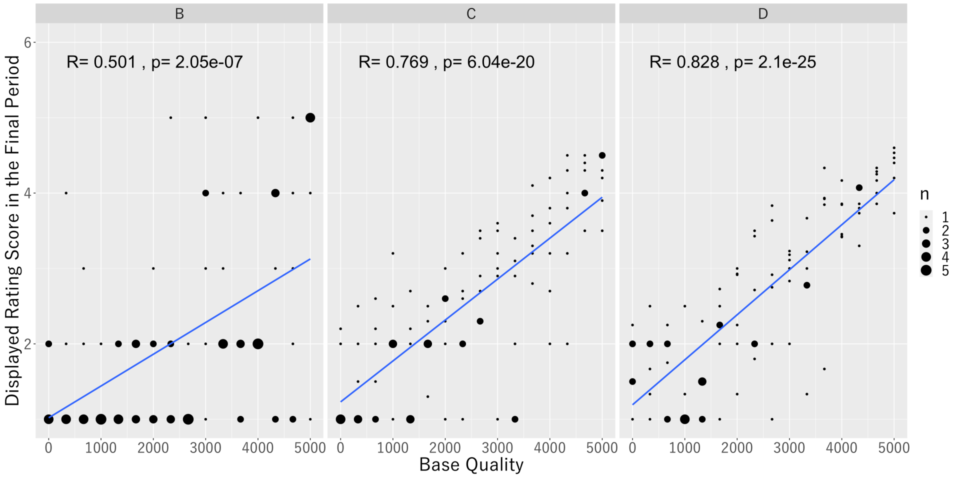


Fig. 3.3: The relationship between base quality and displayed rating score in the final period

Figure 3.3 shows the relation between base quality and displayed rating score at the final period. All treatments present a high positive correlation at the 0.01 level. Mainly in treatments C and D, a clear correlation is observed. However, in treatment B, the low scores can be seen relatively more than the others. Even base quality values with more than 2500 denote low scores such as 1 or 2. This is the main reason why treatment B shows the low perceived score and the low purchasing rate in Table 3.3.

Regarding treatments C and D, it can be considered that the base quality is appropriately reflected on the displayed rating score by aggregating all scores that participants chose. This result indicates that the reason why a review system, in reality, works well is that it fundamentally has this kind of mechanism reflecting base quality on the score. Therefore, even if consumers cannot know the value of base quality, they could make appropriate purchase decisions based on the score. However, in terms of accuracy level, although the correlation coefficient presents a slight difference between C and D, it is not sure which is outperformed. The further accuracy comparison will be conducted in section 3.4.

3.3.4 Effects on rating behavior

Next, how participants make rating decisions and alter their behavior in respective treatments is examined. Figure 3.4 shows the relation of utility values and selected rating scores. These graphs in the figure mean that participants choose higher rating scores as the utility increases. Even in Treatment A, where any rating scores are provided with other participants, an increasing tendency is observed.

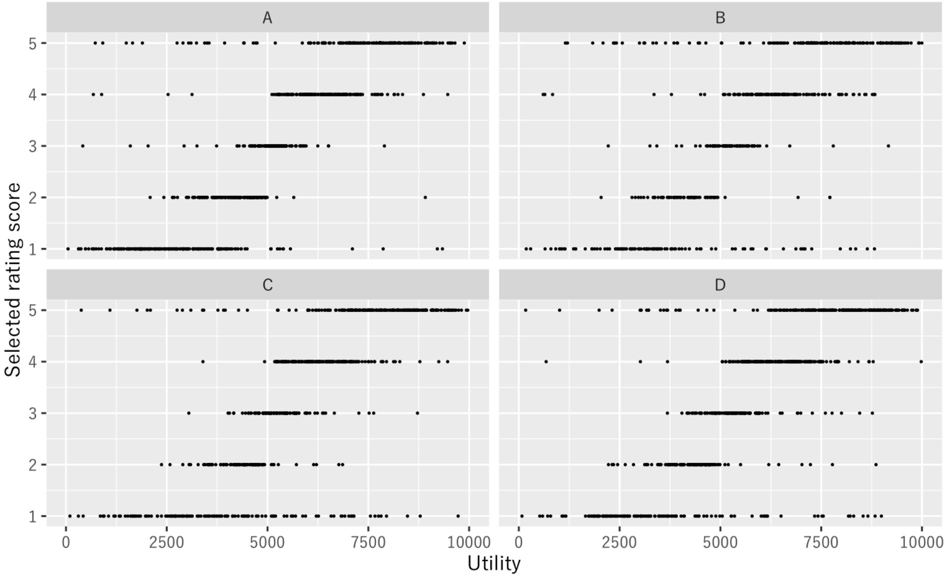


Fig. 3.4: The relationship between utility and selected rating scores in each treatment

The results in the previous subsection demonstrated that in both treatments C and D, there was a strong correlation between displayed rating scores and base quality values. However, participants never know the base quality values and only know the utility value, the total sum of the base quality and heterogeneous quality values. It is interesting that the magnitude of base quality is reflected on the score appropriately.

Tab. 3.5: The result of ordered logistic regression on selected rating scores

	Dependent Variable: Selected Rating Score						
	Treatment A	Treatment B		Treatment C		Treatment D	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Utility	3.63^{***}	2.27^{***}	2.33^{***}	2.91^{***}	2.74^{***}	2.88^{***}	2.69^{***}
	(0.16)	(0.14)	(0.16)	(0.15)	(0.15)	(0.14)	(0.14)
Displayed Rating Score			-0.25^{**}		−0.06		−0.05
			(0.10)		(0.09)		(0.09)
Period	−0.003	0.02	0.01	−0.02	−0.02	0.001	0.01
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Group Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	853	562	466	714	618	774	678
Log Likelihood	−818.92	−665.66	−543.77	−785.97	−671.32	−833.23	−739.54

Note: *p<0.1; **p<0.05; ***p<0.01

Ordered logit estimations. Standard errors in parentheses.

Ordered logistic regression is conducted on selected rating scores to investigate how participants changed their reflection of the observed utilities on rating behavior between treatments. Table 3.5 shows the regression results. From the table, it can be confirmed that in all treatments, the variables of *Utility* are significantly different at the 0.01 level. Treatment A has the largest estimated coefficient of *Utility* and treatment B has the smallest. Even if the variable of displayed rating score is added (models (3), (5), and (7)), the results are not affected at all. On the other hand, *Displayed Rating Score* is not relevant to their behavior of *Selected Rating Score*. Only in treatment B, *Displayed Rating Score* presents an adverse effect, but it is the 0.05 significance level. So, participants in treatment B tend to lower the score if they saw higher scores and conversely raise the score if they saw low scores.

From these results, it can be concluded that if participants were not biased by prior reviews and could rate it based on their utility value, the product/service's quality level could be appropriately reflected on the aggregated rating score. However, only in treatment B, the displayed rating score partly affects raging behavior.

3.4 Discussion

3.4.1 Comparison of treatment C and D prediction performance

In order to compare treatments C and D more precisely, the idea of ROC (Receiver Operating Characteristic) curve is employed, which was widely used originally in medical science and now prevails in the domain of machine learning in information science. The ROC curve is an effective method for evaluating the performance of diagnostic tests. This method can be applied if the displayed rating score is interpreted as an output value obtained by diagnosis tests.

First, in the context of review systems, *True Positive* (TP), *True Negative* (TN), *False Positive* (FP) and *False Negative* (FN) are defined as follows:

- TP: If the rating score exceeds a certain threshold, the product/service quality is good. So, it should be purchased.

- TN: If the rating score is below a certain threshold, the product/service quality is bad. So, it should not be purchased.

- FP: Even if the rating score exceeds a certain threshold, the product/service quality is bad. This indicates that the rating score is wrong and too high.

- FN: Even if the rating score is below a certain threshold, the product/service quality is good. This indicates that the rating score is inversely wrong and too low.

Then, if the definition above is explicitly applied to the model, it can be summarized in Table 3.6. In the model, as explained in Section 3.3.3, a rational decision is to purchase a good with the base quality greater than or equal to 2500, so actual values are separated at the level of 2500.

Tab. 3.6: TP, TN, FP and FN in the model

		Predicted	
		Rating ≥ Threshold (Positive)	Rating < Threshold (Negative)
Actual	Base quality ≥ 2500 (Positive)	TP	FN
	Base quality < 2500 (Negative)	FP	TN

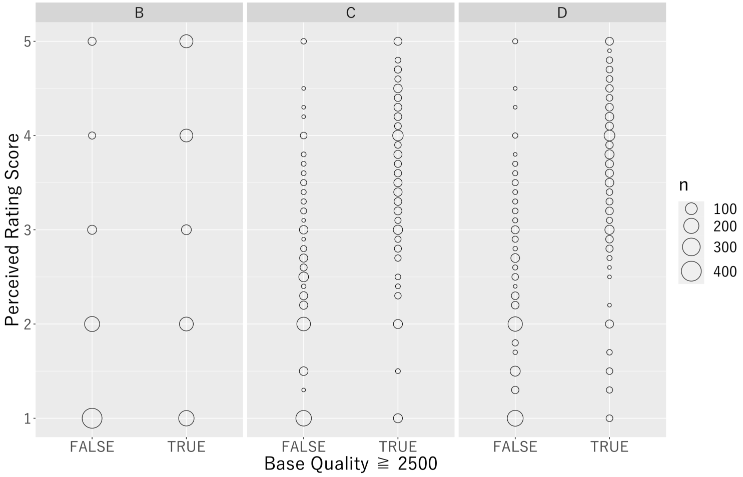


Fig. 3.5: The relationship between base quality and displayed rating score

Figure 3.5 depicts the graphs plotting data according to the concept of TP, TN, FP, and FN. In the graphs, the displayed rating scores are separated into the "≥2500" and "< 2500" base quality group in respective treatments. The circle size alters depending on the frequency of observations. If the threshold is set to around 2.5-3.0 on the vertical axis, the data point in treatments C and D could be

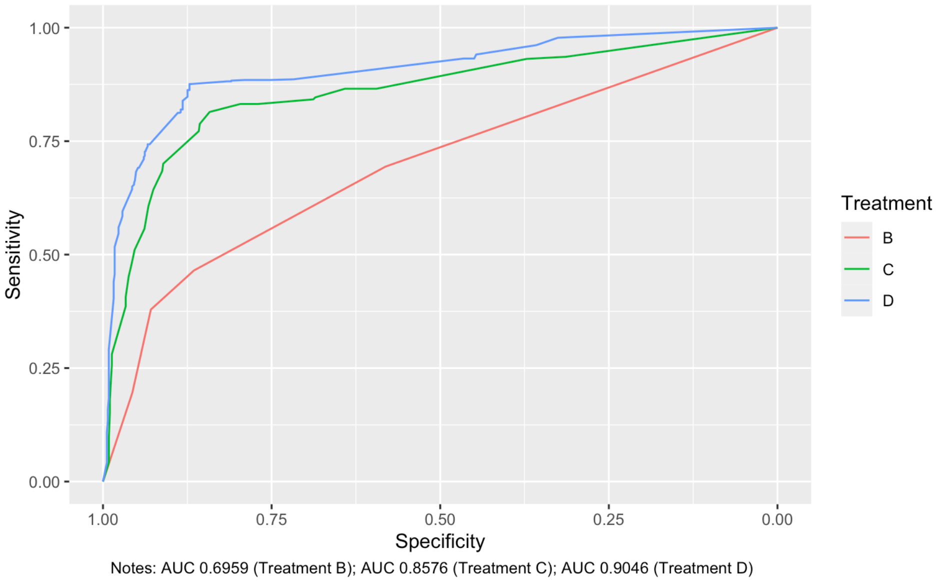


Fig. 3.6: The ROC curve

nicely separated into TP and TN. By using the data, accordingly, ROC curves can be drawn as shown in Figure 3.6. Sensitivity and Specificity are defined as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3.5}$$

Although it is challenging to differentiate C and D in regression analysis in section 3.3, the ROC curves visually demonstrate their difference. Treatment D distinctively outperforms treatment C. AUC (Area Under Curve) also demonstrates treatment D is the highest (AUC = 0.9046), meaning that the treatment D realizes a highly accurate prediction. It can be concluded that a histogram display can predict base quality values more accurately than the others.

The above finding that histograms showed better performance than the average score, in some sense, is consistent with Hu et al. (2006), which reported that the actual distribution of the numerical value of reviews tends to be bimodal and doubt the reliability of the average value. However, the results indicate that the display of average ratings presents relatively better performance (AUC = 0.8576). Since the experiments exclude all unrelated factors and participants make a decision only based on utility value, in the result by Hu et al. (2006), other factors, in reality, might have affected human rating behavior, and a bimodal distribution might have been formed. It would be important what factors could produce such a bimodal distribution as empirical data.

Finally, how the prediction accuracy affects the market efficiency is elucidated. In this study, it is assumed that market efficiency is defined as the following ratio:

$$\text{Efficiency} = \frac{N^{Buyers}}{N^{LatentBuyers}} \tag{3.6}$$

N^{Buyers} shows the number of subjects who purchased a good with >2500 base quality, and $N^{LatentBuyers}$ shows the number of subjects who had the opportunity to purchase a good with >2500 base quality. As a result, Table 3.7 is obtained, which is consistent with the accuracy level. From an experimental point of view, it is concluded that review systems such as the average score and histogram work well to improve market efficiency. Especially the histogram presents the best performance.

Tab. 3.7: The Comparison of market efficiency

Treatment A	Treatment B	Treatment C	Treatment D
0.560	0.497	0.701	0.766

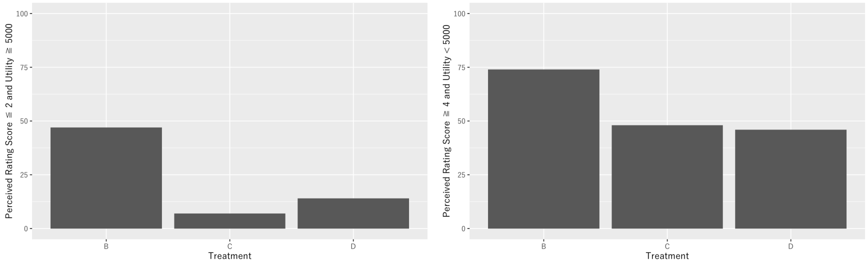
3.4.2 Does treatment B engender a disconfirmation effect?

The experiments confirmed that as information increases from treatment A to D, consumers reach better goods instead of causing information cascade or falling into a bad situation. In the meantime, more participants were negatively affected by displayed rating scores in treatment B, meaning that they tend to rate it conversely. This subsection analyzes such an effect in more detail.

As already shown in Table 3.5, the coefficient of *Displayed Rating Score* was significantly negative in treatment B. It means that if a displayed rating score is high, then participants tend to lower their rating score, and in the meantime, if a displayed rating score is low, they tend to raise their rating score. It could be arisen due to inaccuracy of ratings. If the accuracy is low, participants often are confronted with a situation with FP or FN case, meaning that it is a case with a high rating score and small utility or a case with a low rating score and large utility. This harms rating behavior.

In service marketing, it is broadly said that satisfaction of service is determined according to the disconfirmation of expected quality and experienced quality (Oliver 1980). Based on the theory, consumers may generally reflect their discrepancy between negative expectations by previous negative reviews and actual positive quality in use to their reviews, and vice versa. The result indicates that if consumers see only a recent review (treatment B), it may increase the possibility that the disconfirmation effects happen in online review systems. In general, the effect would be desirable if the discrepancy were unexpectedly positive because consumers feel more satisfaction than they expected (Anderson and Sullivan 1993). However, as shown in Figure 3.7(a), treatment B demonstrates that an unexpectedly negative case is frequently observed. Although Figure 3.7(b) shows many unexpectedly positive cases, there is not a considerable difference between treatment B and the others.

Interestingly, such a significant adverse effect on rating behavior can be observed only in treatment B. That phenomenon can be directly explained from Figure 3.7, but the discrepancy in expectation is mainly caused by heterogeneous quality in experiments. Therefore, if the rating scores are not accumulated, such a disconfirmation effect is likely to happen.



(a) Rating score ≤ 2 and utility ≥ 5000

(b) Rating score ≥ 4 and utility < 5000

Fig. 3.7: The frequency of the disconfirmation effect

3.4.3 Benefit of experiment method

This subsection recaptures and summarizes the benefit of the lab experiment approach in review system studies.

One of the most remarkable features of experimental lab methods is the controllability of preference (Smith 1976). It is impossible to observe people's satisfaction and their utility functions directly. Under an empirical condition, even if one could investigate rating scores with real data, it is naturally challenging to elucidate the fundamental mechanism behind review systems because various factors are intertwined. Against these backdrops, the experimental method enables participants to induce preference (utility function), which experimenters designed beforehand. Therefore, researchers can confine their focus on the mechanism of review systems.

In this study, Figure 3.5 is an exampled. Naturally, data on FP and FN are never observed in reality. Thanks to the experimental method, those data can depict the ROC curve. That enabled us to compare the display styles of average score and histogram. Also, unnecessary factors for analysis could be excluded in this study. Many preceding studies have reported that biases may distort the review results. A typical example is self-selection bias (Li and Hitt 2008, Hu et al. 2009). In the meantime, this experiment could exclude this factor because an experimental situation is designed that subjects must post their reviews always after making a purchase decision. This means that three review systems can be compared under a fair comparison environment.

So far, there have been many studies on online review systems. However, most of them are inclined to empirical aspects using actual data but do not employ experimental methods despite few exceptions (Lafky 2014, Halliday and Lafky 2019). However, this study ascertains that the lab experimental approach on online review systems works well and can complement empirical studies in many respects. This research could have significant contributions to online review system studies.

3.4.4 Risk attitude

In this study, participants' risk attitudes were not evaluated. However, some results are probably related to their risk attitudes. Firstly, in the case of a histogram, risk-averse participants may not buy goods when histogram shows ratings of 1 or 2 even it contains a lot of 5 ratings. In reality, several bad ratings can prevent people from buying goods. However, this experiment showed that histograms generate a higher purchase rate than average scores. Two reasons can explain this deviation from reality. Researchers let participants know that quality is distributed in the sum of uniform distribution to simplify the experimental settings. In reality, consumers are uncertain about the distribution of quality, and it may cause more risk-averse decisions. Another reason is that consumers are informed that there are no incentivized fake reviewers. Consumers suppose that some reviewers write good ratings for rewards, and consumers weigh bad ratings more and ignore good ratings. To consider these features of rating systems in reality, researchers should expand models to include uncertainty of distribution and the possibility of incentivized fake reviews.

Another phenomenon that is related to risk attitude is the relationship between purchase rate and periods in Figure 3.2. Risk-averse participants should avoid buying goods in earlier periods because there are not enough reviews for participants to purchase goods. However, in Figure 3.2, the purchase rate was around 60%, and it seems too high for risk-averse consumers. This result may be because participants expect other participants to buy and write reviews in later periods by showing their cooperative behavior of writing reviews in earlier periods. This behavior happened because the number of participants is lower than that in the real world, and each participant's decision on other participants is influential. Participants may think that sending a signal to others is more beneficial when expecting other participants to reciprocate. Evaluating the effect of the number of participants on purchasing behavior may fill the gap between the experiment and reality.

3.5 Chapter Summary

In this research, an economic experiment is conducted to analyze the effects of information shared on the review platform toward the improvement of consumers purchasing behavior. The result showed that, despite the heterogeneity of preference of goods, average ratings and all history of reviews can make consumers reach better goods. Furthermore, histograms attained better the purchase and the accuracy of rating value compared with other settings.

Further analysis was conducted on the effect of previous reviews on subsequent consumer reviewing behavior. In treatments C and D, participants directly reflect their utility values without the effect of bias on their reviewing behavior. On the contrary, since participants do not have enough reviews in treatment B, they are negatively affected by previous reviews when reviewing. The observed behavior is consistent with disconfirmation theory. In summary, in this experiment, people do not honestly reflect their preferences or experience of goods to their reviews if there is not enough information beforehand. This is mainly caused by the review system's low accuracy, as confirmed in treatment B.

This study contributed to academic research in light of experimental economics to see how the utility or realized experience of goods relates to numerical review values. In the case of real data analysis, preferences are not controlled in the real world, so there is no proof that reviewers who post five stars liked that product or service. This study used economic experiments to control the preference on goods and add some precise results between utility and numerical rating value of reviews.

Toward managers who run online review systems in business, this study implies that showing histograms and average values of numerical reviews could overcome the distortion due to preference heterogeneity and make consumers buy valuable goods. In addition, it is better to improve the websites' user interface so that consumers can fully view all the information before deciding to buy goods, even in a limited time.

However, in the meantime, this research has several limitations. In this experiment, some realistic features of review systems are ignored for simplicity, but they could become limitations. For example, every participant who bought a good has to make a review decision. On the one hand, this could reduce effects such as self-selection bias; and on the other hand, it may miss a notable phenomenon that can be often seen in reality: the distribution of rating tends to be U, or J shaped (Dellarocas and Narayan 2006, Eugene W. Anderson 1998, Hu et al. 2009). Lafky (2014) did an economic experiment and observed J shape when there is a cost of reviewing. However, in this research, because it is not wise to decrease the number of reviews written in the experiment, all buyers must post their reviews. Besides, several other features are ignored that real review systems naturally have: text reviews, the information before purchasing other than reviews, and reviewers' information. These features are not included in this study because the main focus is to see the effect of the mechanism behind review systems and to clarify consumers' fundamental behavior.

In future research, considering the different distribution of preference heterogeneity can understand the difference between niche and mass consumers on the platforms. In this experiment, the distribution of quality is fixed as the sum of two uniform distributions. This assumption is seemed to be simple and suited for the experiment to examine the effect of the rating system. However, this assumption does not fully capture the characteristics of products or services in reality. Some services may have more multimodal distribution for both base and heterogeneous quality. Changes in distributions will affect the variance of utility. If the variance becomes high, then risk-averse participants will avoid buying goods. Comparing the difference of distributions can clarify what kind of service can benefit from rating systems. For example, heterogeneous quality distribution with a peak in low value and a small peak in high value can be represented as niche goods. It may be interesting to compare niche goods with mainstream goods.

Another interesting topic of heterogeneity is predicting heterogeneous preference and assisting customers to buy goods that fit their preferences. Studies in recommendation systems have focused on this topic Kamishima (2016). Generally, recommendation systems define the similarity of preference among customers based on their purchase history. In this study, the similarity among customers is not dealt with since preference heterogeneity is given randomly for simplicity. Thus there is no correlation among consumers' heterogeneous preferences. If there is a correlation among customers' preferences, consumers might predict how their preference is similar or dissimilar from others based on reviews and their purchased experiences. For example, consumers' similarity of preference can be defined as the distance on circumference (Hong and Pavlou 2014). Consumers can predict which consumer is similar to their positions on circumference if the preference is assumed in this simple setting. Further research is required for treating heterogeneity.

Another interesting future challenge is examining the strategy of participating in rating platforms by service providers whose base quality seems low. For example, in reality, fast food chain stores cannot earn high scores in online ratings. One reason for a small rating score is that those rating scores do not consider the convenience of location and only focus on service quality. Since considering such a complex feature of ratings and modeling the advertising effect of online review platforms is challenging, empirical studies are recommended to clarify the strategy. Using the data of profit or other related stores indices may elucidate the effect of participating platforms if the data satisfies some quasi-experimental conditions; for example, comparing the profit of stores before and after they are posted on the rating websites.

Chapter 4

Investigating the Effect of Time Series Reviews on a Service Market

4.1 Introduction

Rating systems have become common sources of information regarding the quality of products and services. Consumers check the rating scores before buying any product or availing any service. There are various types of scores displayed, such as average scores, histograms, the number of good icons, and multidimensional scores, which can help consumers find better goods (Chen et al. 2018). For example, TripAdvisor rates hotels on the basis of their location, cleanliness, services, and value.

This study focuses on time series ratings. For example, YouTube provides time series indices that show the number of viewers in each period. Consumers can ascertain the interesting or important moment and forward a video accordingly, thereby saving time. Considering time series of ratings will be suited for service markets. A service is typically provided in a continuous process. As a result, to spend their time wisely, consumers should understand how a service is designed to provide benefits in its limited time. For example, the organizers of a music festival must decide the order of the performance of various artists. To allow viewers to arrive early in the morning, the festival must invite famous musicians to perform in the morning. However, to maximize consumer satisfaction, the festival requires well-known artists to perform late in the festival to leave a lasting impression on the audience.

This study examines the impact of time series ratings on the service market. Consumers’ subjective assessments influence their rating behavior. Therefore, this study includes an economic experiment that was conducted for further analysis. This study examines whether consumers use various services optimally based on their ratings. Consumers decide when to stop using service. Compared with digital content such as YouTube, interactive services are more constrained by geological issues and time. For example, in sightseeing tours, consumers must follow their guides. It is also difficult to join a sightseeing tour that is already underway. For this reason, and to make the analysis simple, we only allow one service defection and do not consider consumers joining in the middle of a service.

The results show that consumers can make better use of services with time series ratings. However, despite using a time series rating system, consumers cannot fully benefit from U-shaped services. When the psychological aspects of consumers are considered, U-shaped services are the best resource allocations. This study demonstrates that service allocation and rating systems may have trade-offs.

4.2 Experimental Design

4.2.1 Model

N consumers choose whether or not to use a service. Each service $j \in \{1, ..., J\}$ can be used for periods up to T. For each period $t \in \{1, ..., T\}$, service j has an activity level x_{jt}, which has been defined in previous studies (Baucells and Sarin 2010, Gupta et al. 2016). Activity levels represent the instant excitement of the services in each period t. In summary, service j is characterized by an activity-level vector $x_j = (x_{jt})_{t\in\{1,...,T\}}$.

In each round r, consumer $i \in \{1, ..., N\}$ chooses whether or not to use service j in the period $t = 1$. Consumers can observe the rating score d_{jt} of service j if it is available before making a decision. If consumer i chooses to start using service j, consumer i observes the activity level x_{j1} of service j in period $t = 1$, and gains instant utility $v(x_{j1})$. Then, consumer i must consider whether to continue using service j in the period $t = 2$. Suppose that consumer i chooses to discontinue using service j, consumer i can obtain utility equal to the opportunity cost c_t, which consumer i loses when they continue to use service j. Once the consumer chooses to discontinue using service j, they cannot restart using the same service j; nonetheless, they earn opportunity cost c_t for the rest of the periods. This procedure is summarized in Figure 4.1.

We define t_{ij} as the period in which consumer i stops using service j. Then, consumer i's utility is defined as

$$u(t_{ij}) = \begin{cases} \sum_{t=1}^{t_{ij}} x_{jt} + \sum_{t=t_{ij}+1}^{T} c_t & (t_{ij} \in \{1, ..., T-1\}) \\ \sum_{t=1}^{T} x_{jt} & (t_{ij} = T) \\ \sum_{t=1}^{T} c_t & (t_{ij} = 0), \end{cases} \tag{4.1}$$

where $t_{ij} = T$ indicates that consumer i uses the service until the last period, T. $t_{ij} = 0$ indicates that consumer i does not use service j at all.

4.2.2 Treatment

Two rating systems have been compared: (1) an ordinal rating system (control experiment) and (2) a time series rating system (treatment experiment). The illustration is shown in Figure 4.3. The two systems differ in terms of the timing of their ratings and how consumers observe them. In the control experiment, consumers decided on the total satisfaction value of service j when they finished consuming the service as shown in Step 3 in Figure 4.2. The rating is defined as $d^A_{ij} \in \{1, 2, 3, 4, 5\}$, where one is the worst and five is the best. Participants had to rate the service if they decided to use it, regardless of whether or not they consumed

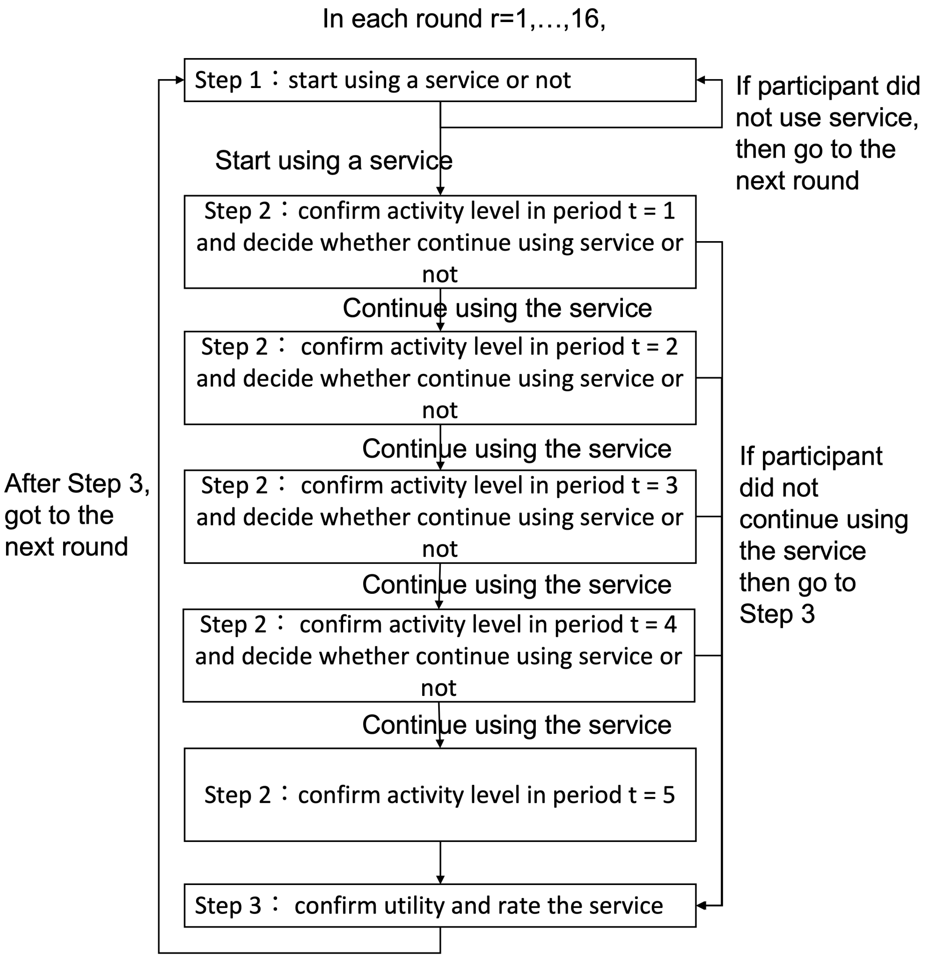

In each round r=1,…,16,
Step 1：start using a service or not
If participant did not use service, then go to the next round
Start using a service
Step 2：confirm activity level in period t = 1 and decide whether continue using service or not
Continue using the service
Step 2： confirm activity level in period t = 2 and decide whether continue using service or not
Continue using the service
Step 2： confirm activity level in period t = 3 and decide whether continue using service or not
Continue using the service
Step 2： confirm activity level in period t = 4 and decide whether continue using service or not
Continue using the service
Step 2：confirm activity level in period t = 5
Step 3： confirm utility and rate the service
After Step 3, got to the next round
If participant did not continue using the service then go to Step 3

Fig. 4.1: The procedure of control experiment

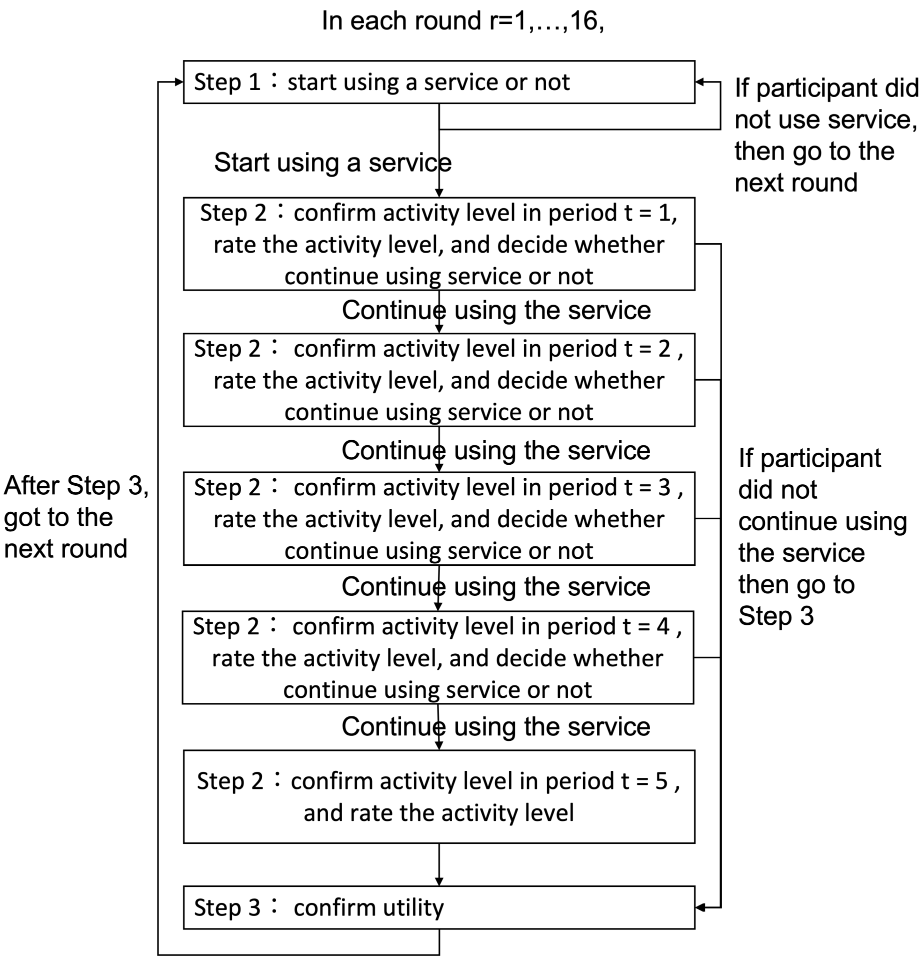
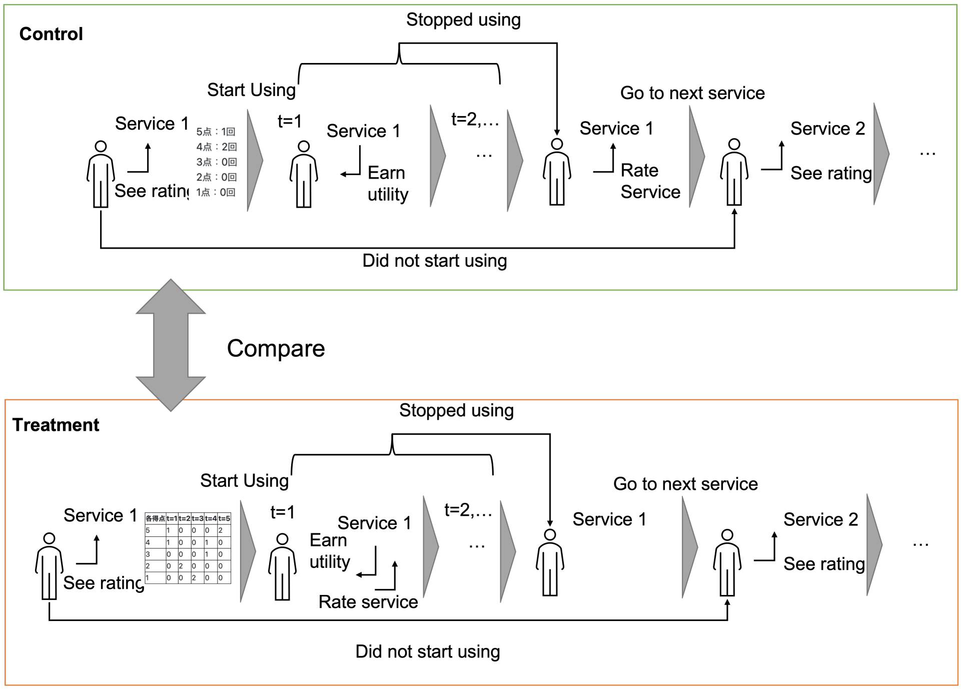

In each round r=1,…,16,
Step 1：start using a service or not
If participant did not use service, then go to the next round
Start using a service
Step 2：confirm activity level in period t = 1, rate the activity level, and decide whether continue using service or not
Continue using the service
Step 2： confirm activity level in period t = 2 , rate the activity level, and decide whether continue using service or not
Continue using the service
Step 2： confirm activity level in period t = 3 , rate the activity level, and decide whether continue using service or not
Continue using the service
Step 2： confirm activity level in period t = 4 , rate the activity level, and decide whether continue using service or not
Continue using the service
Step 2：confirm activity level in period t = 5 , and rate the activity level
Step 3： confirm utility
After Step 3, got to the next round
If participant did not continue using the service then go to Step 3

Fig. 4.2: The procedure of treatment experiment

the entire service until the fifth period. For example, those who used a service for only one period rated it in the same style as those who used the same service for a maximum of five periods. This setting was adopted because, in reality, sometimes consumers post reviews, although they did not use the services fully. However, this rating may not be sufficient to inform other consumers regarding the activity levels of the services. Rating scores consist of ratings posted by users with different lengths of usage. Time series ratings are expected to solve these problems.



各得点	t=1	t=2	t=3	t=4	t=5
5	1	0	0	0	2
4	1	0	0	1	0
3	0	0	0	1	0
2	0	2	0	0	0
1	0	0	2	0	0

Fig. 4.3: An illustration of this study

In the treatment experiment, participants rated the service for each period after experiencing each activity level of service. Rating is defined as $d_{ij}^{B} = (d_{ijt})_{t\in\{1,\ldots,T\}}$, $d_{ijt} \in \{1,2,3,4,5\}$. Compared with the control experiment, d_{ij}^{B} is defined as the vector of d_{ijt}. The number of ratings that participant i provided for j depended on the number of periods in which service j was used. Depending on how long they used the service, some rated only the first period, whereas others rated it from period $t = 1$ to period $t = 5$.

This research also aimed to clarify the differences in activity patterns in the rating system. Multiple patterns of service-activity levels were also prepared. The three main activity-level allocations are crescendo, decrescendo, and U-shaped patterns. The crescendo pattern encompasses x_{ij} that $x_{ijt} < x_{ijt+1}$ for all t. The activity level in this pattern gradually increased during the later periods. Decrescendo pattern encompasses x_{ij} that $x_{ijt} > x_{ijt+1}$. The activity level in this pattern gradually decreased during the later periods. The U-shaped pattern encompasses x_{ij} that $x_{ijt} > x_{ijt+1}$ for $t = 1, 2, 3$ and $x_{ijt} < x_{ijt+1}$ for $t = 4, 5$. These three patterns were selected based on the results of previous studies. Gupta et al. (2016) showed that the crescendo and U-shaped patterns are important upon considering the memory decay and acclimation of consumers. They proved that these patterns maximize consumers' experience of utility in their settings. Several other patterns were also included for concerns that participants might predict all the patterns added in this research and further distort their behavior. These additional patterns, however, were not used in the analysis.

4.2.3 Arrangement for efficient data collection

Since the number of subjects was limited, the order of the services that each participant encountered was determined, as shown in Table 4.1. This order is the same as that described in Section 3.

Tab. 4.1: Subject's decision sequence in experiments (The case of $N = T = J = 8$)

Round r	Service j							
	a	b	c	d	e	f	g	h
$r = 1$	1	2	3	4	5	6	7	8
$r = 2$	2	3	4	5	6	7	8	1
$r = 3$	3	4	5	6	7	8	1	2
$r = 4$	4	5	6	7	8	1	2	3
$r = 5$	5	6	7	8	1	2	3	4
$r = 6$	6	7	8	1	2	3	4	5
$r = 7$	7	8	1	2	3	4	5	6
$r = 8$	8	1	2	3	4	5	6	7

4.2.4 Setting up experiments

Parameters were set as $N = 16, J = 16, T = 5, R = 16$. For simplicity, we defined the opportunity cost in each period as $c_t = c = 200$. Activity levels were defined based on two elements: the total activity level of each service and its patterns. Participants were told that the total value of activity levels was drawn from an unknown distribution with a mean of 1000[1]. The total value of the activity levels was the same as the total opportunity cost for $T = 5$ periods. Therefore, the participant's expected utility for the first-round decision, which should be made without the displayed ratings, was the same.

However, the actual setting of the total activity level was determined using a more complex procedure for two reasons: to reduce the variance of the total activity level between groups and to ensure that participants do not predict the actual parameter settings. To minimize the variance in the total activity level between groups, each group had to have almost the same realized distribution of the total activity level of services. Therefore, the total activity level of each service was allocated from the set {750, 750, 800, 800, 850, 850, 900, 900, 1100, 1100, 1150, 1150, 1200, 1200, 1250, 1250 }, consisting of 16 numbers, with an average number of 1000. However, in this simple setting, participants could have predicted the value of the total activity level from their experience as the experiment proceeded. To prevent this, a randomized value $z \sim U[-50, 50]$ was added for each total activity level. This randomized variable was supposed to ensure that the participants do not predict the value, but rather believed that the total amount is generated from a simpler unknown distribution.

After setting the total activity level for each service, the activity allocation level was chosen for each period. As explained in Section 4.2.2, three main activity level patterns were included: crescendo, decrescendo, and U-shaped. Additionally, four other dummy patterns were added to ensure that the participants did not predict the patterns used in the experiment. These dummy patterns were selected randomly, except for dummy pattern 1, which was a simple flat allocation. The patterns are summarized in Table 4.2, and their graphical images are shown in Figure 4.4. These values indicate the ratio of the activity level in that service. The total activity level of each service was allocated based on Table 4.2. However, to ensure that participants did not predict the patterns, $z' \sim U[-30, 30]$ was added for each activity level. Table 4.3 presents an example of the actual distribution of services and activity levels. All the values used in the experiment are listed in Appendix Section A.5.

4.2.5 Evaluation of Risk Attitude

This study further evaluated the risk attitude of each participant. Risk attitude is determined by the decision to use services when there are no ratings. For example, the decision to start using a service for the first time. Participants had two options. The first option was to begin using services with uncertain activity levels. The

[1] Participants were expected to make decisions with the knowledge of independent and identical unknown distribution with a mean of 1000.

Tab. 4.2: The patterns of activity levels

Pattern	The number of services in the experiment	Allocation of activity levels (ratio) t=1	t=2	t=3	t=4	t=5
Crescendo	4	1	2	3	4	5
Decrescendo	4	5	4	3	2	1
U-shaped	4	4	2	1	3	5
Dummy Pattern 1	1	3	3	3	3	3
Dummy Pattern 2	1	3	2	5	1	4
Dummy Pattern 3	1	1	4	5	2	3
Dummy Pattern 4	1	3	2	1	5	4

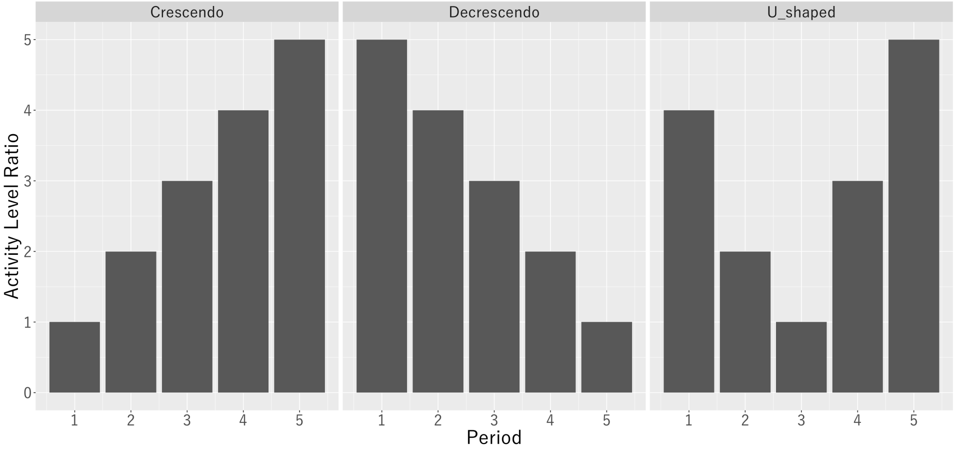


Fig. 4.4: The graphical image of major three activity level patterns

Tab. 4.3: The example of activity levels of services.

Service	Pattern	Allocation of activity levels					
		t=1	t=2	t=3	t=4	t=5	total
1	Crescendo	93	142	223	329	428	1215
2	U-shaped	265	131	83	238	378	1095
3	U-shaped	237	99	48	152	244	780
4	Crescendo	60	190	261	320	444	1275
5	Dummy Pattern 3	37	185	266	128	134	750
6	Decrescendo	235	211	124	109	56	735
7	Decrescendo	373	305	230	170	92	1170
8	U-shaped	314	133	63	267	423	1200
9	U-shaped	249	137	32	148	289	855
10	Dummy Pattern 2	243	125	362	89	306	1125
11	Decrescendo	271	198	162	76	73	780
12	Dummy Pattern 4	171	133	64	274	228	870
13	Decrescendo	369	253	211	129	88	1050
14	Dummy Pattern 1	258	263	212	239	228	1200
15	Crescendo	57	142	151	255	265	870
16	Crescendo	60	96	153	238	308	855

second option was not to start using services and earn profits from opportunity costs. Risk attitude data, therefore, enrich the analysis of consumers.

Participants took the risk attitude test after completing the two experiments. This study adopted the bomb risk-elicitation task introduced by Crosetto and Filippin (2013). This task was implemented using oTree (Holzmeister and Pfurtscheller 2016). In the elicitation task, 64 boxes were displayed on a screen. After the participants pressed the start button, the task started, and boxes were checked one by one every second. The participants could stop checking the box at any time they wanted. Participants' utility is defined as follows:

$$\text{utility} = \begin{cases} 10 \times \text{the number of checked boxes} & \text{(no bomb in the checked boxes)} \\ 0 & \text{(a bomb in checked boxes)} \end{cases} \tag{4.2}$$

The participants' utility was 20 times the number of boxes checked during the task. However, the participants did not know which box contained the bomb. The participants could not obtain utility if the bomb box was in the checked box. The utility ranged from 0 to 1260, and the expected value was 315. Three rounds of risk elicitation were performed for each participant.

4.2.6 Monetary Rewards

Participants received monetary rewards based on the utility they earned in the experiments. The monetary rewards were calculated as follows:

$$\begin{aligned} \text{Monetary Reward} \quad &= \quad \text{Participation Reward} + \text{Control Experiment Reward} \\ &+ \quad \text{Treatment Experiment reward } + \text{Risk attitude task reward} \end{aligned}$$

The participation reward was a fixed amount of 1500 Yen. For the control and treatment experiments the reward depended on the utility value where the reward value was equal to the realized utility value in a randomly chosen round. For example, if round 8 is randomly selected for the control experiment, participant *i*'s control experiment reward will be the utility value earned in round 8. The wealth effect (Charness et al. 2016), which applies to the risk-attitude task reward, is expected to be reduced by randomly selecting one result; the reward is the realized utility in a randomly chosen round during the risk-attitude evaluation.

4.2.7 Experimental procedure

Undergraduate and graduate students of the University of Tokyo were recruited via SNS. A total of 96 students participated in the experiment. This experiment was conducted from August 24, 2021, to August 26, 2021, and from September

14, 2021, to September 16, 2021, for a total of six groups [2]. Each experiment involved 16 participants. As described in Chapter 3, this experiment was conducted online. Participants accessed the website set up by oTree (Chen et al. 2016) from their respective homes. For check-in, the participants were asked to access the Zoom Videoconference URL, 15 minutes before the experiment. Sixteen participants were randomly selected for each group and provided an oTree URL to access the experiment web page. Two experiments were conducted, followed by a risk attitude evaluation.

4.3 Findings

4.3.1 Descriptive statistics

Table 4.4 shows the descriptive statistics of the results of the experiments. The control and treatment experiments' results are summarized in the *All* column. The results of each pattern are summarized in the *crescendo*, *decrescendo*, and $U - shaped$ columns. $StartUsing_{ir} \in \{0, 1\}$ is a dummy variable that takes the value of 1 if the participant decided to start using the service, and 0 otherwise. The average variable in the table represents the proportion of participants who chose to begin using the services. The results of both the control and treatment experiments show that more than half of the decision makings, participants chose to begin using the services.$UsedPeriods_{ir} \in \{1, 2, 3, 4, 5\}$ shows the periods of the service usage. No usage data has been excluded while calculating the average value in the table. The results show that $UsedPeriods$ depends on patterns.

4.3.2 Effects on using decisions

Table 4.5 shows the results of binary logit regression on $StartUsing$. $Treatment$ is a dummy variable that shows whether the data is from a treatment experiment. Column (1) of Table 4.5 shows that more participants decided to start using services in the treatment condition. The results of (1) is consistent with those of (2) when the control variable (total activity level of each service) and fixed effects of round and player are added. The same regression was conducted for each service activity pattern. Figure 4.5 depicts the ratio of decision-makings in which participants choose to begin using services in vivid colors rather than grayscale. The ratio is decreased in crescendo services and increased in the decrescendo and U-shaped services. The results in Table 4.5 show that these differences are significant in (3),

[2] Ethical approval for this study was obtained from the Research Ethics Committee of the School of Engineering, University of Tokyo (KE21-29)

Tab. 4.4: Descriptive statistics

experiment	control							
pattern	All		crescendo		decrescendo		U-shaped	
	mean	sd	mean	sd	mean	sd	mean	sd
Utility	1061	133	1046	117	1123	167	1041	114
StartUsing	0.65	0.48	0.53	0.50	0.83	0.38	0.63	0.48
UsedPeriods	3.61	1.62	4.48	1.24	3.02	1.60	3.36	1.75
observation	1,536		384		384		384	

experiment	treatment							
pattern	All		crescendo		decrescendo		U-shaped	
	mean	sd	mean	sd	mean	sd	mean	sd
Utility	1084	129	1034	105	1184	145	1060	92
StartUsing	0.74	0.44	0.45	0.50	0.98	0.13	0.91	0.28
UsedPeriods	3.00	1.63	4.83	0.72	2.55	0.96	2.11	1.76
observation	1,536		384		384		384	

(5), and (7). Despite adding a control variable and fixed effects, the differences are significant in (4), (6), and (8).

Tab. 4.5: The result of binary logit regression on StartUsing

	Dependent variable:							
	StartUsing							
	All		crescendo		decrescendo		U-shaped	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	0.4^{***}	0.6^{***}	-0.3^{**}	-0.4^{*}	2.4^{***}	3.5^{***}	1.8^{***}	3.3^{***}
	(0.1)	(0.1)	(0.1)	(0.2)	(0.4)	(0.5)	(0.2)	(0.3)
Total Activity Level		0.01^{***}		0.01^{***}		0.01^{***}		0.01^{***}
		(0.00)		(0.00)		(0.00)		(0.00)
Constant	0.6^{***}	-5.6^{***}	0.1	-6.5^{***}	1.6^{***}	-9.5^{***}	0.5^{***}	-11.6^{***}
	(0.1)	(0.6)	(0.1)	(1.3)	(0.1)	(2.1)	(0.1)	(1.7)
Round FEs	No	Yes	No	Yes	No	Yes	No	Yes
Player FEs	No	Yes	No	Yes	No	Yes	No	Yes
Observations	3,072	3,072	768	768	768	768	768	768
Log Likelihood	−1,872.3	−1,383.0	−529.7	−308.5	−209.6	−96.3	−368.9	−201.4
Akaike Inf. Crit.	3,748.5	2,992.0	1,063.5	843.0	423.2	418.6	741.8	628.8
Note:								*p<0.1; **p<0.05; ***p<0.01

Tab. 4.6: The result of ordered logit regression on UsedPeriods

	Dependent variable:							
	UsedPeriods							
	All		crescendo		decrescendo		U-shaped	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	-0.7^{***}	-0.6^{***}	1.2^{***}	2.1^{***}	-0.5^{***}	-0.4^{***}	-1.4^{***}	-1.6^{***}
	(0.1)	(0.1)	(0.4)	(0.5)	(0.1)	(0.1)	(0.2)	(0.2)
Total Activity Level		0.00^{***}		0.01^{***}		0.01^{***}		0.00^{***}
		(0.00)		(0.00)		(0.00)		(0.00)
Round FEs	No	Yes	No	Yes	No	Yes	No	Yes
Player FEs	No	Yes	No	Yes	No	Yes	No	Yes
Observations	2,138	2,138	374	374	696	696	590	590

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

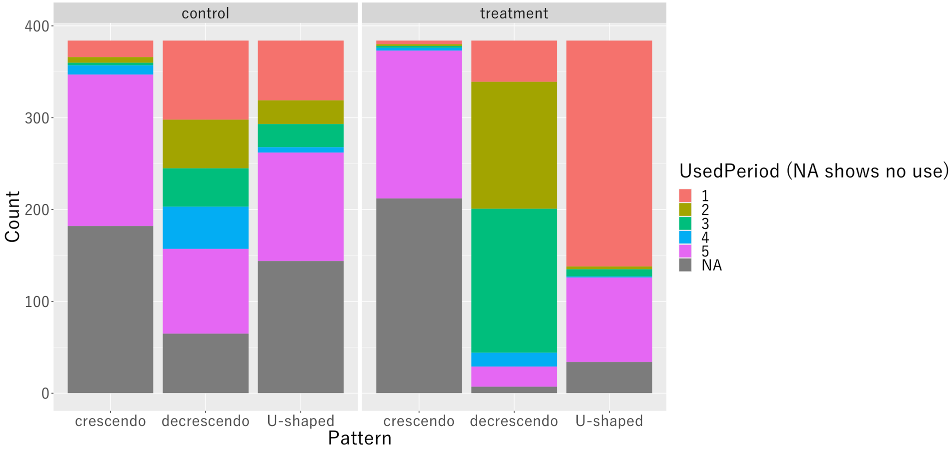


Fig. 4.5: The ratio of StartUsing and UsedPeriods

Table 4.6 shows the results of the ordered logit regression on $UsedPeriods$. (1) shows that the periods of usage in the treatment experiment decreased compared with those in the control experiment. Despite adding a control variable and fixed effects, $UsedPeriods$ is significant in (2). Figure 4.5 shows that more participants used the services till the end in the crescendo pattern. Tables (3) and (4) show that $UsedPeriods$ increased in the treatment setting. However, in Tables (5)–(8) as well as in the decrescendo and U-shaped patterns, $UsedPeriods$ decreased in the treatment setting. In figure 4.5, we can see that decrescendo services increased the usage during periods 1 and 2, whereas U-shaped services increased the usage only during period 1. It is noteworthy that the full usage of U-shaped services did not increase in the treatment.

Table 4.7 shows the results of the regression on using decisions evaluating risk attitude. There is a significant relationship between risk attitudes and decisions (1) to (4). Average number of boxes collected shows the average number of boxes collected by each participant. The average number of boxes collected was 25.36 and the standard deviation was 7.12.

Tab. 4.7: The result of regression considering risk attitude

	Dependent variable:			
	StartUsing		UsedPeriods	
	logistic		*ordered logistic*	
	(1)	(2)	(3)	(4)
The Average Number of Boxes Collected	0.011^{*} (0.006)	0.017^{**} (0.007)	0.021^{***} (0.006)	0.026^{***} (0.006)
Treatment		0.782^{***} (0.106)		-0.536^{***} (0.085)
Total Activity Level		0.009^{***} (0.0004)		0.003^{***} (0.0002)
Decrescendo		3.628^{***} (0.186)		-2.701^{***} (0.117)
Dummy		1.034^{***} (0.138)		-1.829^{***} (0.134)
U-shaped		2.029^{***} (0.149)		-3.259^{***} (0.132)
Constant	0.559^{***} (0.145)	-8.426^{***} (0.476)		
Round FEs	No	Yes	No	Yes
Observations	3,072	3,072	2,138	2,138
Note:		*p<0.1; **p<0.05; ***p<0.01		

4.3.3 Effects on utility

Table 4.8 shows the utility regression resutls. From column (1), the average utility increases significantly, and continues to increase in column (2) despite the addition of a control variable and fixed effects. There are no significant changes in the utility of crescendo services, between the control and treatment groups. While columns (5) and (6) depict a significant increase in the utility of decrescendo services, columns (7) and (8) show an increase in the utility of U-shaped services. However, when compared to decrescendo services, the effect size of U-shaped services. In both settings, it may be difficult to improve the optimal U-shape use. Figure 4.6 shows the optimal usage proportion of the U-shaped pattern. The number in each box shows the optimal length of usage: 0, 1, or 5. In the case of services that are optimal for a 5-period usage, the proportion of actual optimal usage is low.

Tab. 4.8: The result of regression on Utility

	Dependent variable:							
	Utility							
	All		crescendo		decrescendo		U-shaped	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	23.0***	24.3***	−12.6	−3.4	61.0***	52.7***	18.9**	21.4***
	(4.7)	(3.3)	(8.0)	(5.8)	(11.3)	(5.5)	(7.5)	(5.0)
Total Activity Level		0.5***		0.4***		0.7***		0.4***
		(0.01)		(0.02)		(0.01)		(0.01)
Constant	1,060.9***	529.2***	1,046.4***	615.4***	1,122.7***	415.0***	1,040.8***	612.8***
	(3.3)	(19.3)	(5.7)	(34.5)	(8.0)	(32.5)	(5.3)	(30.2)
Round FEs	No	Yes	No	Yes	No	Yes	No	Yes
Player FEs	No	Yes	No	Yes	No	Yes	No	Yes
Observations	3,072	3,072	768	768	768	768	768	768
R^2	0.01	0.5	0.00	0.6	0.04	0.8	0.01	0.6
Adjusted R^2	0.01	0.5	0.00	0.5	0.04	0.8	0.01	0.6

Note: *p<0.1; **p<0.05; ***p<0.01

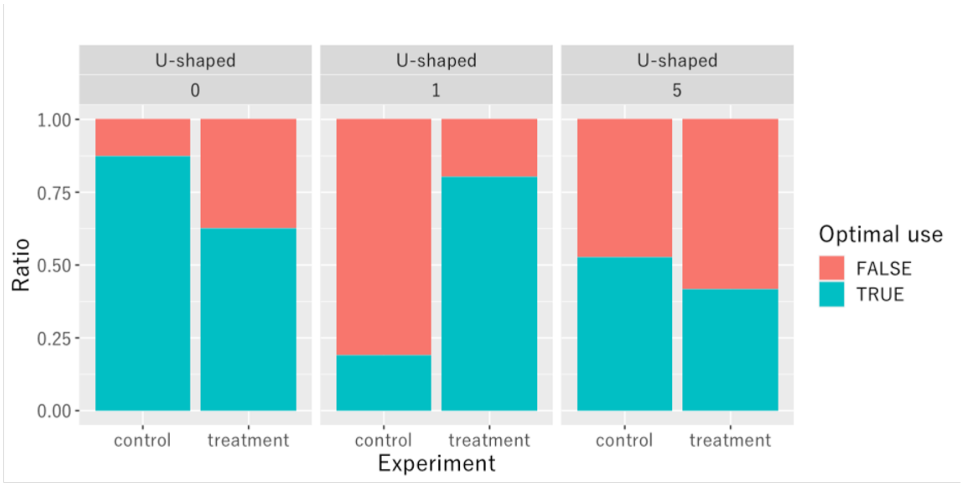


Fig. 4.6: The ratio of optimal usage of U-shaped services

4.4 Discussion

Even under a time-series rating system, U-shaped services are not optimally used. Figure 4.6 shows that U-shaped services that are optimal for five periods are not used for five periods. When considering human psychological characteristics, U-shaped services are regarded as optimal activity-level allocations (Gupta et al. 2016). Although the utility function in this study does not include these features, inadequate information on the latter part of the services contributes to inefficient service use. This inefficiency may also be attributed to the withdrawal of consumers in the earlier periods. To solve this problem, service providers need more consumers to use their services until the end and post ratings. In this study, it was observed that risk-seeking consumers were inclined to use more services. Therefore, service providers may ask such consumers to use their services. Alternatively, service providers can advertise their services as exciting at the end. Another solution is to provide incentives for those who use the services until the end. However, incentivized usage may reduce the intrinsic motivation to use the services and distort reviews. Moreover, deciding the incentive amount would be a complex problem too. Thus, further studies are required to address this issue.

4.5 Chapter Summary

This study examines time series rating systems in the consumption of services. We conducted an economic experiment to include actual service usage and rating behavior to ascertain the difference between ordinal rating and time series rating systems. The time-series rating system increased consumer utility. However, it encouraged consumers to stop using decrescendo and U-shaped services in earlier periods. Particularly, introducing a time series rating system did not increase the optimal usage of U-shaped services. These results indicate that consumers can benefit from a time series rating system. However, service providers may choose their service allocation for a more matched activity-level allocation.

Although time series ratings are still not popular in the real world, it is likely to be adopted in the near future. With an increase in data regarding service consumption, it will become easier to accumulate time series ratings of services. For example, monitoring consumers' facial expressions while experiencing services can create an average exciting moment of services. This information can be used to determine the quality of services during every consumption period. In such cases, platformers must consider the effects on both consumers and the content of the services. This study contributes to the initial suggestions on these effects.

However, this study has several limitations. First, the strategies of the service providers have not been considered. To simplify the experiment, the service patterns provided in the market were fixed. For a more precise analysis of the service activity allocation problem, researchers must model the strategies of service providers and consumers to ascertain the equilibrium of optimal allocation with the withdrawal of consumers and reputation systems. The incentives of service providers should also be modeled in these analyses.

Furthermore, for simplicity, this study does not assume any payment transactions in the model. While consumers0 pay money before using certain services, others may be subscription-based, with consumers paying a fee periodically. This difference may affect the service provider.

Additionally, this study focused only on three extreme but straightforward activity patterns. In reality, it is difficult to provide accurate examples of services that strictly follow these three patterns as most services follow more complex activity level patterns. Further studies are needed to analyze activity-level patterns and reviews more precisely.

This study can also be applied to subscription services. In general, subscription services offer content through fixed payments. After making a payment, consumers face the problem of optimally using the services with the limited period. Assume that music and movies are two examples of content. They should offer varying levels of activity or levels of satisfaction. By assuming services in our model as content, our model is suitable for analyzing the effect of ratings on consumers' content usage in subscription-based services. Further analysis can be conducted in this context.

Part II

Mystery Shopping

Chapter 5

Mystery Shopping Considering Lifestyle Heterogeneity

5.1 Introduction

The increasingly diverse nature of customer preferences has made it necessary for service companies to ascertain how their services satisfy customers. Mystery shopping has become popular and commercially successful with other marketing research methods. The Mystery Shopping Professional Association estimated the turnover of global mystery shopping at over $2 billion in 2017 (MSPA 2018). In general, mystery shopping companies let mystery shoppers experience their client's services and report on service quality by completing a checklist, writing text comments, and giving subjective evaluations.

The widespread use of smartphones has reduced the cost of writing and sending mystery shopping reports, allowing mystery-shopping companies to hire general customers as mystery shoppers. However, improvements in mystery shopping are still required (Blessing and Natter 2019). As the number of mystery shoppers increases, the diversity in their preferences will increase, which may cause a mismatch between a mystery shopper and a service provider. For example, a mystery shopping company sends a mystery shopper with low extroversion to a service provider whose service consists of interactions between customers. The mismatch between service providers and receivers can provide incorrect advice to service providers. One way to prevent mismatches is to understand the difference in mystery shoppers' subjective evaluation of services. However, the number of mystery shopping reports per mystery shopper is insufficient to derive the tendencies of each mystery shopper's service evaluation. Therefore, mystery-shopping companies need additional variables besides reporting data to understand mystery shoppers' tendencies in the subjective evaluation of services.

In this study, the lifestyle factor score of mystery shoppers is used to understand the diversity and consistency of the subjective evaluation of services. The merit of using lifestyle over candidates (e.g. age, gender, and occupation) is that lifestyle is related to the daily consumption of necessities and leisure activity (Green et al. 2006), which are both typical features of services. Several prior studies have focused on the relationship between consumer behavior and lifestyle (Aschemann-Witzel et al. 2021, Palomba 2020). Lifestyle is applied to the subjective evaluation of mystery shoppers.

The objective of this research is to examine the validity of lifestyle as a representative trait of a mystery shopper related to the subjective evaluation of services. Large-scale data from mystery shopping reports of izakaya (Japanese pub) and lifestyle factor scores of mystery shoppers are used. The lifestyle factor score was calculated from the results obtained from questionnaires based on personality and customer consumption patterns. How each lifestyle group of mystery shoppers differs in considering each service attribute (hospitality, atmosphere, or waiting time) when rating revisit intention are examined. It aims to clarify whether lifestyle factor scores can demonstrate the differences in mystery shoppers' subjective service evaluation. Revisit intention is often explained as a result of overall satisfaction with service attributes (Muskat et al. 2019), and understanding its dependencies will enrich mystery shopping reports. It is necessary to consider the complex interaction among service attributes and how different lifestyle mystery shoppers evaluate its interaction. Thus, multi-group structured equation modeling (multi-group SEM) is used to ascertain how lifestyle factors affect a mystery shopper's evaluation of service attributes. Multi-group SEM can evaluate the difference in path coefficients between multiple groups and is often used to determine the differences among different groups of consumers (Matsuo et al. 2018, Truong 2013). In particular, two research questions are proposed: (1) Can lifestyle factors improve the understanding of service evaluation by mystery shoppers? (2) Will mystery shoppers with different lifestyles have different preferences for existing brands based on their service attributes? Answering these questions will demonstrate that mystery-shopping companies can use lifestyle to represent the tendency of mystery shoppers and allow mystery-shopping companies to provide a more advanced mystery shopping service. Using lifestyle data, mystery-shopping companies can optimize or recommend which mystery shoppers are sent.

Analysis of mystery shopping data with mystery shoppers' lifestyle factor scores is explained in the next session. Drawing on these results, this chapter is concluded with the managerial implications of the study.

5.2 Data

In this study, mystery shopping data obtained from MS & Consulting Co. Ltd., a Japanese mystery shopping company is used. The company conducts more than 200,000 mystery shopping per year. These data are suitable for the research purpose of evaluating the effect of lifestyle factors on mystery shoppers' service evaluation. The mystery shopping data of izakaya (Japanese pubs) and lifestyle factor score data of mystery shoppers are used.

5.2.1 Mystery shopping data of izakaya

the izakaya data is used in this study because of the popularity of mystery shopping in this field in Japan. The mystery-shopping company conducted mystery shopping at izakaya (Japanese pubs) from November 2, 2018, through December 22, 2019.

For this study, mystery shoppers evaluated seven service attributes besides their checklists. These attributes are common for both practitioners and academics: (1) revisit intention (Gupta and Zeithaml 2006, Mittal and Kamakura 2001), (2) hospitality (Hau-siu Chow et al. 2007), (3) atmosphere (Lehman et al. 2014), (4) food quality (Namkung and Jang 2007, Sulek and Hensley 2004, Gupta et al. 2012), (5) smile (Otterbring 2017, Gabriel et al. 2015), (6) hygiene (Lehman et al. 2014), and (7) waiting time (Shunko et al. 2018, Sulek and Hensley 2004). Each attribute requires a response on a four-point Likert scale. Each report is answered when they experience the service at the store. Each report has a unique ID of a store and its brand. Thereby, the store that mystery shoppers visited can be identified. Some shoppers conducted multiple mystery shopping sessions during the period.

Tab. 5.1: Descriptive statistics

Variable	Number of Unique ID	
Survey	17448	
Brand	787	
Store	3355	
Mystery shopper	4868	
	Mean	SD
Revisit intention	3.16	0.68
Hospitality	2.97	0.72
Atmosphere	3.18	0.63
Food quality	3.18	0.63
Hygiene	3.10	0.69
Smile	3.16	0.69
Waiting time	3.05	0.73

The descriptive statistics is presented in Table 5.1. The average number of mystery shopping sessions during the data collection period was 22.5 for each brand and 5.3 for each store. The average mystery shopping of each shopper was 3.6. The rows from revisit intention to waiting time shows the average evaluation values of service attributes. Shoppers evaluated them on a four-point Likert scale: 4 is best, and 1 is worst. Hospitality and waiting times were evaluated severely because their average scores were comparatively low.

5.2.2 Lifestyle factor score data of mystery shoppers

The mystery-shopping company collected lifestyle factor score data of mystery shoppers from March 2018 through December 2019. Table 5.2 presents the actual questionnaire of the lifestyle survey method, which are translated from Japanese

Tab. 5.2: Questionnaire items of lifestyle survey method

Question No.	Items (I see myself as someone who ...)
Q1	prefers lively places
Q2	has sociability
Q3	likes to go outside instead of staying home
Q4	prefers trying new things
Q5	does not care about prices for high-quality foods
Q6	does not care about prices for maintaining health
Q7	tends to buy new or trending products
Q8	prefers reading books
Q9	tidies up my room every day
Q10	is scrupulous
Q11	prefers doing housework
Q12	keeps household accounts
Q13	prefers writing
Q14	considers myself fulfilled right now
Q15	is satisfied with daily lives
Q16	buys inexpensive items by collecting information
Q17	does not hesitate to go out further to buy cheaper goods
Q18	has frequent changes in mood
Q19	is careless with money
Q20	feels anxious about health

to English. This questionnaire is based on the results of earlier research (Takenaka et al. 2013, 2016).

Tab. 5.3: Factor scores of the questionnaire items

	Items	Active	Conscious	Planned	Fulfilling life	Economical	Brief
Active	Q1	**0.84**	-0.13	-0.02	0.01	0.00	0.02
	Q2	**0.74**	-0.07	0.08	0.04	-0.04	-0.04
	Q3	**0.61**	0.00	0.02	-0.03	0.02	-0.01
	Q4	**0.46**	0.22	-0.01	-0.01	0.06	-0.10
Conscious	Q5	-0.05	**0.78**	-0.11	0.05	-0.06	0.07
	Q6	-0.09	**0.76**	0.03	0.00	-0.04	0.03
	Q7	0.11	**0.64**	0.00	-0.09	0.07	0.10
	Q8	-0.08	**0.24**	0.22	0.03	0.01	-0.12
Planned	Q9	0.11	-0.09	**0.77**	0.02	-0.13	0.15
	Q10	0.01	-0.09	**0.74**	-0.06	-0.03	0.05
	Q11	0.12	0.10	**0.47**	0.01	0.00	0.01
	Q12	-0.07	-0.05	**0.44**	0.04	0.06	0.03
	Q13	0.05	0.16	**0.36**	-0.03	0.02	-0.12
Fulfilling life	Q14	-0.04	-0.06	0.01	**1.05**	0.00	0.05
	Q15	0.07	0.05	0.02	**0.72**	0.03	-0.05
Economical	Q16	-0.04	0.00	-0.04	0.04	**0.81**	-0.03
	Q17	0.08	-0.04	-0.01	-0.01	**0.65**	0.10
Brief	Q18	-0.10	-0.03	0.13	0.02	0.04	**0.66**
	Q19	0.14	0.22	-0.13	0.08	-0.07	**0.55**
	Q20	-0.05	0.02	0.06	-0.07	0.05	**0.48**

The questionnaire is administrated to 26,646 mystery shoppers to conduct a factor analysis of the lifestyle survey. Some of them participated in mystery shopping of izakaya on Table 5.1. Table 5.3 presents the detailed items of the questionnaire and the value of the factor loading matrix. Maximum likelihood is used for extraction and Promax rotation with Kaiser normalization for factor analysis. Eigenvalues are 4.161, 1.935, 1.714, 1.559, 1.246, 1.065, 0.991, and so on. Six factors are adopted based on the Scree-plot and Kaiser–Guttman methods. The six factors are named as (1) Active consumption type, (2) Planned consumption type, (3) Conscious consumption type, (4) Fulfilling life consumption type, (5) Economical consumption type, and (6) Brief consumption type. Then Cronbach's coefficient alpha is checked to assess the consistency of each factor; they were calculated as (1) 0.730, (2) 0.659, (3) 0.853, (4) 0.681, (5) 0.669, and (6) 0.550. Although there are various qualitative descriptors for the range of acceptable Cronbach's coefficient alpha (Taber 2018), some factors, such as (6) Brief-consumption type, have a slightly low value. However, six factors are still used, considering the ease of interpretation. A factor score (Bartlett score) is normalized for each mystery shopper to 1, 2, 3, 4, or 5 (top 20% highest factor mystery shoppers are rated 5, average 3) and defined them as a lifestyle factor score.

Examining the questionnaire items in Table 5.2 and factor scores in Table 5.3, individuals with Active lifestyle seem to like to try new services. Additionally, they might prefer highly interactive services with employees or other customers because of their high sociability. Individuals with Conscious lifestyle will prefer high-quality services without considering the price. They are assumed to prefer expensive, fancy restaurants. Individuals with Planned lifestyle seem to care about the detailed quality of service. Individuals with Fulfilled life lifestyle are highly satisfied and prefer enjoyable or leisure services. Those with Economical lifestyle will choose inexpensive services and evaluate services in terms of their price. However, as price data is not used in this study, it might be difficult to observe their evaluation tendency in this research. Those with Brief lifestyle might prefer exciting services that change their mood.

5.3 Lifestyle Factor Score and Service Evaluation

Multi-group structural equation modeling (multi-group SEM) is not used to assess the effect of lifestyle factor scores on paths from each service attribute to revisit intention. A path diagram illustrating the contribution of service attributes to revisit intention is constructed. Then, estimating the model coefficients by conducting multi-group SEM, the difference in the estimated path coefficients between the high factor score group and the low factor score group is confirmed.

5.3.1 Constructing the single structural equation model of service evaluation

First, all mystery shopping data is used to construct a path diagram of the contribution of perceived service attributes to the evaluation of revisit intention using structural equation modeling (for the remainder of this paper, this normal structural equation modeling is designated as single-group SEM, in contrast to multi-group SEM). Among these seven evaluations of services used for this research, revisit intention (or repurchase intention) is considered as a total evaluation of service quality and related to revisiting (repurchase) itself (Mittal and Kamakura 2001). For example, ACSI put customer loyalty, which includes repurchase intention, as the dependent variable of the model Fornell et al. (1996). Therefore, revisit intention is set as the dependent variable and examined the relationship of other variables. AMOS is used to conduct the SEM.

Tab. 5.4: Correlation matrix

	Revisit intention	Hospitality	Atmosphere	Food quality	Hygiene	Smile	Waiting time
Revisit intention	1						
Hospitality	0.624	1					
Atmosphere	0.6	0.592	1				
Food quality	0.566	0.414	0.455	1			
Hygiene	0.416	0.373	0.451	0.382	1		
Smile	0.586	0.672	0.697	0.41	0.404	1	
Waiting time	0.437	0.431	0.411	0.376	0.366	0.397	1

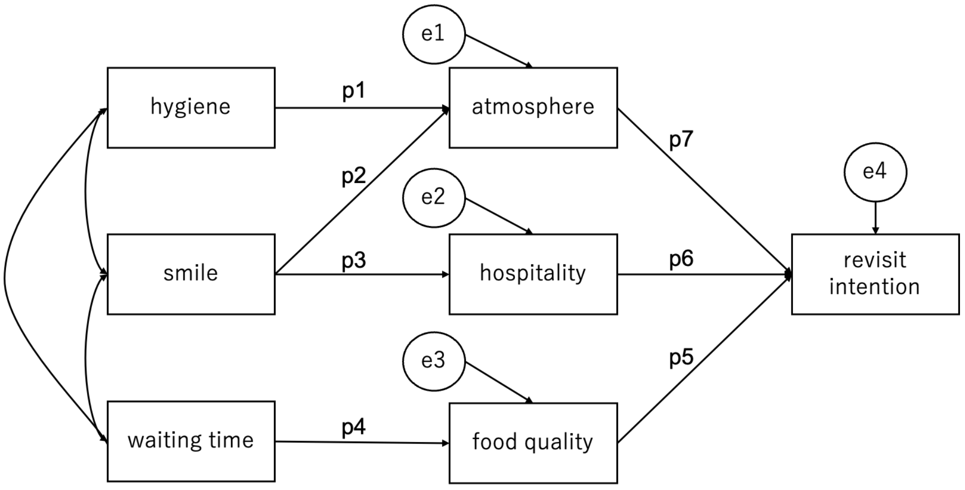


Fig. 5.1: Path diagram.

The correlation matrix is calculated in Table 5.4. A path model is exploratorily constructed for the multi-group SEM based on the result of correlation analysis. Several candidate models are proposed based on the relationships of the variables that are consistent with intuition. These models are tested and adapted the model in Figure 5.1, which has high goodness of fit among other models with the same number of paths. Revisit intention has a high correlation with hospitality, atmosphere, smile, and food quality. However, as smile is also highly correlated with atmosphere and hospitality, the direct path from smile to revisit intention is considered.

The goodness of fit are on the first row of Table 5.6: $GFI = 0.916$, $AGFI = 0.785$, $CFI = 0.887$, $RMSEA = 0.176$. Generally speaking, the model has a good fit if $GFI > 0.9$, $CFI > 0.9$, $RMSEA < 0.1$, and AGFI and GFI have close values. This model did not show high goodness of fit. Therefore, the heterogeneity of mystery shoppers' service evaluation may affects the goodness of fit. If lifestyle factors are related to this heterogeneity, then considering lifestyle factors in the model can be expected to increase the goodness of fit.

5.3.2 Results of the multi-group structural equation modeling

The effect of each lifestyle factor on each path coefficient is examined. It is assumed that differences exist in the path coefficients of the structural model between the high factor score group and the low factor score group. Multi-group structural equation modeling is conducted (Jöreskog 1971, Byrne 2010). Multi-group SEM estimates the coefficient of each path of the path model between two groups and compares the estimated coefficients to observe the difference between the two groups. Multi-group SEM is conducted to consider both the groups of mystery shoppers with the top 20% lifestyle factor scores, or those who scored 5, as a high factor score group, and those who have bottom 20% lifestyle factor scores, or those who scored 1, as a low factor score group.

Table 5.5 presents the results. Differences between the estimated coefficients of each group for each path is tested. Significant differences on p2 (smile → atmosphere) and p3 (smile → hospitality) are found for several lifestyle groups. For example, in the comparison of Active lifestyle groups, p2 (smile → atmosphere) is high in high Active lifestyle group (0.640/0.600 for high/low group) and p3 (smile → hospitality) is high in high Active lifestyle group (0.683/0.645 for high/low group). These two paths are from smile. Since smile is an important nonverbal communication (Choi et al. 2020), Active people, who are likely to have sociability, will take smile into account to rate atmosphere and hospitality. Smiling may be considered an important element of high-quality services for Conscious individuals who prefer high-quality services. For P4 (waiting time → food quality), Tab. 5.5: Results of multi-group SEM

Path			Active	Planned	Con-	Fulfilling	Econo-	Brief
					scious	life	mical	
p1	Hygiene →	high	0.206	0.185	0.201	0.216	0.184	0.204
	Atmosphere	low	0.183	0.208	0.213	0.196	0.210	0.219
p2	Smile →	high	0.640 *	0.648 **	0.650 *	0.631 *	0.646	0.610
	Atmosphere	low	0.600	0.581	0.582	0.597	0.628	0.607
p3	Smile →	high	0.683 **	0.692 **	0.697 *	0.668	0.695	0.674
	Hospitality	low	0.645	0.641	0.675	0.659	0.670	0.662
p4	Waiting time →	high	0.404 **	0.400	0.403 **	0.387	0.404	0.376
	Food quality	low	0.345	0.363	0.379	0.367	0.371	0.390
p5	Food quality →	high	0.363	0.328	0.348	0.338	0.322	0.364
	Revisit intention	low	0.318	0.336	0.316	0.315	0.349	0.348
p6	Hospitality →	high	0.356 **	0.373	0.388	0.350 **	0.388	0.350
	Revisit intention	low	0.406	0.368	0.400	0.407	0.352	0.368
p7	Atmosphere →	high	0.261	0.287	0.278	0.283	0.270	0.301
	Revisit intention	low	0.282	0.284	0.253	0.281	0.300	0.272

* p<0.05 ** p<0.01

Active and Conscious individuals emphasize this path. Since waiting time can affect food temperature, Conscious people may consider it an important element of quality. In contrast, p6 (hospitality → revisit intention) was high in the low Active and Fulfilling life groups. The reason is probably that individuals in the less Active and Fulfilling life lifestyle groups will not expect sophisticated hospitality more easily reflect hospitality from a service provider to their rating. Highly Active and Fulfilling life individuals might demand more sophisticated hospitality because Active people have high sociability, and Fulfilling life people are relatively satisfied with their lives. Any significant differences in terms of economics and brief consumption could not be confirmed. Nevertheless, four factors (Active, Planned, Conscious, and Fulfilling life) significantly affected several path coefficients to revisit intention.

Further the goodness of fit for four lifestyle factors (Active, Planned, Conscious, and Fulfilling life) that are found significant differences in the coefficients of some paths are analyzed. Equality constraints on paths for which a significant difference between the high and low factor score group of each lifestyle factor are assumed. Multi-group SEM is conducted again to ascertain the differences in the goodness of fit. Table 5.6 shows results obtained for multi-group SEM. Although these values are not simply compared to those of a single-group SEM because of different samples, these models for the considered lifestyle factors have a better fit than before. The AGFI improved for all four models (0.813,0.820,0.811, and 0.831 for Active, Planned, Conscious, and Fulfilling life), and RMSEA improved for all four models (0.116,0.114,0.116, and 0.110 for Active, Planned, Conscious, and Fulfilling life). In summary, different lifestyle factor groups have different weights for service attributes and underpin Hypothesis 1.

Tab. 5.6: The goodness of fit of each SEM

Type of SEM	Data	GFI	AGFI	CFI	RMSEA
Single-group SEM	ALL (*N*=17448)	0.916	0.785	0.887	0.176
Multi-group SEM	High Active (*N*=4485) and Low Active (*N*=3082)	0.913	0.813	0.886	0.116
	High Planned (*N*=4074) and Low Planned (*N*=3200)	0.913	0.820	0.886	0.114
	High Conscious (*N*=4371) and Low Conscious (*N*=2936)	0.907	0.811	0.884	0.116
	High Fulfilling life (*N*=3608) and Low Fulfilling life (*N*=3081)	0.916	0.831	0.889	0.110

Note: Sample sizes differ because some mystery shoppers undertook several mystery shopping sessions during the period.

5.4 Lifestyle Factor Score and Brand Evaluation

Earlier SEM analysis clarified that different lifestyle factor groups incorporate different service attributes. Since different brands have their own concepts, different weights are assumed for service attributes. Therefore, mystery shoppers' lifestyle can affect their preference of service brands (Hypothesis 2).

To examine Hypothesis 2, the relationship between the revisit intention of each brand, which could be considered as customer preferences, with their lifestyle is examined in this study. The top five mystery shopped brands in the data collection periods are analyzed. Table 5.7 summarize the descriptive statistics for those five brands. Although the number of surveys conducted in the top five brands is imbalanced, each brand seems to have sufficient data. The table confirms that the average scores of service attribute results differ from the unique strengths and weaknesses

Tab. 5.7: Descriptive statistics of the brands

	All		Brand 1		Brand 2		Brand 3		Brand 4		Brand 5	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Revisit Intention	3.16	0.68	3.15	0.62	3.23	0.66	3.17	0.60	3.16	0.66	3.11	0.68
Hospitality	2.97	0.72	2.94	0.69	3.02	0.71	2.88	0.66	3.10	0.74	2.88	0.71
Atmosphere	3.18	0.63	3.12	0.59	3.27	0.63	3.17	0.61	3.23	0.61	3.12	0.70
Food Quality	3.18	0.63	3.12	0.60	3.29	0.57	3.13	0.62	3.17	0.63	3.17	0.58
Hygiene	3.10	0.69	2.98	0.68	3.22	0.69	3.19	0.65	3.09	0.71	3.23	0.69
Smile	3.16	0.70	3.06	0.66	3.27	0.72	3.21	0.65	3.36	0.62	3.22	0.67
Waiting Time	3.05	0.73	2.94	0.73	3.19	0.73	3.23	0.63	3.06	0.75	3.10	0.67
# of Surveys	17732		5373		1735		651		382		339	

of the brands' services. For example, Brand 1 has a low mean waiting-time score (2.94), and Brand 2 has a high mean waiting-time score (3.19). Therefore, from the multi-group SEM results, the subjective evaluation of mystery shoppers on different service brands can differ according to their lifestyle factors.

Tab. 5.8: Results of ANOVA on revisit intention for each pair of lifestyle factor and brand

		Active		Planned		Conscious		Fulfilling life		Econo-mical		Brief	
Brand 1	high	3.22	***	3.25	***	3.24	***	3.26	***	3.15		3.13	*
	low	3.05		3.09		3.12		3.08		3.16		3.21	
Brand 2	high	3.29		3.23		3.24		3.30	*	3.21		3.27	
	low	3.20		3.19		3.14		3.18		3.26		3.26	
Brand 3	high	3.26	*	3.27	*	3.28	*	3.18		3.10		3.19	
	low	3.10		3.11		3.11		3.10		3.19		3.20	
Brand 4	high	3.23		3.22		3.31		3.25		3.13		3.15	
	low	3.16		3.12		3.26		3.18		3.27		3.14	
Brand 5	high	3.25	*	3.19	*	3.17		3.22		3.11		3.08	
	low	2.92		2.93		2.97		3.03		3.25		3.13	

* $p<0.05$ ** $p<0.01$ *** $p<0.001$

Two groups are compared in multiple SEM analysis: the top 20% factor score group (scored 5) and the bottom 20% factor score group (scored 1) of mystery shoppers. Subsequently, the average score of revisit intention of each brand is calculated. Table 5.8 presents the result. Additionally, ANOVA is applied for each pair of brands and lifestyle factors to assess the significance of the differences.

The results showed significant differences in revisit intention between the high and low lifestyle factor groups. Brand 1 exhibits clear differences in revisit intention for four lifestyle factors: Active (3.22/3.05, high/low, $p < 0.001$), Planned (3.25/3.09 for high/low, $p < 0.001$), Conscious (3.24/3.12 for high/low, $p < 0.001$), and Fulfilling life (3.26/3.08 for high/low, $p < 0.001$). All showed a significant effect on subjective evaluation of service in multiple SEM analysis before. Additionally, significant differences in Brands 3 for Active (3.26/3.10 for high/low, $p < 0.05$), Planned (3.27/3.11 for high/low, $p < 0.05$), and Conscious (3.28/3.11 for high/low, $p < 0.05$), and in Brand 5 for Active (3.25/2.92 for high/low, $p < 0.05$), and Planned (3.19/2.93 for high/low, $p < 0.05$) are found. Presumably, these differences are derived from each brand's strength of service attributes. For example, the high Active factor group does not emphasize hospitality and smile when they evaluate the revisit intention from Table 5.5. Therefore, the high Active factor group has a high revisit intention for Brands 3 and 5 in Table 5.8 because, from Table 5.7, Brands 3 and 5 is confirmed to received lower scores for satisfaction with hospitality and smile than other brands. In summary, it is shown that mystery shoppers with different lifestyles prefer different brands, which supports Hypothesis 2.

5.5 Chapter Summary

This study examined the relationships between the lifestyle of mystery shoppers and their shopping reports. Firstly, the path model of service attributes and revisit intention are constructed. Using mystery shopping data of Japanese pubs (izakaya) and mystery shoppers' lifestyle factor scores, multi-group SEM considering lifestyle factor scores is performed, which showed better goodness of fit for several paths. Especially, Active, Planned, Conscious, and Fulfilling life lifestyle factors affect several relationships of smile, atmosphere, hospitality, waiting time, food quality, and revisit intentions. Moreover, mystery shoppers' revisit intentions for the top five mystery shopped brands, which have different service attribute levels, are checked and it is confirmed that the evaluation of each brand by mystery shoppers is consistent with the result in multi-group SEM. This study's contribution is that unique mystery shopping and lifestyle data of mystery shoppers are used, and results support the idea that lifestyle factor scores can be indices that represent mystery shoppers' subjective service evaluations.

5.5.1 Theoretical implication

Through the multi-group SEM analysis of mystery-shopping data of izakaya and mystery shoppers' lifestyle factor scores, significant differences of service preference between the high and low lifestyle factor score groups for Active, Planned, Conscious, and Fulfilling life factors are confirmed. Additionally, different brand data is used to show that mystery shoppers' lifestyle affects their brand preferences too. These results contribute to prior lifestyle studies that clarify the relationship between lifestyle and consumer behavior. Ishigaki et al. (2010) used lifestyle data with POS data to improve their probabilistic double-latent semantic indexing model for category mining. Takenaka et al. (2016) used log data of smart home appliance users and their lifestyle data. A new perspective of combining lifestyle data with mystery shoppers' subjective service evaluation data are provided. Further research in lifestyle with other data sources will enrich the understanding of the effect of lifestyles on consuming behavior.

The results also contributed to mystery shopping research. As Finn and Kayandé (1999) suggested, demographic traits might affect mystery shoppers. However, Brito and Rambocas (2016) find little evidence of the effect of such traits on mystery shopping. In contrast to their research, it is shown that lifestyle might affect the subjective evaluation of services using large-scale data.

The results have the potential to tackle the problem that Blessing and Natter (2019) presented in their research. They showed that mystery shoppers' subjective evaluation has lower consistency with the satisfaction value of actual consumers. One reason is that there are insufficient mystery shoppers per store to reduce their heterogeneity, which might affect the inconsistency with actual consumers. The result suggests that considering the lifestyle factor of mystery shoppers can reduce the heterogeneity of their subjective evaluation. Mystery shopping companies can use lifestyle data to modify the report to reflect mystery shoppers' tendency of subjective service evaluation by considering their lifestyle factor scores.

5.5.2 Limitation

One of the limitations of this study is that price data, which is important for customer satisfaction, is not used. For example, Ryu and Han (2010) showed that price could be the moderator in the final satisfaction from services. Ishigaki et al. (2010) reported that the lifestyle of consumers affected their buying behavior between high-priced range items and low-priced range items. The data might have been affected by unobserved prices related to Economical and Brief lifestyle factors. For example, statistically significant differences of path coefficient in Table 5.5 for Economical and Brief lifestyle factors are not found in this study. Adding these data in future research may clarify the effect of these two lifestyles on subjective evaluation.

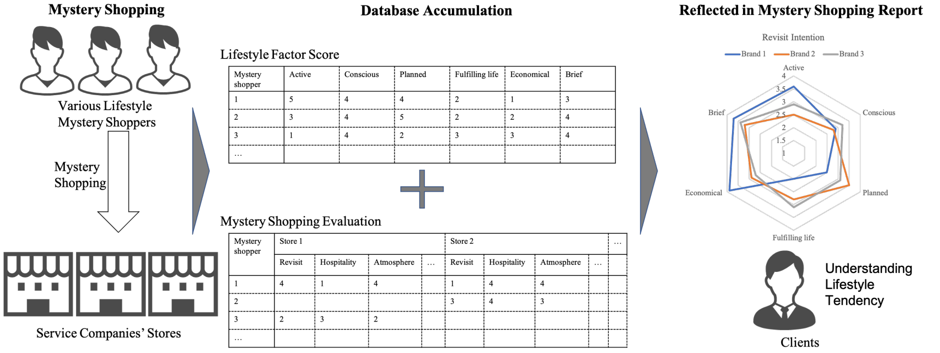
Another limitation is the difference in sample sizes among brands. This research shows that some brands have a considerably smaller sample size than Brand

1. The sample size might be insufficient to detect expected effects. From Table 5.5, mystery shoppers who have a high Active factor score are observed to emphasize hospitality to evaluate revisit intention. Therefore, they may rate higher revisit intention for Brand 4, which has a high hospitality score, according to Table 5.7. However, from Table 5.8, no significant difference between mystery shoppers with a high Active lifestyle factor score and low shoppers are found.

5.5.3 Managerial implications

Combining lifestyle factor scores and mystery shoppers' subjective evaluation is used to create a new mystery shopping research method in several ways. Three managerial implications are proposed: (1) sending particular lifestyle mystery shoppers to focus on a particular sector, (2) sending various lifestyle mystery shoppers to find preferable customer segments, and (3) conducting inter-industry comparisons by using lifestyle.

Mystery-shopping companies could optimize which mystery shoppers to send by considering their lifestyles. If a restaurant wants Active consumers, a mystery shopping company can send such shoppers to the restaurant. Only a 20-question procedure is necessary for mystery shopping companies to understand the lifestyle of a new mystery shopper, which is inexpensive. Mystery shopping companies have to consider both lifestyle distribution of mystery shoppers and client companies' requests and optimize client companies' satisfaction. For example, if a company wants shoppers with Active and Conscious lifestyles, a mystery shopping company assembles shoppers with high Active and Conscious lifestyle factor scores within the clients' request.



Lifestyle Factor Score

Mystery shopper	Active	Conscious	Planned	Fulfilling life	Economical	Brief
1	5	4	4	2	1	3
2	3	4	5	2	2	4
3	1	4	2	3	3	4
…						

Mystery Shopping Evaluation

Mystery shopper	Store 1				Store 2				…
	Revisit	Hospitality	Atmosphere	…	Revisit	Hospitality	Atmosphere	…	
1	4	1	4		1	4	4		
2					3	4	3		
3	2	3	2						
…									

Fig. 5.2: Conceptual image of mystery shopping research method considering lifestyle factor scores.

Client companies can purposely not ask mystery shopping companies to choose which shoppers to come to their stores. This strategy will bring them different benefits. Mystery shopping by various lifestyle mystery shoppers enables client companies to grasp the reaction of each lifestyle customer. Figure 5.2 presents the concept of how clients received mystery shopping reports considering mystery shoppers' lifestyle. Mystery shopping reports considering lifestyle have two elements, in addition to traditional mystery shopping reports: visualizing which lifestyle mystery shoppers prefer clients' services and how their services have strengths and weaknesses compared with those offered by other companies. Client companies can optimize their marketing strategies by considering lifestyle tendencies. They can develop marketing strategies that target those lifestyle customers. This optimization can help companies acquire new customers because some mystery shoppers are not daily users of the services. Managers can observe their reactions to the services directly. Additionally, this optimization will assist client companies in creating services that correspond to diverse customer preferences. Moreover, the client companies can do so better than their competitors.

For client companies with two or more brands in different industries, mystery shoppers with lifestyles can enable inter-industry comparison to grasp a new marketing hypothesis. Mystery shoppers visit various service industries such as restaurants, apparel shops, supermarkets, and mobile phone retailers. The relationship of evaluation among industries can be analyzed by using the lifestyle of mystery shoppers. For example, suppose it is found that the people with the same or similar lifestyle prefer a particular apparel shop and a particular restaurant. In that case, there might be some unobserved variable attracting the lifestyle group. This variable can assist managers and researchers in devising a new marketing hypothesis or a managing strategy.

5.5.4 Future research

Based on the results, future research will explore implementing a new mystery shopping method considering lifestyle to actual service. Investigating its effects on management change is important to certify the benefits of reports considering lifestyle factors. For example, interviewing managers who adopt this method will give us a more detailed image of how they apply it in their daily business.

Another future task is to investigate a more general relationship between lifestyle factor scores and service preference. One approach is to examine the relationship in multiple industries. For example, individuals with Brief lifestyle may show apparent differences in subjective evaluation of transportation services, of which speed is the crucial factor.

Chapter 6

The Effect of a Digital Device that Visualizes Mystery Shoppers' Satisfaction on Service Employees

6.1 Introduction

The widespread use of smartphones and the increasing number of users of reputation systems have provided opportunities for managers and employees to confirm customer satisfaction (CS) with their shops. For example, Google Maps and Yelp now display shop reviews. Because this information sometimes points out problems with a shop and can affect its reputation, managers should focus on it.

The relationship between employees and customer satisfaction is modeled in the service profit chain (SPC) as the effect of employees on customers (Heskett et al. 1994). However, in recent studies, customer feedback has been shown to improve employee satisfaction (ES). Mortimer and Laurie (2016) used cross-lagged data to show that CS affects ES. Although it is not from consumers, several studies have also shown the causal effect of feedback from managers on workers (Huang et al. 2019, Jung et al. 2010). Thus, it is important to investigate how CS ratings can affect service industries.

In this study, mystery shopping data collected before and after the adoption of a new digital system were used. A mystery shopping company created an application (app) for the employees of client companies to observe mystery shopping results directly. Mystery shopping reports comprise checklists for service operations and a subjective evaluation of services. Traditionally, these results are shared directly with client company managers. Adopting this new app may increase employees' opportunities to observe customer feedback.

This study aims to examine the effect of customer feedback on employees. Figure 6.1 illustrates the study’ s methodology. The results show that the app's introduction is related to checklist scores, which indicate the service quality of the service providers. In addition, the results showed that the frequency of app usage is related to checklist scores and revisit intention.

6.2 Data

Mystery shopping data and data logs of new app adoption for each store were collected from April 1, 2018 to March 31, 2020. The data were obtained from MS & Consulting Co. Ltd., a Japanese mystery shopping company. This new app is called “tenpoket.” Employees of the client companies that decided to adopt the app were asked to install it on their smartphones. The timing of installation differs between companies and employees. Through this app, employees were able to check mystery shopping results and review comments from other employees or managers.

The mystery shopping data consist of the total score of the checklists and the

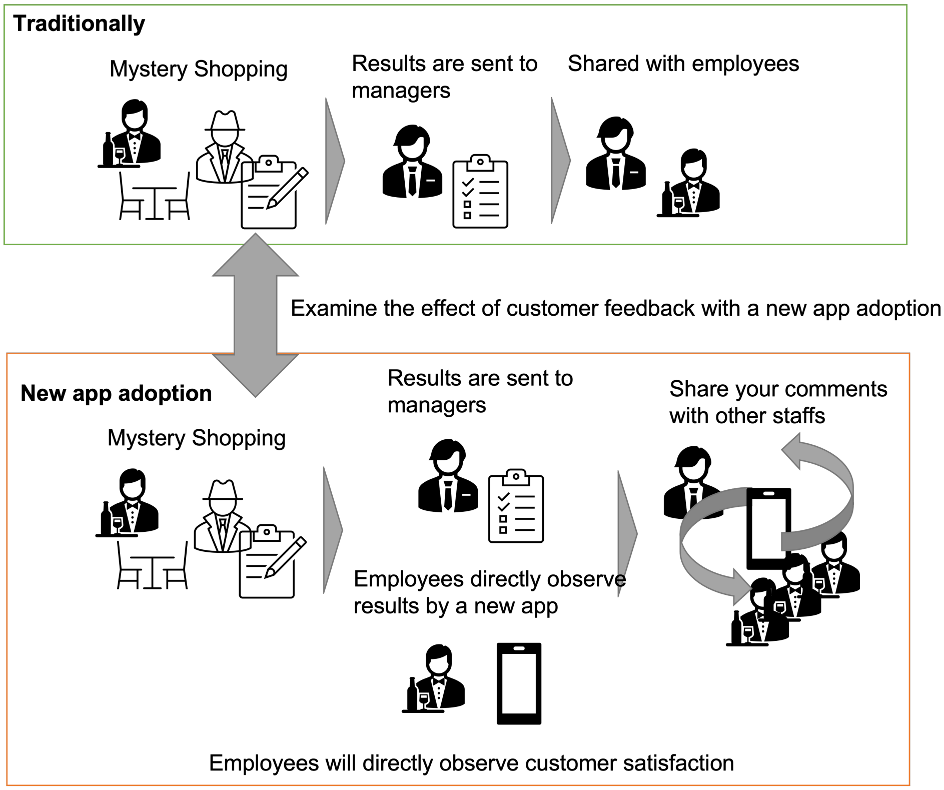

Traditionally
Mystery Shopping
Results are sent to managers
Shared with employees
Examine the effect of customer feedback with a new app adoption
New app adoption
Mystery Shopping
Results are sent to managers
Share your comments with other staffs
Employees directly observe results by a new app
Employees will directly observe customer satisfaction

Fig. 6.1: An illustration of this study

subjective evaluation of the service of each shop. The checklists are customized for each store, so the total scores differ for each store. To compare a particular store's checklist result with other stores, the percentage of the score, *Total Rate*, is used. *Total Rate Brand Scaled*, which is the scaled *Total Rate* within the same brand, is also used. This scale is used because most brands use checklists that are similar to one another. In addition to the checklists, mystery shoppers are asked to indicate their *Revisit Intention* for each shop on a four-point Likert scale, where 4 is the best and 1 is the worst. This is an optional question, so some shops do not have data on *Revisit Intention*.

The data logs of tenpoket have four variables that represent how employees responded to the mystery shopping reports. The four variables are *Read Log*, *Notice Log*, *Good Log*, and *Comment Log*. The *Read Log* is a timestamp of when each employee read each mystery shopping report on their app. The *Notice Log* documents the timestamp of when an employee filled in the Notice Sheet for the mystery shopping report. On the Notice Sheet, employees write their comments on the mystery shopping report that they checked. The Notice Sheet is shared with other employees and managers who use tenpoket. The *Good Log* shows when an employee responded to the mystery shopping report with a like button. The *Comment Log* shows when managers write comments on the Notice Sheet. In this study, the data were screened, and abnormal data were excluded [1].

Each row of data was re-aggregated into monthly and store-specific data. For *Total Rate*, the average value of each store in each month was calculated. For *Revisit Intention*, the average value of each store in each month was calculated and rounded into integers because most of the stores conduct a maximum one mystery shopping report every month, enabling ordered logit regression. For tenpoket logs, the total value of each store in each month was calculated. A total of 75,082 mystery shopping reports were generated over a period of 24 months. However, mystery shopping was conducted once every few months in most shops. The average number of months between two sets of mystery shopping conducted in the same shop was 4.72, with a median of 2. Because the average value is higher than the median, some shops frequently use mystery shopping. A total of 3,566 shops used mystery shopping in the 24-month study period. As only one-third of these shops adopted tenpoket during the data acquisition period, the data include 24,584 *Read Log* data samples, 2,749 *Notice Log* samples, 2,457 *Good Log* samples, and 982 *Comment Log* samples.

The descriptive statistics are summarized in (1) of Table 6.1. The datasets (1) are divided into two groups: (2) datasets of shops that did not adopt tenpoket and (3) datasets of shops that adopted tenpoket. If a shop did not use tenpoket in the

[1] Some data exceeded a 100 % Total Rate.

Tab. 6.1: Descriptive statistics

	(1) All datasets
Total Rate	0.797 (0.16)
Total Rate Brand Scaled	0.000 (0.99)
Revisit Intention	3.237 (0.68)
Read Log	5.996 (24.90)
Notice Log	0.474 (3.09)
Good Log	0.62 (11.07)
Comment Log	0.097 (1.85)
Observation	354640

	(2) Dataset of shops that did not adopt tenpoket	(3) Dataset of shops that adopted tenpoket
Total Rate	0.791 (0.16)	0.801 (0.15)
Total Rate Brand Scaled	-0.011 (0.99)	0.008 (0.99)
Revisit Intention	3.211 (0.69)	3.25 (0.68)
Read Log	0 (0.00)	10.066 (31.62)
Notice Log	0 (0.00)	0.796 (3.97)
Good Log	0 (0.00)	1.041 (14.33)
Comment Log	0 (0.00)	0.163 (2.40)
Observation	143390	211250

	(4) Dataset of shops that adopted tenpoket before data collection period	(5) Dataset of shops that adopted tenpoket during data collection period
Total Rate	0.802 (0.16)	0.800 (0.15)
Total Rate Brand Scaled	0.026 (0.99)	-0.015 (0.99)
Revisit Intention	3.269 (0.68)	3.227 (0.68)
Read Log	12.793 (35.04)	6.658 (26.33)
Notice Log	1.169 (4.83)	0.33 (2.43)
Good Log	1.631 (18.67)	0.303 (5.03)
Comment Log	0.254 (3.13)	0.049 (0.82)
Observation	117372	93878

	(6) Before tenpoket adoption datasets of (5)	(7) After tenpoket adoption datasets of (5)
Total Rate	0.798 (0.15)	0.800 (0.15)
Total Rate Brand Scaled	-0.039 (0.99)	-0.008 (0.99)
Revisit Intention	3.235 (0.67)	3.224 (0.69)
Read Log	0 (0.00)	8.696 (29.79)
Notice Log	0 (0.00)	0.43 (2.77)
Good Log	0 (0.00)	0.395 (5.75)
Comment Log	0 (0.00)	0.064 (0.93)
Observation	21997	71881

data collection period, then the data for this shop are considered (2). If a shop used tenpoket in the data collection period, then the data for the shop are considered (3). *ReadLog* is checked for the data collection period, and if a shop has $ReadLog > 0$ data in the first month of the data acquisition period, then the data are considered (4), which is the dataset of shops that adopted tenpoket before the data collection period. Conversely, if the shop has $ReadLog = 0$ data in the first month of the data acquisition period and it later becomes $ReadLog > 0$, then the data are considered (5), is the dataset of shops that adopted tenpoket during the data collection period. In this study, the effect of adopting tenpoket is compared before and after the app's introduction. Dataset (5) is divided into two datasets: (6) data before tenpoket adoption and (7) data after tenpoket adoption. The *Total Rate* and *Revisit Intention* is high for (4). In addition, the *Total Rate* in (7) is higher than in (6).

6.3 Estimation Technique

6.3.1 Effect of tenpoket introduction on the outcomes

First, to estimate the effect of the tenpoket introduction on the *Total Rate*, the following model is used:

$$TotalRate_{it} = tenpoket_{it} + Shop_i + YearMonth_t + \epsilon_{it}. \tag{6.1}$$

Here, $TotalRate$ is the percentage of the total score. $TotalRateBrandScaled$ is not used as an output because using fixed values is sufficient, and the estimated coefficients are intuitively understandable. *tenpoket* is a dummy variable for whether shops introduced tenpoket in period t. We defined $tenpoket = 1$ if the shop's $ReadLog_t > 0$ before t or at t. In addition, the fixed effects of shops, $Shop_i$, and fixed effects of year and month, $YearMonth_t$, are added. Shop ID dummy variables are used for shop fixed effects, and year and month (e.g., March 2019) are used for the year and month fixed effects. Shop fixed effects may include the size of the shop, number of employees, or location. These elements can affect $TotalRate$, but they are not observed in this study. Year and month fixed effects may include the increasing demand for an izakaya (Japanese pub) at the end of the year. In addition, to verify the effect of tenpoket introduction on *Revisit Intention*, the following model is estimated:

$$RevisitIntention_{it} = tenpoket_{it} + Shop_i + YearMonth_t + \epsilon_{it}. \tag{6.2}$$

6.3.2 Relationship between tenpoket activity and outcomes

Shops may differ in how they use tenpoket in their daily business. Some shops may encourage employees to read and write comments on mystery shopping results, while others may not. How frequently do employees see the effects of mystery shopping on the outcomes? In order to estimate the effect of the use of tenpoket on the outcomes in each store, the following model is used:

$$TotalRate_{it} = ActivityVariables_{it} + tenpoket_{it} + Shop_i + YearMonth_t + \epsilon_{it} \quad (6.3)$$

$$RevisitIntention_{it} = ActivityVariables_{it} + tenpoket_{it} + Shop_i + YearMonth_t + \epsilon_{it}. \quad (6.4)$$

$ActivityVariables_{it}$ includes $ReadLog_{it}$, $NoticeLog_{it}$, $GoodLog_{it}$, and $CommentLog_{it}$. $tenpoket_{it}$ is added to control the effect of the tenpoket introduction, focusing only on the effect of tenpoket usage. The fixed effects of shop and month were also considered.

6.4 Results

6.4.1 Effect of tenpoket introduction on the outcomes

The results of the regression for all data are shown in Table 6.2. The total number of observations was 354,640, and the total number of revisit intentions was 265,965. Columns (1)–(4) show the results of the regression using model 6.1. Most of the results show that tenpoket adoption positively correlates with the total rate. However, when considering both the shop fixed effect and year-month fixed effect, the estimated coefficient of tenpoket decreases and shows low significance. Because the estimated coefficient of tenpoket is low in (2) and (4), shop fixed effects may be important in these estimated relationships. The result for Revisit Intention is similar to the Total Rate. Columns (5)—(8) show the regression results of model 6.2. There are significant effects in (5) and (7), but not in (6) and (8), which consider shop fixed effects. In summary, some positive effects on the Total Rate are confirmed, but not on Revisit Intention. Both results show that the shop fixed effect exists for the estimation. One candidate for the shop fixed effect is how frequently shops use tenpoket. This is examined in subsection 6.4.2

Tab. 6.2: The result of regression on Total Rate and Revisit Intention for Dataset (1)

Dependent variable:	Total Rate				Revisit Intention			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
tenpoket	0.009^{***}	0.005^{***}	0.007^{***}	0.002^{*}	0.037^{***}	0.008	0.033^{***}	0.006
	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)	(0.006)	(0.003)	(0.007)
Constant	0.792^{***}				3.215^{***}			
	(0.0004)				(0.002)			
Shop FEs	No	Yes	No	Yes	No	Yes	No	Yes
YearMonth FEs	No	No	Yes	Yes	No	No	Yes	Yes
Observations	354,640	354,640	354,640	354,640	265,965	265,965	265,965	265,965
R^2	0.001	0.0001	0.0005	0.00001	0.001	0.00001	0.001	0.000003
Adjusted R^2	0.001	−0.268	0.0004	−0.269	0.001	−0.256	0.0005	−0.257

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Tab. 6.3: The result of regression on Total Rate and Revisit Intention for Dataset (5)

Dependent variable:	Total Rate				Revisit Intention			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
After tenpoket	0.056***	0.064***	0.015	0.010	0.0005	0.012	0.006	0.004
	(0.011)	(0.013)	(0.014)	(0.017)	(0.008)	(0.009)	(0.010)	(0.012)
Constant	−0.063***				3.224***			
	(0.010)				(0.008)			
Shop FEs	No	Yes	No	Yes	No	Yes	No	Yes
YearMonth FEs	No	No	Yes	Yes	No	No	Yes	Yes
Observations	55,891	55,891	55,891	55,891	45,130	45,130	45,130	45,130
R^2	0.0004	0.0005	0.00002	0.00001	0.000000	0.00004	0.00001	0.000003
Adjusted R^2	0.0004	−0.054	−0.0004	−0.055	−0.00002	−0.060	−0.001	−0.060

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 6.3 shows the regression results that only use datasets (6) and (7): the data before and after the adoption of tenpoket. Limiting the regression to only these datasets more clearly demonstrates the causality of tenpoket adoption. Columns (1)–(4) show the results of the regression on the Total Rate, and columns (5)–(8) show the results of the regression on Revisit Intention. In columns (1) and (2), the effect of tenpoket on the Total Rate is observed. However, when considering year-month fixed effects, no significant effect was observed. In columns (5)–(8), there is no confirmed significant effect on Revisit Intention.

6.4.2 Relationship between tenpoket activity and outcomes

The results are listed in Table 6.4. Here, (1)—(4) and (6)-–(9) examine the effect of each log. The Notice and Good logs are related to Total Rate and Revisit Intention. Meanwhile, (5) and (10) include all logs and reduce the overlapping effects among the variables. The results show that the Notice and Good logs are still positively related to outputs.

Tab. 6.4: The result of regression on Total Rate and Revisit Intention with tenpoket Activity Variables

Dependent variable:	Total Rate					Revisit Intention				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Read Log	0.00002*				−0.00001	0.00005				−0.0001*
	(0.00001)				(0.00002)	(0.0001)				(0.0001)
Notice Log		0.0005***			0.0005***		0.002***			0.002***
		(0.0001)			(0.0001)		(0.001)			(0.001)
Good Log			0.0001***		0.0001**			0.001***		0.0005***
			(0.00004)		(0.00004)			(0.0002)		(0.0002)
Comment Log				0.0003	0.0001				0.002*	0.001
				(0.0002)	(0.0002)				(0.001)	(0.001)
tenpoket	0.002*	0.002*	0.002*	0.002*	0.002*	0.005	0.005	0.005	0.005	0.005
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Shop FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
YearMonth FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	354,640	354,640	354,640	354,640	354,640	265,965	265,965	265,965	265,965	265,965
R^2	0.00002	0.0001	0.0001	0.00002	0.0001	0.00001	0.0001	0.0001	0.00002	0.0001
Adjusted R^2	−0.269	−0.269	−0.269	−0.269	−0.269	−0.257	−0.256	−0.256	−0.256	−0.256

Note: *p<0.1; **p<0.05; ***p<0.01

6.4.3 Discussion

The descriptive statistics and regressions show that the introduction of tenpoket is related to the total rate. However, after limiting the analysis to include only data for before and after tenpoket adoption and considering the fixed shop effects, no significant effects on the total rate were observed. Fixed shop effects may include the number of employees, sales amount, floor area, and customer unit price. When considering the effect of tenpoket, how frequently shops use tenpoket may matter for these fixed effects. The shop's motivation for using the app is not considered in equations 6.1 and 6.2. The introduction of a new app does not necessarily mean the full benefits of the app are received. When the relationship with logs is considered in equations 6.3 and 6.4, the Notice and Good logs show the relationships with total scores.

In contrast, it is still difficult to confirm the effect of tenpoket on revisit intention. There may be several reasons for this finding. The first reason is that mystery shoppers rate the effect on a 4-point Likert scale. Most of them evaluated the service with a 3. These rough evaluations may make it challenging to observe precise improvements in service quality. Another reason is that it is difficult to directly satisfy new mystery shoppers and earn higher revisit intention by improving the checklist score based on previous mystery shoppers' feedback. Subjective evaluations of service tend to vary more based on individuals.

In most of the results, the Notice and Good logs have a relationship with the output. For the Good logs, it is intuitive that employees will push the like button for mystery shopping reports with a high total rate. In contrast, there may be two reasons for the Notice logs: employees write notice comments, thus improving service quality and resulting in high mystery shopping scores, or employees only write notice comments on mystery shopping reports with a high total score. In our study, the direction of causality remains unclear. However, it is natural for shop managers to make employees write notice comments if a problem is pointed out in the mystery shopping reports. Further research on the causal direction is necessary.

In order to observe the effect of mystery shopping more precisely, employee satisfaction data should be analyzed . The use of tenpoket probably directly affects employee satisfaction. According to the service profit chain (SPC), the effect of employee satisfaction on outputs is moderated by many elements. Prior studies have shown that there is a lag effect in each element of the SPC Evanschitzky et al. (2012). To understand the effect of feedback, directly observing employee satisfaction to obtain more evidence is necessary. Future research should also evaluate other elements to clarify the precise effect of customer satisfaction feedback on service providers.

6.5 Chapter Summary

This study examined how customer satisfaction affects service providers by introducing a digital app for mystery shopping. This digital app, tenpoket, enables employees to observe mystery shopping results more directly and conveniently. Thus, it is an excellent opportunity to examine the effect of customer satisfaction feedback on service providers.

This study shows that the introduction of tenpoket is related to mystery shopping scores. Furthermore, the app log data indicate that shops that post notice sheets on feedback more frequently received higher mystery shopping scores than before the introduction of the tenpoket. This result shows that an active commitment to feedback is related to increased service quality. This differs from a prior study in which the feedback effect was investigated for specific cases only. This study contributes to the field by confirming these effects through a larger data sample and digital implementation.

To better understand the effect of feedback on employees, the following topics should be addressed: (1) the effect of using store-specific data should be examined in great detail, (2) date-specific data should be used to analyze the effect in shorter periods, and (3) text analysis should be used to examine the effect of comments written by mystery shoppers on employees.

Using store-specific data can enable us to examine variables that are expected to be affected by employee performance. For example, the sales amount may be increased after reading tenpoket feedback and service operations may be improved. As most of these variables are objective measures, there is no fear regarding the subjective effect caused by the heterogeneity of service raters.

Using date-specific data is another possible method to observe the feedback effect on employees. This study used the satisfaction of mystery shoppers and total checklist scores as outputs. However, mystery shopping is conducted several times a year in most shops. Hence, more sensitive changes are challenging to analyze right after an employee sees a report by tenpoket. Using objective shop-specific data to analyze these changes or conducting customer surveys on the days before and after mystery shopping reports are sent may be a better way to determine how customer satisfaction changes over short periods.

Text analysis may enrich our understanding of the effects of mystery shoppers' feedback. In most cases, mystery shoppers wrote text comments for each shop. Written text may be more effective than objective checklists. For example, writing comments positively or negatively affects employee behaviors. This study may suggest that mystery shopping companies revise their manual for mystery shoppers to allow them to express their thoughts on services more positively or negatively.

Part III

Discussion and Conclusion

Chapter 7

Discussion and Conclusion

In this chapter, the two research questions presented in the Introduction are discussed, and four studies from Chapters 3 to 6 are summarized. The two main research questions are: (1) Does current customer satisfaction display truly benefit consumers and employees? (2) Is there an optimal rating system for services?

7.1 Discussion

7.1.1 Do current customer satisfaction rating systems truly benefit consumers and employees?

Consumer utility increases in many settings in the experiments conducted in Chapters 3 and 4. Chapter 3 shows that consumers gained higher utility in all three settings (i.e., B, C, and D) than that in control experiment A. In Chapter 4, although no control experiment was conducted, compared with the first round, consumer utility increased in the later rounds (see Figure A.1 in the Appendix). In general, consumer utility increases as the amount of displayed information increases. In setting D in Chapter 3 and the treatment setting in Chapter 4, the system showed rating scores in a detailed style and achieved the highest consumer utility in each study. These findings are consistent with previous theoretical studies (Acemoglu et al. 2017, Besbes and Scarsini 2018).

One concern is that consumers may be negatively affected by ratings in the first round of experiments. Herd behavior (Banjeree 1992) and information cascades (Duan et al. 2008) can occur in decision making. In Chapter 3, some consumers did not purchase high-quality services because of their low rating in the former rounds in the treatment setting. In Chapter 4, consumers do not optimally use several U-shaped services. This is consistent with research question 3. These phenomena improved in well-informed rating settings: setting D in Chapter 3 and treatment in Chapter 4. This study contributes to the literature by providing empirical evidence from controlled experiments. However, in the treatment setting in Chapter 4, U-shaped services that are optimal for five-period usage are not used well by consumers. For example, a simple increase in the amount of information is insufficient to improve consumers' irrational behavior. Further research is required to obtain more optimal rating systems.

Another critical concern is the effect of the rating system on employees. Customers' opinions have gained attention in the marketing field (Griffin and Hauser 1993). Currently, service companies have many opportunities to determine customer opinions. This feedback can improve employee productivity by showing them the weaknesses of their services, and consequently, the service quality. However, criticism from customers may hurt employees, and decrease their job satisfaction and psychological security. This study uses mystery shopping data, and the results of the data analysis in Chapter 6 show that a significant relationship exists between customer feedback app usage and mystery shopping results.

In summary, the studies in Chapters 3, 4, and 6 confirm that satisfaction display in current service markets benefits both consumers and employees.

7.1.2 Is there a more optimal rating system for services?

Chapter 4 also focuses on the difference between activity level patterns of services. Prior studies focused on these patterns and showed the optimal pattern under the effect of psychological aspects of human decision-making, such as acclimation, satiation, and memory decay Baucells and Sarin (2010), Gupta et al. (2016). In the experiment described in Chapter 4, the optimal use of services by consumers was examined under time-series rating systems. Although consumers achieved optimal use of services in most activity level patterns, they could not fully increase the optimal use of U-shaped items. This result shows that whether and how services benefit from rating systems depends on their contents, even though they are beneficial for consumers if they use them adequately. Future research should examine the trade-off between service diversity and rating systems to construct an optimal rating system for the service markets.

In Chapter 5, the relationship between lifestyle and service evaluation and that between lifestyle and brand preference are examined using mystery shopping data and mystery shoppers' lifestyle factor data. These relationships provide the possibility of using lifestyle to optimize acquiring data on service evaluation by service receivers.

In Chapter 5, we propose three future implications for lifestyle factors. First, mystery shopping companies can send a mystery shopper with a specific lifestyle who is optimized for a target store. Second, mystery shopping companies can send mystery shoppers with various lifestyles with a balanced frequency to collect reports on services received by them. Finally, for client companies with service brands in multiple service fields, lifestyle can integrate the findings from apparently different service fields. These implications are expected to improve ways to gather and summarize service evaluations that are more optimal for service markets.

In summary, the studies in Chapters 4 and 5 clarified perspective on service evaluation that the improvement in it would contribute to the better use by service receivers of evaluations.

7.2 Conclusion

In this study, we conducted four analyses of customer service ratings. Two studies were related to customer ratings, whereas the other two were related to professional ratings by mystery shoppers. Through both topics, we show how these systems now positively affect the service industry and discuss how they can be improved. In summary, this study empirically clarified how consumers rate services and how they are affected in turn by these ratings through two major methods: rating systems and mystery shopping.

References

Acemoglu D, Makhdoumi A, Malekian A, Ozdaglar AE (2017) Fast and slow learning from reviews. *SSRN Electronic Journal* .

Ameri M, Honka E, Xie Y (2019) Word of mouth, observed adoptions, and anime-watching decisions: The role of the personal vs. the community network. *Marketing Science* 38(4):567–583.

Andersen ET, Simester DI (2004) Long-Run effects of promotion depth on new versus established customers: Three field studies. *Marketing Science* 23(1).

Anderson EW, Sullivan MW (1993) The antecedents and consequences of customer satisfaction for firms. *Marketing Science* 12(2):125–143.

Aschemann-Witzel J, de Hooge IE, Almli VL (2021) My style, my food, my waste! consumer food waste-related lifestyle segments. *Journal of Retailing and Consumer Services* 59.

Banjeree AV (1992) A simple model of herd behavior. *Q. J. Econ.* 107(3):797–817.

Barber CS, Tietje BC (2004) A distribution services approach for developing effective competitive strategies against "big box" retailers. *Journal of Retailing and Consumer Services* 11(2):95–107.

Baucells M, Sarin RK (2010) Predicting utility under satiation and habit formation. *Manage. Sci.* 56(2):286–301.

Bavafa H, Hitt LM, Terwiesch C (2018) The impact of E-Visits on visit frequencies and patient health: Evidence from primary care. *Manage. Sci.* 64(12):5461–5480.

Beck J, Miao L (2003) Mystery shopping in lodging properties as a measurement of service quality. *Journal of Quality Assurance in Hospitality & Tourism* 4(1-2):1–21.

Bernhardt KL, Donthu N, Kennett PA (2000) A longitudinal analysis of satisfaction and profitability. *J. Bus. Res.* 47(2):161–171.

Besbes O, Scarsini M (2018) On information distortions in online ratings. *Oper. Res.* 66(3):597–610.

Blessing G, Natter M (2019) Do mystery shoppers really predict customer satisfaction and sales performance? *Journal of Retailing* 95(3):47–62.

Brito PQ, Rambocas M (2016) Assessing the impact of mystery client traits on service evaluation. *Journal of Services Marketing* 30(4):411–426.

Byrne BM (2010) *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (New York: Routledge/Taylor & Francis Group), 2nd edition.

Cervellon MC, Poujol JF, Tanner JF (2019) Judging by the wristwatch: Salespersons' responses to status signals and stereotypes of luxury clients. *Journal of Retailing and Consumer Services* 51:191–201.

Charness G, Gneezy U, Halladay B (2016) Experimental methods: Pay one or pay all. *J. Econ. Behav. Organ.* 131:141–150.

Chen DL, Schonger M, Wickens C (2016) oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance* 9:88–97.

Chen PY, Hong Y, Liu Y (2018) The value of multidimensional rating systems: Evidence from a natural experiment and randomized experiments. *Manage. Sci.* 64(10):4629–4647.

Chen R, Barrows C (2015) Developing a mystery shopping measure to operate a sustainable restaurant business: The power of integrating with corporate executive members' feedback. *Sustainability: Science Practice and Policy* 7(9):12279–12294.

Chevalier JA, Mayzlin D (2006) The effect of word of mouth on sales: Online book reviews. *J. Mark. Res.* 43(3):345–354.

Chintagunta PK, Gopinath S, Venkataraman S (2010) The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Science* 29(5):944–957.

Choi S, Choi C, Mattila AS (2020) Are all smiles perceived equal? the role of service provider's gender. *Service Science* 12(1):1–7.

Crosetto P, Filippin A (2013) The "bomb" risk elicitation task. *J. Risk Uncertain.* 47(1):31–65.

Dahana WD, Miwa Y, Morisada M (2019) Linking lifestyle to customer lifetime value: An exploratory study in an online fashion retail market. *Journal of business research* 99:319–331.

de Langhe B, Fernbach PM, Lichtenstein DR (2016) Navigating by the stars: Investigating the actual and perceived validity of online user ratings. *J. Consum. Res.* 42(6):817–833.

Dellarocas C, Narayan R (2006) A statistical measure of a population's propensity to engage in post-purchase online word-of-mouth. *Stat. Sci.* 21(2):277–285.

Duan W, Gu B, Whinston AB (2008) Do online reviews matter? - an empirical investigation of panel data. *Decis. Support Syst.* 45(4):1007–1016.

Dutt CS, Hahn G, Christodoulidou N, Nadkarni S (2019) What's so mysterious about mystery shoppers? understanding the qualifications and selection of mystery shoppers. *Journal of Quality Assurance in Hospitality & Tourism* 20(4):470–490.

Eger L, Mičík M (2017) Customer-oriented communication in retail and net promoter score. *Journal of Retailing and Consumer Services* 35:142–149.

Eugene W Anderson (1998) Customer satisfaction and word of mouth. *J. Serv. Res.* 1(1):1–14.

Evanschitzky H, Wangenheim Fv, Wünderlich NV (2012) Perils of managing the service profit chain: The role of time lags and feedback loops. *J. Retail.* 88(3):356–366.

Finn A (2001) Mystery shopper benchmarking of Durable-Goods chains and stores. *Journal of Service Research* 3(4):310–320.

Finn A, Kayandé U (1999) Unmasking a phantom: a psychometric assessment of mystery shopping. *Journal of Retailing* 75(2):195–217.

Forman C, Ghose A, Wiesenfeld B (2008) Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research* 19(3):291–313.

Fornell C, Johnson MD, Anderson EW, Cha J, Bryant BE (1996) The american customer satisfaction index: Nature, purpose, and findings. *Journal of marketing* 60(4):7–18.

Gabriel AS, Daniels MA, Diefendorff JM, Greguras GJ (2015) Emotional labor actors: a latent profile analysis of emotional labor strategies. *The Journal of applied psychology* 100(3):863–879.

Gao Gg, Greenwood BN, Agarwal R, McCullough JS (2015) Vocal minority and silent majority: How do online ratings reflect population perceptions of quality. *Miss. Q.* 39(3):565–590.

Gilmore JH (1997) The four faces of mass customization. *Harvard Business Review* .

Godes D, Silva JC (2012) Sequential and temporal dynamics of online opinion. *Marketing Science* 31(3):448–473.

Goldberg LR (1992) The development of markers for the Big-Five factor structure. *Psychological assessment* 4(1):26–42.

Green GT, Cordell HK, Betz CJ, Distefano C (2006) Construction and validation of the national survey on recreation and the environment's lifestyles scale. *Journal Of Leisure Research* 38(4):513–535.

Griffin A, Hauser JR (1993) The voice of the customer. *Marketing science* 12(1):1–27.

Gupta AD, Karmarkar US, Roels G (2016) The design of experiential services with acclimation and memory decay: Optimal sequence and duration. *Manage. Sci.* 62(5):1278–1296.

Gupta S, Dasgupta S, Chaudhuri R (2012) Critical success factors for experiential marketing: Evidences from the indian hospitality industry. *International Journal of Services and Operations Management* 11(3):314–334.

Gupta S, Zeithaml V (2006) Customer metrics and their impact on financial performance. *Marketing Science* 25(6):718–739.

Halliday SD, Lafky J (2019) Reciprocity through ratings: An experimental study of bias in evaluations. *Journal of Behavioral and Experimental Economics* 83(May):101480.

Hau-siu Chow I, Lau VP, Wing-chun Lo T, Sha Z, Yun H (2007) Service quality in restaurant operations in china: Decision- and experiential-oriented perspectives. *International Journal of Hospitality Management* 26(3):698–710.

Heskett JL, Jones TO, Loveman GW, Sasser WE, Schlesinger LA (1994) Putting the Service-Profit chain to work. *Harvard business review* 72(2):164–174.

Hitt LM, Tambe P (2016) Health care information technology, work organization, and nursing home performance. *ILR Review* 69(4):834–859.

Ho YCC, Wu J, Tan Y (2017) Disconfirmation effect on online rating behavior: A structural model. *Information Systems Research* 28(3):626–642.

Holzmeister F, Pfurtscheller A (2016) otree: The “bomb” risk elicitation task. *Journal of Behavioral and Experimental Finance* 10(July):105–108.

Hong Y, Pavlou PA (2014) Product fit uncertainty in online markets: Nature, effects, and antecedents. *Information Systems Research* 25(2):328–344.

Hu N, Pavlou PA, Zhang J (2006) Can online reviews reveal a product’s true quality? empirical findings analytical modeling of online word-of-mouth communication. *Proceedings of the ACM Conference on Electronic Commerce* 2006:324–330.

Hu N, Zhang J, Pavlou PA (2009) Overcoming the j-shaped distribution of product reviews. *Communications of the ACM* 52(10):144–147.

Huang G, Sudhir K (2020) The causal effect of service satisfaction on customer loyalty. *Manage. Sci.* .

Huang N, Burtch G, Gu B, Hong Y, Liang C, Wang K, Fu D, Yang B (2019) Motivating user-generated content with performance feedback: Evidence from randomized field experiments. *Manage. Sci.* 65(1):327–345.

Ishigaki T, Takenaka T, Motomura Y (2010) Category mining by heterogeneous data fusion using PdLSI model in a retail service. *2010 IEEE International Conference on Data Mining*, 857–862 (IEEE).

Jiang Y, Guo H (2015) Design of consumer review systems and product pricing. *Information Systems Research* 26(4):714–730.

Jöreskog KG (1971) Simultaneous factor analysis in several populations. *Psychometrika* 36(4):409–426.

Jung JH, Schneider C, Valacich J (2010) Enhancing the motivational affordance of information systems: The effects of real-time performance feedback and goal setting in group collaboration environments. *Manage. Sci.* 56(4):724–742.

Kamishima T (2016) Algorithms of recommender systems. `http://www.kamishima.net/archive/recsys.pdf`.

Kumar V, Pozza ID, Ganesh J (2013) Revisiting the satisfaction-loyalty relationship: Empirical generalizations and directions for future research. *J. Retail.* 89(3):246–262.

Lafky J (2014) Why do people rate? theory and evidence on online ratings. *Games and economic behavior* 87(December 2013):554–570.

Latham GP, Ford RC, Tzabbar D (2012) Enhancing employee and organizational performance through coaching based on mystery shopper feedback: A quasi-experimental study. *Human resource management* 51(2):213–229.

Lehman DW, Kovács B, Carroll GR (2014) Conflicting social codes and organizations: Hygiene and authenticity in consumer evaluations of restaurants. *Management science* 60(10):2602–2617.

Li G, Li G, Kambele Z (2012) Luxury fashion brand consumers in china: Perceived value, fashion lifestyle, and willingness to pay. *Journal of business research* 65(10):1516–1522.

Li X, Hitt LM (2008) Self-Selection and information role of online product reviews. *Information Systems Research* 19(4):456–474.

Liu CHS, Su CS, Gan B, Chou SF (2014) Effective restaurant rating scale development and a mystery shopper evaluation approach. *International Journal of Hospitality Management* 43:53–64.

Lowndes M, Dawes J (2001) Do distinct SERVQUAL dimensions emerge from mystery shopping data? a test of convergent validity. *Canadian Journal of Program Evaluation* 16(2):41–54.

Lu SF, Rui H (2018) Can we trust online physician ratings? evidence from cardiac surgeons in florida. *Manage. Sci.* 64(6):2557–2573.

Luria G, Yagil D (2008) Procedural justice, ethical climate and service outcomes in restaurants. *International Journal of Hospitality Management* 27(2):276–283.

Magnani M (2020) The economic and behavioral consequences of online user reviews. *Journal of economic surveys* 34(2):263–292.

Massara F, Scarpi D, Porcheddu D (2020) Can your advertisement go abstract without affecting willingness to pay? product-centered versus lifestyle content in luxury brand print advertisements. *Journal of advertising research* 60(1):28–37.

Matsuo M, Minami C, Matsuyama T (2018) Social influence on innovation resistance in internet banking services. *Journal of Retailing and Consumer Services* 45:42–51.

Mattsson J (2012) Strategic insights from mystery shopping in B2B relationships. *Journal of strategic marketing* 20(4):313–322.

Mendes AB, Cardoso MGMS (2006) Clustering supermarkets: the role of experts. *Journal of Retailing and Consumer Services* 13(4):231–247.

Mittal V, Anderson EW, Sayrak A, Tadikamalla P (2005) Dual emphasis and the long-term financial impact of customer satisfaction. *Marketing Science* 24(4):544–555.

Mittal V, Kamakura WA (2001) Satisfaction, repurchase intent, and repurchase behavior: Investigating the moderating effect of customer characteristics. *JMR, Journal of marketing research* 38(1):131–142.

Moe WW, Trusov M, Smith RH (2011) The value of social dynamics in online product ratings forums. *J. Mark. Res.* 48(3):444–456.

Mortimer K, Laurie S (2016) A Cross-Lagged test of the association between customer satisfaction and employee job satisfaction in a relational context. *J. Appl. Psychol.* 101(5):743–755.

MSPA (2018) Mystery shopping – how big is the market. [online] https://www.mspa-ea.org/news/newsitem/58-mystery-shopping-how-big-is-the-market.html, (Accessed 6 April 2021).

Muskat B, Hörtnagl T, Prayag G, Wagner S (2019) Perceived quality, authenticity, and price in tourists' dining experiences: Testing competing models of satisfaction and behavioral intentions. *Journal of Vacation Marketing* 25(4):480–498.

Namkung Y, Jang S (2007) Does food quality really matter in restaurants? its impact on customer satisfaction and behavioral intentions. *Journal of Hospitality & Tourism Research* 31(3):387–409.

Oliver RL (1980) A cognitive model of the antecedents and consequences of satisfaction decisions. *J. Mark. Res.* 17(4):460.

Otterbring T (2017) Smile for a while: the effect of employee-displayed smiling on customer affect and satisfaction. *Journal of Service Management* 28(2):284–304.

Palomba A (2020) Consumer personality and lifestyles at the box office and beyond: How demographics, lifestyles and personalities predict movie consumption. *Journal of Retailing and Consumer Services* 55.

Peterman K, Young D (2015) Mystery shopping: An innovative method for observing interactions with scientists during public science events. *Visitor studies* 18(1):83–102.

Pierce L, Snow DC, McAfee A (2015) Cleaning house: The impact of information technology monitoring on employee theft and productivity. *Manage. Sci.* 61(10):2299–2319.

Plummer JT (1974) The concept and application of life style segmentation. *Journal of marketing* 38(1):33–37.

Porter MC, Heyman JE (2018) We've shopped before: Exploring instructions as an influence on mystery shopper reporting. *Journal of Retailing and Consumer Services* 45(May):12–20.

Resnick P, Zeckhauser R, Swanson J, Lockwood K (2006) The value of reputation on ebay: A controlled experiment. *Exp. Econ.* 9(2):79–101.

Rust Anthony J, Roland T (1993) Customer satisfaction, customer retention, and market share. *J. Retail.* 69(2):193–215.

Ryu K, Han H (2010) Influence of the quality of food, service, and physical environment on customer satisfaction and behavioral intention in Quick-Casual restaurants: Moderating role of perceived price. *Journal of Hospitality & Tourism Research* 34(3):310–329.

Schlosser AE (2005) Posting versus lurking: Communicating in a multiple audience context. *J. Consum. Res.* 32(2):260–265.

Shunko M, Niederhoff J, Rosokha Y (2018) Humans are not machines: The behavioral impact of queueing design on service time. *Management science* 64(1):453–473.

Smith V (1976) Experimental economics: Induced value theory. *The American economic review* 66(2):274–279.

Staats BR, Dai H, Hofmann D, Milkman KL (2017) Motivating process compliance through individual electronic monitoring: An empirical examination of hand hygiene in healthcare. *Management Science* 63(5).

Sulek JM, Hensley RL (2004) The relative importance of food, atmosphere, and fairness of wait: The case of a full-service restaurant. *The Cornell hotel and restaurant administration quarterly* 45(3):235–247.

Sunder S, Kim KH, Yorkston EA (2019) What drives herding behavior in online ratings? the role of rater experience, product portfolio, and diverging opinions. *J. Mark.* 83(6):93–112.

Taber KS (2018) The use of cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education* 48(6):1273–1296.

Takenaka T, Ishigaki T, Motomura Y (2011) Demand forecasting method for service industries focusing on human behavior. *The 25th Annual Conference of the Japanese Society for ArtificialIntelligence, 2011*, 1–4, 25.

Takenaka T, Koshiba H, Motomura Y, Ueda K (2013) Product/service variety strategy considering mixed distribution of human lifestyles. *CIRP Annals* 62(1):463–466.

Takenaka T, Nishikoori H, Nishino N, Watanabe K (2020) Re-design of service systems based on employee satisfaction, customer satisfaction and labour productivity. *European Review of Service Economics and Management* 10:17–47.

Takenaka T, Yamamoto Y, Fukuda K, Kimura A, Ueda K (2016) Enhancing products and services using smart appliance networks. *CIRP Annals - Manufacturing Technology* 1(65):397–400.

Tambe P, Hitt LM (2012) The productivity of information technology investments: New evidence from IT labor data. *Information Systems Research* 23(3-part-1):599–617.

Tan T, Netessine S (2017) At your service on the table: Impact of tabletop technology on restaurant performance. *Manage. Sci.* (August 2020).

Tarantola C, Vicard P, Ntzoufras I (2012) Monitoring and improving greek banking services using bayesian networks: An analysis of mystery shopping data. *Expert systems with applications* 39(11):10103–10111.

Težak Damijanić A (2019) Wellness and healthy lifestyle in tourism settings. *Tourism Review* 74(4):978–989.

Truong Y (2013) A cross-country study of consumer innovativeness and technological service innovation. *Journal of Retailing and Consumer Services* 20(1):130–137.

Tupes EC, Christal RE (1992) Recurrent personality factors based on trait ratings. *Journal of personality* 60(2):225–251.

Wells WD, ed. (1974) *Life Style and Psychographics* (Chicago, IL: American Marketing Association).

Wiele TVd, Hesselink M, Van Iwaarden J (2005) Mystery shopping: A tool to develop insight into customer service provision. *Total Quality Management & Business Excellence* 16(4):529–541.

Wilson A (2002) Attitudes towards customer satisfaction measurement in the retail sector. *International Journal of Market Research* 44(2):1–9.

Wilson AM (2001) Mystery shopping : Using service performance. *Psychology & Marketing* 18(7):721–734.

Wu C, Che H, Chan TY, Lu X (2015) The economic value of online reviews. *Marketing Science* 34(5):739–754.

Yaoyuneyong G, Whaley JE, Butler RA, Williams JA, Jordan KL Jr, Hunt L (2018) Resort mystery shopping: A case study of hotel service. *Journal of Quality Assurance in Hospitality & Tourism* 19(3):358–386.

Ye Q, Law R, Gu B (2009) The impact of online user reviews on hotel room sales. *Int. J. Hosp. Manage.* 28(1):180–182.

Ye Q, Li G, Gu B (2011) A cross-cultural validation of the web usage-related lifestyle scale: An empirical investigation in china. *Electronic commerce research and applications* 10(3):304–312.

Zeithaml Va (1981) How consumer evaluation processes differ between goods and services. *Marketing of Services* 9(1):186–190.

Zhang J, Tucker H, Albrecht JN (2020) The reflexive self-project of "lifestyle entrepreneurial migrants". *Journal of Travel and Tourism Marketing* 37(5):535–546.

Zhao Y, Yang S, Narayan V, Zhao Y (2013) Modeling consumer learning from online product reviews. *Marketing Science* 32(1):153–169.

Zhu F, Zhang Xm (2010) Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *J. Mark.* 74(2):133–148.

Zorica MB, Ivanjko T, Spiranec S (2014) Mystery shopping in libraries – are we ready? *Qualitative and Quantitative Methods in Libraries* 2:433–442.

Acknowledgment

本研究を進めるにあたり，多くの方々にご指導・ご協力を頂きました.
まず，著者の指導教員である

東京大学大学院 工学系研究科 技術経営戦略学専攻 准教授　西野成昭先生

に深く感謝申し上げます．西野先生には，学部４年次の研究室配属から６年間ご指導を頂きました.
また

産業技術総合研究所　竹中毅氏

にも深く感謝申し上げます．竹中氏には，産業技術総合研究所のリサーチアシスタントとして受け入れていただき，共同研究に携わりながら，修士1年時からほぼ５年間ご指導いただきました.
また

早稲田大学 国際学術院 国際教養学部　石川竜一郎先生

にも深く感謝申し上げます．石川先生には，修士1年次から博士1年次を通して，特に意思決定理論の観点からご指導をいただきました.
また

東京大学大学院 工学系研究科 技術経営戦略学専攻 講師 木見田康治先生

にも深く感謝申し上げます．木見田先生には，博士2年次から博士3年次まで，筆者の研究に対して貴重なご教示をいただきました.
また MS & Consulting Co. Ltd. の錦織浩志氏には、ミステリーショッピングの研究に関して、様々な形でご助力いただきました。深く感謝致します。
研究室秘書の江えいぎさん，田中珠恵さんにも様々な形でお世話になりました．深く感謝致します.
西野研究室の皆様には，研究生活を通じて，様々な形で刺激を受けました．厚く感謝致します.
最後に, いつでも私を応援してくださった両親に, 何より心から感謝致します.

令和4年3月髙橋 裕紀

Publication and Presentation

Publications and Presentations Related to Doctoral Dissertation

Journal

- (Revision and Resubmit) 徐亦陶, 高橋裕紀, 木見田康治, 西野成昭：「長期間の継続的サービス利用を促す最適な情報提供タイミングの分析」サービソロジー論文誌.

- (Under Review) Takahashi H, Nishino N, More Information Leads to Better Purchases? Analysis of Review System – Effects on Purchasing and Rating Behaviors through Economic Experiments. PLOS ONE.

- (Accepted) Takahashi H, Kawasaki S, Takenaka T, Nishikoori H, Mystery Shopping Considering Lifestyle Heterogeneity. International Journal of Services and Operations Management.

Conference Papers

- Takahashi H, Nishino N, Ishikawa R (2019) Service switching in case-based decisions following bad experiences: Online reviews data of Japanese hairdressing salons. The 27th International Conference on Case-Based Reasoning (ICCBR 2019) Workshop Proceedings, 74-84, 8-12.

- Inoue Y, Takenaka T, Takahashi H (2018) Effect of service recommendation methods on platform ecosystem development. Proceedings of the ICSSI2018/ICServ2018.

Presentation(International)

- Takahashi H, Nishino N (2021) How does a reputation system affect your purchasing and reviewing behavior? An analysis through economic experiments, International Workshop for Lab and Field Experiments, 17-18, March 2021.

- Takahashi H, Nishino N, Shinmura T (2019) Meal Specific Consumption Cycles On Choice Behavior: In Case Of Japanese Meal. INFORMS Annual Meeting 2019, 20-23, November 2019.

Presentation (in Japan)

- 高橋裕紀，西野成昭：「サービス評価プラットフォームが購買選択とレビューに与える影響：経済実験による分析」サービス学会第 9 回国内大会，9-10, 3 月 2021.

- 高橋裕紀，西野成昭，新村猛：「飲食メニュー選択における選択肢固有の飽きに関する実証分析」サービス学会第 8 回国内大会，12-13，3 月 2020.

Publications and Presentations not Related to Doctoral Dissertation

Journal

- Nishino N, Takenaka T, Takahashi H (2017) Manufacturer’s strategy in a sharing economy. CIRP Annals 66(1):409–412.

Conference Papers

- Takahashi H, Nishino N, Takenaka T, Ishikawa R (2020) Interpreting value creation model by case-based decision theory. Procedia CIRP 88:584–588.

- Nishino N, Takenaka T, Takahashi H, Inoue Y (2020) Platform in manufacturing for enhancement of product value by sharing. Procedia CIRP 88:574–579.

- Takahashi H, Nishino N, Takenaka T (2018) Multi-agent simulation for the manufacturer’s decision making in sharing markets. Procedia CIRP 67:546–551.

Presentation(International)

- Takahashi H, Nishino N, Takenaka T, Ishikawa R (2019) Interpreting value creation model by case-based decision theory. 13th CIRP Conference on Intelligent Computation in Manufacturing Engineering, CIRP ICME '19, 17-19, July 2019.

- Takahashi H, Nishino N, Takenaka T, Ishikawa R (2018) Applicability of case-based decision theory to service contexts. Joint International Conference of Service Science and Innovation (ICSSI 2018) and Serviceology (ICServ2018), 13-15, November 2018.

- Takahashi H, Nishino N, Takenaka T (2017) Multi-agent simulation for the manufacturer’s decision making in sharing markets. 11th CIRP Conference on Intelligent Computation in Manufacturing Engineering, CIRP ICME '17, 17-19, July 2017.

Presentation (in Japan)

- 高橋裕紀，西野成昭，竹中毅，石川竜一郎：「潜在的印象に基づく類似度を用いた意思決定モデルの検証」サービス学会第7回国内大会, 2-3, 3月2019.

- 高橋裕紀，西野成昭，竹中毅，石川竜一郎：「事例ベース意思決定理論よるサービスの満足の記述と応用に関する検討」サービス学会第6回国内大会, 東京, 3月2018.

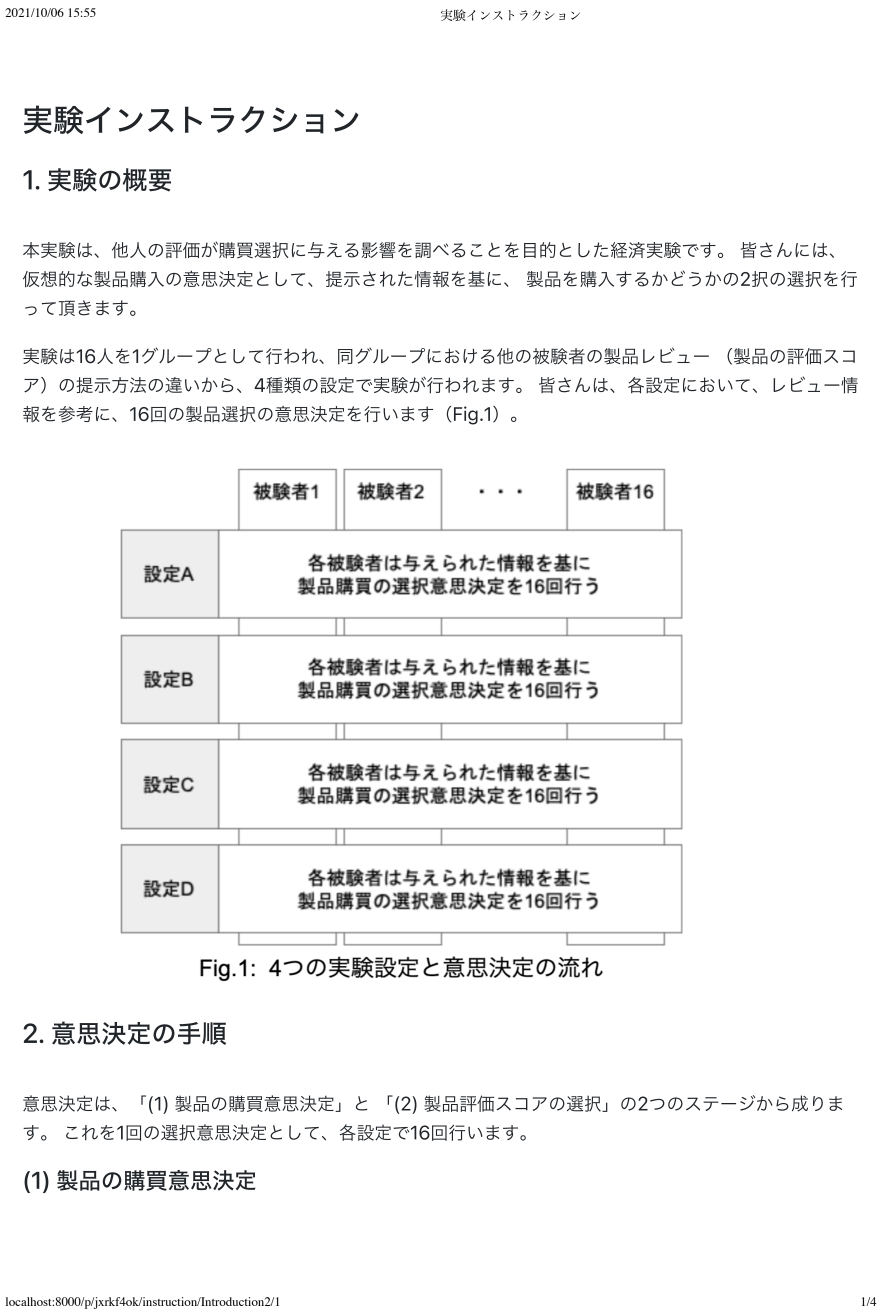
- 高橋裕紀，西野成昭，竹中毅：「マルチエージェントシミュレーションを用いたシェアリングサービスにおける生産者の意思決定分析」サービス学会第5回国内大会，27-29, 3月2017.

Appendix A

Appendix

A.1 Instructions of Experiments in Chapter 3

Instructions were written in Japanese. These instructions are read out by using Text-to-speach of Google Cloud.



実験インストラクション

1. 実験の概要

本実験は、他人の評価が購買選択に与える影響を調べることを目的とした経済実験です。 皆さんには、仮想的な製品購入の意思決定として、提示された情報を基に、 製品を購入するかどうかの2択の選択を行って頂きます。

実験は16人を1グループとして行われ、同グループにおける他の被験者の製品レビュー （製品の評価スコア）の提示方法の違いから、4種類の設定で実験が行われます。 皆さんは、各設定において、レビュー情報を参考に、16回の製品選択の意思決定を行います（Fig.1）。

	被験者1	被験者2	・・・	被験者16
設定A	各被験者は与えられた情報を基に 製品購買の選択意思決定を16回行う			
設定B	各被験者は与えられた情報を基に 製品購買の選択意思決定を16回行う			
設定C	各被験者は与えられた情報を基に 製品購買の選択意思決定を16回行う			
設定D	各被験者は与えられた情報を基に 製品購買の選択意思決定を16回行う			

Fig.1: 4つの実験設定と意思決定の流れ

2. 意思決定の手順

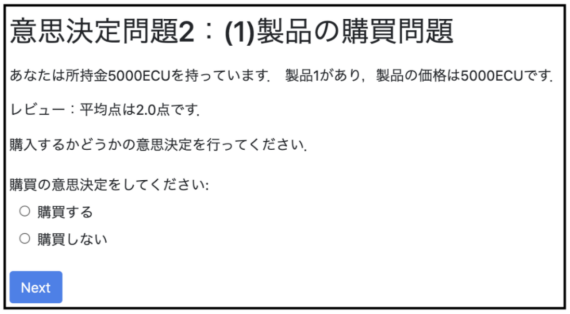
意思決定は、「(1) 製品の購買意思決定」と 「(2) 製品評価スコアの選択」の2つのステージから成ります。 これを1回の選択意思決定として、各設定で16回行います。

(1) 製品の購買意思決定

Fig.2のような画面で、製品を購入するかどうかを決定してください。 選択後、Nextボタンを押してください。ただし、一度ボタンを押すと修正ができませんので、慎重に行ってください。

製品の種類は「製品1」〜「製品16」の16種類で、順番はランダムです。ただし、同じ設定の実験内では同じ製品が出てくることはなく、各製品について1度だけ購買意思決定を行うことになります。

なお、画面に表示されている、所持金、製品価格、レビューについては、後で説明します。また、ECUは実験の中の仮想的な通貨の単位で、全ての実験設定を通じてECUを用います。



意思決定問題2：(1)製品の購買問題

あなたは所持金5000ECUを持っています. 製品1があり，製品の価格は5000ECUです.

レビュー：平均点は2.0点です.

購入するかどうかの意思決定を行ってください.

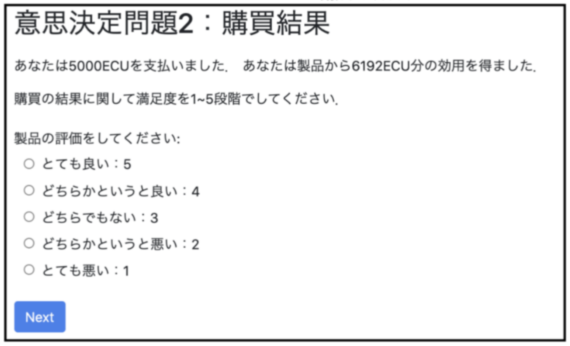
購買の意思決定をしてください:

○ 購買する
○ 購買しない

Fig.2: 製品の購買意思決定の画面

(2) 製品評価スコアの選択

(1) 製品の購買意思決定で、「購入する」を選択した場合は. Fig.3のような結果画面が表示されます。 この結果の満足度として、「とても良い（5点）」から「とても悪い（1点）」の5段階評価で選択してください。 選択後、Nextボタンを押してください。ただし、一度ボタンを押すと修正ができませんので、慎重に行ってください。 なお、画面に表示されている「効用」については、次節で説明します。



意思決定問題2：購買結果

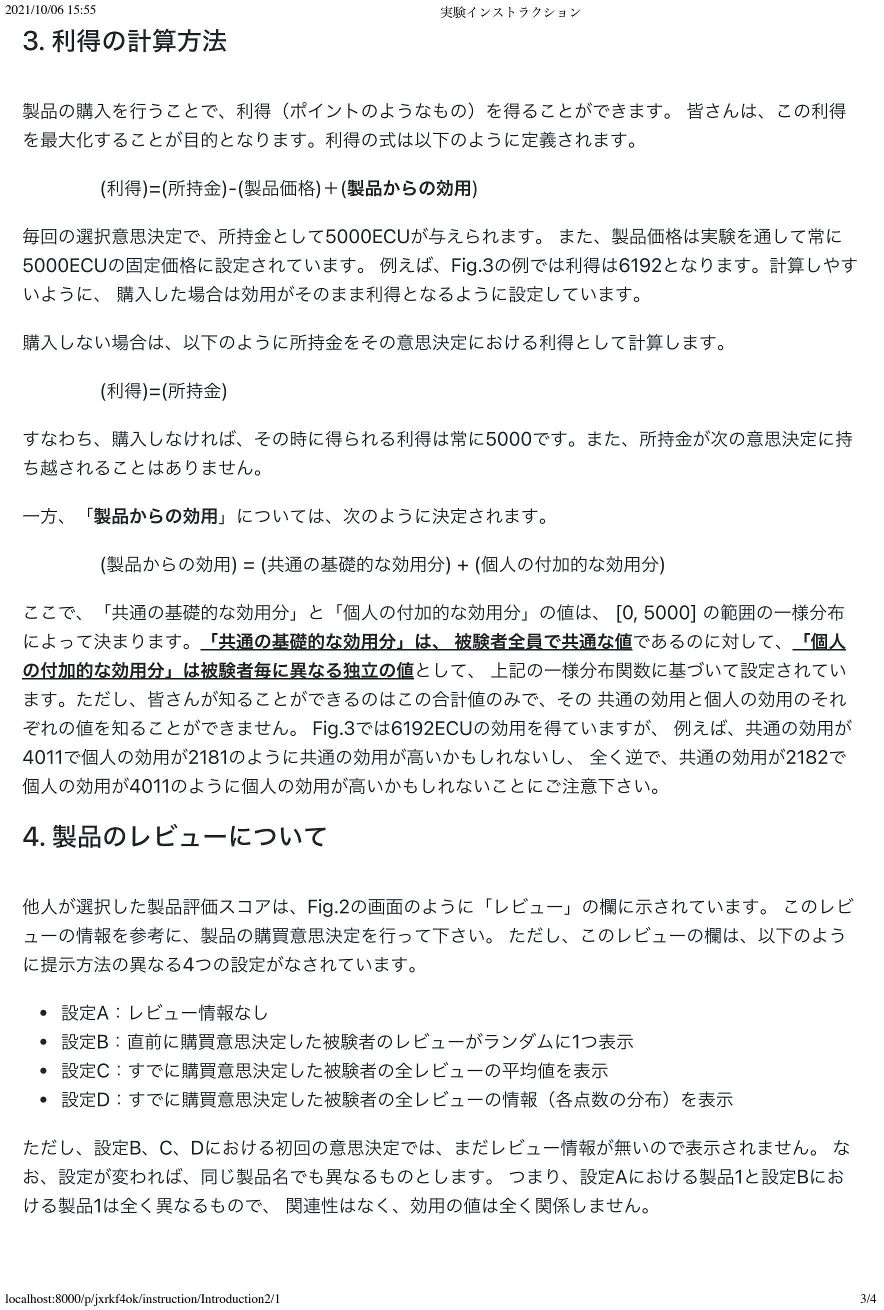
あなたは5000ECUを支払いました. あなたは製品から6192ECU分の効用を得ました.

購買の結果に関して満足度を1~5段階でしてください.

製品の評価をしてください:

○ とても良い：5
○ どちらかというと良い：4
○ どちらでもない：3
○ どちらかというと悪い：2
○ とても悪い：1

Fig.3: 製品の評価スコアの選択画面



3. 利得の計算方法

製品の購入を行うことで、利得（ポイントのようなもの）を得ることができます。 皆さんは、この利得を最大化することが目的となります。利得の式は以下のように定義されます。

(利得)=(所持金)-(製品価格)＋**(製品からの効用)**

毎回の選択意思決定で、所持金として5000ECUが与えられます。 また、製品価格は実験を通して常に5000ECUの固定価格に設定されています。 例えば、Fig.3の例では利得は6192となります。計算しやすいように、 購入した場合は効用がそのまま利得となるように設定しています。

購入しない場合は、以下のように所持金をその意思決定における利得として計算します。

(利得)=(所持金)

すなわち、購入しなければ、その時に得られる利得は常に5000です。また、所持金が次の意思決定に持ち越されることはありません。

一方、「**製品からの効用**」については、次のように決定されます。

(製品からの効用) = (共通の基礎的な効用分) + (個人の付加的な効用分)

ここで、「共通の基礎的な効用分」と「個人の付加的な効用分」の値は、 [0, 5000] の範囲の一様分布によって決まります。**「共通の基礎的な効用分」は、 被験者全員で共通な値**であるのに対して、**「個人の付加的な効用分」は被験者毎に異なる独立の値**として、 上記の一様分布関数に基づいて設定されています。ただし、皆さんが知ることができるのはこの合計値のみで、その 共通の効用と個人の効用のそれぞれの値を知ることができません。 Fig.3では6192ECUの効用を得ていますが、 例えば、共通の効用が4011で個人の効用が2181のように共通の効用が高いかもしれないし、 全く逆で、共通の効用が2182で個人の効用が4011のように個人の効用が高いかもしれないことにご注意下さい。

4. 製品のレビューについて

他人が選択した製品評価スコアは、Fig.2の画面のように「レビュー」の欄に示されています。 このレビューの情報を参考に、製品の購買意思決定を行って下さい。 ただし、このレビューの欄は、以下のように提示方法の異なる4つの設定がなされています。

- 設定A：レビュー情報なし
- 設定B：直前に購買意思決定した被験者のレビューがランダムに1つ表示
- 設定C：すでに購買意思決定した被験者の全レビューの平均値を表示
- 設定D：すでに購買意思決定した被験者の全レビューの情報（各点数の分布）を表示

ただし、設定B、C、Dにおける初回の意思決定では、まだレビュー情報が無いので表示されません。 なお、設定が変われば、同じ製品名でも異なるものとします。 つまり、設定Aにおける製品1と設定Bにおける製品1は全く異なるもので、 関連性はなく、効用の値は全く関係しません。

設定A〜Dがどの順番で行われるかは、実験毎にランダムに決まっています。 開始時にどの設定になっているかは、必ず確認して下さい。

5. 実験参加に対する謝金について

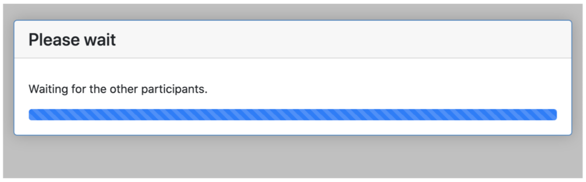
謝金は、固定報酬1600円と変動報酬の合計額で計算されます。 変動報酬は、各設定における16回の利得のうちからそれぞれランダムで1つ取り出し、 取り出された4つの利得の合計値を使います。 その合計値に0.1を乗じた値（小数点以下は四捨五入(注１)）とします。すなわち、以下の式となります。

> (変動報酬) = 0.1×{ (設定Aで選ばれた利得) + (設定Bで選ばれた利得) + (設定Cで選ばれた利得) + (設定Dで選ばれた利得) }

例えば、設定Aで7000、設定Bで4000、設定Cで8500、設定Dで3500が選ばれた場合、 変動報酬は0.1×(7000+4000+8500+3500)=2300円 となります。

6. Waiting画面について

実験中，Fig.4の画面が表示されることがあります.



Please wait

Waiting for the other participants.

Fig.4: Waiting画面

こちらは他の被験者の選択を待っているときに表示されます。 他の被験者の選択が終了すると自動的に次の画面に移りますので、そのままお待ち下さい。

次へ

A.2 Realized Heterogeneous quality in Previous Periods in Chapter 3

Define *Average Prior Heterogeneous Quality* as the average value of service *j*'s realized heterogeneous quality until the decision-making period. Regressions in Table 3.4 and 3.5 are reconducted with adding *Average Prior Heterogeneous Quality* in control variables. Table A.1 shows that the coefficients of *Base Quality* are still significant even adding *Average Prior Heterogeneous Quality*. Additionally, the coefficients of *Average Prior Heterogeneous Quality* are significant. This is because if *Average Prior Heterogeneous Quality* is high, then *Displayed Rating Score* should become higher. However, there are no significant effect of *Average Prior Heterogeneous Quality* on *Selected Rating Score*.

Tab. A.1: The results of binary logistic regression on purchase decision

	Dependent Variable: Purchase Decision			
	Treatment A	Treatment B	Treatment C	Treatment D
	(1)	(2)	(3)	(4)
Base Quality	0.005	0.15^{***}	0.27^{***}	0.30^{***}
	(0.01)	(0.01)	(0.01)	(0.01)
Average Prior Heterogeneous Quality	0.01	0.06^{***}	0.06^{***}	0.10^{***}
	(0.01)	(0.01)	(0.01)	(0.01)
Period	-0.02^{*}	-0.05^{***}	-0.05^{***}	−0.01
	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.48^{***}	0.28^{***}	0.40^{***}	0.48^{***}
	(0.03)	(0.03)	(0.03)	(0.02)
Group Dummies	Yes	Yes	Yes	Yes
Observations	1,440	1,440	1,440	1,440
Log Likelihood	−1,018.44	−862.95	−758.40	−662.42
Akaike Inf. Crit.	2,054.88	1,743.90	1,534.80	1,342.84

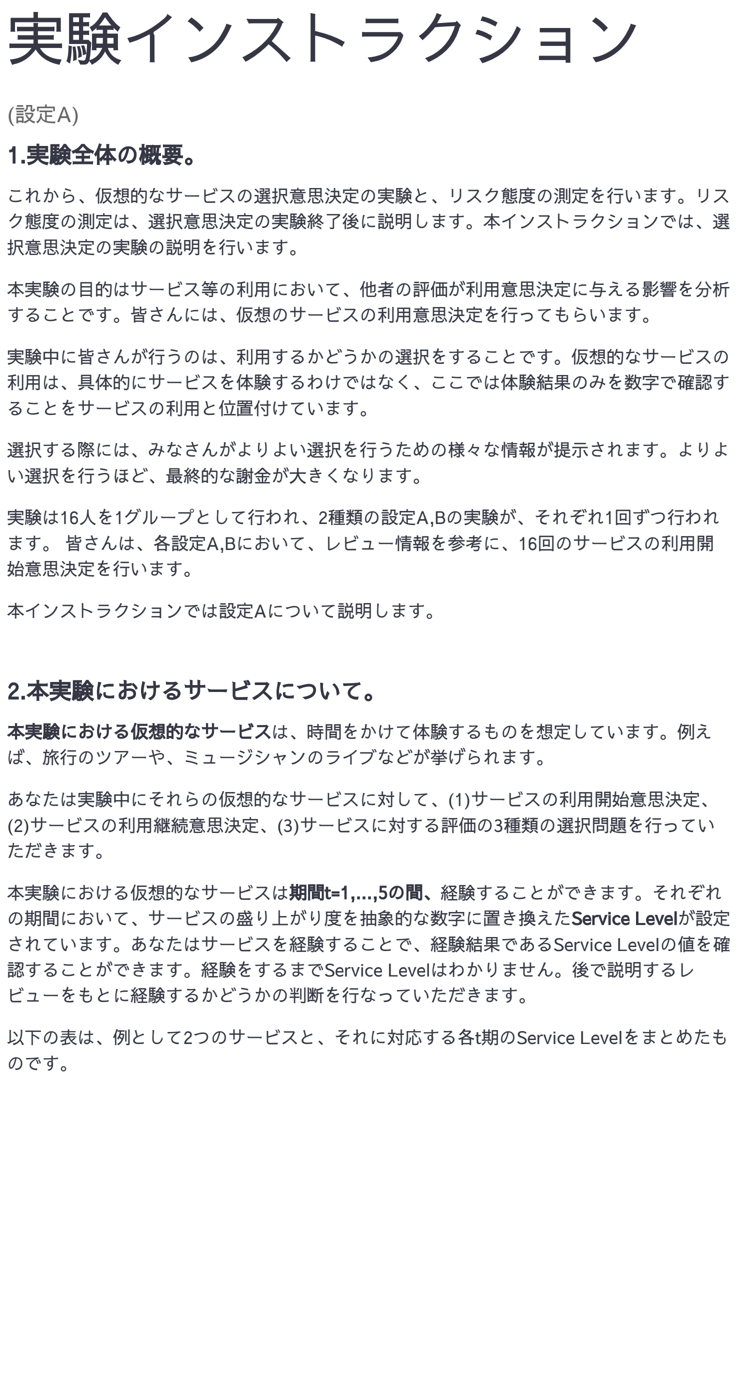
Note: *p<0.1; **p<0.05; ***p<0.01
Binary logit estimations. Standard errors in parentheses.

Tab. A.2: The result of ordered logistic regression on selected rating scores

	Dependent Variable: Selected Rating Score						
	Treatment A	Treatment B		Treatment C		Treatment D	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Utility	3.62***	2.32***	2.42***	2.90***	3.01***	2.86***	2.86***
	(0.16)	(0.15)	(0.16)	(0.15)	(0.17)	(0.14)	(0.15)
Displayed Rating Score			−0.17**		−0.08		−0.06
			(0.07)		(0.13)		(0.11)
Average Prior Heterogeneous Quality	−0.02	−0.07	−0.06	−0.05	−0.003	−0.04	−0.01
	(0.07)	(0.08)	(0.10)	(0.08)	(0.10)	(0.08)	(0.09)
Period	−0.01	0.03	0.01	−0.03	−0.02	0.01	0.01
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Group Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	792	507	466	662	618	720	678
Log Likelihood	−758.92	−595.82	−543.59	−726.96	−671.32	−782.33	−739.54

Note: *p<0.1; **p<0.05; ***p<0.01
Ordered logit estimations. Standard errors in parentheses.

A.3 Instructions of Experiments in Chapter 4



実験インストラクション

(設定A)

1.実験全体の概要。

これから、仮想的なサービスの選択意思決定の実験と、リスク態度の測定を行います。リスク態度の測定は、選択意思決定の実験終了後に説明します。本インストラクションでは、選択意思決定の実験の説明を行います。

本実験の目的はサービス等の利用において、他者の評価が利用意思決定に与える影響を分析することです。皆さんには、仮想のサービスの利用意思決定を行ってもらいます。

実験中に皆さんが行うのは、利用するかどうかの選択をすることです。仮想的なサービスの利用は、具体的にサービスを体験するわけではなく、ここでは体験結果のみを数字で確認することをサービスの利用と位置付けています。

選択する際には、みなさんがよりよい選択を行うための様々な情報が提示されます。よりよい選択を行うほど、最終的な謝金が大きくなります。

実験は16人を1グループとして行われ、2種類の設定A,Bの実験が、それぞれ1回ずつ行われます。 皆さんは、各設定A,Bにおいて、レビュー情報を参考に、16回のサービスの利用開始意思決定を行います。

本インストラクションでは設定Aについて説明します。

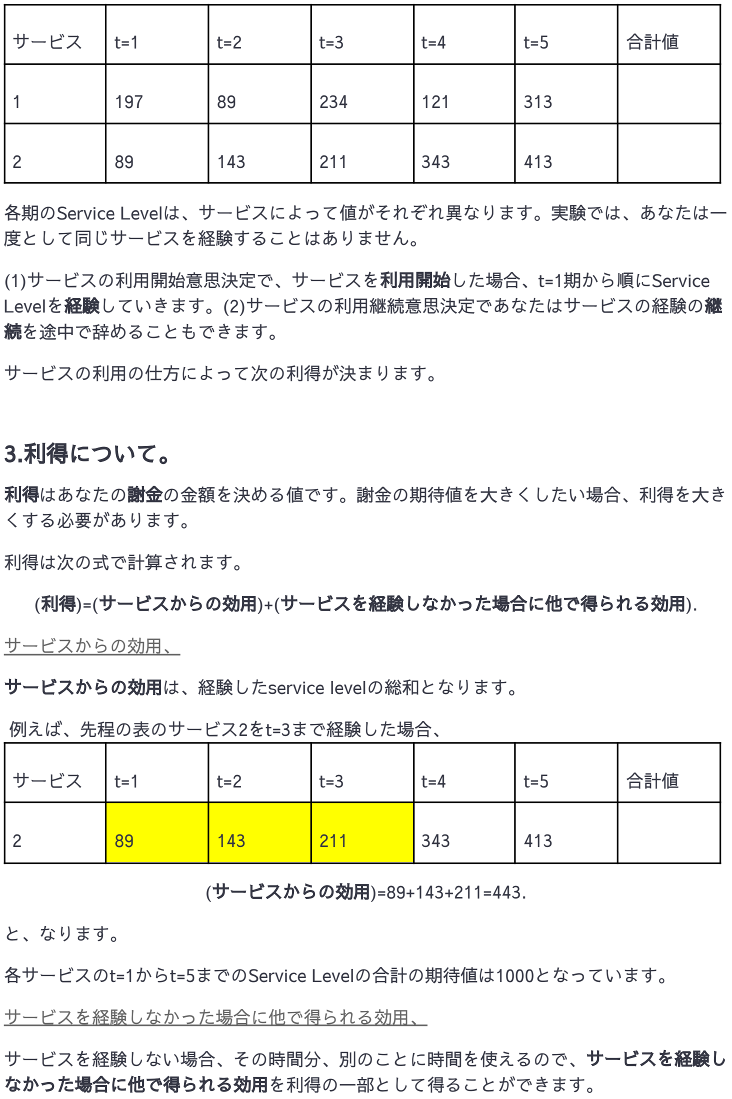
2.本実験におけるサービスについて。

本実験における仮想的なサービスは、時間をかけて体験するものを想定しています。例えば、旅行のツアーや、ミュージシャンのライブなどが挙げられます。

あなたは実験中にそれらの仮想的なサービスに対して、(1)サービスの利用開始意思決定、(2)サービスの利用継続意思決定、(3)サービスに対する評価の3種類の選択問題を行っていただきます。

本実験における仮想的なサービスは**期間t=1,...,5の間、**経験することができます。それぞれの期間において、サービスの盛り上がり度を抽象的な数字に置き換えた**Service Level**が設定されています。あなたはサービスを経験することで、経験結果であるService Levelの値を確認することができます。経験をするまでService Levelはわかりません。後で説明するレビューをもとに経験するかどうかの判断を行なっていただきます。

以下の表は、例として2つのサービスと、それに対応する各t期のService Levelをまとめたものです。



サービス	t=1	t=2	t=3	t=4	t=5	合計値
1	197	89	234	121	313	
2	89	143	211	343	413	

各期のService Levelは、サービスによって値がそれぞれ異なります。実験では、あなたは一度として同じサービスを経験することはありません。

(1)サービスの利用開始意思決定で、サービスを**利用開始**した場合、t=1期から順にService Levelを**経験**していきます。(2)サービスの利用継続意思決定であなたはサービスの経験の**継続**を途中で辞めることもできます。

サービスの利用の仕方によって次の利得が決まります。

3.利得について。

利得はあなたの**謝金**の金額を決める値です。謝金の期待値を大きくしたい場合、利得を大きくする必要があります。

利得は次の式で計算されます。

(**利得**)=(**サービスからの効用**)+(**サービスを経験しなかった場合に他で得られる効用**).

サービスからの効用、

サービスからの効用は、経験したservice levelの総和となります。

例えば、先程の表のサービス2をt=3まで経験した場合、

サービス	t=1	t=2	t=3	t=4	t=5	合計値
2	89	143	211	343	413	

(**サービスからの効用**)=89+143+211=443.

と、なります。

各サービスのt=1からt=5までのService Levelの合計の期待値は1000となっています。

サービスを経験しなかった場合に他で得られる効用、

サービスを経験しない場合、その時間分、別のことに時間を使えるので、**サービスを経験しなかった場合に他で得られる効用**を利得の一部として得ることができます。

サービスを経験しなかった場合に他で得られる効用は、固定で1期あたり200となります。

例えば、先程の表のサービス2をt=3まで経験し、t=4以降で継続しなかった場合、

サービス	t=1	t=2	t=3	t=4	t=5	合計値
2	89	143	211	343	413	

t=4とt=5の2期分の、**サービスを経験しなかった場合に他で得られる効用**が得られるので、

(**サービスを経験しなかった場合に他で得られる効用**) = 200×2 = 400.

と、なります。

利得の合計は、先程のサービスからの効用と合わせて

(**利得**)=(**サービスからの効用**)+(**サービスを経験しなかった場合に他で得られる効用**)

=(89+143+211)+200×2

=443+400=843.

と、なります。

全期間、経験しなかった場合(サービスを利用開始しなかった場合)に、**サービスを経験しなかった場合に他で得られる効用**の合計は200×5=1000となっており、各サービスのService Levelの合計値の期待値の1000と同じ数値になっています。

4. 設定Aの実験のプロセス。

本実験は、それぞれ全16ラウンド行います。

各ラウンドは図1に従って進みます。

まず画面1で利用開始の意思決定を行った後、利用開始する場合は画面2に移ります。もし利用開始しない場合は、次のラウ

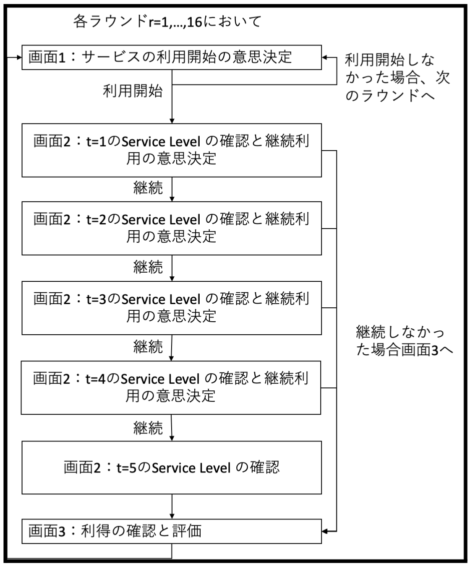

各ラウンドr=1,...,16において
画面1：サービスの利用開始の意思決定
利用開始しなかった場合、次のラウンドへ
利用開始
画面2：t=1のService Level の確認と継続利用の意思決定
継続
画面2：t=2のService Level の確認と継続利用の意思決定
継続
画面2：t=3のService Level の確認と継続利用の意思決定
継続
画面2：t=4のService Level の確認と継続利用の意思決定
継続
画面2：t=5のService Level の確認
継続しなかった場合画面3へ
画面3：利得の確認と評価

図1: 実験のプロセス

ンドに進み、再度、画面1から始まります。

画面2ではt=1から5で毎回、経験したService Levelの値を確認しつつ、継続するかどうかの意思決定を行っていただきます。継続する場合は、そのまま次の期に進みます。継続しない場合、もしくは最後まで経験した場合は画面3に移ります。

画面3ではこれまで得た利得を確認し、サービスに対する評価をつけていただきます。評価をつけた後は、次のラウンドに進み、画面1に移ります。

4.1 画面１。

まずサービスの利用開始意思決定を行っていただきます。図2は実際に表示される画面です。

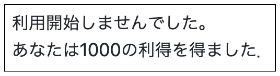
利用開始の意思決定の際には、今まで他の人がそのサービスにつけたレビューを確認することができます。同じサービスのService Levelは実験参加者で共通ですので、レビューを参考にした上で、利用開始するか決めてください。

利用開始する場合、

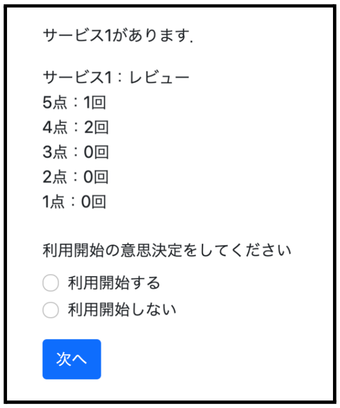
画面2へ進み、t=1のService Levelを経験します。

利用開始しない場合、

下のメッセージが表示され、次のラウンドに進み、画面１から再度始めます。



利用開始しませんでした。
あなたは1000の利得を得ました.



サービス1があります.

サービス1：レビュー
5点：1回
4点：2回
3点：0回
2点：0回
1点：0回

利用開始の意思決定をしてください

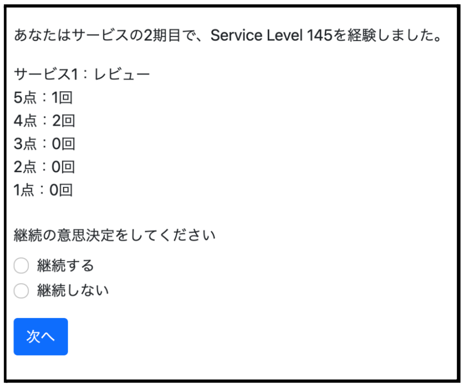
○ 利用開始する
○ 利用開始しない

図2: 画面1

4.2 画面2。

まず各t期において、そのサービスのservice levelの値を確認していただきます。図3は実際に表示される画面です。

例えば、画面1でサービスを利用開始した場合、まずt=1のservice levelの値を確認することとなります。



あなたはサービスの2期目で、Service Level 145を経験しました。

サービス1：レビュー
5点：1回
4点：2回
3点：0回
2点：0回
1点：0回

継続の意思決定をしてください

○ 継続する
○ 継続しない

図3: 画面2

その後t+1以降のサービスを継続するかどうかの意思決定を行います。

継続する場合、

継続する場合、t+1期に進み、再度画面2から行います。

例えばt=1期のService Levelを経験していた場合、継続すると次のt=2期のService Levelを経験することになります。

継続しない場合、

もし継続しないを選んだ場合、t+1期以降のservice levelは得られません。その分t+1期以降の**サービスを経験しなかった場合に他で得られる効用**を得ることができます。

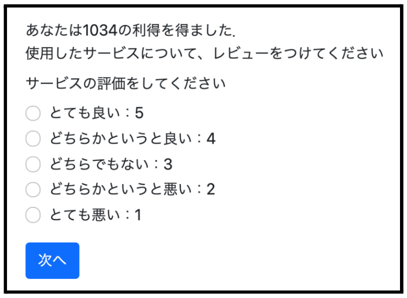
画面3に進んで、本ラウンドの利得の確認を行います。

例えばt=1期のService Levelを経験していた場合、継続しないと、t=2以降のService Levelを経験せずに、t=2からt=5までの**サービスを経験しなかった場合に他で得られる効用**を得ます。その後画面3の利得の確認に移ります。

4.3 画面3。

まずこのラウンドで得た利得を確認します。図4は実際に表示される画面です。

ここまで経験したservice levelに基づいて、サービス自体のレビューをつけてください。このレビューは集計され、次に同一のサービスの利用開始意思決定をする実験参加者に提示されます。次のラウンドに進み、次のサービスに対する意思決定を画面1から行います。



あなたは1034の利得を得ました.

使用したサービスについて、レビューをつけてください

サービスの評価をしてください

○ とても良い：5

○ どちらかというと良い：4

○ どちらでもない：3

○ どちらかというと悪い：2

○ とても悪い：1

図4: 画面3

5. 謝金の計算について。

あなたの謝金は、**固定報酬**と二種類の**変動報酬**の合計からなります。

謝金=固定報酬＋本実験における変動報酬+リスク態度の測定における変動報酬.

固定報酬は1500円。本実験における変動報酬は以下のルールで決まります。リスク態度の測定における変動報酬については、リスク態度の測定の際に詳しく説明します。

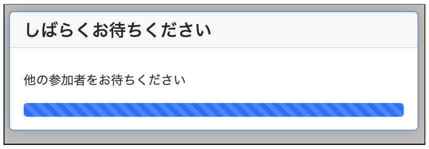
各設定ABそれぞれにおいて、全ラウンドのサービスの選択結果の中から、実験終了後に、実験参加者共通でランダムで1つのラウンドの選択結果が選ばれ、それをもとに、本実験における変動報酬を以下の式で計算します。

本実験における変動報酬=(設定Aで選ばれたラウンドの利得)+(設定Bで選ばれたラウンドの利得).

他の実験参加者の謝金や利得の大きさは、あなたの謝金に影響しません。各ラウンドの自分の利得を最大化するのが、自分の謝金の期待値を最大化することに繋がります。

6. 実験の進行について。

インストラクション終了後は、各自で実験を進めていただきます。各ページで先に進むときは画面の下部に表示されている[次へ]ボタンを押してください。また実験を進めていく間、図5の待機画面が表示される場合があります。



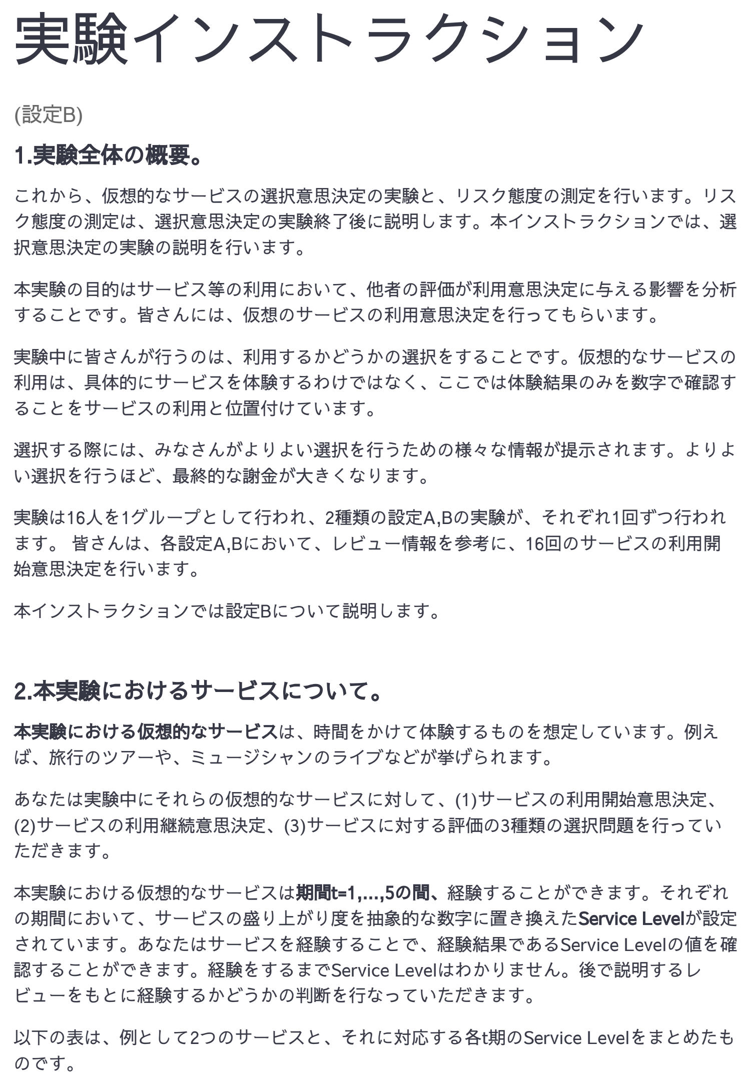
しばらくお待ちください

他の参加者をお待ちください

図5: 待機画面

これは他のプレイヤーの選択を待っているときに表示される画面です。そのまましばらくお待ち下さい。

質問がある場合はzoomのチャットに「質問です」とお書きください。ブレイクアウトルームで口頭で対応いたします。



実験インストラクション

(設定B)

1.実験全体の概要。

これから、仮想的なサービスの選択意思決定の実験と、リスク態度の測定を行います。リスク態度の測定は、選択意思決定の実験終了後に説明します。本インストラクションでは、選択意思決定の実験の説明を行います。

本実験の目的はサービス等の利用において、他者の評価が利用意思決定に与える影響を分析することです。皆さんには、仮想のサービスの利用意思決定を行ってもらいます。

実験中に皆さんが行うのは、利用するかどうかの選択をすることです。仮想的なサービスの利用は、具体的にサービスを体験するわけではなく、ここでは体験結果のみを数字で確認することをサービスの利用と位置付けています。

選択する際には、みなさんがよりよい選択を行うための様々な情報が提示されます。よりよい選択を行うほど、最終的な謝金が大きくなります。

実験は16人を1グループとして行われ、2種類の設定A,Bの実験が、それぞれ1回ずつ行われます。 皆さんは、各設定A,Bにおいて、レビュー情報を参考に、16回のサービスの利用開始意思決定を行います。

本インストラクションでは設定Bについて説明します。

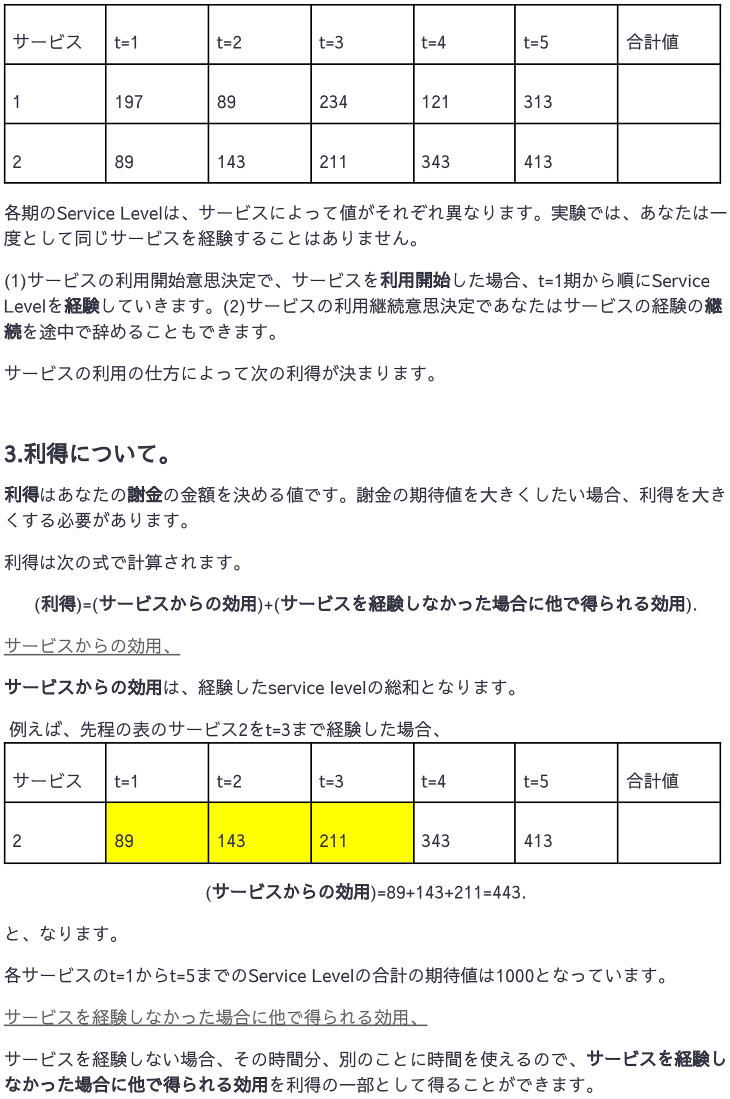
2.本実験におけるサービスについて。

本実験における仮想的なサービスは、時間をかけて体験するものを想定しています。例えば、旅行のツアーや、ミュージシャンのライブなどが挙げられます。

あなたは実験中にそれらの仮想的なサービスに対して、(1)サービスの利用開始意思決定、(2)サービスの利用継続意思決定、(3)サービスに対する評価の3種類の選択問題を行っていただきます。

本実験における仮想的なサービスは**期間t=1,...,5の間、**経験することができます。それぞれの期間において、サービスの盛り上がり度を抽象的な数字に置き換えた**Service Level**が設定されています。あなたはサービスを経験することで、経験結果であるService Levelの値を確認することができます。経験をするまでService Levelはわかりません。後で説明するレビューをもとに経験するかどうかの判断を行なっていただきます。

以下の表は、例として2つのサービスと、それに対応する各t期のService Levelをまとめたものです。



サービス	t=1	t=2	t=3	t=4	t=5	合計値
1	197	89	234	121	313	
2	89	143	211	343	413	

各期のService Levelは、サービスによって値がそれぞれ異なります。実験では、あなたは一度として同じサービスを経験することはありません。

(1)サービスの利用開始意思決定で、サービスを**利用開始**した場合、t=1期から順にService Levelを**経験**していきます。(2)サービスの利用継続意思決定であなたはサービスの経験の**継続**を途中で辞めることもできます。

サービスの利用の仕方によって次の利得が決まります。

3.利得について。

利得はあなたの**謝金**の金額を決める値です。謝金の期待値を大きくしたい場合、利得を大きくする必要があります。

利得は次の式で計算されます。

(**利得**)=(**サービスからの効用**)+(**サービスを経験しなかった場合に他で得られる効用**).

サービスからの効用、

サービスからの効用は、経験したservice levelの総和となります。

例えば、先程の表のサービス2をt=3まで経験した場合、

サービス	t=1	t=2	t=3	t=4	t=5	合計値
2	89	143	211	343	413	

(**サービスからの効用**)=89+143+211=443.

と、なります。

各サービスのt=1からt=5までのService Levelの合計の期待値は1000となっています。

サービスを経験しなかった場合に他で得られる効用、

サービスを経験しない場合、その時間分、別のことに時間を使えるので、**サービスを経験しなかった場合に他で得られる効用**を利得の一部として得ることができます。

サービスを経験しなかった場合に他で得られる効用は、固定で1期あたり200となります。

例えば、先程の表のサービス2をt=3まで経験し、t=4以降で継続しなかった場合、

サービス	t=1	t=2	t=3	t=4	t=5	合計値
2	89	143	211	343	413	

t=4とt=5の2期分の、**サービスを経験しなかった場合に他で得られる効用**が得られるので、

(**サービスを経験しなかった場合に他で得られる効用**) = 200×2 = 400.

と、なります。

利得の合計は、先程のサービスからの効用と合わせて

(**利得**)=(**サービスからの効用**)+(**サービスを経験しなかった場合に他で得られる効用**)

=(89+143+211)+200×2

=443+400=843.

と、なります。

全期間、経験しなかった場合(サービスを利用開始しなかった場合)に、**サービスを経験しなかった場合に他で得られる効用**の合計は200×5=1000となっており、各サービスのService Levelの合計値の期待値の1000と同じ数値になっています。

4. 設定Bの実験のプロセス。

本実験は、それぞれ全16ラウンド行います。

各ラウンドは図1に従って進みます。

まず画面1で利用開始の意思決定を行った後、利用開始する場合は画面2に移ります。もし利用開始しない場合は、次のラウンドに進み、再度、画面1から始まります。

画面2ではt=1から5で毎回、経験したService Levelの値を確認しつつ、t期のService Level

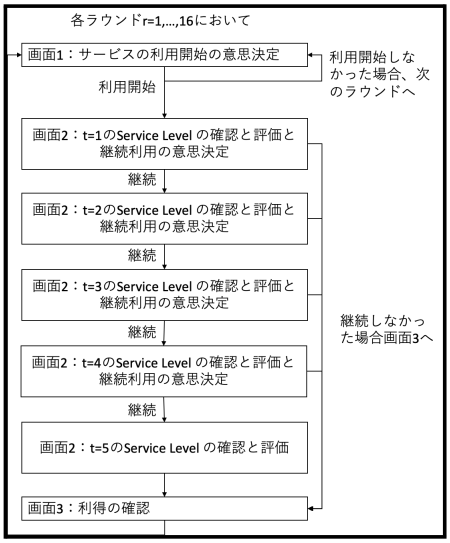

各ラウンドr=1,...,16において
画面1：サービスの利用開始の意思決定
利用開始しなかった場合、次のラウンドへ
利用開始
画面2：t=1のService Level の確認と評価と継続利用の意思決定
継続
画面2：t=2のService Level の確認と評価と継続利用の意思決定
継続
画面2：t=3のService Level の確認と評価と継続利用の意思決定
継続
画面2：t=4のService Level の確認と評価と継続利用の意思決定
継続
画面2：t=5のService Level の確認と評価
継続しなかった場合画面3へ
画面3：利得の確認

図1: 実験のプロセス

の評価をし、更に継続するかどうかの意思決定を行っていただきます。継続する場合は、そのまま次の期に進みます。継続しない場合、もしくは最後まで経験した場合は画面3に移ります。

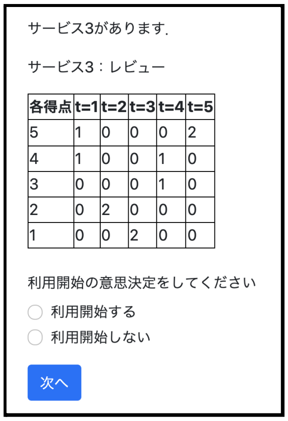
画面3ではこれまで得た利得を確認します。確認した後は、次のラウンドに進み、画面1に移ります。

4.1 画面１。

まずサービスの利用開始意思決定を行っていただきます。図2は実際に表示される画面です。

利用開始の意思決定の際には、今まで他の人がそのサービスにつけたレビューを確認することができます。同じサービスのService Levelは実験参加者で共通ですので、レビューを参考にした上で、利用開始するか決めてください。

レビューは各t期のService Levelに対してこれまでつけられた評価の数が載っています。例えば図2では、サービス3のt=1のService Levelに対して5点が1回、4点が1回、となります。



サービス3があります.

サービス3：レビュー

各得点	t=1	t=2	t=3	t=4	t=5
5	1	0	0	0	2
4	1	0	0	1	0
3	0	0	0	1	0
2	0	2	0	0	0
1	0	0	2	0	0

利用開始の意思決定をしてください

○ 利用開始する
○ 利用開始しない

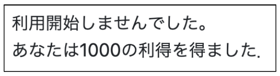
図2: 画面1

利用開始する場合、

画面2へ進み、t=1のService Levelを経験します。

利用開始しない場合、

下のメッセージが表示され、次のラウンドに進み、画面１から再度始めます。



利用開始しませんでした。
あなたは1000の利得を得ました.

4.2 画面2。

まず各t期において、そのサービスのservice levelの値を確認していただきます。図3は実際に表示される画面です。

例えば、画面1でサービスを利用開始した場合、まずt=1のservice levelの値を確認することとなります。

次にそのService Levelに対して5点満点で評価をつけてもらいます。このレビューは集計され、次に同一のサービスの利用開始意思決定をする実験参加者に提示されます。

その後t+1以降のサービスを継続するかどうかの意思決定を行います。

継続する場合、

継続する場合、t+1期に進み、再度画面2から行います。

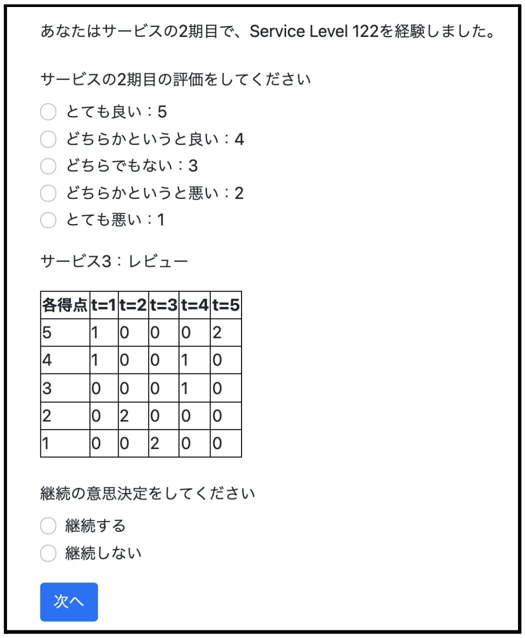
例えばt=1期のService Levelを経験していた場合、継続すると次のt=2期のService Levelを経験することになります。

継続しない場合、

もし継続しないを選んだ場合、t+1期以降のservice levelは得られません。その分t+1期以降の**サービスを経験しなかった場合に他で得られる効用**を得ることができます。

画面3に進んで、本ラウンドの利得の確認を行います。

例えばt=1期のService Levelを経験していた場合、継続しないと、t=2以降のService Levelを経験せずに、t=2からt=5までの**サービスを経験しなかった場合に他で得られる効用**を得ます。その後画面3の利得の確認に移ります。



あなたはサービスの2期目で、Service Level 122を経験しました。

サービスの2期目の評価をしてください

○ とても良い：5
○ どちらかというと良い：4
○ どちらでもない：3
○ どちらかというと悪い：2
○ とても悪い：1

サービス3：レビュー

各得点	t=1	t=2	t=3	t=4	t=5
5	1	0	0	0	2
4	1	0	0	1	0
3	0	0	0	1	0
2	0	2	0	0	0
1	0	0	2	0	0

継続の意思決定をしてください

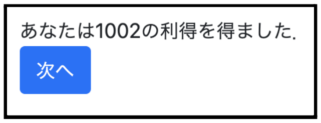
○ 継続する
○ 継続しない

図3: 画面2

4.3 画面3。

このラウンドで得た利得を確認します。図4は実際に表示される画面です。

その後、次のラウンドに進み、次のサービスに対する意思決定を画面1から行います。



あなたは1002の利得を得ました.

図4: 画面3

5. 謝金の計算について。

あなたの謝金は、**固定報酬**と二種類の**変動報酬**の合計からなります。

謝金=固定報酬＋本実験における変動報酬+リスク態度の測定における変動報酬.

固定報酬は1500円。本実験における変動報酬は以下のルールで決まります。リスク態度の測定における変動報酬については、リスク態度の測定の際に詳しく説明します。

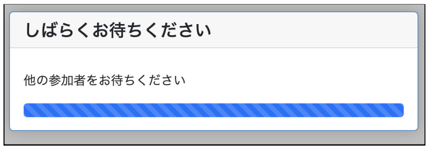
各設定ABそれぞれにおいて、全ラウンドのサービスの選択結果の中から、実験終了後に、実験参加者共通でランダムで1つのラウンドの選択結果が選ばれ、それをもとに、本実験における変動報酬を以下の式で計算します。

本実験における変動報酬=(設定Aで選ばれたラウンドの利得)+(設定Bで選ばれたラウンドの利得).

他の実験参加者の謝金や利得の大きさは、あなたの謝金に影響しません。各ラウンドの自分の利得を最大化するのが、自分の謝金の期待値を最大化することに繋がります。

6. 実験の進行について。

インストラクション終了後は、各自で実験を進めていただきます。各ページで先に進むときは画面の下部に表示されている[次へ]ボタンを押してください。また実験を進めていく間、図5の待機画面が表示される場合があります。



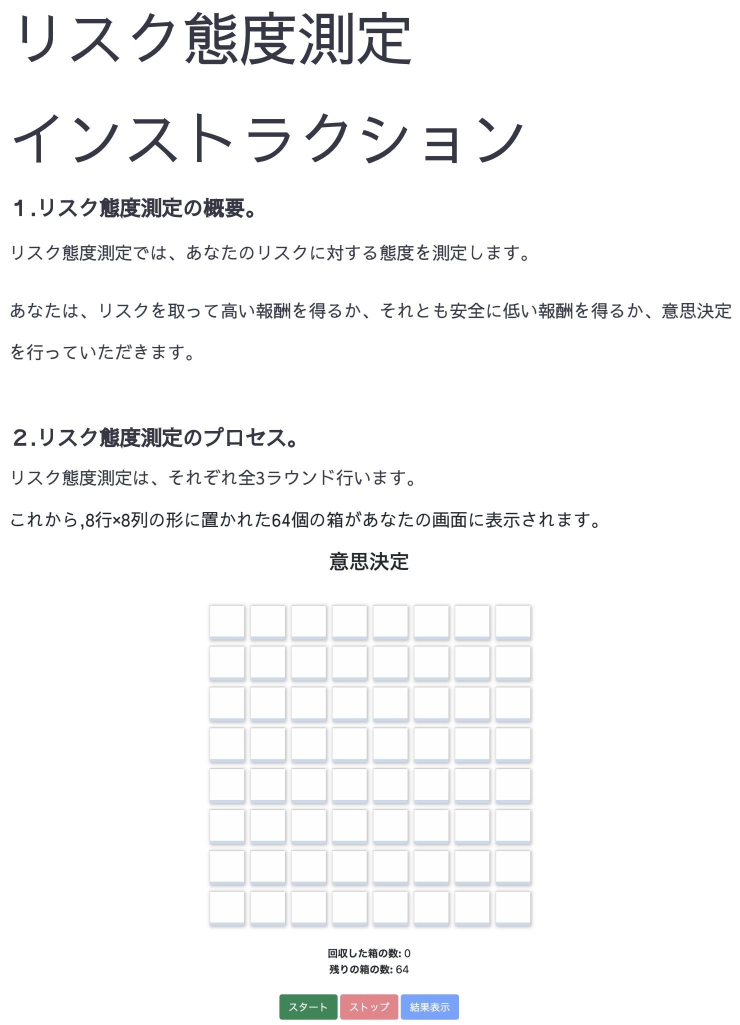
しばらくお待ちください

他の参加者をお待ちください

図5: 待機画面

これは他のプレイヤーの選択を待っているときに表示される画面です。そのまましばらくお待ち下さい。

質問がある場合はzoomのチャットに「質問です」とお書きください。ブレイクアウトルームで口頭で対応いたします。



リスク態度測定

インストラクション

１.リスク態度測定の概要。

リスク態度測定では、あなたのリスクに対する態度を測定します。

あなたは、リスクを取って高い報酬を得るか、それとも安全に低い報酬を得るか、意思決定を行っていただきます。

２.リスク態度測定のプロセス。

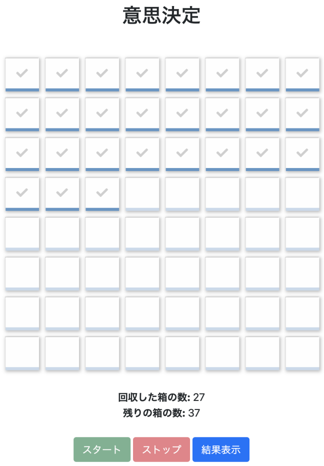
リスク態度測定は、それぞれ全3ラウンド行います。

これから,8行×8列の形に置かれた64個の箱があなたの画面に表示されます。

意思決定

回収した箱の数: 0
残りの箱の数: 64

スタートボタンを押すとタスクが始まり、1秒につき1つの箱が自動で回収されます。左上から自動で回収されていき、回収された箱にはチェックマークが付きます。一つの箱につき、20点得ることができます。



意思決定

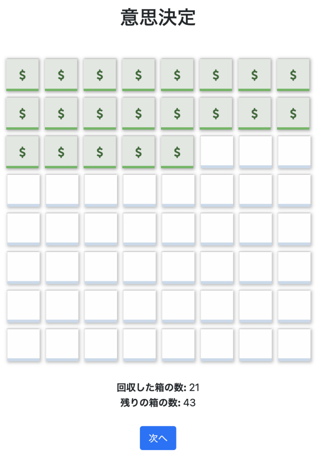
回収した箱の数: 27
残りの箱の数: 37

この内ある一つの箱には爆弾が入っています。残りの爆弾の入っていない63個の箱はそれぞれ20点となります.あなたはどこに爆弾が入っているか事前に知ることはできません。爆弾がどの箱に入っているかは、等確率で決まります。一部の箱で爆弾が出やすいことはありません。

あなたのやることは、どのタイミングでこの回収を止めるかを決めることです。いつでもStopボタンを押して箱の回収を止めることができます。もし爆弾の入った箱を回収してしまっていた場合、あなたの今回のラウンドでの合計点数は0となります。爆弾の入った箱を回収する前に、箱を回収するのをやめれば、そこまでに自動で集められた箱の合計点数を得ることができます。

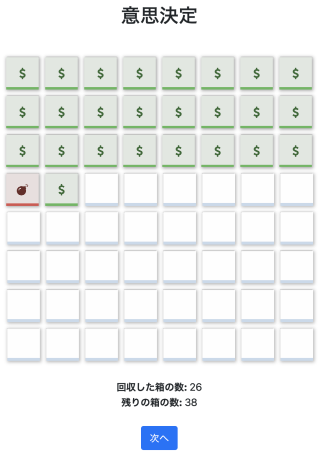
各ラウンドは結果表示ボタンを押すことで終了します。結果表示ボタンを押すと結果が確認できます。

以下の図は結果表示ボタンを押すと表示される画面です。爆弾の入っていなかった箱にはお金のマークが、爆弾の入っていた箱には爆弾のマークが表示されます。画像左は420点、画像右は0点となります。



意思決定

回収した箱の数: 21
残りの箱の数: 43



意思決定

回収した箱の数: 26
残りの箱の数: 38

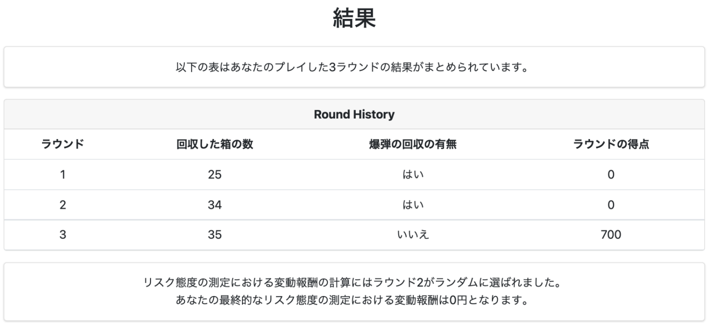
３.謝金の計算について

あなたの謝金は、**固定報酬**と二種類の**変動報酬**の合計からなります。

謝金=固定報酬＋実験における変動報酬+リスク態度の測定における変動報酬

本文章ではリスク態度の測定における変動報酬について説明します。

あなたはこれから3ラウンド、箱の回収タスクを行います。すべてのラウンドが終了後、ある一つのラウンドの結果がランダムに選ばれ、それがそのままリスク態度の測定における変動報酬となります。



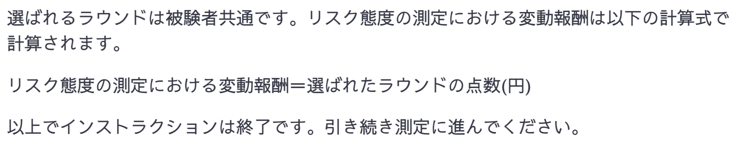
結果

以下の表はあなたのプレイした3ラウンドの結果がまとめられています。

Round History

ラウンド	回収した箱の数	爆弾の回収の有無	ラウンドの得点
1	25	はい	0
2	34	はい	0
3	35	いいえ	700

リスク態度の測定における変動報酬の計算にはラウンド2がランダムに選ばれました。
あなたの最終的なリスク態度の測定における変動報酬は0円となります。



選ばれるラウンドは被験者共通です。リスク態度の測定における変動報酬は以下の計算式で計算されます。

リスク態度の測定における変動報酬＝選ばれたラウンドの点数(円)

以上でインストラクションは終了です。引き続き測定に進んでください。

A.4 Additional Graphs in Chapter 4

Following graphs shows

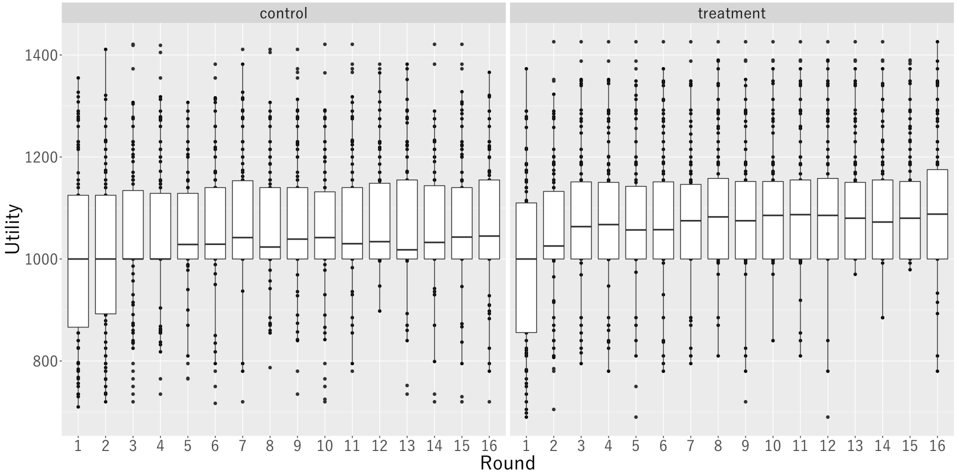


Fig. A.1: The relationship between utility and round

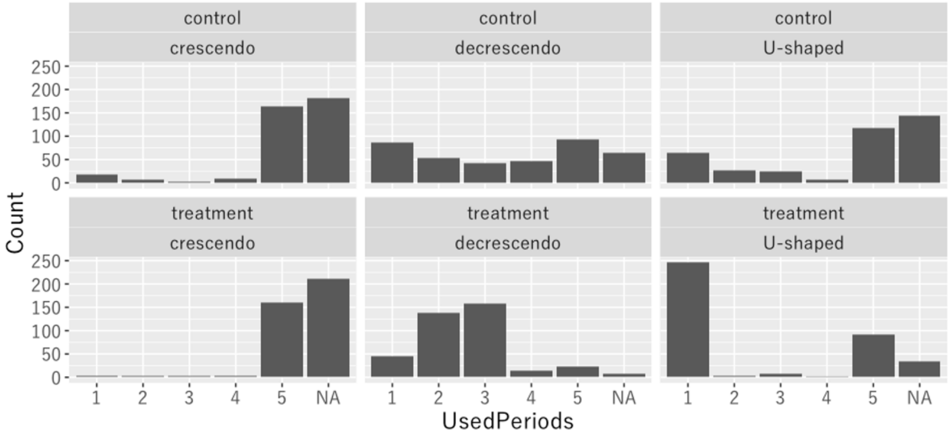


Fig. A.2: The histogram of used periods (NA represents that participants did not use service)

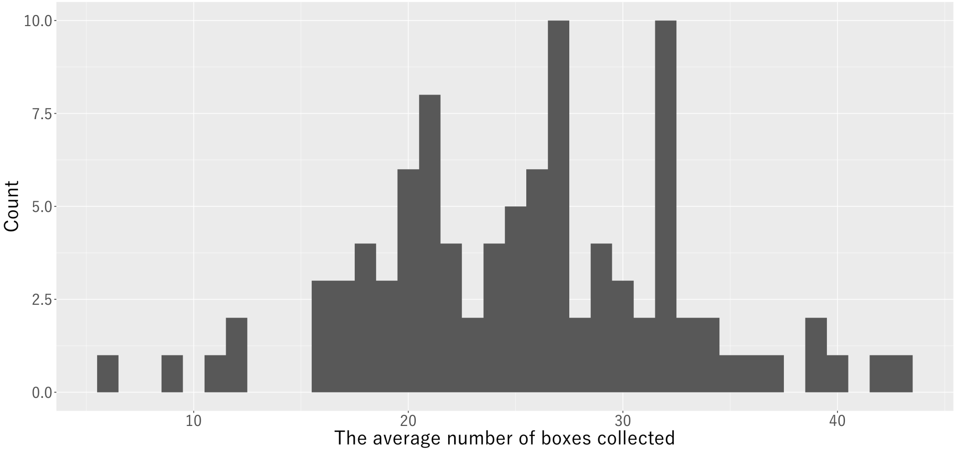


Fig. A.3: The histogram of each participant's average number of collected box on evaluation of risk attitude

A.5 The list of activity Levels in Chapter 4

In this section, all parameters of services in experiments of Chapter 4 are listed in Table A.3.

Tab. A.3: The actual settings of activity levels in the experiment.

Group	Experiment	Service	Pattern	Allocation of activity levels					
				t=1	t=2	t=3	t=4	t=5	total
1	control	1	crescendo	93	142	223	329	428	1215
1	control	2	U-shaped	265	131	83	238	378	1095
1	control	3	U-shaped	237	99	48	152	244	780
1	control	4	crescendo	60	190	261	320	444	1275
1	control	5	dummy	37	185	266	128	134	750
1	control	6	decrescendo	235	211	124	109	56	735
1	control	7	decrescendo	373	305	230	170	92	1170
1	control	8	U-shaped	314	133	63	267	423	1200
1	control	9	U-shaped	249	137	32	148	289	855
1	control	10	dummy	243	125	362	89	306	1125
1	control	11	decrescendo	271	198	162	76	73	780
1	control	12	dummy	171	133	64	274	228	870
1	control	13	decrescendo	369	253	211	129	88	1050
1	control	14	dummy	258	263	212	239	228	1200
1	control	15	crescendo	57	142	151	255	265	870
1	control	16	crescendo	60	96	153	238	308	855
1	treatment	1	dummy	177	127	27	237	212	780
1	treatment	2	U-shaped	205	121	58	140	256	780
1	treatment	3	decrescendo	370	336	243	138	98	1185
1	treatment	4	decrescendo	416	357	253	157	107	1290
1	treatment	5	dummy	58	262	272	141	152	885
1	treatment	6	dummy	222	201	211	255	251	1140
1	treatment	7	decrescendo	297	233	202	152	46	930
1	treatment	8	decrescendo	293	214	156	126	81	870
1	treatment	9	crescendo	63	94	162	180	281	780
1	treatment	10	U-shaped	262	97	48	197	311	915
1	treatment	11	U-shaped	351	139	84	238	433	1245
1	treatment	12	U-shaped	265	131	83	238	378	1095
1	treatment	13	crescendo	56	139	215	306	379	1095
1	treatment	14	crescendo	95	177	203	307	373	1155
1	treatment	15	crescendo	81	99	152	204	244	780

1	treatment	16	dummy	252	140	380	85	343	1200
2	control	1	decrescendo	281	231	131	81	56	780
2	control	2	dummy	162	141	141	169	182	795
2	control	3	decrescendo	405	322	278	168	87	1260
2	control	4	dummy	240	171	342	73	284	1110
2	control	5	U-shaped	227	115	57	197	319	915
2	control	6	crescendo	75	89	158	242	291	855
2	control	7	U-shaped	293	145	90	259	413	1200
2	control	8	crescendo	71	110	186	212	321	900
2	control	9	decrescendo	242	213	153	82	45	735
2	control	10	dummy	164	103	75	323	220	885
2	control	11	dummy	68	269	353	175	245	1110
2	control	12	U-shaped	304	126	53	224	388	1095
2	control	13	decrescendo	424	297	244	151	99	1215
2	control	14	U-shaped	199	126	27	173	225	750
2	control	15	crescendo	50	170	231	280	394	1125
2	control	16	crescendo	63	167	226	293	406	1155
2	treatment	1	U-shaped	350	189	55	248	403	1245
2	treatment	2	dummy	64	302	421	158	240	1185
2	treatment	3	U-shaped	300	124	52	221	383	1080
2	treatment	4	U-shaped	261	143	35	157	304	900
2	treatment	5	decrescendo	351	305	214	131	94	1095
2	treatment	6	U-shaped	226	89	41	201	313	870
2	treatment	7	crescendo	71	81	146	226	271	795
2	treatment	8	dummy	219	149	92	429	341	1230
2	treatment	9	crescendo	65	191	234	317	408	1215
2	treatment	10	decrescendo	233	166	120	102	69	690
2	treatment	11	crescendo	42	95	181	248	259	825
2	treatment	12	dummy	137	164	115	161	113	690
2	treatment	13	crescendo	93	121	223	285	388	1110
2	treatment	14	decrescendo	296	218	177	86	78	855
2	treatment	15	dummy	145	101	266	68	230	810
2	treatment	16	decrescendo	355	324	234	132	95	1140
3	control	1	decrescendo	432	331	256	163	108	1290
3	control	2	dummy	246	251	200	227	216	1140
3	control	3	U-shaped	303	115	72	202	373	1065
3	control	4	crescendo	32	134	177	208	304	855
3	control	5	dummy	218	185	69	383	315	1170
3	control	6	decrescendo	227	177	160	124	32	720
3	control	7	U-shaped	243	102	42	190	323	900

3	control	8	dummy	74	235	240	83	163	795
3	control	9	crescendo	66	100	171	192	296	825
3	control	10	dummy	159	83	258	82	243	825
3	control	11	U-shaped	218	75	37	164	256	750
3	control	12	U-shaped	270	155	66	198	376	1065
3	control	13	crescendo	60	136	251	300	378	1125
3	control	14	decrescendo	363	297	244	175	91	1170
3	control	15	crescendo	98	152	238	349	453	1290
3	control	16	decrescendo	248	205	155	120	67	795
3	treatment	1	crescendo	57	90	144	226	293	810
3	treatment	2	decrescendo	417	304	223	167	104	1215
3	treatment	3	crescendo	60	181	219	297	383	1140
3	treatment	4	decrescendo	271	223	125	77	54	750
3	treatment	5	crescendo	97	181	209	315	383	1185
3	treatment	6	dummy	80	299	388	147	256	1170
3	treatment	7	U-shaped	265	126	96	210	353	1050
3	treatment	8	crescendo	29	85	134	198	244	690
3	treatment	9	dummy	195	93	282	73	242	885
3	treatment	10	U-shaped	199	84	53	113	256	705
3	treatment	11	dummy	147	126	136	180	176	765
3	treatment	12	U-shaped	243	85	57	157	298	840
3	treatment	13	decrescendo	260	248	177	94	76	855
3	treatment	14	U-shaped	330	185	81	243	451	1290
3	treatment	15	dummy	198	135	85	394	313	1125
3	treatment	16	decrescendo	399	353	221	182	75	1230
4	control	1	U-shaped	227	94	38	178	303	840
4	control	2	U-shaped	340	144	62	251	433	1230
4	control	3	U-shaped	284	179	66	218	378	1125
4	control	4	dummy	183	131	29	247	220	810
4	control	5	crescendo	95	177	203	307	373	1155
4	control	6	dummy	78	291	378	143	250	1140
4	control	7	crescendo	61	138	254	304	383	1140
4	control	8	crescendo	32	91	143	210	259	735
4	control	9	decrescendo	308	207	161	130	64	870
4	control	10	crescendo	60	88	153	168	266	735
4	control	11	decrescendo	244	173	133	127	73	750
4	control	12	U-shaped	267	97	63	175	328	930
4	control	13	dummy	149	194	180	154	193	870
4	control	14	decrescendo	391	337	238	147	102	1215
4	control	15	dummy	195	133	380	93	309	1110

4	control	16	decrescendo	419	369	233	190	79	1290
4	treatment	1	decrescendo	288	210	153	124	80	855
4	treatment	2	U-shaped	324	136	58	239	413	1170
4	treatment	3	decrescendo	374	341	225	149	81	1170
4	treatment	4	U-shaped	229	127	27	133	264	780
4	treatment	5	crescendo	97	181	209	315	383	1185
4	treatment	6	dummy	236	143	388	69	334	1170
4	treatment	7	U-shaped	273	130	98	216	363	1080
4	treatment	8	decrescendo	243	201	152	118	66	780
4	treatment	9	crescendo	74	87	155	238	286	840
4	treatment	10	dummy	59	182	246	68	165	720
4	treatment	11	U-shaped	218	85	39	195	303	840
4	treatment	12	dummy	196	91	60	284	254	885
4	treatment	13	crescendo	34	138	183	216	314	885
4	treatment	14	decrescendo	401	345	244	151	104	1245
4	treatment	15	dummy	192	204	229	238	232	1095
4	treatment	16	crescendo	69	185	215	338	393	1200
5	control	1	crescendo	77	137	223	291	397	1125
5	control	2	crescendo	35	109	196	267	323	930
5	control	3	dummy	240	243	236	267	229	1215
5	control	4	U-shaped	329	158	112	258	433	1290
5	control	5	decrescendo	399	353	221	182	75	1230
5	control	6	dummy	28	232	271	86	178	795
5	control	7	decrescendo	284	185	160	95	71	795
5	control	8	decrescendo	263	190	138	114	75	780
5	control	9	U-shaped	190	71	32	174	268	735
5	control	10	U-shaped	288	181	67	221	383	1140
5	control	11	U-shaped	242	87	43	182	286	840
5	control	12	dummy	154	107	281	71	242	855
5	control	13	decrescendo	387	280	205	155	98	1125
5	control	14	crescendo	72	125	128	181	259	765
5	control	15	crescendo	93	173	197	299	363	1125
5	control	16	dummy	217	167	72	370	329	1155
5	treatment	1	crescendo	75	133	217	283	387	1095
5	treatment	2	decrescendo	253	209	178	131	69	840
5	treatment	3	dummy	77	320	393	184	226	1200
5	treatment	4	U-shaped	325	156	111	255	428	1275
5	treatment	5	U-shaped	293	177	51	246	373	1140
5	treatment	6	U-shaped	199	132	62	148	299	840
5	treatment	7	crescendo	73	81	163	205	288	810

5	treatment	8	decrescendo	278	202	147	120	78	825
5	treatment	9	dummy	135	67	218	74	211	705
5	treatment	10	decrescendo	354	325	213	141	77	1110
5	treatment	11	crescendo	73	85	152	234	281	825
5	treatment	12	dummy	166	115	57	319	258	915
5	treatment	13	decrescendo	382	276	202	153	97	1110
5	treatment	14	U-shaped	217	121	24	124	249	735
5	treatment	15	dummy	243	248	197	224	213	1125
5	treatment	16	crescendo	68	177	241	313	431	1230
6	control	1	dummy	86	276	350	153	245	1110
6	control	2	crescendo	79	164	239	352	396	1230
6	control	3	decrescendo	371	321	226	139	98	1155
6	control	4	U-shaped	364	147	105	227	417	1260
6	control	5	decrescendo	263	217	164	126	70	840
6	control	6	dummy	171	91	278	86	259	885
6	control	7	dummy	181	81	55	259	234	810
6	control	8	U-shaped	246	137	29	170	273	855
6	control	9	U-shaped	207	84	33	163	278	765
6	control	10	decrescendo	220	216	153	78	68	735
6	control	11	crescendo	31	101	184	251	303	870
6	control	12	crescendo	67	150	272	328	413	1230
6	control	13	U-shaped	312	193	73	239	413	1230
6	control	14	dummy	227	212	223	216	247	1125
6	control	15	decrescendo	374	333	206	172	70	1155
6	control	16	crescendo	73	85	152	234	281	825
6	treatment	1	decrescendo	223	185	160	119	63	750
6	treatment	2	U-shaped	285	173	49	240	363	1110
6	treatment	3	decrescendo	355	282	248	148	77	1110
6	treatment	4	dummy	169	123	151	139	168	750
6	treatment	5	U-shaped	226	79	39	170	266	780
6	treatment	6	crescendo	61	183	222	301	388	1155
6	treatment	7	decrescendo	428	299	261	135	77	1200
6	treatment	8	decrescendo	255	244	174	92	75	840
6	treatment	9	dummy	37	209	266	122	176	810
6	treatment	10	U-shaped	282	161	69	207	391	1110
6	treatment	11	crescendo	52	89	157	259	313	870
6	treatment	12	dummy	260	159	428	77	366	1290
6	treatment	13	U-shaped	254	141	31	176	283	885
6	treatment	14	crescendo	73	152	221	328	366	1140
6	treatment	15	crescendo	68	92	148	229	303	840

6	treatment	16	dummy	270	148	88	436	318	1260