Assignment 2

Conducting Experiment

Group 1

Thy Cao 673415dc

Dan Gong 727292rg

Felix Masselter 735890fm

Baichuan Ji 717660bj

Long Lê 735090al

Course: BIM Research Methods

MSc Business Information Management

Rotterdam School of Management, Erasmus University

Table of Contents

Exercise 1

1. Research Question

How do space norms affect the behavior of writing quality reviews on online consumer review sites?

2. Biases

Self-selection bias: the reason why individuals who choose to write reviews may be systematically different. For instance, those with extreme experiences, so very positive or very negative, are more likely to write a review. It could also be the case that more knowledgeable reviewers may write more detailed reviews while the less knowledgeable reviewers may be reluctant to post a review in the first place, ultimately creating a non-representative sample.

Omitted variable bias: unable to control for all factors affecting the behavior of writing a quality review. For instance, characteristics on a product level such as product quality or product defects can create an omitted variable bias.

Reverse causality: it can be difficult to determine if *Space Norms* influence *Writing Quality Review*, it could be the other way around that *Writing Quality Review* impacts *Space Norms*.

Simultaneity: *Space Norms* could influence a new reviewer's *Writing Quality Review* but at the same time a reviewer's *Writing Quality Review* could shape the *Space Norms* that influence future reviewers which creates a feedback loop, ultimately creating simultaneity.

3. Independent Constructs & Dependent Construct

The key constructs in this article examine how social behavioral norms in digital spaces influence user content generation behavior.

The independent construct focuses on space norms which can be defined as the social expectations and standards of behavior that emerge from observing others' actions within a specific digital environment. Space norms differ from traditional social norms by being specific to and embedded within the digital space

The dependent construct centers on user-generated content quality which encompasses the characteristics and attributes of content produced by users in these digital spaces.

The theoretical framework also incorporates *Social Presence* as a moderating construct, which represents the psychological experience of being with others in a mediated environment. It can act as a psychological mechanism that can strengthen or weaken normative influences. Together, these constructs form a theoretical model for understanding how social dynamics and environmental cues in digital spaces shape user behavior and content creation.

4. Measurements & Weaknesses

The researchers measured their independent construct (*Space Norm*) through 2 manipulated conditions (high-quality, low-quality) and 1 baseline condition (control) that established perceived review standards. In the high-quality norm, participants were shown 4 high-quality treatment reviews whereas in the low-quality norm, they were shown 4 low-quality treatment reviews. In the control condition, no review was shown.

The moderator (*Social Presence*) was measured using a Likert-type scale adapted from Kumar and Benbasat (2006), with reliability and factor analyses confirming the validity of this measure. Participants rated each item on a Likert scale (1-7). These responses were then averaged to create an overall *Social Presence* score for each participant.

The dependent construct (*Quality Review Writing Behavior*) was measured using 2 different approaches. The first approach was Textual Analysis Composite Measure (Quality 1), which provided an objective measurement of review quality using automated text analysis and evaluated by machine learning. This composite measure includes four dimensions: length, objectivity, readability, and number of relevant points. Length refers to the word and sentence count, indicating the amount of detail provided. Objectivity assesses the factual content of the review by comparing it to product descriptions. Readability is calculated using indices like the Flesch-Kincaid score, measuring how easy the text is to understand. Finally, the number of relevant points captures the distinct product-related features discussed, indicating informativeness. The second approach was the Human-Coded Composite Measure (Quality 2), used to evaluate the quality of reviews based on the judgment of 2 human coders. This composite measure includes four dimensions: relevance, thoughtfulness, informativeness, and helpfulness. Relevance assesses how closely the review content pertains to the product, while thoughtfulness reflects the care and effort invested in writing. Informativeness measures the amount of useful information provided, and helpfulness gauges the review's value for other consumers. Each dimension is rated on a Likert scale (1-7), and the scores are averaged to create an overall quality score.

The authors ensured independent construct validity by pre-testing the high-quality and low-quality treatment reviews to confirm their respective quality levels through human raters. Raters assessed the reviews on relevance, thoughtfulness, informativeness, and helpfulness, with strong inter-rater reliability (α > 0.88). Regarding the moderator, they used a well-established scale (adapted from Kumar and Benbasat, 2006) and confirmed its reliability and validity through factor analysis and high internal consistency (Cronbach's α = 0.89). For the dependent construct, 2 approaches of measurement (Textual Analysis and Human-Coded) used in the study enhance validity by capturing both objective and subjective dimensions of *Writing Quality Review Behavior*.

However, the measurements used in this study have some weaknesses. First, Human-Coded ratings could suffer from subjective biases or limited generalizability due to the specific rater pool used in the study. Second, the moderator was measured but not manipulated, making the study not definitively establish a causal relationship between *Social Presence* and its moderating effects on *Space Norms* and *Review Writing Behavior*. Finally, as the Textual Analysis used machine learning, the training data might be biased because all participants were students, affecting the final results.

5. Visualization

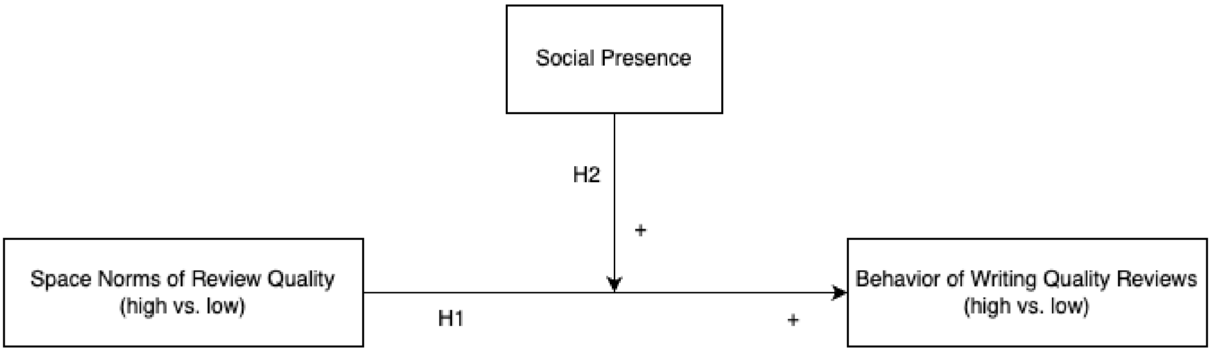

Social Presence
H2
+
Space Norms of Review Quality
(high vs. low)
H1
+
Behavior of Writing Quality Reviews
(high vs. low)

Figure 1. Hypotheses from the Study of Hou & Ma (2022)

The diagram illustrates how *Social Presence* moderates the relationship between *Space Norms of Review Quality* (high vs. low) and the *Behavior of Writing Quality Reviews* (high vs. low). A low norm refers to an environment with poor-quality behaviors, reflecting low expectations for effort or quality, whereas a high norm signifies an environment with high-quality behaviors, reflecting strong expectations for effort and quality. Correspondingly, the behavior of writing low-quality reviews involves minimal effort and poor-quality content while writing high-quality reviews reflects significant effort and well-crafted, informative content.

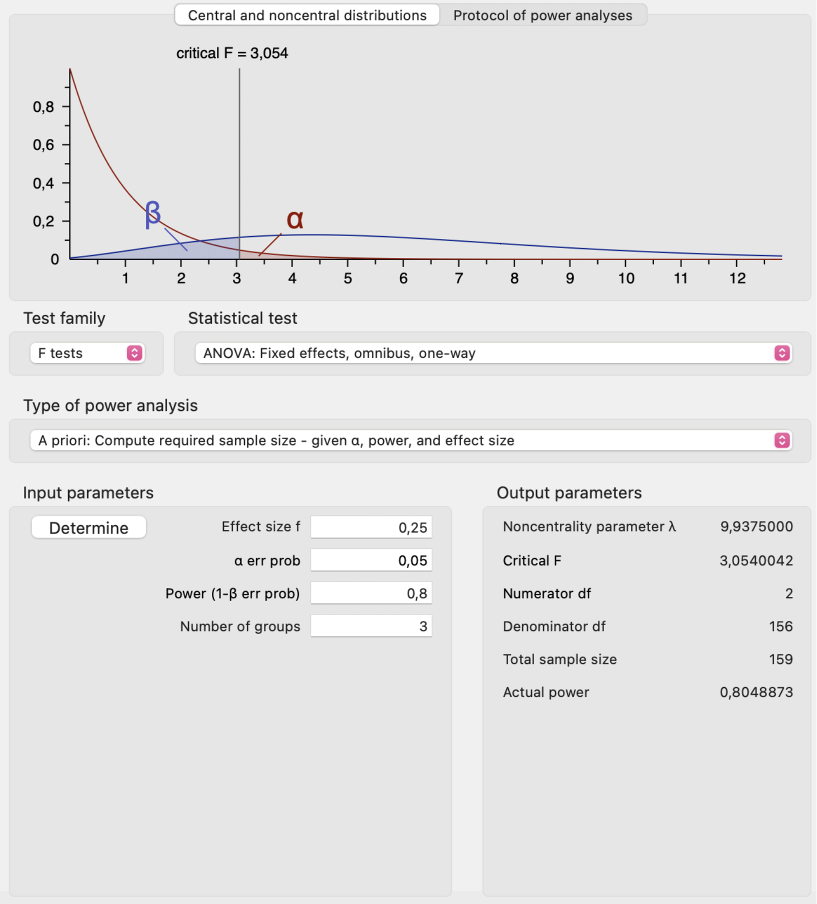
In this model, *Space Norms* serve as an independent variable, representing the quality expectations set within an online space. *Behavior of Writing Quality Reviews* is the dependent variable, reflecting the quality of reviews participants produce.

The arrows with plus signs indicate that *Space Norms* positively influence *Review Writing Behavior*. However, this effect is strengthened by the level of *Social Presence*. Thus, *Social Presence* enhances the impact of Space Norms on *Writing Quality Review Behavior*.

6. Sample Size

The authors did not provide specific details on how the sample size was determined, which is a limitation of this study. To address this, we conducted a power analysis to estimate an appropriate sample size based on the study design. Given that the study includes three groups (two treatment groups and one control group), we assumed a medium effect size of 0.25, measured by Cohen's f (Kotrl, 2003). We set the significance level α at 0.05 to control for Type I error, and chose a power level of 0.80, meaning there's an 80% probability of detecting an effect if it exists.

Using G*Power, we selected the F-test family and the ANOVA: Fixed effects, omnibus, one-way statistical test, as shown in Figure 1. Based on these parameters, our analysis recommended a sample size of 159 participants.


Central and noncentral distributions
Protocol of power analyses
critical F = 3,054
0,8
0,6
0,4
0,2
0
β
α
1
2
3
4
5
6
7
8
9
10
11
12
Determine

Test family: F tests

Statistical test: ANOVA: Fixed effects, omnibus, one-way

Type of power analysis: A priori: Compute required sample size - given α, power, and effect size

Input parameters

Effect size f 0,25

α err prob 0,05

Power (1-β err prob) 0,8

Number of groups 3

Output parameters

Noncentrality parameter λ 9,9375000

Critical F 3,0540042

Numerator df 2

Denominator df 156

Total sample size 159

Actual power 0,8048873

Figure 2. Calculating Sample Size with G*Power

The study of Hou & MA (2022)'s sample size of 168 is slightly larger than our recommendation, which suggests the study is sufficiently powered to detect a medium effect size. This additional sample size could slightly increase the power, potentially providing a more robust test of the hypotheses. However, if the authors had conducted a similar power analysis, they could have transparently justified their sample size decision.

7. Incentive Scheme & Downsides

Participants were incentivized by course credits. However, this approach of using only course credit incentives has several downsides. First, there's a possibility that participants view the experiment as coercive (Miller & Kreiner, 2008), which might affect the behavior of writing reviews. Second, relying on course credit restricts the participant pool to students who are highly educated, reducing the diversity of the sample and possibly biasing results toward student demographics (Sears, 1986).

8. Randomization, Manipulation & Attention Check

The authors conducted a manipulation check by requiring participants to complete a questionnaire after they submitted the reviews. The questionnaire included ratings of perceived *Social Presence* using a validated scale, and questions assessing how relevant, thoughtful, informative, and helpful they believed others had been in writing reviews. If participants in different treatment conditions rated *Social Presence* and *Review Quality* (relevance, thoughtfulness, informativeness, helpfulness) differently, the manipulations worked as intended. This is the case in Study 1 as the high Cronbach’s $\alpha = 0.91$ indicated excellent internal consistency (UCLA Institute for Digital Research and Education, 2024).

Meanwhile, the randomization check was also conducted after they completed the questionnaire by asking them to report age, gender, and past review-writing experience. If there were no significant differences in demographics or past review experience between groups, it confirms successful randomization. If significant differences were found, it would suggest a flaw in the randomization process, potentially introducing confounds. A randomization check confirmed that participants' review-writing experience and demographics did not influence the results

Regarding the attention check, participants were asked to answer some dummy questions about the products before being debriefed and dismissed (Hou & Ma, 2022). Based on the results, they could know whether participants spent a reasonable amount of time reading the instructions and reviews. Participants who spent an unusually short or long time might be flagged as inattentive.

9. External Validity

9.1. Study 1

This study used a randomized lab experiment with participants who were mostly students, resulting in a low external validity. The controlled environment and the lack of population generalization might not accurately reflect real-world behaviors.

Attempts to ensure external validity: Firstly, the study used a fictional website modelled after Amazon.com, which creates a realistic e-commerce interface, to control other extraneous variables. Secondly, the study used common consumer goods (calculator and headphones) that the participants were familiar with to reduce confounding variables. Thirdly, by incorporating both search-type (evaluated based on attributes) and experience-type (evaluated after use) products, the findings can generalize across different product categories. Lastly, the study tried to imitate a structured experience similar to typical product use when buying a product in real life. The participants were allowed to test the product, read the product manual and experience the product by doing math problems or listening to a song clip.

9.2. Study 2

This study consisted of an observation field study which was conducted by collecting review data from e-commerce sites Philips.com and Aritzia.com, making it high in external validity. Compared to study 1, this creates a real-world setting where one can observe the natural review-writing behavior and where customers have authentic product experiences which consequently increases external validity. Furthermore, these two sites were chosen to maximize the variation in product types (59 subcategories), demographics and price levels. This data diversity enhances the external validity by making the findings more generalizable.

One the other hand, one could argue that there are still some limitations to the external validity as the researchers only tested the writing behavior of the 4th, 5th and 6th reviewer for each product but later reviews for example the behavior of the 150th reviewer could indicate a different pattern. Lastly, they used Philips and Aritzia which are not considered general marketplaces such as Amazon. To further strengthen the external validity the study could have included a general marketplace.

10. Suggested Improvements for the Study

According to the authors, the *Space Norm* theory should not only apply to product reviews but also service reviews which was not tested in this study. To alleviate this potential gap, one could conduct a parallel experiment focused on service reviews to test the generalizability of *Space Norm* effects beyond product reviews. Using hotel stays as the context, the researchers could maintain the same experimental design (high vs. low vs control treatment reviews) but adapt the review content to focus on service attributes like room cleanliness and staff responsiveness. The quality metrics (length, readability, informativeness, etc.) would remain consistent, but their application would be validated specifically for service contexts. Participants would read reviews and write their own reviews after experiencing a standardized hotel service scenario through detailed descriptions or video content. This would directly test whether *Space Norms* influence review writing behavior similarly for services as they do for products.

Another key limitation of the study is that *Social Presence* serves as a moderator variable but was only measured post-hoc rather than being manipulated. While the study appropriately tests the moderating effect of *Social Presence* on the relationship between *Space Norms* and *Writing Quality Review*, manipulating different levels of *Social Presence* (High vs. Low) while maintaining its moderator role would strengthen the findings. This could be implemented through an experimental design where *Space Norms* (High Quality vs. Low Quality vs. Control) remain the main independent variable, but *Social Presence* cues are systematically varied to create different levels of the moderator. For instance, interface elements like user presence indicators and social cues would allow for a more controlled test of how different levels of *Social Presence* moderate the space norm effect, while maintaining the theoretical framework of the original study.

Lastly, a pre-study design could be proposed to aim to calibrate the treatment size. The researchers would run a pre-study where they create multiple experimental conditions varying the number of treatment reviews. So instead of having just 4 in our current study, a number the researchers came up with through theoretical assumptions, we would create a condition for 2, 4, 6, 8 and reviews. Through this pre-study, we use attention metrics to measure whether attention diminishes with more reviews and, most importantly, we can measure the strength of norm perception. Thanks to these measurements we would be able to identify the inflection point where additional reviews no longer strengthen norm perception. This would make the choice of treatment size an empirical-based decision and create a stronger methodological foundation for the main study.

Exercise 2

1. Experiment

1.1. Hypotheses

H1: A lower level of product functional condition leads to a higher likelihood of reviewing from customers
H2: The negative effect of product condition on the likelihood of reviewing is mediated by customer satisfaction

1.2. Theoretical Model

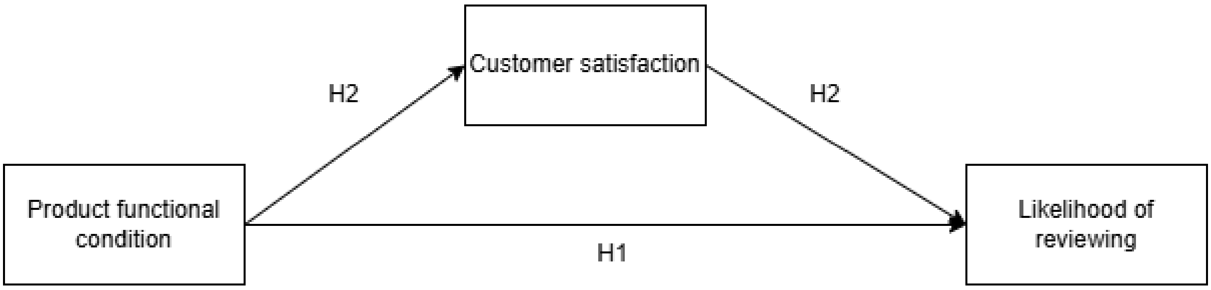


Figure 3. Hypotheses

1.3. Experiment Design

Incentive scheme: monetary incentive ($20.00 at the end of the study)

Experimental setting:
We will carry out a randomized online experiment. To address the research question, we choose to simulate an experience similar to that when customers carry out online shopping on large e-commerce platforms (e.g. Amazon, Aliexpress, etc.). Compared to other shopping scenarios, such as buying goods at a physical store, we believe this can most effectively help us answer the research questions because customers are more used to online review systems.

We will create a webpage that simulates the product listing page of an e-commerce site. Customers can view different options for a product on this page. Subsequently, they can choose an option and go to the product page, where they can view information such as product descriptions, prices, etc. and they can perform actions like adding to a cart and checking out.

Participants' tasks and experiment treatment:

Participants will be asked to use the webpage and choose a product available on the page. The product we choose for the experiment is a desk lamp. This product is suitable for this experiment as it is a simple and generic product that is familiar to most people, and thus they can evaluate its functionality. On this page, they will have 2 options for the same type of product (i.e. they can choose between desk lamp A and desk lamp B), stimulating a scenario with multiple sellers. The 2 options will provide the exact same functionalities, only differences in aesthetic features to simulate varying product choices.

We will have 3 versions of the product based on usability levels: Completely Malfunctioning, Partially Malfunctioning, and Fully Functional. Conceptually, "Completely Malfunctioning" and "Partially Malfunctioning" will be our treatment groups, and "Fully Functional" will be our control group. A summary description of each version is as below:

Completely Malfunctioning	**Partially Malfunctioning**	**Fully Functional**
The lamp cannot emit light	The lamp can emit unstable, blinking, dim light	The lamp can emit stable, bright light

This stratification helps to better understand how customers would react to different levels of product condition and to examine the mediation effect. For example, customers with partially malfunctioning products might still be satisfied with their product and are not as motivated to leave a review as customers with a completely broken product. Note that the 3 versions differ only in their usability, and are completely similar in size, shape and other features.

The experiment will be split into 2 phases: the product experience phase and the product review phase. The participants will receive their monetary compensation **upon completion of both phases.** In the product experience phase, at the beginning of the experiment, participants will receive an email that has 2 items: a survey containing questions related to control variables, and a link to access the webpage to choose between 2 options of desk lamp. Participants click on the product they want to choose to go to the product page, where they can view product information and then choose the product. After that, we will randomly send them 1 of the 3 versions of that option. After receiving the product, they will have 2 days to experience the product. "Experience the product" means that they have to at least use the product's main functionality (in this case, plug in the lamp and turn on the light). To ensure this, we will carry out an attention check. For the second phase, after 2 days of experience, we will send the participants an email to inform them of the end of the first phase, ask them to rate the product and **provide them with an option to leave a review for the product**. Specifically, the email will contain a simple form with 2 questions:

1. How was your experience with the product (required)? (Participants rate on a Likert scale of 1-7)
2. Would you like to leave a review about the product (optional)? (A text box will be provided alongside the question so that participants can directly leave a review if they want to)

Explanation for the question choices:

Question 1: a required question, serves as a signal that the participants have completed both phases. Additionally, the answers for this question will be used to analyze customer satisfaction in our mediation effect hypothesis.

Question 2: an optional question, answers to this question will be our input data to measure likelihood of review. Note that when we indicate the question as **optional**, there's a chance that participants will choose to not leave a review (minimal effort to complete the experiment), even though naturally they want to. Therefore, we will also communicate that "Your review will be valuable for us and other customers". This is to reduce the effect of participants only making minimal effort to complete the task.

Variables and measurement:

The likelihood of reviewing will be calculated as the proportion of participants who have given a text review. Additionally, we include other control variables: demographic (age, gender, education), and review experience. These control variables are used for checking randomization and ensuring internal validity.

2. Sample Population

In this experiment, we recruit participants from the online platform Prolific.

We employ a power analysis to calculate an appropriate sample size. We hold two treatment groups and one control group, so the F-test family and the ANOVA are more suitable to apply here. Assuming a medium effect size of $f = 0.25$, $\alpha = 0.05$, and $\beta = 0.2$, we determined that a total sample size of 159 is needed to observe a significant difference.

3. Attention & Manipulation Check

On the first day after receiving the product, participants will receive an email with questions designed to check their engagement and validate the experimental manipulations.

An attention check confirms that participants genuinely interacted with the product. In the email, there will be dummy questions that can only be answered accurately by those who used it. For example, *"What is the color of the product's light during operation?"* and the answers are presented as multiple-choice options: blue, green, red, yellow, and N/A. Participants providing inconsistent or irrelevant answers will be flagged as irrelevant and excluded from the analysis.

Then in the next question, we will ask them to assess the product's condition. As a manipulation check, the main question will be "*In your opinion, what is the condition of the product?*" and the answers are presented as multiple-choice options: Completely Malfunctioning, Partially Malfunctioning, Fully Functional.

4. Internal & External Validity

4.1. External Validity

Because the question does not explicitly specify online or offline shopping (customers can review online for both types of shopping), we choose to replicate an online shopping experience as customers of this group are more familiar with online reviews, making it more generalizable.

We try to replicate as close as possible the experience to that in the real world. We build a website for online shopping. The website has a product listing page that displays options. In the real world, this is the page where users, after searching for a specific product, have different options to choose from. This simulates the scenario in which customers can choose from many sellers for a specific product. We also simulate a product page, where customers can read product descriptions for the product they are choosing. We choose a generic and familiar product (desk lamp) that customers can easily buy online.

To further ensure external validity, we would carry out a quasi-experimental field experiment, using the Difference-in-Difference method. We rule out the possibility of conducting a randomized field experiment, as it would be infeasible and unethical to assign a real customer to buy a malfunctioning product. We collaborate with a company (e.g. Philips) to collect historical data on customer reviews of a product. It is common when a company has a batch of products with some malfunctioning features that were not detected during production, and this batch was sold on the market. We would compare the data of customers' reviews from this malfunctioning batch and a good batch during the same period.

4.2. Internal Validity

Participants will be instructed and communicated clearly about the purpose of the experiment to avoid the threat of protocol adherence (Slack & Draugalis, 2001). We also make sure that in the product review phase, participants will not be confused that leaving a text review is mandatory for completing the task.

We use randomization to prevent inconsistencies between the groups within the study (Dunbar-Jacob, 2012). We randomly assign each version of the product to the participants. Additionally, we include control variables demographics (age, gender, education) and review experience for randomization check. Review experience is an important variable, as participants might be more inclined to write a review simply because they are more used to it.

Blinding of treatment will also be used to prevent participants from being aware of their treatment group assignment (Dunbar-Jacob, 2012). The participants will be told their participation is for understanding customer product adoption and contributing to product improvement. We have different treatment groups as different product versions corresponding to different malfunction levels. This helps us isolate the effect of the independent variable and better understand the mechanism.

We ensure that our treatment groups and control group are clearly different by carrying out manipulation checks. We ensure the control of extraneous factors by carefully monitoring the experiment on our own website and limiting the product type to only 1 product (desk lamp), which makes the result consistent and comparable. We also carry out a pretest to make sure our randomization and manipulation work.

5. Result analysis

5.1. Main Results

Hypothesis 1 Test:

Logistic regression will be employed to answer the research question. Categorical data of *Product Condition* collected from the experiment and binary data of *Likelihood of Reviewing* (review or no review) will be used in the logistic regression model, which directly models binary outcomes and provides interpretable coefficients (odds ratios).

Hypothesis 2 Test:

To test this hypothesis, we employ Bootstrap to measure the mediation effect of *Customer Satisfaction* on the relationship between *Product Condition* and *Likelihood of Reviewing*. Categorical data of *Customer Satisfaction* (1-7) and *Product Condition* together with binary data of *Likelihood of Reviewing* will be used in the model. Bootstrap does not impose the restrictive assumptions of normality and linearity inherent in traditional mediation analysis methods, making it a suitable option for this one.

5.2. Randomization Check

ANOVA will be used to test whether the treatment groups differ significantly from the control group while controlling for variables like gender, age, review experience, etc. Most of them are categorical data.

5.3. Manipulation Check

ANOVA will also be used to verify the manipulation effectiveness. Data collected from manipulation checks are categorical and mean differences between groups will be analyzed.

References

Dunbar-Jacob, J. (2012). Minimizing threats to internal validity. *Intervention research: Designing, conducting, analysing, and funding,* 91–106.

Hou, J., & Ma, X. (2022). Space norms for constructing quality reviews on online consumer review sites. *Information Systems Research, 33*(3), 1093–1112.

Kotrl, J. W. (2003). The incorporation of effect size in information technology, learning, and performance research. Retrieved from https://www.semanticscholar.org/paper/The-Incorporation-of-Effect-Size-in-Information-%2C-%2C-Kotrl/e5c3504ca4baef11c1cda8ec085833dbccb63259

Miller, W. E., & Kreiner, D. S. (2008). Student perception of coercion to participate in psychological research. *North American Journal of Psychology, 10*(1), 53. Retrieved from https://link-gale-com.eur.idm.oclc.org/apps/doc/A178452279/AONE?u=erasmus&sid=oclc&xid=232df121

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology, 51*(3), 515.

Slack, M. K., & Draugalis Jr, J. R. (2001). Establishing the internal and external validity of experimental studies. *American Journal of Health-System Pharmacy, 58*(22), 2173–2181.

UCLA Institute for Digital Research and Education. (2024). What does Cronbach's alpha mean? *UCLA Statistical Consulting Group.* Retrieved November 12, 2024, from https://stats.oarc.ucla.edu/spss/faq/what-does-cronbachs-alpha-mean/