6-18-2022

TAG ME IF YOU CAN – (HOW) SHOULD PLATFORMS TAG FAKE REVIEWS AND FAKE USERS?

Stefanie Erlebach
Ulm University, stefanie.erlebach@uni-ulm.de

Alexander Kupfer
University of Innsbruck, alexander.kupfer@uibk.ac.at

Andrea Wrabel
Ulm University, andrea.wrabel@uni-ulm.de

Steffen Zimmermann
Ulm University, steffen.zimmermann@uni-ulm.de

TAG ME IF YOU CAN – (HOW) SHOULD PLATFORMS TAG FAKE REVIEWS AND FAKE USERS?

Research in Progress

Stefanie Erlebach, Ulm University, Ulm, Germany, stefanie.erlebach@uni-ulm.de

Alexander Kupfer, University of Innsbruck, Innsbruck, Austria, alexander.kupfer@uibk.ac.at

Andrea Wrabel, Ulm University, Ulm, Germany, andrea.wrabel@uni-ulm.de

Steffen Zimmermann, Ulm University, Ulm, Germany, steffen.zimmermann@uni-ulm.de

Abstract

Although fake online consumer reviews (OCRs) and fake users are an increasing problem for online review systems, there is no consensus on how digital platforms should handle them after their detection. Therefore, platforms use different displaying strategies for detected fake OCRs and fake users, such as "doing nothing", censoring, or tagging them. It is, however, still unclear how these different strategies affect trust in multiple dimensions (i.e., trust in OCRs, reviewers and platform) and willingness to pay of consumers. We therefore propose in this research in progress paper an incentive-compatible experimental design for examining how different displaying strategies for fake OCRs and fake users influence consumers' trust dimensions and willingness to pay. By conducting the proposed experiment, we expect to provide relevant insights for researchers and practitioners on how platforms should display fake OCRs and fake users to mitigate negative implications on consumers, producers and platforms.

Keywords: Fake Reviews, Fake Users, Trust, Willingness to Pay.

1 Introduction

Online consumer reviews (OCRs) represent a trusted source of peer-generated information to inform consumers' purchase decisions (Ba and Pavlou, 2002; Zhang et al., 2017; Hu et al., 2011). In recent years, however, OCR manipulation became a phenomenon (Lappas et al., 2016) affecting almost all digital platforms. A recent report classified 7.1% of all analyzed OCRs on Yelp and 10.7% on Google as suspicious (Uberall, 2021). Learning about the consequences of fake OCRs[1] is essential as they affect all relevant stakeholders of digital platforms. Over two third of consumers, for instance, say they generally do not trust[2] OCRs because of the existence of fake OCRs and fake users (Pitman, 2022). For platforms and producers, a recent industry report suggests that e-commerce revenue in the United States would be more than $10 billion higher without malicious content (CHEQ, 2021).

As a consequence, researchers and practitioners have made serious efforts to prevent the dissemination of malicious content (Wu et al., 2020) by continuously developing algorithms to detect fake OCRs (e.g., Kumar et al., 2018, Shan et al. 2021) and fake users (e.g., Mukherjee et al., 2012; Akoglu et al., 2013). However, little is known about how digital platforms should display fake OCRs and fake users *after* detecting them. While indicating malicious content as fake increases trust in the platform itself

[1] We define fake OCRs according to Ansari and Gupta (2021) as "[deliberately manipulated] online reviews [...] to deceive customers by knowingly fostering incorrect information and inducing an action that the customer would unlikely take without the manipulation" (p. 100). Further and for the sake of this study, we define a fake user as the author of a fake OCR.

[2] For the sake of this study, we define trust according to Mayer et al. (1995) as the "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" (p. 712).

(Ananthakrishnan et al., 2020), it is unknown how trust in OCRs and trust in reviewers is affected. Hence, even if trust in platform can be increased by indicating the malicious content, it is yet impossible to conclude the actual effect on consumer decisions and economic outcomes without simultaneously knowing the effects on trust in OCRs and trust in reviewers.

These ambiguities are also reflected in the different ways platforms deal with detected fake OCRs or, more precisely, how they display them to consumers: While some platforms do not disclose their strategy for dealing with malicious content (e.g., Best Buy), others inform consumers about censoring fake OCRs (e.g., Amazon). Again others display fake OCRs publicly but tag them as fake, for instance with a note that they are "not recommended" (e.g., Yelp). Given this lack of understanding on how to deal with malicious content on digital platforms, our study addresses the following research question:

RQ: *What is the impact of tagging fake OCRs and fake users on consumers' trust dimensions and willingness to pay?*

To answer this research question, we plan to conduct an incentive-compatible laboratory experiment. To capture the overall economic effects from (potentially opposing) effects on different trust dimensions, we use willingness to pay (WTP) which represents a proxy for demand of a reviewed good in an experimental setting (Brynjolfsson et al., 2019). In the experiment, participants have to solve a puzzle for which they can additionally purchase a tutorial video as support. Participants can learn about the quality of the video by reading OCRs. We vary across treatments (i) the valence of fake content and (ii) the way fake content is handled. For the latter, we either tag OCRs as fake, tag users as fake, tag both OCRs and users as fake or simply censor (i.e., delete) the manipulated OCRs. This experimental setting allows us to compare the effects of different forms of displaying on the three trust dimensions (i.e., trust in OCRs, trust in reviewers, trust in platform) as well as on WTP. Since the video can be valuable for the participants (i.e., support in solving the puzzle), we ask participants to state their WTP according to the incentive-compatible Becker-DeGroot-Marschak (BDM) method (Becker et al., 1964).

We expect the results of our study to provide relevant theoretical contributions on existing research as follows: By considering trust from a holistic perspective, we examine different trust dimensions relevant for the consideration of fake OCRs and fake users and capture the overall economic effects. Further, we aim at a better understanding of the role of fake OCR valence on consumers' trust dimensions. Finally, and to the best of our knowledge, we will be the first to examine consumers' WTP by applying the incentive-compatible BDM method in the context of fake OCRs and fake users. In addition, by examining WTP, we expect our results to provide platforms with important insights on how to deal with fake OCRs and fake users in order to avoid negative effects on their economic outcomes.

2 Related Literature

While an extensive stream of research on OCR manipulation examines the detection of fake OCRs and fake users (see, e.g., Wu et al., 2020, Ansari and Gupta, 2021 or Paul and Nikolaev, 2021 for literature reviews), the actual impact of fake OCRs and fake users on consumers is less investigated. These studies also vary in the extent of information that consumers receive about the existence of malicious content. We identify three different levels of information that are shown to consumers: the lowest level of information does not give any information about the dissemination or presence of fake OCRs and fake users (i.e., *level 1*). In the next level, consumers (typically participants of an experiment) are primed on fake OCRs and fake users by giving them information about their existence and impact (i.e., *level 2*). In the highest level, the individual malicious content is explicitly highlighted as fake (i.e., *level 3*).

Level 1: Studies in this category analyze the impact of malicious content without any information about its existence. Hence, only prior knowledge of consumers that OCRs can be manipulated and user identities can be fake influences their product evaluation, trust and purchase intention (Bambauer-Sachse and Mangold, 2013; Ahmad and Sun, 2018). Other studies (Filieri, 2016; Carbonell et al., 2019) investigate OCR characteristics that make consumers suspicious regarding potential malicious content and find that writing style is an important determinant for consumers to identify suspicious content. These studies further highlight that the presence of a suspicious writing style decreases trust in OCRs.

Level 2: In this category, studies prime participants (typically in an experimental setting) about the fact that OCRs can be manipulated. While some studies find that raising participants' awareness of OCR manipulation decreases trust in OCRs (Ma and Lee, 2014) and trust in reviewers (DeAndrea et al., 2018), Munzel (2016) does not observe a change in the trustworthiness of the OCR source. Priming towards OCR manipulation, on the other hand, decreases (increases) purchase intention in case of positive (negative) OCR valence (Ma and Lee, 2014).

Level 3: The study by Ananthakrishnan et al. (2020) is – to the best of our knowledge – the only one that examines *level 3* information (i.e., malicious content is explicitly highlighted as fake). In their experiment, the authors analyze different possibilities on how to deal with fake OCRs. In particular, the authors compare the case of censoring fake OCRs with the case of tagging malicious OCRs as fake and examine its effect on trust in the platform as the OCR provider. They observe that trust in platform is higher when malicious OCRs are highlighted as fake instead of deleted. Moreover, their findings reveal that trust in platform is even higher when a platform additionally provides a heuristic summary (in form of a trust score) to reduce consumers' cognitive burden when confronted with fake OCRs. The authors also observe that consumers are not effective in processing motivational differences (e.g., self-promotion vs. harming a competitor) behind fake OCRs with positive or negative valence.

Our study is a *level 3* study: First, while the study by Ananthakrishnan et al. (2020) indicates that tagging OCRs as fake increases consumers' trust in platform, it is unclear how trust in OCRs and reviewers is affected. This is important since both *level 1* and *level 2* studies already highlight that fake OCRs and fake users decrease trust in OCRs (e.g., Ma and Lee, 2014; Carbonell et al., 2019) and reviewers (e.g., DeAndrea et al., 2018). Second, no study has yet examined how tagging fake OCRs and fake users simultaneously affects the different trust dimensions and it remains unclear whether these effects differ. Third, varying fake OCR valence introduces additional complexity and has not been investigated in the context of fake users and all relevant trust dimensions. Fourth, even though purchase intention has been addressed in the context of fake OCRs or fake users (e.g., Ma and Lee, 2014; Ahmad and Sun, 2018; Carbonell et al., 2019), the effect of *level 3* information on WTP has – to the best of our knowledge – not yet been investigated and there is a lack of understanding on how WTP is affected by its presence.

3 Theoretical Background and Hypotheses Development

In this section, we discuss the theoretical background on how tagging OCRs and users as fake[3] affects consumers' WTP. More specifically, we expect that different dimensions of consumers' trust (i.e., trust in OCRs, trust in reviewers, trust in platform) mediate the effects between the independent variables and the dependent variable WTP. Finally, we also expect the valence of fake OCRs to moderate the relation between our independent variables and the mediating variables. Figure 1 outlines our research model.

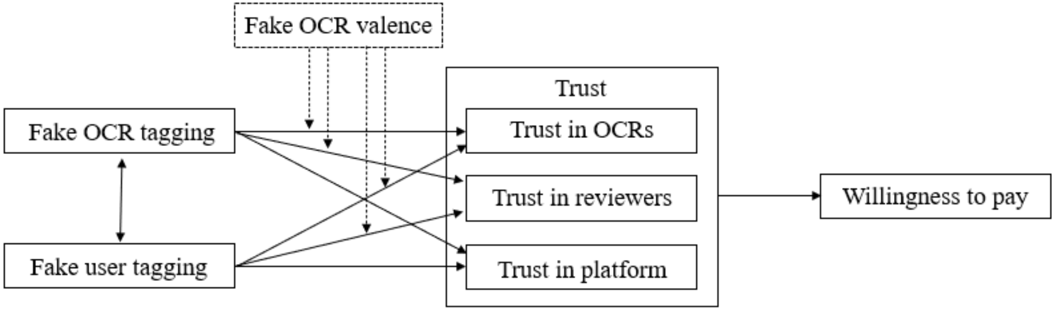


Figure 1. Research Model

We expect OCR manipulation to be a persuasion attempt directed at consumers (Ansari and Gupta, 2021) and that tagging OCRs and users as fake consequently activates persuasion knowledge. Thus, we draw on persuasion knowledge model (Friestad and Wright, 1994) to develop our hypotheses regarding

[3] Our understanding of tagging OCRs or users as fake can be seen in Table 1. It is inspired by the displaying style of Yelp and implies that we add a note to the content which indicates that it is suspected to be fake.

trust in OCRs and reviewers. The model explains how consumers use persuasion knowledge to cope with persuasion attempts by agents (e.g., salespeople, advertisers or brands) in order to maintain control over the outcome. This so-called persuasion coping behavior comprises several consumer behaviors such as interpreting, evaluating or adaptively responding and is shaped by the interaction of three knowledge structures: First, persuasion knowledge itself which includes knowledge about persuasion tactics and consumers' persuasion coping behavior. Second, agent knowledge which comprises beliefs about the objectives, traits and competencies of an agent who sends the message. And third, topic knowledge which includes the beliefs about the object or topic of the respective message. Existing research outlined that certain situations cause consumers to activate persuasion knowledge (Campbell and Kirmani, 2000; Kirmani and Zhu, 2007) and that persuasion knowledge is more likely to be activated if the persuasion attempt is easily accessible (Campbell and Kirmani, 2000). Hence, a platform that solely informs consumers about deleting malicious content on an ongoing basis activates only little persuasion knowledge. Persuasion knowledge is more activated, on the other hand, if a platform explicitly tags malicious OCRs or users as fake by making the persuasion attempt highly visible. The consequence of the latter, however, is that consumers might activate their deception-aware mindset (Boush et al., 2009) and become more skeptical towards the truthfulness of OCRs (Zhang et al., 2016). Thus, we expect that tagging fake OCRs or fake users decreases trust in OCRs and trust in reviewers. Accordingly, we state our first hypothesis as follows:

Hypothesis 1: *Tagging OCRs or users as fake decreases trust in OCRs and trust in reviewers.*

Since tagging OCRs as fake directly relates to the persuasion message, topic knowledge aspects of persuasion knowledge are mainly addressed. As a consequence, we expect that trust in OCRs is more negatively affected than trust in reviewers. Hence, we hypothesize:

Hypothesis 2: *Tagging OCRs as fake decreases trust in OCRs more than trust in reviewers.*

In the same vein, we expect that tagging users as fake directly relates to objectives of the message sender which implies that agent knowledge aspects of persuasion knowledge are mainly addressed. As a consequence, we expect that trust in reviewers is more negatively affected than trust in OCRs. Accordingly, we hypothesize:

Hypothesis 3: *Tagging users as fake decreases trust in OCRs less than trust in reviewers.*

When OCRs and users are tagged as fake simultaneously, both topic knowledge and agent knowledge are addressed at the same time. Drawing again on Campbell and Kirmani (2000) who argue that consumers activate more persuasion knowledge the easier accessible the manipulative intent, we expect that the effect on trust in OCRs and trust in reviewers is strongest when both OCRs and users are tagged as fake. Hence, we state our fourth hypothesis as follows:

Hypothesis 4*: Simultaneously tagging OCRs and users as fake leads to the strongest decrease in trust in OCRs and trust in reviewers.*

Fake OCRs can origin from self-promotional or competitive aspects (Mukherjee et al., 2012). While self-promotional content typically has positive valence with the aim to increase sales, competitors' content is commonly negative with the aim to harm the respective firm. Since Campbell and Kirmani (2000) argue that salespeople's (hidden) motive of selling is easier accessible than other motives, we expect that persuasion knowledge is more activated when the tagged OCRs have a positive valence as this represents the case of self-promotion. In other words, the persuasion attempt is more accessible for positive than for negative fake OCRs. We therefore expect that the valence of OCRs that are tagged as fake or whose authors are tagged as fake moderates the relation in *H1* and *H2*. Further, we also expect the same moderating effect when OCRs and users are simultaneously tagged as fake (i.e., *H4*). As we consider the platform to be an unbiased provider of information, we do not consider similar effects for trust in platform. Therefore, we hypothesize:

Hypothesis 5: *Effects in H1-H4 are stronger if the valence of OCRs that are tagged as fake or their authors are tagged as fake users is positive.*

From the platform perspective, tagging OCRs and users as fake serves as a signal for consumers. Therefore, we draw on signaling theory (Spence, 1973) to develop our hypothesis regarding trust in platform. Signaling theory describes how information asymmetries in decisions under uncertainty can be reduced by the use of signals. The reduction of information asymmetries for the signal receiver is in turn accompanied by the ability to evaluate the signal provider accordingly (Spence, 1973). Signals are widely applied in management contexts (see, e.g., Connelly et al., 2011 for a review). For the case of fake OCRs and fake users, a platform can either inform consumers that they are steadily deleting the malicious content[4] or explicitly tagging the malicious content as fake. Informing about censoring without providing more information represents a low signal quality as consumers have no information about the actual amount of malicious content. On the other hand, explicitly highlighting fake OCRs or fake users represents an obvious and direct signal and reduces the (perceived) information asymmetries for consumers. Reducing information asymmetries, in turn, comes along with an increase of the signal provider's trustworthiness (e.g., Siegfried et al., 2020). Thus, and in line with the findings by Ananthakrishnan et al. (2020), we state our sixth hypothesis as follows:

Hypothesis 6: *Tagging OCRs and users as fake increases trust in platform.*

After having hypothesized the effects of tagging OCRs and users as fake on the respective trust dimensions, we examine the effect of trust on consumers' WTP. Numerous studies examine the effect of trust on WTP in the context of sharing platforms (e.g., Otto et al., 2018), e-commerce (e.g., Ba and Pavlou, 2002) or social network sites (e.g., Han and Windsor, 2011). Since all these studies observe a positive relationship between the respective trust construct and WTP, we hypothesize:

Hypothesis 7: *Trust increases WTP.*

4 Experimental Design and Analysis Plan

We plan to address our research questions by applying an experimental approach that allows us to examine the effect of tagging fake OCRs and fake users on different trust dimensions and on WTP, respectively. In the following, we outline the experimental design in detail.

4.1 Experimental Task and Treatment Variations

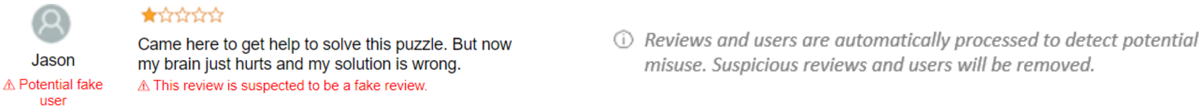
In our experiment, the participants' main task is to solve a puzzle. Participants get an initial budget of $1.70 which they can use to purchase a tutorial video that explains how to proceed with the puzzle. If the puzzle is solved correctly, they receive the remaining budget as a bonus in addition to a completion reward of $0.40. Otherwise, they only receive the completion reward. To support participants in making their purchase decision, we provide OCRs for the respective tutorial video.

As we aim to examine whether positive and negative fake OCR valence affect trust differently, we need to vary the quality of the tutorial video: To investigate the effect of positive (negative) fake OCR valence, the video needs to have low (high) quality. If the fake OCR valence is not opposed to the quality of the video, non-fake OCR valence would be the same as fake OCR valence and we would not be able to identify the effect. While the OCRs are the same for both videos, the three most negative (positive) OCRs represent malicious content for the high-quality (low-quality) video. Hence, for the high-quality video, we include manipulated OCRs with negative valence (that could come from a competitor). In case of the low-quality video, we include manipulated OCRs with positive valence (that could represent self-promotion).

Consequently, and depending on the treatment, the tutorial video is either of high quality and contains an important hint to solve the puzzle correctly or the video is of low quality without including this hint. For both video qualities, we then display the manipulated content in four different ways. In particular, these OCRs either (i) have a fake OCR tag, (ii) have a fake user tag, (iii) have a fake OCR tag and a fake user tag or (iv) are censored (i.e., deleted) and only a note informing participants about a potential removal of OCRs in general is displayed. Overall, we have two video quality treatments (i.e., high vs.

[4] Censoring malicious content and informing consumers about it is common practice. On Google Maps, for instance, the information reads as follows: "Reviews are automatically processed to detect inappropriate content like fake reviews and spam. We may take down reviews that are flagged in order to comply with Google policies or legal obligations."

low quality) and for each video quality, we have four displaying treatments (i.e., fake OCR vs. fake user vs. fake OCR + fake user vs. censoring). Figure 2 shows an exemplary manipulated OCR for tagging fake OCRs and users (a) as well as the note provided for censoring fake OCRs (b). For displaying variations where either fake OCRs or fake users are tagged, only the respective tag in (a) is displayed.



Jason

⚠ Potential fake user

★☆☆☆☆

Came here to get help to solve this puzzle. But now my brain just hurts and my solution is wrong.

⚠ This review is suspected to be a fake review.

ⓘ *Reviews and users are automatically processed to detect potential misuse. Suspicious reviews and users will be removed.*

(a) Tagging (b) Censoring

Figure 2. Treatment variation.

By reading OCRs, participants can identify if the respective video is of low or high quality and should respectively state their WTP. In particular, participants are asked to state their WTP between $0.00 and $1.00 for the tutorial video according to the BDM method (Becker et al., 1964). By using the BDM method, we want participants to state their true WTP, as they can only watch the tutorial video, if their stated price is higher than a randomly drawn price. Further, we create an incentive-compatible measurement for WTP, as their potential payment is reduced by the price for the video.

4.2 Experimental Procedure

The experimental procedure is illustrated in Figure 3. We use the web-based survey software SoSci Survey to implement the experiment.[5] Before conducting the experiment, we plan to do pretesting to check if participants understand the task and questions and to eliminate potential issues and ambiguities (Reynolds and Diamantopoulos, 1998). Given the complexity of the experiment, we decided to not recruit crowd workers but to conduct a laboratory experiment with students. To estimate the required number of participants, we conducted a power analysis using G*power (Faul et al., 2007). Assuming a medium effect size of around 0.6 and aiming at a power of 0.95 in our experiment, the minimum sample size for each of two groups that are compared using a t-test is 61. Considering that some of the participants might fail to correctly answer the attention checks, we round the number up to 70. Since we have a total of eight groups (i.e., four displaying styles for each fake OCR valence), we require approx. 560 participants who will be randomly assigned to one of our treatments.

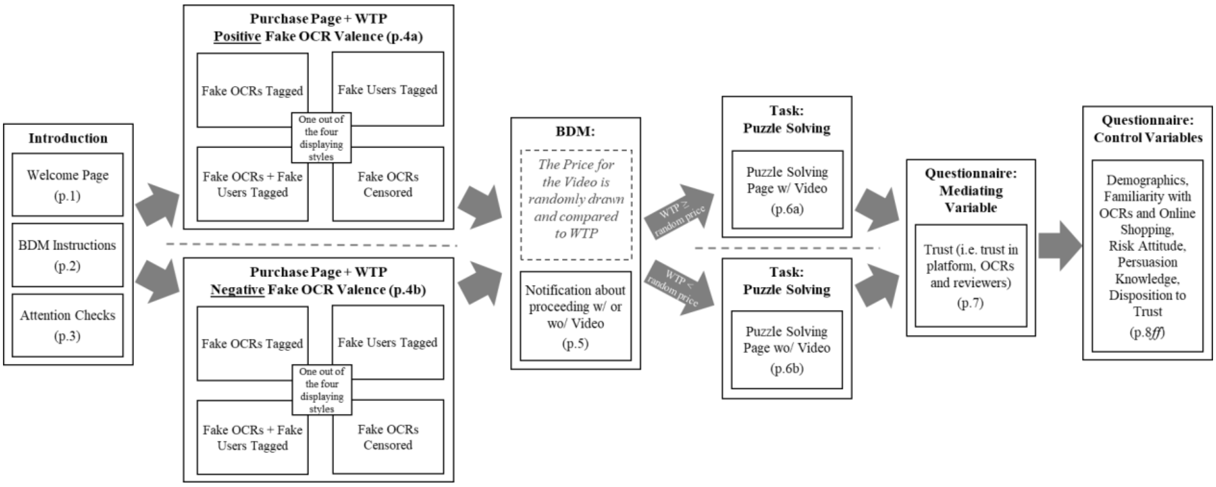

Introduction
Welcome Page (p.1)
BDM Instructions (p.2)
Attention Checks (p.3)
Purchase Page + WTP Positive Fake OCR Valence (p.4a)
Fake OCRs Tagged
Fake Users Tagged
One out of the four displaying styles
Fake OCRs + Fake Users Tagged
Fake OCRs Censored
Purchase Page + WTP Negative Fake OCR Valence (p.4b)
Fake OCRs Tagged
Fake Users Tagged
One out of the four displaying styles
Fake OCRs + Fake Users Tagged
Fake OCRs Censored
BDM:
The Price for the Video is randomly drawn and compared to WTP
Notification about proceeding w/ or wo/ Video (p.5)
WTP ≥ random price
WTP < random price
Task: Puzzle Solving
Puzzle Solving Page w/ Video (p.6a)
Task: Puzzle Solving
Puzzle Solving Page wo/ Video (p.6b)
Questionnaire: Mediating Variable
Trust (i.e. trust in platform, OCRs and reviewers) (p.7)
Questionnaire: Control Variables
Demographics, Familiarity with OCRs and Online Shopping, Risk Attitude, Persuasion Knowledge, Disposition to Trust (p.8ff)

Figure 3. Experimental Procedure.

[5] An outline of the experiment is available in the supplementary material under the following link: https://cutt.ly/XDlbYAJ. It also includes a detailed description of the puzzle, a visual representation of the displaying variations, a table including all OCRs and item descriptions.

At first, participants receive basic information about the experiment including our incentive-compatible payment structure (p.1) and are introduced to the BDM method (p.2). To ensure the quality of our data, we ask participants BDM-related questions to check if they have understood how they can state their WTP accordingly (p.3). Only if they correctly answer these questions, they are forwarded to the purchase page (p.4a/b). On the purchase page, participants are asked to state their WTP between $0.00 and $1.00 for a tutorial video that can help them with the given task of solving a puzzle. Besides some basic information (i.e., screenshot of the video and a short, incomplete puzzle description), ten OCRs are shown. To avoid an order bias, the position of the OCRs is randomized. Participants are either assigned to the high-quality video (p.4a) or the low-quality video (p.4b) and to one of the four displaying styles. Thus, three of the OCRs either have (i) a fake OCR tag, (ii) a fake user tag, (iii) a fake OCR and a fake user tag or (iv) are censored (i.e., deleted) and only a note informing participants about a potential removal of malicious content is displayed. After participants state their WTP, a random price is drawn individually for each participant. If the random price is equal or lower than the stated WTP, the tutorial video is purchased and can be watched (p.5). On the next page, participants have to solve the puzzle either with (p.6a) or without the video (p.6b). After submitting their solution, participants are informed whether they solved it correctly or not. Finally, we ask participants about their trust in OCRs, reviewers and platform (p.7) as well as about the control variables described in the next subsection (p.8*ff*).

4.3 Variables

Independent variables: Our independent variables are fake OCR tagging and fake user tagging. Thus, we either tag fake OCRs, fake users or both. Participants' decisions are compared against the case of censoring (i.e., deleting) the contents that were otherwise tagged as fake. For this case, we only display a note that generally informs participants about a potential removal of OCRs.

Mediating variables: As outlined above, we conjecture trust in OCRs, trust in reviewers and trust in platform to mediate the effects of fake OCR tagging and fake user tagging on WTP. For measuring trust in OCRs and trust in reviewers, we draw on the measurement scale for electronic word-of-mouth skepticism by Zhang et al. (2016). In particular, we measure trust in OCRs by modifying three of their items and trust in reviewers by modifying six of their items. To measure trust in platform, we adapt seven items from the measurement scale for trust in e-commerce by McKnight et al. (2002) and complement them by modifying one item from Ananthakrishnan et al. (2020).

Moderating variable: To analyze a potential moderating effect of fake OCR valence, we only tag the three most positive (negative) OCRs for the low-quality (high-quality) video.

Dependent variable: Our variable of interest is WTP. As outlined above, we measure WTP by using the incentive-compatible BDM method (Becker et al., 1964).

Control variables: We include several control variables that might affect participants' decisions and are important to control for. Thus, we ask about participants' sociodemographic factors (i.e. age, gender, education, country of origin, income), their familiarity with online shopping, reading OCRs and solving puzzles. In this vein, we also ask about participants' general attitude towards OCRs adapting four items from Park et al. (2007). Further, we check if they acknowledge tutorial videos or help in general. We also ask for their general risk attitude (Dohmen et al., 2011) as this might influence their WTP as well. Further, we measure participants' disposition to trust using four items described in Gefen (2000). Finally, in order to check if the treatment variations had the desired effects, we ask about participants' persuasion knowledge adapting items from Bambauer-Sachse and Mangold (2013).

4.4 Analysis Plan

We plan to analyze the data collected with the laboratory experiment as follows. We will remove participants that fail to correctly answer the attention check questions and check the dataset for outliers.

As a first step of analysis, we will test the effects of tagging fake OCRs and/or users on the three trust dimensions (i.e., trust in OCRs, reviewers and platform) as hypothesized in *H1-H3* and *H6*, respectively. For this purpose, we will use independent sample t-tests to assess statistical significance of the difference between the respective treatment and the case of censoring (i.e., deleting) the contents that were otherwise tagged as fake. Further, the different tagging variations will also be compared using t-tests in order to examine the effect simultaneously tagging fake OCRs and users (i.e., *H4*).

Next, to test whether fake OCR valence moderates the effects of tagging fake OCRs and/or users on the three trust dimensions (i.e., *H5*), we will compare trust dimensions for each tagging variation with negative and positive fake OCR valence. We will again assess statistical significance of the respective differences with t-tests.

Finally, to test the effect of trust on WTP (i.e., *H7*), we will estimate an OLS model for each of the three trust dimensions to examine if there is a significant effect on WTP for any of them. Further, we plan to investigate potential mediating effects of tagging fake OCRs and/or users on WTP via each of the three trust dimensions. In more detail, we plan to apply a mediation analysis for each of the trust dimensions using the PROCESS macro (Model 4) for SPSS (Hayes, 2017). In particular, we want to choose a bootstrapping procedure instead of the traditional test by Sobel (1982), since, according to Hayes (2017), the bootstrapping confidence interval tends to have a higher power.

5 Conclusion

On digital platforms, OCRs represent an important information source for consumers prior to making their purchase decisions, especially when they do not have an on-hand experience about the product prior to their purchase (Kwark et al., 2014; Manes and Tchetchik, 2018). However, fake OCRs negatively affect consumers' trust in OCRs (e.g., Ma and Lee, 2014; Carbonell et al., 2019) and reviewers (e.g., DeAndrea et al., 2018). Although a lot of research exists in the context of detecting fake OCRs and users, little is known about how platforms should deal with fake OCRs and fake users after detecting them. With our experimental study, we aim at closing this gap and thus providing a relevant contribution in the following ways: First, we want to extend existing research by considering effects of tagging fake OCRs and fake users on the relevant trust dimensions (i.e., trust in OCRs, reviewers, platform) and WTP. Second, we aim at a better understanding of the impact of fake OCR valence. Third, we want to examine an incentive-compatible WTP leading to direct economic consequences for the participants. Further, we expect our proposed experiment to provide important implications on how to deal with fake OCRs and fake users after detecting them. In more detail, by explicitly accounting for three different trust dimensions, we can examine which of the dimensions affects WTP most. This allows us to give implications for digital platforms that have more objectives than just increasing their own trust. Furthermore, we can also give recommendations to producers (competitors) who might post positive (negative) fake OCRs what to do when fake OCRs and users are displayed in a particular way.

Finally, our study has certain limitations which might, however, serve as starting points for future research. First, we do not plan to give participants an explanation why OCRs or users are tagged as fake. Nevertheless, it could be interesting to examine the effects of this on participants. In this vein, it might also be interesting to investigate how indicating different types of fake OCRs affects trust or WTP. In particular, besides manipulated (i.e., "fake") OCRs, one could also try to tag biased OCRs (i.e., OCRs written in exchange for a discount or free product by the producer). Second, we only consider either positive or negative fake OCRs for the low-quality or high-quality video, respectively. Thus, to extend our research, it might be interesting to consider fake OCR valences that are mixed or match the video quality. Third, we only focus on OCRs that consist of ratings and textual content. Future research might also investigate potential different effects if OCRs without textual content are tagged as fake. Fourth, our proposed experiment focuses on a specific product (i.e., the tutorial video) that is only available for a limited time and that has a low price. To address this issue, further research could still examine other products (e.g., products with longer consumption times or higher prices) or services. Fifth, it might be interesting to examine the effect of tagging OCRs or users as fake on various platform types (e.g., sharing platforms). Finally, although the online shopping situation in our experiment is linked to solving a puzzle, it remains an artificial situation by conducting a laboratory experiment with students. Thus, future research could try to transfer our experiment to a real online shopping situation.

References

Ahmad, W. and J. Sun (2018). “Modeling consumer distrust of online hotel reviews” *International Journal of Hospitality Management* 71, 77–90.

Akoglu, L., R. Chandy and C. Faloutsos (2013). “Opinion Fraud Detection in Online Reviews by Network Effects”. In: *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, Cambridge, Massachusetts: 2–11.

Ananthakrishnan, U. M., B. Li and M. D. Smith (2020). “A Tangled Web: Should Online Review Portals Display Fraudulent Reviews?” *Information Systems Research* 31 (3), 950–971.

Ansari, S. and S. Gupta (2021). “Review Manipulation: Literature Review, and Future Research Agenda” *Pacific Asia Journal of the Association for Information Systems* 13 (1), 97–121.

Ba, S. and P. A. Pavlou (2002). “Evidence of the Effect of Trust Building Technology in Electronic Markets: Price Premiums and Buyer Behavior” *MIS Quarterly* 26 (3), 243–268.

Bambauer-Sachse, S. and S. Mangold (2013). “Do consumers still believe what is said in online product reviews? A persuasion knowledge approach” *Journal of Retailing and Consumer Services* 20 (4), 373–381.

Becker, G. M., M. H. DeGroot and J. Marschak (1964). “Measuring Utility by a Single-Response Sequential Method” *Behavioral Science* 9 (3), 226–232.

Boush, D. M., M. Friestad and P. Wright (2009). *Deception in the Marketplace: The Psychology of Deceptive Persuasion and Consumer Self-Protection.* 1st Edition. New York: Routledge.

Brynjolfsson, E., A. Collis and F. Eggers (2019). “Using massive online choice experiments to measure changes in well-being” *Proceedings of the National Academy of Sciences* 116 (15), 7250-7255.

Campbell, M. C. and A. Kirmani (2000). “Consumers’ Use of Persuasion Knowledge: The Effects of Accessibility and Cognitive Capacity on Perceptions of an Influence Agent” *Journal of Consumer Research* 27 (1), 69–83.

Carbonell, G., C.-M. Barbu, L. Vorgerd and M. Brand (2019). “The impact of emotionality and trust cues on the perceived trustworthiness of online reviews” *Cogent Business & Management* 6 (1), 1586062.

CHEQ (2021). *The cost of Lost Revenue Opportunities: How bots and invalid users are disrupting online customer acquisition*. URL: https://irp.cdn-website.com/9d8f1a2e/files/uploaded/CHEQ%20Report%20-%20Paid%20Marketing%20-%20Cost%20of%20Lost%20Revenue%20Opportunities%202021.pdf (visited on 15 November 2021).

Connelly, B. L., S. T. Certo, R. D. Ireland and C. R. Reutzel (2011). “Signaling Theory: A Review and Assessment” *Journal of Management* 37 (1), 39–67.

DeAndrea, D. C., B. van der Heide, M. A. Vendemia and M. H. Vang (2018). “How People Evaluate Online Reviews” *Communication Research* 45 (5), 719–736.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp and G. G. Wagner (2011). “Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences” *Journal of the European Economic Association* 9 (3), 522–550.

Faul, F., E. Erdfelder, A.-G. Lang and A. Buchner (2007). “G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences” *Behavior Research Methods* 39 (2), 175–191.

Filieri, R. (2016). “What makes an online consumer review trustworthy?” *Annals of Tourism Research* 58, 46–64.

Friestad, M. and P. Wright (1994). “The Persuasion Knowledge Model: How People Cope with Persuasion Attempts” *Journal of Consumer Research* 21 (1), 1–31.

Gefen, D. (2000). “E-commerce: the role of familiarity and trust” *Omega* 28 (6), 725–737.

Han, B. O. and J. Windsor (2011). “User’s Willingness to Pay on Social Network Sites” *Journal of Computer Information Systems* 51 (4), 31–40.

Hayes, A. F. (2017). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*. 2nd ed. New York, NY: Guilford Press.

Hu, N., I. Bose, Y. Gao and L. Liu (2011). “Manipulation in digital word-of-mouth: A reality check for book reviews” *Decision Support Systems* 50 (3), 627–635.

Kirmani, A. and R. Zhu (2007). "Vigilant Against Manipulation: The Effect of Regulatory Focus on the Use of Persuasion Knowledge" *Journal of Marketing Research* 44 (4), 688–701.

Kumar, N., D. Venugopal, L. Qiu and S. Kumar (2018). "Detecting Review Manipulation on Online Platforms with Hierarchical Supervised Learning" *Journal of Management Information Systems* 35 (1), 350–380.

Kwark, Y., J. Chen and S. Raghunathan (2014). "Online Product Reviews: Implications for Retailers and Competing Manufacturers" *Information Systems Research* 25 (1), 93–110.

Lappas, T., G. Sabnis and G. Valkanas (2016). "The Impact of Fake Reviews on Online Visibility: A Vulnerability Assessment of the Hotel Industry" *Information Systems Research* 27 (4), 940–961.

Ma, Y. J. and H.-H. Lee (2014). "Consumer responses toward online review manipulation" *Journal of Research in Interactive Marketing* 8 (3), 224–244.

Manes, E. and A. Tchetchik (2018). "The role of electronic word of mouth in reducing information asymmetry: An empirical investigation of online hotel booking" *Journal of Business Research* 85, 185–196.

Mayer, R. C., J. H. Davis and F. D. Schoorman (1995). "An Integrative Model of Organizational Trust" *Academy of Management Review* 20 (3), 709-734.

McKnight, D. H., V. Choudhury and C. Kacmar (2002). "Developing and Validating Trust Measures for e-Commerce: An Integrative Typology" *Information Systems Research* 13 (3), 334–359.

Mukherjee, A., B. Liu and N. Glance (2012). "Spotting Fake Reviewer Groups in Consumer Reviews". In: *Proceedings of the 21st International Conference on World Wide Web*. Lyon, France: 191–200.

Munzel, A. (2016). "Assisting consumers in detecting fake reviews: The role of identity information disclosure and consensus" *Journal of Retailing and Consumer Services* 32, 96–108.

Otto, L., P. Angerer and S. Zimmermann (2018). "Incorporating External Trust Signals on Service Sharing Platforms". In: *Proceedings of the 26th European Conference on Information Systems*. Portsmouth, United Kingdom: 78.

Park, D.-H., J. Lee and I. Han (2007). "The Effect of On-Line Consumer Reviews on Consumer Purchasing Intention: The Moderating Role of Involvement" *International Journal of Electronic Commerce* 11 (4), 125–148.

Paul, H. and A. Nikolaev (2021). "Fake review detection on online E-commerce platforms: a systematic literature review" *Data Mining and Knowledge Discovery* 35 (5), 1830–1881.

Pitman, J. (2022). *Local Consumer Review Survey 2022*. URL: https://www.brightlocal.com/research/local-consumer-review-survey/# (visited on 27 March 2022).

Reynolds, N. and A. Diamantopoulos (1998). "The effect of pretest method on error detection rates: Experimental evidence" *European Journal of Marketing*, 32 (5/6), 480–498.

Siegfried, N., J. Löbbers and A. Benlian (2020). "The Trust-Building Nature of Identity Verification in the Sharing Economy: An Online Experiment". In: *Proceedings of the 15th International Conference on Wirtschaftsinformatik*, Potsdam, Germany: 1506–1521.

Shan, G., L. Zhou and D. Zhang (2021). "From conflicts and confusion to doubts: Examining review inconsistency for fake review detection" *Decision Support Systems* 144, 113513.

Sobel, M. E. (1982). "Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models" *Sociological Methodology* 13, 290–312.

Spence, M. (1973). "Job Market Signaling" *The Quarterly Journal of Economics* 87 (3), 355–374.

Uberall (2021). *The State of Online Review Fraud: An Analysis of 4 Million Reviews on Google,* Facebook, Yelp and Tripadvisor. URL: https://join.momentfeed.com/hubfs/2021%20Fake%20Reviews/FakeReviews_Report.pdf (visited on 14 November 2021).

Wu, Y., E. W. T. Ngai, P. Wu and C. Wu (2020). "Fake online reviews: Literature review, synthesis, and directions for future research" *Decision Support Systems* 132, 113280.

Zhang, T., G. Li, T. C. Cheng and K. K. Lai (2017). "Welfare Economics of Review Information: Implications for the Online Selling Platform Owner" *International Journal of Production Economics* 184 (1), 69–79.

Zhang, X. J., M. Ko and D. Carpenter (2016). "Development of a scale to measure skepticism toward electronic word-of-mouth" *Computers in Human Behavior* 56 (1), 198–208.