ABSTRACT

Title of Document: TWO ESSAYS ON USER-GENERATED CONTENTS IN MARKETING

Jin-Hee Huh, Ph.D., 2020

Directed By: Dr.David Godes, Professor and Department Chair Marketing Department

This dissertation studies two topics regarding user-generated content (UGC) in social media context. Due to its rapid proliferation, UGC has received much attention among both practitioners and researchers. Although past studies have investigated diverse aspects of UGC and their influences on consumer decision process, few studies have examined (1) status bias in online peer-feedback on UGC with the presence of reputation system, and (2) the impact of video UGC, a relatively new type of UGC, on product purchase and usage decisions.

In Essay 1, we investigate how a reputation system affects peer-evaluations in an online community. In contrast to previous research on reputation systems, which has predominantly shown that reputation systems can induce posting activity and quality contributions, we study how reputation markers ("badges," for example) may change peer-evaluations. We rely on a unique and detailed data set and employ a difference-in-differences approach, combined with propensity score matching, that suggests that, all else equal, posters receive disproportionately-higher evaluations of their posts

after they earn a reputation marker, irrespective of post quality. This suggests a "rich-get-richer" process induced by a reputation system that may have the unintended effect of favoring some participants at the expense of equally-skilled others.

In Essay 2, we examine the impact of a relatively new type of user-generated content, video user-generated content (VUGC), on consumer's purchase and product usage decision. Due to its highly visual and auditory characteristics, VUGC has been considered as an effective marketing tool. However, as VUGC can have the risk of revealing too much information about the entertainment products, firms need to limit the amount of original content in VUGC. Under these circumstances, game developers need to have an understanding on how VUGC can impact sales and usage. To examine the impact, we collected VUGC activities and game characteristics from diverse sources. Our results indicate that VUGCs can affect sales and game usage, and the impact can vary by VUGC and game product characteristics. Our findings suggest that game developers should consider both VUGC characteristics and their product characteristics jointly when developing and implementing policies on users' VUGC uploading, sharing, and monetizing.

TWO ESSAYS ON USER-GENERATED CONTENTS IN MARKETING

By

Jin-Hee Huh

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:
Professor David Godes, Chair
Professor P.K. Kannan
Professor Guido Kuersteiner
Associate Professor Michael Trusov
Assistant Professor Lingling Zhang

Acknowledgements

I would like to express my sincere gratitude to R.H. Smith School of Business for helping me fulfill my dream of becoming a doctor of Marketing. I would like to extend my sincere gratitude to my advisor Dr. David Godes for his continued support and encouragement of my Ph.D study and research. It was a great privilege and honor to work and study under his guidance, and I cannot imagine having a better advisor and mentor for my Ph.D study.

Aside from my mentor, I would like to thank the rest of my dissertation committee, Dr. P.K. Kannan, Dr. Lingling Zhang, Dr. Michael Trusov, and Dr. Guido Kuersteiner for their encouragement and insight. I would especially like to thank Dr. P.K. Kannan and Dr. Lingling Zhang for their invaluable guidance throughout our research work.

I also thank my fellow Marketing Ph.D students: Nicole Kim, Min Kim, Inhye Kang, Xian Gu, Yuechen Wu, Kalinda Ukanwa, Jared Watson, Seungwoo Lee, Cindy Zhao, and Chutian Wang. I deeply appreciate that I had a chance to grow into a researcher with these wonderful fellows.

I am extremely grateful to my parents for their love, prayers, caring, and sacrifices for educating and raising me. I would like to thank them for their unconditional love and support in whatever I pursue. I also express my thanks to my brother and sister in law for their support and prayers. Lastly, I thank my boyfriend, Jared Gongloff, for his love and understanding, and my best friends, Yeonmi Kim, Hyesun Lee, Sandra Kang, Minjung Kim for their emotional support from afar in Korea.

Table of Contents

The Impact of Reputation Systems on Peer Feedback in Social Media

Jin-Hee Huh* David Godes† Seshadri Tirunillai‡

July 21, 2020

Abstract

Understanding how opinions are received and evaluated is important for an online community because the volume and quality of user opinions are key determinants of success. In this paper, we investigate how a reputation system affects peer evaluations in an online community. In contrast to previous research on reputation systems, which has predominantly shown that reputation systems can induce posting activity and quality contributions, we study how reputation markers (achievement badges, for example) may affect changes in peer evaluations. We rely on a unique and detailed data set and employ a difference-in-differences approach, combined with propensity score matching, that suggests that, all else equal, posters receive disproportionately-higher evaluations of their posts after they earn a reputation marker, irrespective of post quality. This suggests a "rich-get-richer" process induced by a reputation system that may have the unintended effect of favoring some participants at the expense of equally or more highly skilled others.

Keywords: quasi-experimental methods, difference-in-differences approach, reputation system

*Robert H. Smith School of Business, University of Maryland, jhhuh@umd.edu
†Robert H. Smith School of Business, University of Maryland, dgodes@umd.edu
‡C.T.Bauer College of Business, University of Houston, seshandri@bauer.uh.edu

1 Introduction

"Your company's social web efforts will succeed only to the extent that you are able to attract good individuals, motivate them to perform good work, and empower them to get to know and trust one another enough to collaborate toward the end goals of the community. The question is, How do you do that? The answer: By capitalizing on the motivational power of reputation."

(Dellarocas, 2010)

For online communities, the volume and quality of user contributions, such as user-generated content (UGC), are the key determinants of success. To provide the community members with a source of information and help in achieving their goals, online communities need to attract a sufficient volume of contributions from their members. However, the value of user contributions does not depend only on the sheer volume but also on their quality. Contribution quality in such online communities has been found to be improved by peer feedback based reputation systems (Ba and Pavlou 2002; Dellarocas 2004; Dellarocas 2010; Resnick 2000). Thus, many online communities, including e-commerce sites such as eBay, review sites like Yelp and TripAdvisor, and online advice communities such as Quora and Reddit, have adopted reputation systems. Those systems attempt to align incentives such that a high quality post can benefit both the user and the community, as it endows reputation and credibility to users who post content and valuable information for the rest of the community.

Despite the widespread adoption and generally-accepted importance of reputation systems, very little work has addressed the accuracy of such systems or the extent to which they yield the desired impact. In this paper, we propose that this may not always be the case. In fact, as we show, peer feedback might encourage a biased portrayal of the "quality" online content. One possible adverse effect of such phenomenon would be that a poster with a high reputation may gain better peer feedback than a poster with a low reputation, even though these two users are equally skilled. Even worse, the user with a low reputation may post better quality, yet receive lower feedback than the other. Hence, the reputation system could unintentionally distort peer feedback and *reduce* the informativeness of the reputation system. Such a distortion may also discourage newcomers' active participation if they find that there exists under-recognition of content quality.

To examine reputation systems' effect on peer feedback in an online community, we collect data

from one of the biggest question-and-answer online communities, StackOverflow. The site employs a well-developed reputation system. We examine whether, all else equal, a post is more highly rated after a badge is awarded as compared with before. To address endogeneity concerns, we use a difference-in-differences model after matching treatment and control posts using several matching methods with propensity score. Our analysis suggests that after a user earns a reputation badge, there is roughly a 10% increase in positive peer feedback on the user's posts.

Most of the studies on reputation system focus on the users' content contributions, such as volume and quality of contents (Ba and Pavlou 2002; Bolton et al. 2004; Grand and Betts 2013; Zhong 2016). This study contributes to the online reputation system literature by showing the bias in peer feedback introduced by reputation systems where peer evaluation remains strictly anonymous. This study also contributes to the status bias literature in that it shows the status bias in a real-world setting without explicitly measuring the actual quality.

The rest of this paper is organized as follows. Following our review of the literature in Section 2, we describe our research setting and data in Section 3, and our empirical strategy in Section 4. In Section 5, we test our theory and discuss the results. In Section 6, we present the results of several robustness checks. We conclude in Section 7 with a discussion of the limitations of our study, as well as the rich opportunities our results offer for future research.

2 Literature Review

Our work draws on two streams of research: (1) the importance of reputation system in the users' behaviors in online communities, and (2) online users' dependency on heuristic cues in their decision making.

2.1 The Impact of Reputation System on Peer Feedback

Beginning with Resnick et al. (2000), the effect of reputation systems on users' actions in online communities has been of significant academic interest. A reputation system is an information system that mediates and facilitates the process of assessing one's credibility, and which is based on the past actions within the context of a specific community. According to Dellarocas (2003), "The best-known application of online feedback mechanisms to date has been their use as a technology

for building trust in electronic markets.... The application of feedback mechanisms in online marketplaces is particularly interesting, because many of these marketplaces would probably not have come into existence without them."

The importance of reputation systems in online communities has been established by several academic studies (Ba and Pavlou 2002; Bolton et al. 2004; Grand and Betts 2013; Zhong 2016). These studies have shown that the enhanced credibility and trust engendered by reputation systems facilitate users' content consumption and creation. Ba and Pavlou (2002) find that feedback-based reputation system can induce trust in sellers' credibility. In turn, this may benefit the sellers via price premium. Bolton et al. (2004) show that the reputation system induces a substantial improvement in transaction efficiency, benefiting both sellers and buyers on eBay. Grand and Betts (2013) demonstrate that users at StackOverflow increase their posting activities related to a visualized reputation indicator, a reputation badge, immediately before the badge is awarded, suggesting that the incentive effects of the system are credible. Zhong (2016) shows that buyers' demand and sellers' pricing strategies are affected by a reputation system at Taobao. Because a reputation system usually incorporates a user's past behaviors within a community into a visualized reputation indicator, such as reputation scores or badges, content consumers can use the indicator to infer the source's credibility. For content creators, the reputation system promotes the contribution of high-quality content by rewarding them with increased social status within the community.

Whereas peer feedback plays a critical role in the success of reputation systems, to our knowledge little work has addressed the potential bias introduced by reputation systems. The majority of reputation system studies have assumed that users' peer feedback is unbiased, and that reputation scores represent an aggregation of those unbiased opinions. However, a user's peer feedback can be affected by diverse factors, even possibly resulting in a biased reputation system. The findings of Dellarocas and Wood (2008) and Bolton et al. (2013) support such conjecture. They find that the reputation system of eBay can affect members' voluntary feedback submissions through reciprocity or retaliation from the partner member in a transaction. Because only the partners of a transaction are allowed to give feedback to each other in eBay, there is no anonymity in peer evaluation. In addition, users are able to make feedback submissions and rating decisions after observing the transaction partner's actual feedback or having the anticipated probability of whether the partner will submit feedback to them. Under these circumstances, users tend to make

strategic voting decisions. Specifically, Bolton et al. (2013) find that buyers and sellers become more likely to provide feedback when the transaction partner has given feedback. Additionally, eBay users reciprocate positive feedback and retaliate negative feedback. Such reciprocal and retaliative pattern can be a problem for online communities because they tend to reduce the informativeness of the feedback given and likely hurt market efficiency.

While Bolton et al. (2013) shed light on how the design or feature of reputation systems like eBay's may affect peer feedback, our study is quite different. In their setting, users need to know 1) the identity of their evaluator, and 2) the valence of feedback they received (e.g., positive or negative rating). However, in most online communities other than eBay, such information is not usually available and anonymity is guaranteed. In fact, it is typically the case that feedback is aggregated rather than available at the individual level. To the best of our knowledge, no one has investigated the impact of reputation system on peer feedback when users do not have access to such identity information about the evaluator.

2.2 Reputation Marker as a Heuristic Cue

Previous research has shown that online users have limited attention (Zhong 2016), and depend on heuristic cues when making many consumption decisions. This occurs particularly when users are exposed to difficult decisions such as many options or constrained resources. Examples of such cues include reviewer self-disclosure of identity-descriptive information (Forman et al. 2008), readability of content (Ghose and Ipeirotis 2011), content depth, and extremity (Mudambi and Schuff 2010). We believe that, as Zhong (2016) suggests, reputation markers may also be one of the cues affecting content consumption and evaluation. This may be the case because such markers are designed to summarize the feedback on a user's past behavior (Grants and Betts 2013). Also, as Shenkar and Yuchtman-Yaar (1997) note, reputation for past performance serves as a substitute for direct knowledge about product quality where it is difficult to ascertain directly. Thus, when user-generated content is displayed with a positive reputation marker associated with the poster, then it could be possible that the "positive glow" emanating from the poster influences a user's overall attitude towards the content. This, in return, may suggest that the content could be consumed more, or rated more favorably, than other content of equal quality.

In economic sociology, such self-reproducing advantages for high-status individuals is often

described as the Matthew Effect (Merton 1968). Merton (1968) assumes that individuals are biased to positively evaluate high-status individuals irrespective of quality, and that this bias, rather than actual quality differences, perpetuates inequality between high-status and low-status actors. Scholars in status characteristics theory have shown that this status bias results from individual's expectations and prior beliefs about competence and quality (Anderson et al. 2001; Ridgeway and Correll 2006; Berger et al. 1977; Ridgeway and Berger 1986). The Matthew Effect has been demonstrated in a range of diverse areas. For example, Sine et al. (2003) show how overall university prestige increases the licensing rate over that predicted by past performance, even after controlling for the amount of technology produced by the university, the source of research funding, the presence or absence of a medical school, the geographic location of the university, and the resources of the technology licensing office. Kim and King (2014) find that Major League Baseball (MLB) umpires are more likely to call a strike when a pitch was actually a ball if the pitcher has high status.

Much of the past research on the Matthew Effect has examined status bias in experimental studies (e.g., Berger et al. 1972, 1977; Ridgeway 1991; Wagner and Berger 2002; Correll et al. 2007). In order to identify the status bias effect, it is necessary to measure the true quality. However, it is challenging to measure the true quality with observational data. Kim an Kang (2014) show self-reproducing advantage for high-status individuals by comparing judgments made by evaluators to precise measurement of actual quality. However, such precise measurement of actual quality are not easily obtainable in field studies. In this study, by using a difference-in-differences approach, we demonstrate the presence of the Matthew Effect in online communities without explicitly measuring the actual quality. Most importantly, we document the role of the explicit reputation management policies of platforms like StackOverflow in driving these effects.

3 Research Setting and Data

To assess the impact of a reputation marker on peer evaluations in online communities, we collect data from StackOverflow.com (SO), an online community focusing on discussing and answering questions posted on a wide range of topics in computer programming and related areas. Our data set consists of three types of data: (1) badge award data, (2) post activity data, and (3) peer

feedback data. Before describing each, we first highlight the main characteristics of SO.

3.1 StackOverflow

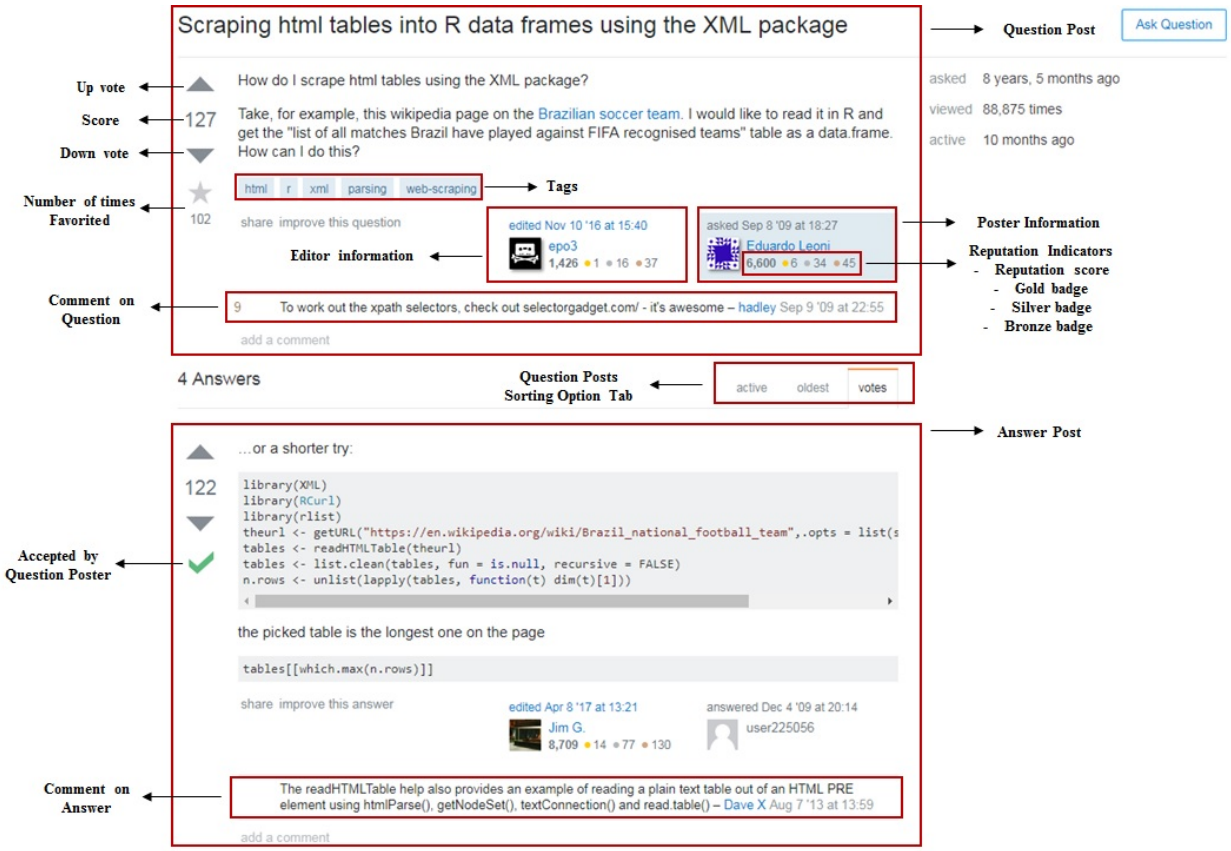
SO is one of the most-active online communities among computer programmers. As of June 2020, SO had more than 12 million registered users and more than 19 million questions and 29 million answers[1]. The core of SO is a collection of questions and answers (Figure 1). Building a reputation within SO is a key motivator that drives users to contribute to the community (Anderson et al., 2012). A user's reputation in SO is determined by two components: objective measures of user activity quantity and subjective measures of user activity quality. The objective measures include the volume of posts uploaded, the frequency of visits to SO, and membership duration. The subjective measures of user activity quality are based on feedback from other users on SO, which will be described in more detail in Section 3.2.

SO follows a common model of question-and-answer (Q&A) site design. Users can post questions, answer questions, comment on posts, and give diverse types of feedback (e.g., vote, comment, edit, or bounty award) on posts. Users can register on the site to obtain reputation scores and badges determined by their own activities and evaluations by peers. For example, a user is awarded a "Nice Answer" badge if one of his posts receives a score of at least 10, where the score is calculated as the number of up votes less the number of down votes. Thus, a high reputation score or large volume of badges can be an indicator of one's expertise, possibly suggesting that the user is highly trusted by his peers.

Our data set for the empirical analysis consists of three distinct types of data, and spans the period from August 4, 2008, the date when the website was launched, to January 19, 2014. We emphasize that the data are not a random sample of the entire SO users or posts. Our use of the total population data enables us to eliminate any potential bias occurring through sampling (Gorard 2013).

[1] https://sostats.github.io/ (accessed on June 7, 2020

Figure 1: StackOverflow Post Example



Source: https://stackoverflow.com/questions/1395528/scraping-html-tables-into-r-data-frames-using-the-xml-package (accessed on June 23, 2019).

3.2 Badge System at StackOverflow

Badges in SO are classified into three levels: bronze, silver, and gold. The badges are awarded based on a user's contributions to SO, and are categorized into seven groups: question, answer, participation, moderation, documentation, tag, and others. Table 1 presents the list of bronze-, silver-, and gold-level badges in each category as of June 7, 2020. The overall attributes of bronze, silver, and gold badges are as follows: [2]

(1) Bronze badges are awarded for the basic usage of SO. These badges encourage users to employ typical, routine functions of the site: posting questions, answering questions, voting up or down, tagging posts, editing, and filling out their user profile. They are relatively easy to get and provide immediate positive feedback to new users. Bronze badges have been awarded about 27,426,400k times as of June 7, 2020.

[2]https://stackoverflow.com/help/badges (accessed on June 7, 2020).

Table 1: List of Badges at StackOverflow.com

	Bronze	Silver	Gold
Question Badges	Altruist, Benefactor, Curious, Investor, Nice Question, Popular Question, Promoter, Scholar, Student	Inquisitive, Favorite Question, Good Question, Notable Question	Socratic, Stellar Question, Great Question, Famous Question
Answer Badges	Explainer, Nice Answer, Revival, Self-Learner, Teacher	Enlightened, Refiner, Generalist, Guru, Lifejacket, Good Answer, Necromancer, Tenacious	Illuminator, Lifeboat, Great Answer, Populist, Unsung Hero
Participation Badges	Autobiographer, Caucus, Commentator, Mortarboard, Precognitive, Quorum, Talkative	Constituent, Pundit, Enthusiast, Epic, Beta, Convention, Outspoken, Yearling	Fanatic, Legendary
Moderation Badges	Citizen Patrol, Cleanup, Critic, Custodian, Disciplined, Editor, Excavator, Organizer, Peer Pressure, Proofreader, Suffrage, Supporter, Synonymizer, Tag Editor, Vox Populi	Deputy, Civic Duty, Reviewer, Strunk and White, Archaeologist, Sportsmanship, Research Assistant, Taxonomist	Marshal, Constable, Sheriff, Steward, Copy Editor, Electorate
Tag Badges	Gold Tag Badges	Silver Tag Badge	Gold Badge
Other Badges	Announcer, Informed	Booster, Census, Not a Robot	Publicist

Table 2: The Number of Times for Five Most Frequently Awarded Gold, Silver, and Bronze Badges

Gold		Silver		Bronze	
Name	No.Awards	Name	No.Awards	Name	No.Awards
Famous Question	717,396	Notable Question	2,480,269	Popular Question	5,043,307
Great Answer	83,879	Yearling	2,249,293	Editor	2,523,903
Great Question	39,278	Necromancer	643,035	Student	2,455,477
Fanatic	34,940	Good Answer	455,388	Informed	2,366,632
Unsung Hero	23,151	Enlightened	414,018	Scholar	2,005,880

(2) Silver badges encourage continued participation and visits to the site. These badges are less common, but attainable if users show sustained interest in SO. Silver badges have been awarded about 7,664.7k times June 7, 2020.

(3) Gold badges signify outstanding dedication or achievement. They require not only consistent participation but also high-level skills and knowledge about certain topics. Gold badges have been awarded about 990.2k times June 7, 2020.

To appreciate the differences among the badges, note that the "Nice Answer" bronze badge requires users to have at least one answer post with at least 10 scores. The "Good Answer" silver badge requires an answer post with at least 25 scores. The "Great Answer" gold badge is awarded to users who have posted an answer with at least 100 scores. Table 2 shows how many times the top five most-frequent gold, silver, and bronze badges have been awarded to users as of June 7, 2020. As expected, gold badges are relatively less frequent than silver or bronze badges.

Table 3 shows the award condition of the five most frequently awarded gold, silver, and bronze badges. The conditions are mostly related to the community members' peer evaluations on posts, such as the number of times the members have marked the posts as their favorite posts or the number of votes that posts have received from the peers. In addition, they can be awarded to users multiple times. Thus, it is reasonable to infer that users with hundreds or thousands of badges can be seen as users who have demonstrated high levels of expertise and whose past posts have been recognized favorably by peer members.

3.3 Badge Award Data

After excluding Tag badges, we observe that 78 badges[3] were awarded 8,246,889 times to

[3]As of Jan 19, 2014, 80 badges were available on SO. However, two badges, Precognitive and Constable, were not

Table 3: List of Badges Awarded Multiple Times

Level	Name	Descriptions
Gold	Famous Question	Question with 10,000 views
	Great Answer	Answer score of 100 or more
	Great Question	Question score of 100 or more
	Fanatic	Visit the site each day for 100 consecutive days
	Unsung Hero	Zero score accepted answers: more than 10 and 25% of total
Silver	Favorite Question	Question bookmarked by 25 users
	Notable Question	Question with 2,500 views
	Yearling	Active member for a year, earning at least 200 reputation
	Necromancer	Answer a question more than 60 days later with score of 5 or more
	Good Answer	Answer score of 25 or more
	Enlightened	First to answer and accepted with score of 10 or more
Bronze	Nice Question	Question score of 10 or more
	Popular Question	Question with 1,000 views
	Editor	First edit
	Student	First question with score of 1 or more
	Informed	Read the entire tour page
	Scholar	Ask a question and accept an answer

1,376,043 users from August 4, 2008 to January 19, 2014. We exclude Tag badges for the data description and analysis as identical Tag badge names were used across bronze, silver and gold badges, so we cannot identify the level of each Tag badge. In addition, since Tag badges were awarded much less often than other types of badges in our data, we believe that ignoring Tag badges does not affect main results. To illustrate, the Tag badges account for only 0.77% of our data. After excluding the Tag badges, the total number of badges in our data is 78. The data include 6,633,482 bronze badge awards, 1,454,251 silver badge awards, and 159,156 gold badge awards. In our data, the average user earned 5.993 badges, comprised of 0.116 gold, 1.057 silver and 4.821 bronze badges (Table 4). We assume gold badges are the reputation markers that would affect peer evaluation the most because they are more conspicuous to users than silver or bronze badges due to scarcity. One interesting design choice of SO is that the gold (or silver) badge icon is not displayed in the profile section if the poster does not own at least one such badge, which could make the first gold (or silver) badge award salient. Thus, our empirical specification treats the first gold badge award as the key treatment event.

awarded to any users on SO, thus, we observe 78 badges were awarded.

Table 4: Summary Statistics of the Number of Badge Awards

	Mean	Std. dev.	Min.	Max.
Gold	0.116	0.927	0	204
Silver	1.057	6.644	0	3778
Bronze	4.821	11.17	0	4951
Gold + Silver + Bronze	5.993	18.287	1	8933

3.4 Post Activity Data

From July 31, 2008, to January 19, 2014, we observe 17,564,597 posts submitted by 1,293,213 users. Among the 17,564,597 posts, there are 6,224,558 question posts and 11,340,039 answer posts. Each question post has 1.822 answer posts on average. In our data, we find that average number of total votes an answer post gets (2.274) is significantly higher than the question post (1.998), which indicates that SO users pay more attention to the quality of answers than that of questions. Thus, only the answer posts are considered for our analysis.

3.5 Peer Feedback Data

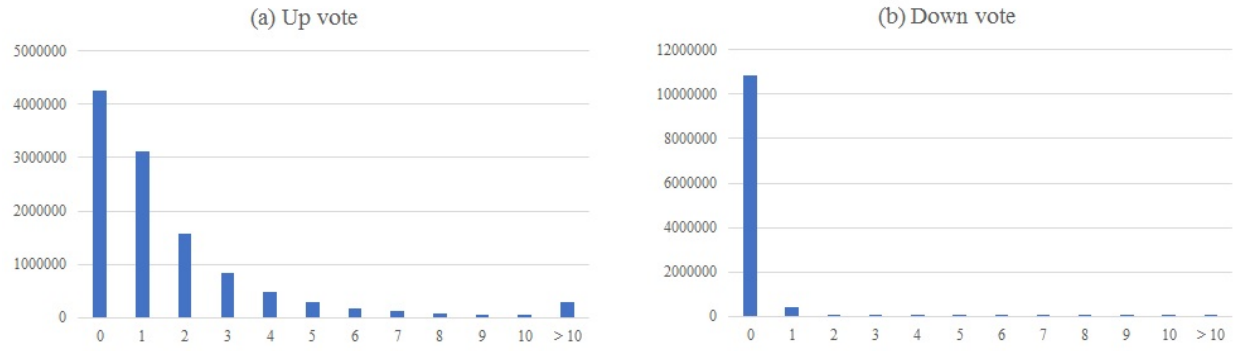
The 11,340,039 answer posts obtained 25,788,474 up or down votes. Of 11,340,039 answer posts, 4,120,050 did not receive any votes. Figure 2 presents the distribution of the total number of up or down votes. As with many other platforms, the peer evaluation at SO is overwhelmingly positive. [4] Therefore, we emphasize the impact of reputation system on up vote throughout the paper. However, for the purpose of completeness, we also examine the impact on both down votes and the combination of up and down votes (score).

3.6 Model-Free Evidence

One of the simplest ways to examine the impact of a reputation marker on peer feedback is to compare the number of votes on answer posts with a gold badge to the number of votes on answer posts without the gold badges. Figure 3 shows the comparison of the average number of up votes,

[4] In our data, only 2.79% of votes are negative. The relatively rare occurrence of peer evaluations has been observed in prior research. In the data set of Resnick and Zeckhauser (2002), only 1.2% of buyers and 0.5% of sellers at eBay left neutral or negative comments. In Dellarocas et al. (2005), the corresponding numbers are 0.7 and 0.5, respectively.

Figure 2: Distribution of Up or Down Votes (Answer Posts Only)



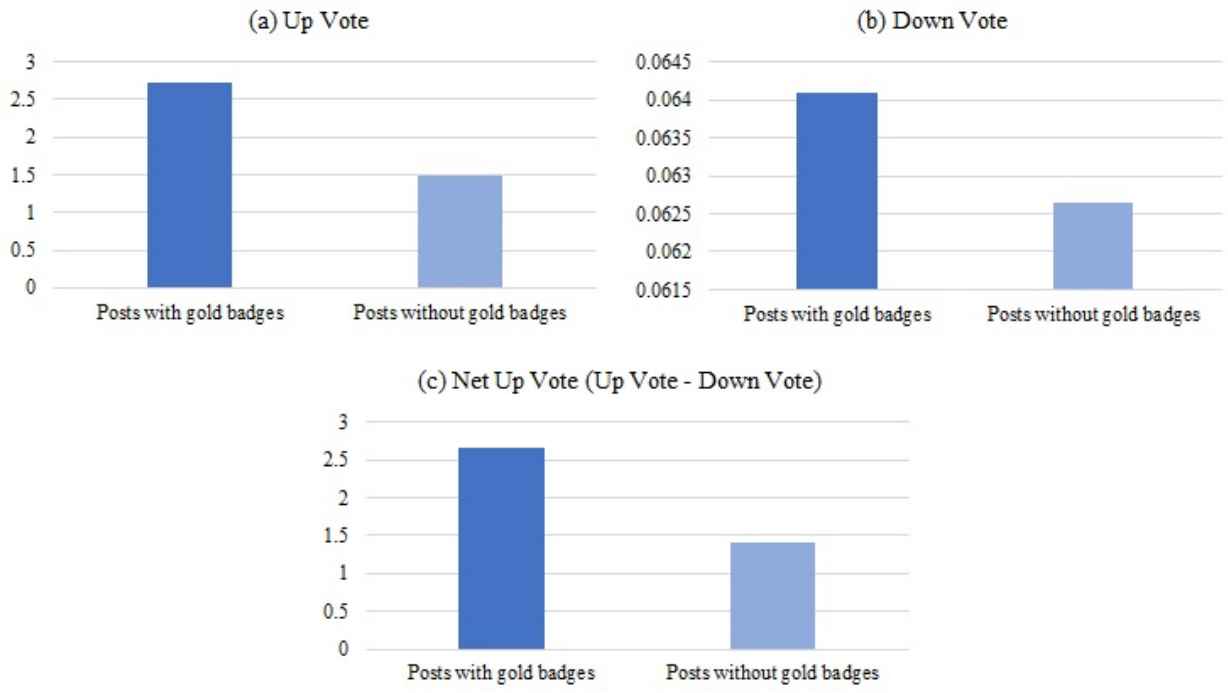
down votes, and score between posts by users with a gold badge and users without any gold badges. As can be inferred from the Figure 3, posts by users with a gold badge get significantly higher volume of net up votes as they get significantly (at the 0.01 level) more up votes from peer members in SO.

Obviously, the comparison does not control for the differences in post characteristics that could possibly affect peer evaluations. Such post characteristics may include quality, popularity, category, or textual characteristics of the answer post or those of the associated question post. For example, if a question post belongs to a popular category, the answer post can get more attention compared to the answer to a question in a less popular category, which possibly leads to high volume of peer evaluations, regardless of the gold badge presence. We must also account for the textual characteristics of the answer post because they are highly relevant to the post quality. Additionally, poster characteristics other than the gold badge presence, such as number of silver or bronze badges, because they may also affect peer evaluations. As we discuss in Section 4, the impact of such post and poster characteristics will be addressed by our difference-in-differences (DiD) model coupled with matching.

4 Difference-In-Differences Model With Matching

We examine the impact of reputation markers on peer evaluations using a difference-in-differences (DiD) model with matching. In our study, treatment posts are those with a gold badge and control

Figure 3: Average Number of Votes of Posts with a Gold Badge vs. Posts without Gold Badges



posts are those without any gold badges. The treatment date is the first gold badge award date. To ensure comparability between the treatment posts and control posts, we employed a propensity matching method coupled with the exact matching.

4.1 Difference-In-Differences Model

In our study, the DiD model compares the peer evaluations of posts in the treatment group (those that received a gold badge) with those posts in the control group over a two week period centered around the treatment date.

We define the pre- and post- treatment window as one week before - and after - the treatment date for the following reasons. First, the window has to be narrow enough to allow identification of the treatment effect amid the other poster or post characteristics that could possibly affect the peer evaluations. Second, to reliably estimate the impact of the reputation marker on peer evaluations, we need sufficient volume of evaluations during the time window. Because the daily volume of votes a post receives on SO is usually small,[5] we aggregate the number of votes a post gets during the

[5]In our data, the average daily number of up and down votes a post gets are 1.083 and 0.032, respectively.

pre-treatment week for comparison with the post-treatment week. This aggregation approach is similar to that employed in previous quasi-experiment studies. For example, Dagger and Danaher (2014) compare three months of aggregate weekly sales data in the pre-treatment period with three months of aggregate weekly sales data in the post-treatment period to examine the effect of store remodeling on new and existing customers.

Figure 4: Timeline of Pre- and Post- Treatment Window of a treatment post

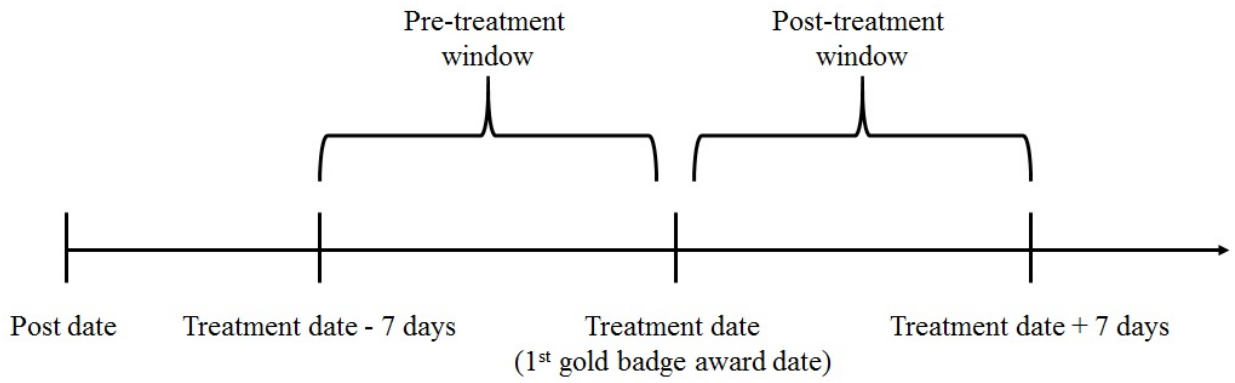


Figure 4 illustrates the time window for each treatment post in our DiD analysis. Note that, unlike in DiD model studies, our treatment date is individual-post specific. Each treatment post i is associated with pre-treatment time window interval $[b_i - 7, b_i - 1]$ and post-treatment time window interval $[b_i + 1, b_i + 7]$, where b_i represents the treatment date of post i.[6] In our data, there exist 1,681 first gold badge award dates from 1,283,173 treatment posts. In section 4.3, we address in detail how we define the treatment date for control posts.

4.2 Treatment Data Construction

Our treatment data consists of answer posts by posters who received their first gold badge at least a week after those posts were posted. This ensures they have a full one-week pre-treatment window. Note that posts whose first gold badges belong to the Answer badges are excluded from

Additionally, 36.94% of answer posts did not get any up or down votes in our data.

[6] Most DiD model studies examine treatment effects by comparing outcomes before and after a single treatment date or period. For example, Wangenheim and Bayón (2007) define a single period, June 2002, as the treatment period. They define the treatment period as June 2002, and assign the treatment period to the control units for the matching and the treatment effect estimation.

the treatment group because peer evaluations on answer posts during the pre-treatment window could directly affect the gold answer badge award. We also exclude the posts by posters who earned additional gold badges before b_i+7 (i.e., the end of the post-treatment window). Out of 73,941 users with at least one gold badge in our data, we obtained 61,824 users satisfying the aforementioned conditions. These members posted 3,988,215 answer posts in our data set. In addition, we exclude posts that did not receive at least one vote before $b_i - 7$. We impose this condition because we believe that such posts are less likely to have changes in the number of votes during the pre- and post-treatment window. Due to the answer-post sorting system at SO, the low-ranked answers may receive fewer peer evaluations over time. The site provides users with three sorting options for answer posts: by recent activity (Active), by posting date (Oldest), or by score (Votes). Although users can choose which sorting option they want to use by clicking a tab in the Answers section (Figure 1), sorting by score is set by default, so we expect that most users in SO will be exposed to high score answers prior to low score answers. Thus, the posts which have not obtained any votes have less probability of attaining a meaningful volume of votes during the pre- and post- treatment window.

Finally, to ensure that post quality remains constant over the pre- and post-treatment window, we impose the following three conditions to the treatment set: no bounty, no editing, and no comments changes during $\left[b_i - 7, b_i - 1\right]$. SO users are allowed to leave edits or comments, and edit or comment histories are shown with the post content (Figure 1). We believe that such edits or comments could also affect users' voting decision. For instance, if an answer post has been edited several times by multiple users, a user might feel that the original answer post quality is not good, thus, he might be less likely to give an up vote. To induce active answer posting, the site allows a question poster to place a bounty, a reputation point award, on its question post. The bounty is funded by points taken from the personal reputation of the question poster who offers it. After placing the bounty, the corresponding reputation score is deducted from the question poster's reputation score. When the question poster finds a satisfactory answer post and decides to offer the bounty to the answer poster, the answer poster is rewarded with the reputation score. The bounty could affect the overall size of attention drawn to the thread as the questions with an open bounty are featured by SO. Therefore, to control the effects of additional comments, edits, and bounty changes, we exclude the posts with any changes in those. After imposing all such conditions, the

final treatment data set is comprised of 1,283,173 answer posts. See Appendix A for a precise cataloguing of the treatment observations removed after each of the aforementioned steps.

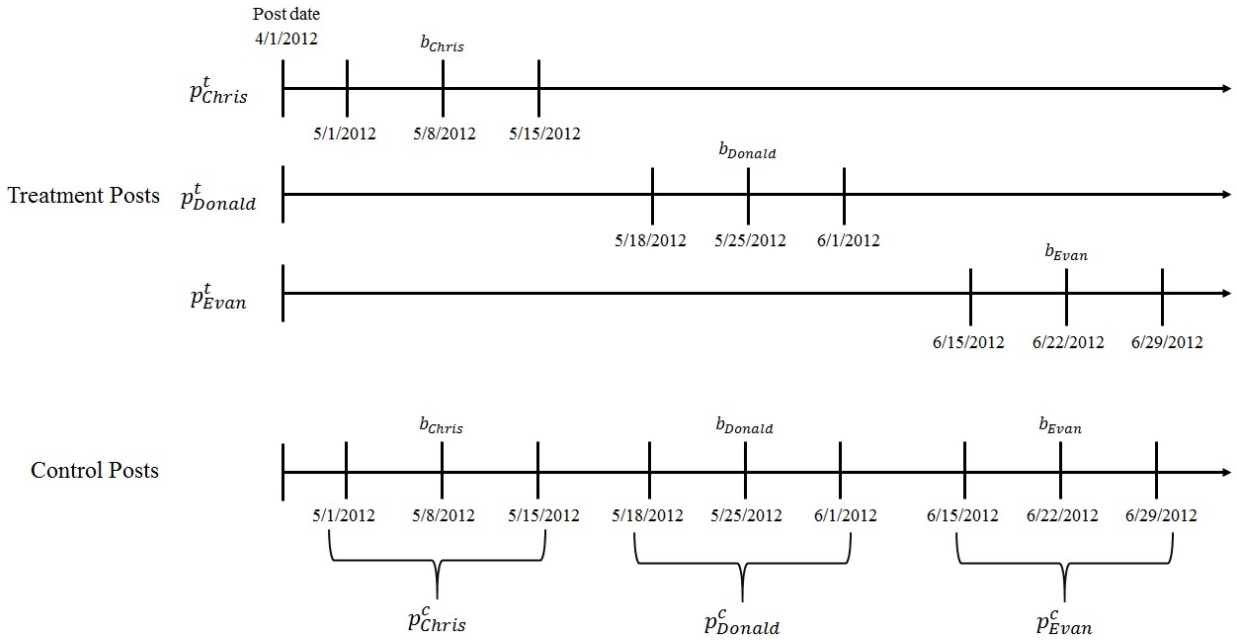
4.3 Control Data Construction

To ensure the comparability between the treatment and control groups, the construction of the control group proceeds through the following steps. First, control posts are restricted to users who were not awarded any gold badges at all. We choose this criterion because it seems to be the best counterfactual for an otherwise-identical treatment post had the latter not received a gold badge. Second, we consider only posts which were posted on the same dates as the treatment posts. This is because some potential temporal trends in the number of votes could confound our estimated treatment effect.

The main challenge associated with our quasi-experimental design arises from the fact that there exist multiple treatment dates. One question this raises is how to assign the treatment date to each control post. We first form clusters of treatment and control posts by original posting date. Then, within each posting date cluster, the list of treatment dates is gathered from all treatment posts, and we replicate the control pool by associating each control post with each treatment date of the cluster group, and treating it as a separate observation with the respective specific treatment date. In our data, there are 1,990 posting date clusters, and there are 540,284 unique combinations of posting date and treatment date. On average, posting date clusters have 345.5 treatment dates, where the number of unique treatment dates within each cluster ranges from 1 to 566.

To illustrate such control post group construction approach, we present a hypothetical example (Figure 5). For example, imagine that there are three treatment posts with the same post date of Apr 1, 2012 by three posters, Chris, Donald, and Evan. Assume that Chris, Donald, and Evan got their first gold badge on three different days, May 8, May 25, and June 22, 2012, respectively. Accordingly, the pre- and post-treatment window of Chris's post is between May 1, 2012, and May 15, 2012, that of Donald's post is between May 18, 2012, and June 1, 2012, and that of Evan's post is between June 15, 2012, and June 29, 2012. Suppose that we find another answer that was posted on Apr 1, 2012 by a user without any gold badge. Then, by mapping the three pre- and post-treatment windows from the three treatment posts on the time span of the control post, we form three control posts, $p^c_{Chris}, p^c_{Donald}$ and p^c_{Evan} in this posting date cluster. The details

Figure 5: Control post group construction example



of the procedure are described in Appendix B. Note that the same post can be used multiple times when constructing control post group if the post belongs to a posting date clusters with multiple treatment dates. Also, if multiple treatment posts were posted on the same date and their posters were awarded the first gold badge on the same date, the same control post - treatment date combination can be matched to the multiple treatment posts. As discussed later in the Section 6, the main results are not affected when we restrict all control candidates to a single unique control in our estimation.

For the same reason explained with respect to the treatment data, we require control posts to have received at least one vote prior to $b_i - 7$. They must also not have any edits, comments, or bounty changes during $[b_i - 7, b_i - 1]$. After the implementation of the aforementioned steps, we begin with the set of 630,391,387 candidate posts in the control pool based on 2,620,025 unique posts. In the following section, we discuss how we use matching methods to identify the right controls for each treatment from the pool of candidates. Again, we present in Appendix A the number of control poster and posts after each of the steps.

4.4 Matching Method

The challenge in successfully estimating treatment effect using DiD model lies in identifying a control group that is as comparable as possible to the treatment group (Avery et al. 2012) to minimize the endogenous treatment effects in non-experimental settings. To find such a control group, matching has frequently been used in social science and marketing studies (Bell et al. 2017; Xu et al. 2017; Avery et al. 2011; Bronnenberg 2010; Garnerfeld 2013; Montagunti et al. 2016; Wangenheim and Bayón 2007; Warren and Sorescu 2017). We use a combination of two commonly-used matching methods: exact matching and propensity score matching (Caliendo and Kopeinig 2008; Rubin and Thomas 2000).

The idea behind exact matching is to match each treatment unit with control units for which all the values of matching variables are exactly identical. Exact matching has the advantage of ensuring that treatment objects are paired on key variables of interest. The most common issue when implementing exact matching is that increasing the number of matching variables to improve the precision of matching increases the chance of excluding treatment observations for which one cannot find a match, thereby reducing the sample size as well as variation among the treatment objects (Stuart 2010). However, we believe that this is unlikely to be a major concern in our case because there are sufficient control posts for each treatment post. In fact, implementing exact matching is advantageous in our study because it helps to identify a manageable set of control posts for each treatment post. Moreover, exact matching can mitigate the possible limitations of propensity score matching (PSM) in our study. As suggested by Caliendo and Kopeinig (2008), it may be best to restrict the PSM model to variables that influence simultaneously the treatment decision and the outcome variable. according to this criterion, as the treatment is the gold badge award for a poster, in our empirical setting, the propensity score model covariates should, be poster-level characteristics only. However, due to the fact that we believe that both poster- and post-level characteristics can affect peer evaluations, including only the poster-level characteristics in the PSM may not reflect all possible factors affecting the outcome variable, which is a post-level variable. As it is not uncommon to use the propensity score approach within sub-populations generated by exact matching (Heckman et al. 1997; Heckman et al. 1998), we use post-level characteristics for exact matching, then find the most-comparable control posts based on PSM.

Before implementing exact matching and PSM, we first need to pair treatment posts with posts in the control pool. When there are multiple treatment posts with the same posting and treatment date, a post in the control pool with the same posting and treatment date is paired with each of the treatment posts for the matching. In addition, when there are multiple posts in the control pool with the same posting and treatment date, then treatment post is paired with each of the posts in control pool for the matching.

To reduce the possibility of violation of the stable unit treatment value assumption (SUTVA) (Rubin 1978), which states that the outcome of one unit is not affected by the treatment assignment of any other individuals, we exclude treatment - control post pairs whose control post belongs to the same thread as the treatment post. That is, the treatment and control must not be answers to the same question. The number of control posts paired up with a treatment post ranges from 3 to 3,155, and the average number is 1,215.

We next match exactly treatment and control posts based on the following three post-level covariates: tag, question post score, and the order of a post in the thread. A tag is a word or phrase that describes the topic of the question, which can be used to help users to identify post content category[7]. We match each treatment to control posts for which the first five tags are identical to prevent post content category characteristics from affecting treatment impact estimate. Question post score and the order of a post in the thread as of one day prior to $b_i - 7$ were also included to rule out potential effects of popularity or quality on treatment impact estimate. After the exact matching, 9,102,376 control posts (treatment - control pair) matched up with 822,912 treatment posts. Each treatment post is paired up with 11.06 number of control posts on average, and the 9,102,376 control posts are from 1,443,813 unique control posts.

Even after reducing potential differences across the treatment and control posts with exact matching, we further rely on PSM to identify control posts which are as identical as possible to treatment posts in terms of observed poster characteristics. As mentioned in 3.2, SO awards gold badges to users based on the volume and quality of their participation. The resultant selection effects appear in Table 5, which shows basic descriptive statistics of poster characteristics. Table 5 demonstrates that the treatment posts exhibit poster-level characteristics that are substantially different from those of the control posts even after exact matching. The poster-level characteris-

[7]https://stackoverflow.com/help/tagging (Oct 13, 2016)

Table 5: Descriptive Statistics after Exact Matching

		Treatment				Control			
	Difference in means	Mean	Std. dev.	Min.	Max.	Mean	Std. dev.	Min.	Max.
Activity level	0.670**	2.756	1.398	0	4	2.086	1.555	0	4
SO membership duration	133.728**	739.911	425.708	1	1971	606.183	407.657	0	1994
Super user indicator	0.161**	0.339	0.474	0	1	0.179	0.383	0	1
Active member indicator	0.058**	0.985	0.123	0	1	0.927	0.261	0	1
The number of silver badges	6.362**	8.921	7.31	0	72	2.558	3.758	0	81
The number of bronze badges	15.070**	26.02	15.498	0	175	10.95	8.088	0	171
The number of answer posts	226.931**	306.002	390.627	1	3012	79.071	136.3	1	2238
The number of question posts	21.891**	28.852	45.185	0	986	6.961	13.031	0	433
No. obs.		822,912				9,102,376			

Note. There is a significant difference ($p < 0.05$) between treatment and control with **

tics possibly affecting the treatment membership, gold badge award, are significantly greater on average for treatment posts compared to the control posts. Thus, we further refine the controls by implementing PSM.

PSM begins with an estimation of a model of treatment group membership. Then, using the fitted probability of membership, the researcher creates a control group in which each treatment unit is matched to similar control unit(s). This is meant to reduce the impact of confounding variables that could bias an estimate of the treatment effect obtained from simply comparing treatment versus control groups (Rosenbaum and Rubin 1983). PSM has been widely adopted by diverse marketing studies (Avery et al. 2012; Bell et al. 2017; Bronnenberg et al. 2010; Garnerfeld et al.2013; Montaguti et al. 2016; Wangeheim and Bayón 2007 ; Xu et al. 2017). Avery et al. (2012) use PSM to understand how retail store openings affect direct channel sales. They compare the observations in four treatment regions where the retail stores opened with observations in control regions before and after the stores opened. To examine how offline-showrooms benefit demand generation and operational efficiency of online-first retailers, Bell et al. (2017) define their treatment group as a 30 mile radius from the location of the showroom, and estimate a DiD model by including weights calculated using PSM. PSM has been used in field experiment studies as well. Bronnenberg et al. (2010) use the method to estimate each household's probability of receiving the TiVo treatment. Garnerfeld et al. (2013) calculate the propensity for participation in a firm's customer referral program, and Montaguti et al. (2016) estimate a customer's likelihood of becoming a multi-channel customer to account for self-selection. Wangeheim and Bayón (2007) study the long-term behavioral and monetary effects of upgrades, downgrades, and denied boarding after matching recipients of these three events with nonparticipants according to a propensity score.

In Xu et al. (2017), treatment users are those who adopt a tablet and control users are those who did not adopt a tablet, and PSM is used to identify the causal impact of tablets on e-commerce sales. For binary treatments like ours, the logit or probit models are most commonly used to estimate the propensity score (Caliendo and Kopeinig 2008). We estimate a logit model of gold-badge award on the following poster-level characteristics; the number of user profile information sections completed, SO membership duration, super user indicator, indicator of whether the user was active as of the treatment date, the number of silver or bronze badges as of the treatment date (badge award date), and the number of answer or question posts submitted by the poster up to the treatment date. [8]

The propensity score logit regression model estimates are shown in Table 6. We consider the model estimated both on main effects only and with interaction effects included. The substantive findings are robust to these alternative approaches. As the former model has a better fit, we report only the matching and DiD estimation results based this on model. [9] [10] The estimated coefficients overall have intuitive interpretations because most of them have a positive effect on the treatment membership.

With the propensity score estimated from the logit model, we first check the common support region of the score between treatment and control pools. After removing those outside the common support, our treatment data has 822,912 posts and the control data has 9,923,174 posts. Then, conditional on the propensity score, we compare a diverse set of matching algorithms and select the best set of matches. The matching algorithms we consider are nearest neighbor matching with or without replacement (Cochran and Rubin 1973) and genetic matching (Diamond and Sekhon 2005; Abadie and Imbens 2006). Also, to minimize the risk of inferior matches when the score of a selected control is far away from that of a matched treatment, we impose a tolerance level on the maximum propensity score distance - known as a "caliper" - on the nearest neighbor and genetic

[8] If a post has a small order of a post variable, it means that the post is ranked highly in the belonging thread. For example, if the order of a post is 1, the post is on the top of the list of its answer thread. This may seem counter-intuitive measure. However, considering the number of answer posts varies across threads, it may be preferable to assign smaller value to the highly ranked posts. Otherwise, a top answer post with a large number of competing answer posts could have a much bigger value compared to a top answer post with a small volume of competing posts.

[9] Our DiD model results are robust to the inclusion of the interaction terms in the propensity score model. These are available from the authors, upon request.

[10] For the logit model estimation, we ignored the treatment - control post pairing to eliminate duplicate observations from the control posts. When there are multiple treatment posts with same posting and treatment date, a post in the control pool can be used to make several treatment - control post pairs. After ignoring treatment control post pairing, the number of observations from control posts are reduced, thus, the number of observations used for logit model estimation (8,899,781) is smaller than the total number of observations after exact matching (9,925,288). There is little or no difference in propensity score and DiD model results when such duplicated rows are not eliminated.

Table 6: Results of Propensity Score Logistic Regression Analysis

	(1) Main effects only	(2) Interaction effects included
Intercept	-4.4561***	-7.7161***
	(0.0103)	(0.0352)
Activity level	0.1340***	0.0000
	(0.0010)	(0.0081)
SO membership duration	-0.0011***	0.0023***
	(0.0000)	(0.0000)
Active member indicator	0.7252***	2.4067***
	(0.0101)	(0.0350)
Super user indicator	0.6047***	0.7027***
	(0.0033)	(0.0371)
The number of silver badges	0.0561***	0.2055***
	(0.0006)	(0.0047)
The number of bronze badges	0.0571***	0.1956***
	(0.0003)	(0.0026)
The number of answer posts	0.0014***	-0.0027***
	(0.0000)	(0.0001)
The number of question posts	0.0231***	0.0273***
	(0.0001)	(0.0006)
Num. obs.	8,899,781	
RMSE	0.675	0.642

Note. ***p < 0.01. For ease of exposition, coefficients of the interaction terms are not reported here. The full list of coefficients is available from the authors, upon request.

matching (Cochran and Rubin 1973; Diamond and Sekhon 2005). Although there is no universal agreement on the caliper width selection, 0.25 standard deviation of estimated propensity score is selected in our study because it is most commonly recommended in the literature (Hoe et al. 2007). The nearest neighbor matching algorithm aims to find posts in the control group that are closest to a post in the treatment group with respect to the probability of a post getting a gold badge. The genetic matching is a generalization of the Mahalanobis distance, and it searches a range of distance metrics to find the particular measure that optimizes post-matching covariate balance. Formally, the genetic matching distance (GMD) between the X covariates for two units i and j is

$$GMD(X_i, X_j, W) = \sqrt{(X_i - X_j)^T (S^{-1/2})^T W S^{-1/2} (X_i - X_j)} \tag{1}$$

where $S^{-1/2}$ is the Cholesky decomposition of the sample covariance matrix of X, X^T is the transpose of the matrix X, and W is $k \times k$ positive definite weight matrix. Each potential distance metric considered corresponds to a particular assignment of weights W for all matching variables. The algorithm weights each variable according to its relative importance for achieving the best overall balance (Diamond and Sekhon 2005). Generally, if a good propensity score model is available, it is recommended that the genetic matching should start with propensity score (Diamond and Sekhon 2005), or the score is included as one of the covariates in the matching model (Sekhon 2011). Hence, three sets of covariates are considered in our genetic matching model: (1) propensity score only, (2) propensity score and the poster-level covariates, and (3) poster-level covariates only. However, after imposing the caliper, genetic matching based on (2) and (3) results in the exclusion of a majority of treatment posts. Thus, we do not present results using this method.[11]. For each of nearest neighbor and genetic matching model, we also consider a range of ratios between control posts and each treatment post, including 1:1, 2:1, or 3:1. The substantive findings are robust to the diverse ratios between control posts and treatment posts.

To find the best matching algorithm, we evaluate each of these by computing the PRB, percentage reduction in bias (Rosenbaum and Rubin 1984; Wangenheim and Bayón 2007). The PRB is the percentage decrease in the mean difference of each covariate between a treatment unit and its

[11]The results of matching and DiD model of these two matching algorithms are available from the authors, upon request.

matching controls. Cochran and Rubin (1973) suggest that the most satisfactory matching algorithm leads to a PRB value of 80% or higher for the covariates, and we follow the guideline to find the best matching method. After comparing the PRBs of each matching algorithm, we find that the matching algorithms with the caliper produce a reasonably good fit as expected.[12] In Table 7, we report the PRB of those matching algorithms. Among the matching algorithms with caliper, 1:3 nearest neighbor matching was selected for the main analysis for the following reasons: 1) The typical disadvantage of caliper matching, reducing the size of matched treatment posts, is minimal; 2) The PRBs for most covariates from the 1:3 nearest neighbor matching are, over or close to, 80%. [13]

4.5 DiD Model Specification

We estimate a linear DiD model with peer evaluation of an answer post as the dependent variable. We index posts by i and the time by $t = 0, 1$ where $t = 0$ for the pre-treatment week and $t = 1$ for the post-treatment week. We denote the outcome variable, peer evaluation, as $Y_{it} \in \{UpVote_{it}, DownVote_{it}, Score_{it}\}$. As mentioned in Section 3.5, among 10 types of peer evaluations available at SO, up and down vote are the most common ones. Thus, to examine how a reputation marker affects peer evaluation, we consider the number of up votes ($UpVote_{it}$) and down votes ($DownVote_{it}$) as the outcome variables. In addition to up and down votes, we also consider score ($Score_{it}$), the number of up votes minus that of down votes, as another outcome variable. This is because the increase or decrease in the number of up or down votes separately may not necessarily indicate clearly the direction of the impact of the system. For example, when there exists a significant increase in up vote among treatment posts comparing to control posts, the increase in up vote may not support the idea of the Matthew Effect if down votes also increase. Thus, we use score as an additional dependent variable to test whether the reputation system induces the

[12] For matching algorithms without caliper imposed, PRBs of activity level, SO membership duration, super user indicator, and active member indicator are mostly around, or greater than 80%. However, PRBs of the number of silver badges, bronze badges, answer posts, and question posts are only about 20 to 30 %. The PRB values of these matching algorithms are available from the authors, upon request.

[13] Thoemmes and West (2011) find that, for clustered data, propensity score models that consider the clustered nature both in estimation of propensity score and matching conditioning on the propensity score perform the best in simulation study. Hence, after grouping each treatment post and its matching control posts by exact matching in a cluster, we estimated fixed effects regression models in which single level logistic regression is used for each individual cluster. However, after imposing caliper, only 42 treatment posts can be matched to control posts, and thus, we do not report the results.

Table 7: Percentage Reduction in Bias of Matching with Caliper

Ratio (Treatment : Control)	Nearest Neighbor			Nearest Neighbor Without Replacement			Genetic		
	1:1	1:2	1:3	1:1	1:2	1:3	1:1	1:2	1:3
Activity level	0.597	0.669	0.720	0.629	0.753	0.849	0.597	0.669	0.720
SO membership duration	0.855	0.756	0.693	0.785	0.664	0.592	0.855	0.753	0.690
Active member indicator	0.884	0.924	0.953	0.894	0.956	0.981	0.884	0.925	0.954
Super user indicator	0.536	0.648	0.722	0.588	0.728	0.817	0.536	0.650	0.723
The number of silver badges	0.834	0.789	0.762	0.822	0.772	0.743	0.834	0.789	0.762
The number of bronze badges	0.897	0.856	0.829	0.888	0.836	0.804	0.900	0.856	0.8289
The number of answer posts	0.886	0.843	0.816	0.883	0.835	0.806	0.886	0.844	0.817
The number of question posts	0.966	0.930	0.907	0.957	0.917	0.892	0.966	0.929	0.906
Number of Matched Treatment Posts	295,501	295,501	295,501	263,582	249,332	240,396	295,501	295,501	295,501

Matthew Effect on peer evaluation.

For post i at time t, the model is as follows:[14]

$$Y_{it} = \alpha_0 + \alpha_1 \cdot GB_i + \alpha_2 \cdot After_t + \lambda \cdot GB_i \cdot After_t + X_{it} \cdot \beta + \varepsilon_{it} \tag{2}$$

where α_0 represents the average volume of peer evaluations for the posts, μ_i is a post and poster fixed effect accounting for the unobservable post or poster characteristics, and ε_{it} is an idiosyncratic error. GB_i (for "Gold Badge") is 1 if post i belongs to the treatment group. $After_t$ is 1 when the time t indicates the post treatment week. This variable accounts for potential temporal factors that may simultaneously affect peer evaluations across various posts. Notably, on average, we expect the volume of feedback to decline over time. The main focus of the DiD model is the interaction between the group and time indicator, $GB_i \cdot After_t$, capturing how peer evaluations in the treatment group change after the gold badge award in contrast to that of control group during the same period. Thus, the significance and the sign of λ allow us to examine whether the reputation marker impacts the peer evaluations. X_{it} represents a vector of control variables including average number of silver or bronze badges presented during each week, average question score, average number of answer posts in the thread, and average ranking of the post in its question thread during week t. One may argue that the effect of four control variables (the number of silver/bronze badges, question post score, and order of a post) should be eliminated in the process of matching and differencing out the equation. However, as the variables used in the matching process do not reflect the changes in these variables

[14] As the key implicit assumption in linear regression is that the dependent variable is continuous, the Poisson may seem more appropriate functional forms for the discrete outcome variables in our study. However, since one of the outcome variables, score, can range from negative to positive integers, we use linear regression for the main analysis in order to have comparable results using the same model assumptions.

during the pre- and post- treatment window, even after matching, there could be remaining effects of the variables on peer evaluation.[15] These four control variables are measured by calculating the average of the observed value on the first and last day of pre- or post-treatment window. One thing to note is that we account for the matched pairs of a treatment and control post(s) by estimating the 2nd difference of the matching treatment and control instead of (2) directly. Specifically, we first difference out (2) within each post i , so that we have the first difference equation for each i. Then, we take the difference between the first difference of a treatment post and that of a matching control post. Using such differenced equations is also advantageous in that time-invariant factors within the individual post do not need to be estimated.

5 Results

The central question of this study is whether the observed existence of a reputation marker (a gold badge) has a causal effect on the evaluation of the individual post. Before showing the results of the DiD model, we present model-free evidence (see Table 8) by comparing mean values of outcome variables between treatment and control posts and between pre- and post-treatment period. The most notable finding is that the average volume of peer evaluations is small even after the weekly aggregation. Thus, we expect that the size of the 2nd difference estimator will be small even if it is statistically significant. Overall, we find that the number of up votes treatment posts got during pre- and post-treatment period is higher than the number of up votes control posts got. Although the up votes received by treatment posts stayed roughly the same in the pre- and post-treatment period, the up votes that the control posts received during the post-treatment period declined in the pre-treamtent window. Accordingly, the overall DiD value of up votes is statistically significant and positive. These sets of results may indicate that posts with reputable badges suffer less from the temporal decline in peer evaluation. Next, we present the formal results of the DiD model which accounts for the impact of control variables on the peer evaluations.

Table 9 shows the estimation results of the DiD model of up vote (Column 1), down vote (Column 2), and score (Column 3) under nearest neighbor matching with a 1:3 ratio. The key

[15] The numeric reputation score (Figure 1) might also be one of the control variables as the reputation score is shown next to the number of silver and bronze badges. However, we did not consider the reputation score in our DiD model as the reputation score is highly correlated with the number of gold, silver, and bronze badges by its definition.

Table 8: Comparison of Peer Evaluations Between Treatment and Control Posts

Outcome	Groups	Pre-treatment	Post-treatment	Difference	DiD
Up vote	Treatment posts	7.38e-03	7.56e-03	1.86e-04 (2.38e-04)	7.17e-04***
	Control posts	6.64e-03	6.11e-03	-5.31e-04*** (1.50e-04)	
Down vote	Treatment posts	1.30e-04	1.50e-04	2.03e-05 (3.134e-05)	4.58e-05
	Control posts	1.70e-04	1.50e-04	-2.55e-05 (2.28e-05)	
Score	Treatment posts	7.25e-03	7.41e-03	1.66e-04 (2.40e-04)	6.71e-04**
	Control posts	6.47e-03	5.96e-03	-5.06e-04*** (1.51e-04)	

Note. **p < 0.05; ***p < 0.01.

result in Table 9 is λ, the 2nd difference estimators, and λ for up votes and score are positive and significant. This suggests that, all else being equal, posters receive disproportionately-higher evaluations of their posts after they earn a reputation marker, irrespective of post quality. In addition, the positive and significant coefficient for the score indicates that a post will be evaluated more favorably by peer-members when its poster has a gold badge. Even though the magnitude of the 2nd difference estimators seems small, given that our model is a DiD linear model, the significance and sign of the estimator has a more meaningful interpretation.[16] In fact, our analysis suggests that after a user earns a reputation badge, there is a 10.06% and 9.59% increase in up votes and score on user's posts, respectively.[17]

From Table 9, we can also find some other factors affecting peer evaluations of posts. As expected, the average question score affects peer evaluations positively since more attention from SO users would be drawn to the entire thread as the question post gets popular. The negative coefficient of the number of answer posts suggests when it has more competition for peer evaluations among answer posts in a thread, then answer posts will get less positive peer evaluations.

Having established the impact of reputation badges on peer evaluation, we next examine whether

[16] Our DiD model does not take into account the possibility that multiple posts were posted by the same user, thus, poster level fixed or random effects were not considered in the model. This is because the 1st difference removes the impact of time-invariant individual poster level. With respect to the impact of time-varying individual poster effect, we believe that, as we used the short period of window, the impact would be trivial.

[17] To calculate the increase percentage (i.e. mean value of dependent variable divided by estimated value of λ), we estimate (2) instead of the the differenced equation.

Table 9: Difference-in-Differences Model Estimation Results

	1:3 Nearest Neighbor with Caliper		
	(1)	(2)	(3)
	Up	Down	Score
λ(2nd Difference Estimator)	7.43e-04***	2.53e-05	7.17e-04***
	(2.12e-04)	(3.12e-05)	(2.15e-04)
The number of silver badges	-3.61e-04	1.74e-05	-3.78e-04
	(4.22e-04)	(6.19e-05)	(4.26e-04)
The number of bronze badges	2.70e-04	-9.99e-06	2.80e-04
	(1.84e-04)	(2.70e-05)	(1.86e-04)
Question score	1.19e-01***	2.38e-03***	1.17e-01***
	(2.16e-03)	(3.17e-04)	(2.18e-03)
The number of answer posts	-2.68e-02***	2.86e-03***	-2.96e-02***
	(4.30e-03)	(6.32e-04)	(4.35e-03)
Ranking of the post	3.64e-01***	5.45e-02***	3.09e-01***
	(7.14e-03)	(1.05e-03)	(7.21e-03)
Num. obs.		626,934	
RMSE	0.167	0.027	0.169

Note. **p < 0.05; ***p < 0.01.

the impact varies with observable poster or post characteristics. In particular, we interact treatment effect with the number of silver badge and bronze badges, question score, number of answer posts and ranking of the post as of a day before treatment date. Table 10 presents the estimates of the treatment effect and the interaction of the poster and post characteristics.

For Mod_i,

$$
\begin{aligned}
&Y_{it} \\
&= \alpha_0 + \alpha_1 \cdot GB_i + \alpha_2 \cdot After_t + \alpha_3 \cdot Mod_i + \beta_1 \cdot GB_i \cdot After_t + \beta_2 \cdot GB_i \cdot Mod_i + \beta_3 \cdot After_t \cdot Mod_i \\
&\qquad + \lambda \cdot GB_i \cdot After_t \cdot Mod_i + X_{it} \cdot \Gamma + \varepsilon_{it}
\end{aligned} \tag{3}
$$

In Table 10, when the moderating effect of poster characteristics (silver and bronze badge) taken into account, we find that the main treatment effect is significant for up vote and score. The moderating negative effect suggests that the magnitude of the first gold badge effect is weakened by the increase in other reputation markers. The linear combination of the main and interaction

Table 10: Moderating Effect of Post and Poster Characteristics

			(1) Up	(2) Down	(3) Score
Poster characteristics	Silver badge	Main effect	1.08e-03***	3.05e-05	1.05e-03***
			(3.80e-04)	(5.59e-05)	(3.84e-04)
		Interaction effect	-6.35e-05	-5.31e-06	-5.82e-05
			(7.39e-05)	(1.09e-05)	(7.37e-05)
	Bronze badge	Main effect	1.52e-03***	8.44e-05	1.43e-03**
			(5.56e-04)	(8.17e-05)	(5.62e-04)
		Interaction effect	-4.75e-05	-4.51e-06	-4.30e-05
			(3.44e-05)	(5.06e-06)	(3.48e-05)
Post characteristics	Question score	Main effect	3.69e-04	3.31e-05	3.36e-04
			(2.43e-04)	(3.70e-05)	(2.47e-04)
		Interaction effect	7.45e-04***	-6.87e-06	7.52e-04***
			(1.64e-04)	(2.48e-05)	(1.66e-04)
	Answer posts	Main effect	5.04e-04	-1.00e-04	6.04e-04
			(4.86e-04)	(7.14e-05)	(4.91e-04)
		Interaction effect	1.42e-04	5.21e-05**	8.94e-05
			(1.80e-04)	(2.65e-05)	(1.82e-04)
	Ranking	Main effect	1.26e-03**	7.98e-05	1.19e-03**
			(5.26e-04)	(7.73e-05)	(5.31e-04)
		Interaction effect	-3.43e-04	-4.25e-05	-3.00e-04
			(3.87e-04)	(5.69e-05)	(3.91e-04)

Note. **p < 0.05; ***p < 0.01.

effect shows that the first gold badge has a significant and positive effect on peer evaluations with a small number of silver or bronze badges. This is in line with the findings by Azoulay et al. (2013), which states that the magnitude of status shock is greater for the ones who is less renowned in the field.

When interaction effect of post characteristics is included, most of the treatment and interactions are statistically insignificant. However, the linear combination of the main and interaction effects shows the heterogeneity in the treatment effect.

6 Robustness Check

To validate our main results, we conducted a series of robustness checks. Specifically, we check whether the results are robust to alternative matching methods, caliper sizes, DiD model assumption, window lengths, clustered errors, and the exclusion of replicated controls. Additionally, by conducting a lab experiment, we could not only replicate our results in a controlled setting but also understand the underlying psychological mechanism driving our results.

As the first robustness check, we assess the sensitivity of results to different matching methods and matching ratio. Table 11 shows the DiD model 2nd difference estimator results under diverse matching algorithms and ratios. Although some simulation studies (Barabas 2004; Heckman et al. 1998) report that various matching algorithms all yield comparable results, the composition of constructed control groups may vary across algorithms as each differs in the weights attached to control group members. Thus, we want to ensure that the results are not sensitive to the choice of matching algorithms. Under different matching algorithms abd ratios, we find that the 2nd difference estimators are generally positive and significant. When score is the dependent variable, nearest neighbor matching with replacement produces estimates in the range of 6.13e-04 to 7.17e-04, nearest neighbor matching without replacement produces estimates in the range of 6.30e-04 to 8.82e-04, and genetic matching produces estimates in the range of 5.26e-04 to 7.22e-04.[18]

As the second robustness check, we impose various sizes of caliper on the matching to assess the robustness of the results with respect to different matched samples. Our baseline matching utilizes one-to-three nearest neighbor matching to derive the closest matched control post(s), under

[18] All matching algorithms reported in Table 10 are the ones with caliper imposed. The matching algorithms without caliper imposed show qualitatively similar results, which are available from the authors, upon request.

Table 11: 2nd Difference Estimator under Diverse Matching Algorithms

	Matching algorithm	Outcome variable	Ratio (Treatment: Control) 1:1	1:2	1:3
(1)	Nearest Neighbor With Replacement	Up	6.83e-04**	7.30e-04***	7.43e-04***
			(3.17e-04)	(2.44e-04)	(2.12e-04)
		Down	7.00e-05	2.00e-05	2.53e-05
			(4.52e-05)	(3.56e-05)	(3.12e-05)
		Score	6.13e-04**	7.10e-04***	7.17e-04***
			(3.20e-04)	(2.46e-04)	(2.15e-04)
(2)	Nearest Neighbor Without Replacement	Up	9.11e-04***	7.80e-04***	7.02e-04***
			(3.41e-04)	(2.74e-04)	(2.49e-04)
		Down	2.88e-05	2.38e-05	7.15e-05**
			(4.87e-05)	(3.96e-05)	(3.59e-05)
		Score	8.82e-04***	7.57e-04***	6.30e-04***
			(3.44e-04)	(2.77e-04)	(2.52e-04)
(3)	Genetic Matching	Up	5.92e-04*	7.17e-04***	7.49e-04***
			(3.17e-04)	(2.44e-04)	(2.12e-04)
		Down	6.64e-05	7.17e-04***	2.68e-05
			(4.53e-05)	(2.44e-04)	(3.11e-05)
		Score	5.26e-04*	6.99e-04***	7.22e-04***
			(3.20e-04)	(2.46e-04)	(2.15e-04)

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

a caliper size 0.25 times the standard deviation of the propensity score. However, as Smith and Todd (2005) note, a possible drawback of caliper matching is that it is difficult to know a priori what choice for the tolerance level is reasonable. Imposing an appropriate caliper enables researchers to avoid bad matches and, thus, the match quality rises. The decrease in caliper size can improve the matching quality, but at the same time, it can decrease the number of treatment units paired with control units (Wagenheim and Bayon 2007). In our case, only 36% treatment posts were matched to control posts under the 0.25 of standard deviation caliper. Thus, we examine the robustness of our findings under a wide range of caliper size, specifically, from 0.01 to 0.99 of standard deviation of propensity score with an increment of by 0.005. The DiD model estimation results using this wide range of calipers show that our main findings are robust to diverse caliper sizes. The 2nd difference estimator for score ranges from 4.95e-05 to 1.76e-03 and the estimator is significant under all range of caliper size. For details of the 2nd difference estimator, see Appendix C.

As the third robustness check, we assess the sensitivity of the results with respect to model assumption. In (2), the error term of DiD models are assumed to follow normal distribution.

Table 12: Difference-in-Differences Poisson Model Estimation Results

	(1) Up	(2) Down
(Intercept)	-4.32***	-8.72***
	(3.66e-02)	(2.09e-01)
GB_i	1.31e-01***	-2.16e-01
	(2.68e-02)	(1.90e-01)
$After_t$	-8.37e-02***	-1.56e-01
	(2.24e-02)	(1.41e-01)
$GB_i \cdot After_t$	1.13e-01***	3.13e-01
	(3.75e-02)	(2.60e-01)
The number of silver badges	-1.29e-02***	-2.92e-02
	(4.36e-03)	(3.06e-02)
The number of bronze badges	-9.92e-03***	-1.93e-02
	(2.13e-03)	(1.45e-02)
Question score	1.60e-01***	1.20e-01***
	(2.00e-03)	(1.72e-02)
The number of answer posts	-1.89e-02***	5.35e-02
	(5.78e-03)	(3.28e-02)
Ranking of the post	-5.35e-01***	1.07e-01
	(2.30e-02)	(9.61e-02)
Num. obs.	1,844,870	
Log Likelihood	-73,293.91	-2,769.847

Note. *p ¡ 0.1; **p < 0.05; ***p < 0.01.

However, as the dependent variables, the number of up and down votes, can be interpreted as discrete variables, traditional linear regression model would not be appropriate. Thus, a Poisson regression model is estimated for up and down votes. We are not able to estimate a Poisson regression model for score since the score ranges from -2 to 9 in our data. Also, in order to estimate the Poisson regression, dependent variables should be non-negative and integer-valued variables, we directly estimate (2) instead of the differenced equation. Table 12 reports the Poisson model estimation results, and λ, the coefficient of $GB_i \cdot After_t$, suggests the effect of first gold badge on peer evaluation. The table indicates that the main results still hold under Poisson distribution assumption.

As the fourth robustness check, we assess the sensitivity of our estimation results to assumptions on time windows. Specifically, we repeat the regressions using samples that fall within three, five, ten, or fourteen days before and after the first gold badge award date. We report the 2nd difference estimates in Table 13. The results of this analysis produce positive and significant estimates,

Table 13: Robustness Check on Time Window Length

	Outcome variable		
	(1) Up	(2) Down	(3) Score
3 days	1.48e-03***	5.63e-05**	1.43e-03***
	(1.79e-04)	(2.73e-05)	(1.81e-04)
5 days	1.21e-03***	3.58e-05	1.18e-03***
	(1.95e-04)	(2.93e-05)	(1.98e-04)
10 days	3.82e-04*	-8.86e-06	3.91e-04*
	(2.34e-04)	(3.50e-05)	(2.37e-04)
14 days	-5.11e-04**	-2.20e-05	-4.89e-04*
	(2.66e-04)	(3.77e-05)	(2.69e-04)

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

suggesting that the main results are not affected by the length of time window.

As the fifth robustness check, we cluster errors at the post level to account for autocorrelation in our data (Bertrand et al. 2004). Even though the use of two periods (before and after) and the use of differenced equation mitigate potential serial correlation problems, we assess whether the DiD estimator itself produces spurious significant results due to the possible biases in the estimated standard error by clustering our standard errors by the treatment post. Table 14 reports the DiD model estimation results with the clustered error. Although the coefficients of some control variables for down vote model are not significant, the 2nd difference estimators for up vote and score are still significant and positive with the clustered error term, which are consistent with our main findings.

Next, we check whether the results are impacted by our use of control observations multiple times via replication. As previously mentioned, the same control post may be used multiple times when the post belongs to a posting date clusters with multiple treatment dates, and even a same control post of a treatment date (a control post - treatment date combination) can be matched to multiple treatment posts if multiple treatment posts were posted on the same date and their posters were awarded the first gold badge on the same date. Note that in the final treatment and control data after matching, 28,499 control posts in 14,176 post date - treatment date combination clusters are replicated into 65,805 control posts matched to the treatment posts. After excluding the replicated control posts, we have 591,015 controls matched to 295,501 treatment posts. We

Table 14: Difference-in-Differences Model Estimation with Clustered Standard Errors

	1:3 Nearest Neighbor with Caliper		
	(1)	(2)	(3)
	Up	Down	Score
λ(2nd Difference Estimator)	7.43e-04***	2.53e-05	7.17e-04***
	(2.78e-04)	(4.00e-05)	(2.80e-04)
The number of silver badges	-3.61e-04	1.74e-05	-3.78e-04
	(5.55e-04)	(7.44e-05)	(5.60e-04)
The number of bronze badges	2.70e-04	-9.99e-06	2.80e-04
	(2.58e-04)	(3.24e-05)	(2.61e-04)
Question score	1.19e-01***	2.38e-03	1.17e-01***
	(1.15e-02)	(1.81e-03)	(1.16e-02)
The number of answer posts	-2.68e-02**	2.86e-03	-2.96e-02**
	(1.36e-02)	(2.32e-03)	(1.38e-02)
Ranking of the post	3.64e-01***	5.45e-02***	3.09e-01***
	(4.72e-02)	(1.64e-02)	(5.09e-02)
Num. obs.	626,934		
RMSE	0.167	0.025	0.169

Note. **p < 0.05; ***p < 0.01.

report the DiD model estimation results in Table 15. All substantive results are the same, and numerical differences are negligible. Thus, we are satisfied that our replication approach is not critical to our results.

7 Lab Experiment

Finally, we conducted a lab experiment to ensure the internal validity of our findings. In our experiment, we recruited 779 participants on MTurk. Participants were randomly assigned to one of two conditions: 'High reputation badge' or 'Low reputation badge'. 248 participants were assigned to the 'High reputation badge condition', and 255 Participants were assigned to the 'Low reputation badge condition'.

Prior to manipulating the level of reputation badges, participants were asked to imagine that they are searching for a possible solution for their smartphone battery drain issue in an online question answering community. We then gave participants a brief description of the community as well as its reputation system. Specifically, participants were told that the online community deploys a reward system which awards points and assigns users to different levels, and that the

Table 15: Difference-in-Differences Model Estimation with Unique Control

	1:3 Nearest Neighbor with Caliper		
	(1)	(2)	(3)
	Up	Down	Score
λ(2nd Difference Estimator)	7.17e-04***	2.85e-05	6.89e-04***
	(2.20e-04)	(3.25e-05)	(2.22e-04)
The number of silver badges	-4.36e-04	3.82e-06	-4.40e-04
	(4.36e-04)	(6.45e-05)	(4.41e-04)
The number of bronze badges	2.96e-04	-3.78e-06	2.99e-04
	(1.91e-04)	(2.82e-05)	(1.93e-04)
Question score	1.15e-01***	2.23e-03***	1.13e-01***
	(2.22e-03)	(3.28e-04)	(2.24e-03)
The number of answer posts	-2.24e-02***	3.05e-03***	-2.54e-02***
	(4.43e-03)	(6.55e-04)	(4.48e-03)
Ranking of the post	3.47e-01***	5.59e-02***	2.91e-01***
	(7.27e-03)	(1.07e-03)	(7.35e-03)
Num. obs.		591,015	
RMSE	0.168	0.025	0.170

Note. **p < 0.05; ***p < 0.01.

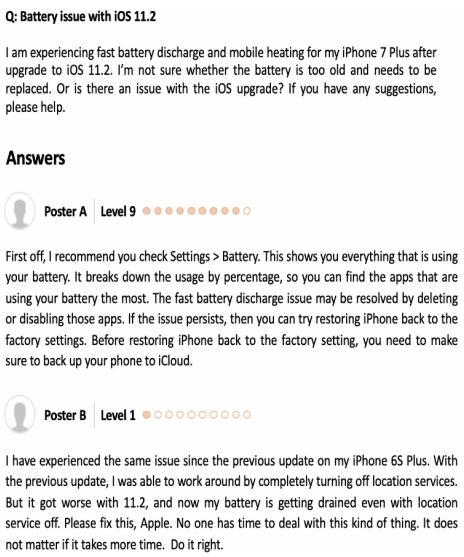
users with level 7 or higher are considered as highly reputable ones. Next, participants were shown a thread of posts containing one question post and two answer posts of the question post.

To manipulate the presence of reputation badge, for the 'High reputation badge' condition, the poster of the first answer post has a high level badge, level 9, and the second poster has the lowest level badge, level 1. For the participants in the 'Low reputation badge' condition, the poster of the first post has a relatively low level badge, level 3, and the second poster has again the lowest level badge, level 1. The stimuli for each of these conditions are presented in Figure 6. As for the manipulation check, we asked how high they think the first poster's badge level on the platform at the end of the experiment. Compared to the 'Low reputation badge' group, the 'High reputation badge' group indicated that the first poster had a higher level of badge ($M_{High} = 5.91$, $M_{Low} = 4.55$; $t = 10.89$, $p < 0.01$).

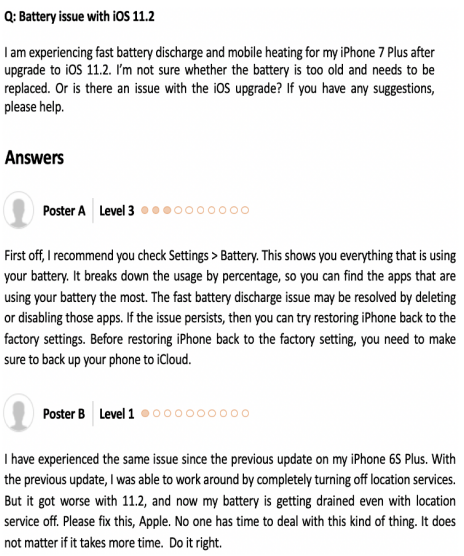
After the manipulation, we asked participants their likelihood of giving an up vote or down vote to the first post on a 7-point scale. The results presented in Figure 7 (a) and (b) replicate the main findings in Section 5. As illustrated by Figure 7(a), a post with a high reputation badge are more likely to get an up vote ($M_{High} = 5.95$) than a post with a low reputation badge ($M_{Low} = 5.68$; $t = 2.39$, $p = 0.02$). However, as shown in Figure 7 (b), the level of a reputation badge does not

Figure 6: Lab experiment stimuli

(a) High reputation badge



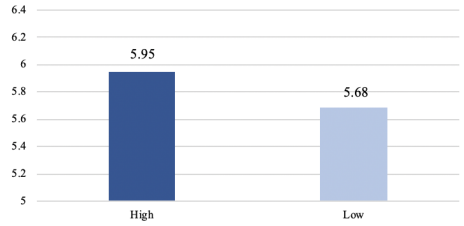
(b) Low reputation badge



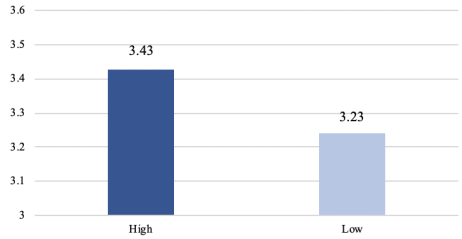
significantly affect the likelihood of getting a down vote ($M_{High} = 3.43$, $M_{Low} = 3.24$; $t = 0.94$, $p = 0.35$).

Figure 7: Lab experiment results

(a) Up vote



(b) Down vote



Additionally, participants were asked to indicate their attitudes (1 = negative, dislike, unfavorable; 7 = positive, like, favorable) towards the first post and its poster. We averaged responses to create a composite attitude index for post and poster attitude each, and these items demonstrated high internal consistency ($\alpha_{Post} = 0.93$; $\alpha_{Poster} = 0.95$). The attitudes towards the post does not differ by the badge level($M_{High} = 5.94$, $M_{Low} = 5.78$; $t = 1.57$, $p = 0.12$). However, participants have more positive attitudes toward the the poster with a high reputation badge than the one with low reputation badge($M_{High} = 5.90$, $M_{Low} = 5.68$; $t = 1.87$, $p = 0.06$). Participants were also

asked to indicate how helpful, useful, and credible they think the first post is (1 = not at all; 7 = very). We could find that the post with the high reputation badge are likely to be more credible than the post with the low reputation badge ($M_{High} = 5.89$, $M_{Low} = 5.61$; $t = 2.64$, $p = 0.01$). However, there are not significant differences between the two posts in terms of helpfulness and usefulness. The results may indicate that participants' more positive attitude towards the poster might enhance the credibility of the post. Additionally, the enhanced credibility could induce the Matthew Effect on peer evaluation of the post with high reputation badge.

8 Conclusion

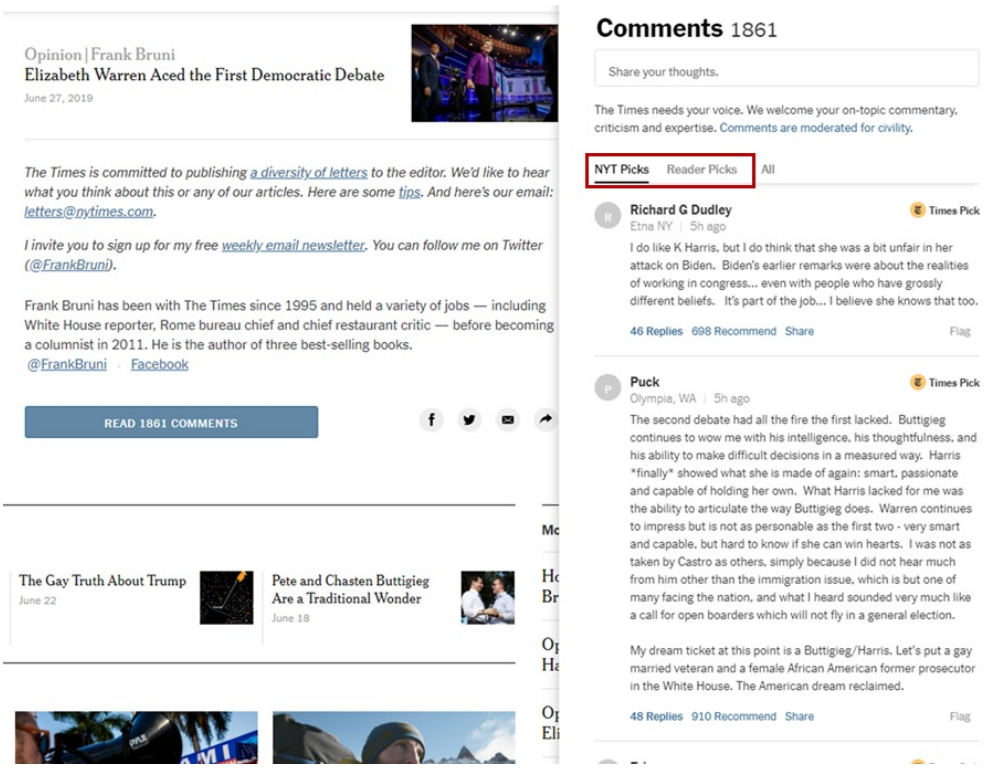
Using a DiD model with diverse matching methods, we find a significant effect of reputation badges on peer evaluations. After earning a prominent reputation badge, a poster receives more positive peer evaluations as compared with before he earns the badge. This effect suggests that online community members rely on heuristic cues while consuming and assessing posts. If the cue implies the poster's status or respect, it can even induce the Matthew Effect.

Our analysis contributes to the literature on status bias and Matthew Effect in several ways. First, our study provides a systematic test of status bias in a real-world situation by utilizing user activity data that are detailed and comprehensive. Second, although past researchers have shown the importance of reputation system and its effect on community members' welfare, little has been studied as to how the system can affect its key ingredient, peer evaluations. As far as we know, there are only a few exceptional studies, such as Dellarocas and Wood (2008) and Bolton et al. (2013). However, while these studies focus on strategic behavior by members, these studies assume that users are aware of the evaluator's identity and its evaluation. Our study is the first to empirically examine the impact of reputation system on the bias in online peer evaluations in more general setting where anonymity of an evaluator is guaranteed.

The results could help in modeling optimal reputation mechanisms and content recommendations for platforms. As reputation systems are regarded as a successful credentialing mechanism, and its use is quite prevalent among online communities, it may not be plausible or wise for platforms to remove the reputation system only because of the fear of this bias. However, platforms could consider some ways to lessen the bias in peer feedback induced by reputation systems. Specif-

ically, platforms might be able to reduce the degree of bias by redesigning reputation systems. For instance, if the Matthew Effect is largely driven by visual display or the name of the reputation badge in the poster profile information section (Figure 1), a platform can award guaranteed reputation indicators to users whose content quality is strictly assessed only by the platform or by the group of selected users; Incorporating or putting more weights on experts' evaluations could decrease the bias in peer feedback. Additionally, platforms might be able to diminish the degree of bias by disclosing the presence of bias in peer feedback. Since peer feedback can affect content recommendations, and vice versa, the bias in peer feedback may also be reduced if platforms implement supplementary content recommendations alongside the content recommendation strictly based on peer feedback. Such supplementary content recommendations include featuring high quality contents by low reputation users or featuring high quality contents measured by text analysis. Or platforms could provide users with platform-choice as well as peer-choice contents. A prime example of this is the New York Times as shown in Figure 8.

Figure 8: New York Times user posting recommendation



The future study may examine the effectiveness of the aforementioned reputation system and

content recommendation suggestions in terms of reducing bias in peer feedback.

While our study offers contributions academically and practically, we acknowledge that our study is burdened with several limitations. We have focused on a single online community platform, StackOverflow. While we believe the results to be relatively general, it would be important to replicate the results in other platforms, for instance, ones characterized by different types of content. Especially, it would be interesting to investigate the peer evaluation bias in platforms whose contents have more subjective nature, such as Yelp or TripAdvisor. Second, our matching might not capture all post or poster characteristics that have affected peer evaluations. In Table 8, we could observe that there exist differences in up vote volume between treatment posts and their matched control posts during pre-treatment period. This might indicate that the treatment and control post pairs could not be highly compatible. However, when using observational data, matching can be limited in that it cannot address the unobserved characteristics. Besides, our lab experiment, where we compare the peer evaluations of the identical post under two different conditions, shows consistent results as the DiD model results, we believe that it is not a major problem. Last, we cannot ascertain how much attention SO users give to reputation badge when they read posts.

9 References

Abadie, Alberto and Guido Imbens (2006), "Large sample properties of matching estimators for average treatment effects," Econometrica, 74 (1), 235-267.

Anderson, Cameron, Oliver P. John, Dacher Keltner and Ann M. Kring (2001), "Who attains social status? Effects of personality and physical attractiveness in social groups," Journal of Personality and Social Psychology, 81 (1), 116-132.

Anderson, Ashton, Daniel Huttenlocher, Jon Kleinberg and Jure Leskovec (2012), "Discovering value from community activity on focused question answering sites: a case study of stack overflow," KDD '12 Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 850-858.

Angrist, Joshua D. and Jorn-Steffen Pischke (2009), "Mostly harmless econometrics: An empiricist's comparison," Princeton: Princeton University Press.

Avery, Jill, Thomas J. Steenburgh, John Deighton and Mary Caravella (2012), "Adding bricks to clicks: Predicting the patterns of cross-channel elasticities over time," Journal of Marketing, 76 (3), 96-111.

Azoulay, Pierre, Toby Stuart and Yanbo Wang (2013), "Matthew: Effect or Fable?," Management Science, 60 (1), 92-102.

Ba, Sulin (2001), "Establishing online trust through a community responsibility system," Decision Support Systems, 31 (3), 323–336.

Ba, Sulin and Paul 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.

BBC (2012), "TripAdvisor Rebuked over 'Trust' Claims on Review Site," http://www.bbc.com/news/technology-16823012 (accessed on April 20, 2019).

Bell, David, Santiago Gallino and Antonio Moreno (2017), "Offline showrooms in omnichannel retail: demand and operational benefits," Management Science, 64 (4), 1629-1651.

Bertrand, Marianne, Esther Duflo, Sendhil Mullainathan (2004), "How much should we trust differences-in-differences estimates?," The Quarterly Journal of Economics, 119(1), 249–275.

Bolton, Gary, Elena Katok and Axel Ockenfels (2004), "How effective are electronic reputation mechanisms? An experimental investigation," Management Science, 50(11), 1587-1602.

Bolton, Gary, Ben Greiner and Axel Ockenfels (2013), "Engineering trust: reciprocity in the production of reputation information," Management Science, 59(2), 265-285.

Bronnenberg, B.J., Jean-Pierre Dudé, and Carl F. Mela, (2010) "Do digital video recorders influence sales?," Journal of Marketing Research, 47 (6), 998-1010.

Caliendo, Macro and Sabine Kopeinig (2008), "Some practical guidance for the implementation of propensity score matching," Journal of Economic Surveys, 22 (1), 31-72.

Cochran, William and Donald B. Rubin (1973), "Controlling bias in observational Studies: A review." Sankhyā: The Indian Journal of Statistics, 35 (4), 417-46.

Dagger, Tracey S. and Peter J. Danaher (2014), "Comparing the effect of store remodeling on new and existing customers," Journal of Marketing, 78 (3), 61-80.

Danaher, Peter J., Janghyuk Lee and Laoucine Kerbache (2010), "Optimal internet media selection," Marketing Sicence, 29 (2), 336 - 347.

Darley, John M. and Bibb Latane (1968), "Bystander intervention in emergencies: Diffusion of responsibility," Journal of Personality and Social Psychology, 8 (4), 377 - 383.

Dellarocas, Chrysanthos (2003), "The digitization of word of mouth: Promise and challenges of online feedback mechanisms," Management Science, 49(10), 1407-1424.

Dellarocas, Chrysanthos (2010), "Designing reputation systems for the social web," Boston U. School of Management Research Paper No. 2010-18, https://ssrn.com/abstract=1624697 (accessed on Apr 20, 2019).

Dellarocas, Chrysanthos and Charles A. Wood (2008), "The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias," Management Science, 43 (3), 460 - 476.

Diamond, Alexis and Jasjeet Sekhon (2013), "Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies," Review of Economics and Statistics, 95 (3), 932-945.

Forman, Chris, Anindya Ghose and Batia Wiesenfeld (2008), "Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets," Information Systems Research, 19 (3), 291- 313.

Forman, Chris, Anindya Ghose and Avi Goldfarb (2009), "Competition between local and electronic markets: How the benefit of buying online depends on where you live," Management Science, 55 (1), 47-57.

Garnefeld, Ina, Andreas Eggert, Sabrina V. Helm and Stephen S. Tax (2013), "Growing existing customers' revenue streams through customer referral programs," Journal of Marketing, 77 (4), 17-32.

Ghose, Anindya and, Panagiotis Ipeirotis (2011), "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics," IEEE Transactions on Knowledge and Data Engineering, 23(10), 1498-1512.

Grant, Scott and Buddy Betts (2013), "Encouraging user behaviour with achievements: An empirical Study," MSR '13 Proceedings of the 10th Working Conference on Mining Software Repositories, 65–68.

Heckman, J., Hidehiko Ichimura and Petra E. Todd (1997), "Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme," The Review of Economic Studies, 64 (4), 605- 654.

Heckman, J., Hidehiko Ichimura and Petra E. Todd (1998), "Matching as an econometric estimator," The Review of Economic Studies, 65 (2), 261 - 294.

Ho, Daniel, Kosuke Imai, Gary King and Elizabeth Stuart (2007), "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference," Political Analysis, 15 (3), 199–236.

Kim, Jerry W. and Brayden G. King (2014), "Seeing Stars: Matthew Effects and Status Bias in Major League Baseball Umpiring," Management Science, 60 (11), 2619 - 2644.

Latane, Bibb and John M. Darley (1968), "Group inhibition of bystander intervention in emergencies," Journal of Personality and Social Psychology, 10 (3), 215 - 221.

Latane, Bibb and Steve Nida (1981), "Ten years of research on group size and helping," Psychological Bulletin, 89 (2), 308 - 324.

Lechner, Michael (1999), "Earnings and employment effects of continuous off-the-job training in east Germany after unification," Journal of Business and Economic Statistics, 17 (1), 74-90.

Lechner, Michael (2002), "Some practical issues in the evaluation of heterogenous labour market programmes by matching methods," Journal of the Royal Statistical Society, A, 165, 59 - 82.

Merton, Rober K. (1968) "The Matthew effect in science," Science, 159 (3810), 56–63.

Montaguti, Elisa, Scott A. Neslin and Sara Valentini (2016), "Can marketing campaigns induce multichannel buying and more profitable customers? A field experiment," Marketing Science, 35

(2), 201-217.

Mudambi, Susan M. and David Schuff (2010), "What makes a helpful online review? A study of customer reviews on Amazon.com," MIS Quarterly, 34 (1), 185-200.

NeimanLab (2015), "What happened after 7 news sites got rid of reader comments," http://www.niemanlab.org/2015/09/what-happened-after-7-news-sites-got-rid-of-reader-comments (Accessed on April 20, 2019).

Pavlou, Paul and David Gefen (2004), "Building effective online marketplaces with institution-based trust," Information Systems Research, 15 (1), 37-59.

Resnick, Paul, Richard Zeckhauser, Eric Friedman, and Ko Kuwabara (2000), "Reputation systems," Commu nications of the ACM, 45 – 48.

Resnick, Paul and Richard Zeckhauser (2002), "Trust among strangers in internet transactions: Empirical analysis of eBay' s reputation system," The Economics of the Internet and E-commerce (Advances in Applied Microeconomics), 11, 127 – 157.

Ridgeway, Cecilia L. and Shelley J.Correll (2006), "Consensus and the Creation of Status Beliefs," Social Forces, 85 (1), 431 – 453.

Ridgeway, Cecilia L. and Joseph Berger (1986), "Expectations, legitimation, and dominance behavior in task groups," 51 (5), 603 – 617.

Rosenbuam, Paul R. and Donald B. Rubin (1983), "The central role of the propensity score in observational studies for causal effects," Biometrika, 70 (1), 41–55.

Rosenbuam, Paul R. and Donald B. Rubin (1984), "Reducing bias in observational studies using subclassification on the propensity score," 79, 516-524.

Rubin, Donald B. (1978), "Bayesian Inference for Causal Effects: The Role of Randomization," Annals of Statistics, 6 (1), 34–58.

Rubin, D., and N. Thomas, (2000), "Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates," Journal of the American Statistical Association, 95, 573-585.

Sekhon, Jaseet S. (2011), "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R," Journal of Statistical Software, 42 (7).

Shenkar, Oded and Ephraim Yuchtman-Yaar (1997), "Reputation, Image, Prestige, and Good-

will: An Interdisciplinary Approach to Organizational Standing," Human Relations, 50 (11), 1361-1381.

Sine, Wesley D., Scott Shane and Dante D.Gregorio, "The Halo Effect and Technology Licensing: The Influence of Institutional Prestige on theLicensing of University Inventions," Management Science, 49 (4), 478 – 496.

Smith, Jeffrey A. and Petra E. Todd (2005), "Does matching overcome LaLonde's critique of nonexperimental estimators?," Journal of Econometrics, 125 (1-2), 305-353.

Stuart, Elizabeth A. (2010), "Matching methods for causal inference: A review and a look forward," Statistical Science, 25 (1), 1-21.

The Times (2011), "Can you still trust TripAdvisor?," https://www.thetimes.co.uk/article/can-you-still-trust-tripadvisor-hr2hc35kbng (accessed on April 20, 2019).

Warren, Nooshin L. and Alina Sorescu (2017), "When $1 + 1 > 2$: How investors react to new product releases announced concurrently with other corporate news," Journal of Marketing, 81 (2), 64 - 82.

Wangenheim, Florian and Tomás Bayón (2007), "Behavioral consequences of overbooking service capacity," Journal of Marketing, 71 (4), 36-47.

Zhong, Zemin (2017), "Chasing Diamonds and Crowns: Consumer Limited Attention and Seller Response," Working paper, https://ssrn.com/abstract=2529306 (accessed on April 20, 2019).

Xu, Kaiquan, Jason Chan, Anindya Ghose and Sang Pil Han (2017), "Battle of the channels: The impact of tablets on digital commerce," Management Science, 63(5), 1469-1492.

Appendix A. Treatment and control data construction steps and the number of observations

	Steps	Number of treatment posts
Posters	Posters with at least one gold badge	62,533
	Posters who did not earn additional gold badge until $b_i + 7$	61,824
Posts	Posts by the treatment posters (posted before $b_i - 7$)	2,142,199
	Posts received at least one vote before $b_i - 7$	1,288,450
	Posts with no bounty $\left[b_i - 7, b_i - 1\right]$	1,288,367
	Posts with no editing $\left[b_i - 7, b_i - 1\right]$	1,286,996
	Posts with no comments $\left[b_i - 7, b_i - 1\right]$	1,283,173
	Exact matching - Tag matching	1,209,741
	Exact matching - Question score matching	978,835
	Exact matching - Post ranking matching	822,912

Appendix B. Construction of control post group

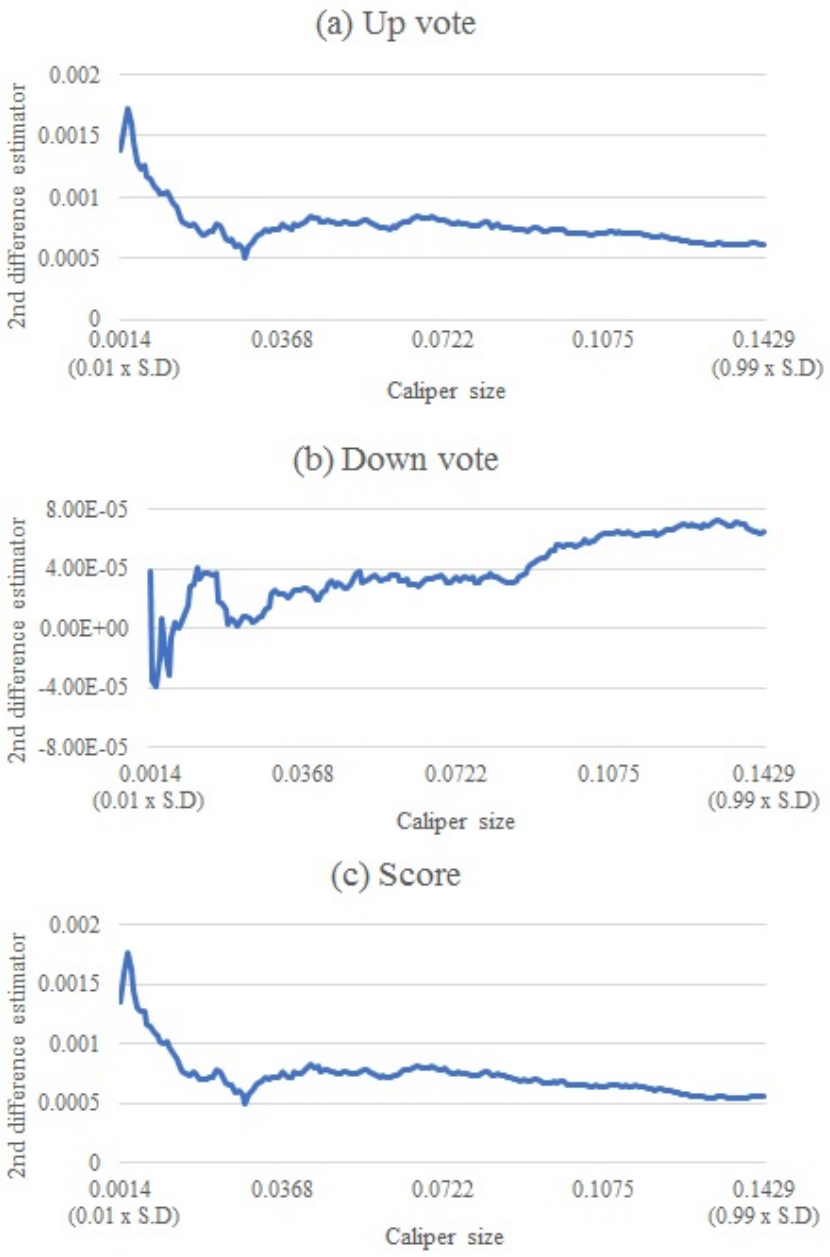
From treatment post set T, we have the posting date set, $P = \{p_1, ..., p_j, ..., p_{1990}\}$ where p_j is an unique posting date. Denote T^{p_j} as the set of treatment posts posted on p_j. In addition, $B^{p_j} = \{b_1, ..., b_k, ..., b_K\}$ represents the group of unique treatment dates of T^{p_j} when b_k is an unique treatment date.

(1) Create $C_1 = \{c_i \,|\, \sum_{d=d_0}^{d_1} GB_{it} = 0\}$ where GB_{it} is 1 if poster of post c_i earned at least one gold badge at day t, and 0 otherwise. d_0 and d_1 are the observed first and last entry of badge award data.

(3) Create $C_2 = \{c_i \in C_1 \,|p_i \in P\}$ where p_i is posting date of c_i.

(5) For $C_2^{p_j}$, the set of control posts posted on p_j, each $c_i \in C_2^{p_j}$ will be reproduced into $\{c_{i1}, ..., c_{ik}, ..., c_{iK}\}$ where treatment date of c_{ik} is $b_k \in B^{p_j}$.

Appendix C. Second difference estimator with diverse caliper sizes



Friend or Foe:
The Impact of Video UGC on Video Game Sales and Usage

Jin-Hee Huh* Lingling Zhang† P.K. Kannan‡

July 21, 2020

Abstract

Consumers are increasingly using online videos to post product reviews and share consumption experience. Compared with traditional text-based content, video user-generated content (VUGC) is more visual and auditory and hence can be more engaging. However, the impacts of VUGC on the product performance is debatable, especially for entertainment products. While VUGC can generate awareness and attention, they also run the risk of cannibalizing and spoiling the original content. In this research, we examine the impact of VUGC on sales and consumer usage for online video games. By leveraging a unique and comprehensive data set compiled from multiple sources, we specify a dynamic panel data model and estimate the heterogeneous effects of VUGC on sales and consumer usage. Our results indicate that VUGC affects both game sales and user engagement. We find that VUGCs generated by popular and high influential users have positive impacts on game sales, however, the promotion effect is mitigated by the substitution effect for high-priced games. We also find that the VUGCs generated by high influential users affect game usage positively for strategy games by small developers, perhaps due to the high informational value of such VUGCs. With the video affective content analysis, we could show that audio and visual feature differences between high- and low-influencer VUGC can affect viewers' responses differently. These findings suggest that game developers should consider both VUGC characteristics and their product characteristics jointly when developing and implementing policies on users' VUGC uploading, sharing, and monetizing.

Keywords: video user-generated content, video games, dynamic panel model

*Robert H. Smith School of Business, University of Maryland, jhhuh@umd.edu
†Robert H. Smith School of Business, University of Maryland, lzhang28@umd.edu
‡Robert H. Smith School of Business, University of Maryland, pkannan@umd.edu

"You can buy attention (advertising). You can beg for attention from the media (PR). You can bug people one at a time to get attention (sales). Or you can earn attention by creating something interesting and valuable and then publishing it online for free."

- David Meerman Scott

"For games with more expansive or re-playable gameplay, Let's Play (video user-generated content) can directly benefit developers. We've watched the playthrough videos and we see the value that this community is adding to our work through sharing themselves. ... But we underestimated how many people would be satisfied with only watching the game instead of playing it themselves"

- Ryan Green, "That Dragon, Cancer" developer

1 Introduction

When Minecraft became one of the most popular video games in history, many people were surprised why a sandbox construction game created so much buzz. Even more astonishingly, gamers were so engaged that they created numerous videos to showcase their ways to build the constructions, or to put in Minecraft language, their worlds. On YouTube alone, around 42 million videos were created by Minecraft gamers. One of the most influential uploaders, Stampy, has attracted more than 9.0 million subscribers and nearly 6.9 billion views by 2018 and ranked the fourth most popular YouTube channel in 2014 (Wikipedia 2018).

This is just one of the many examples in the explosion of user-contributed videos in recent years. As a new form of user-generated content (UGC), video user-generated content (VUGC) has been increasingly prominent for businesses and marketers. VUGC refers to the videos that are created by consumers and uploaded onto the video sharing platforms such as YouTube, Vimeo, and Twitch. The videos can vary in length, content, and quality, but are typically about how the uploader interacts with a particular product or brand.

VUGC is particularly important for the video game industry. The number of gaming video viewers reached 666 million in 2017 globally (SuperData Research 2017). Without a doubt, gaming VUGC has become a money-earning venue on its own. In 2017, the revenue —primarily from advertising, donations, and subscription —reached 5 billion U.S. dollars and is expected to grow

stronger in the years to come (Statista 2019). As for the viewership, young gamers aged between 18 and 25 worldwide spend an average of 3.3 hours each week watching video-game VUGCs (Inc 2018). More than 30% of the gamers in the U.S. reported that they regularly watched video gaming content online (Statista 2017).

Watching other gamers playing video games is not a new phenomenon. In the early days of arcade games, people gathered around and watched gamers playing video games. Spectators cheered for the game player to win or watched how other gamers play their favorite games. Such face-to-face interactions between game players and spectators have been transitioned into broadcasting video game plays on TV or online. In the early days of online video game playing broadcasting, the video contents were rather firm-generated contents than user-generated contents. Video game developers or video productions organized competitions or tournaments among professional game players and broadcasted the competitions live or recorded them to play later on TV or online. Since it was costly to organize the competitions, big title strategy games such as Starcraft and Warcraft were mainly played for these competitions. Since these games were already well known among gamers, viewers watched the videos for fun or to socialize within the gaming community, rather than trying to gain new information from the videos. Gros et al. (2017) find that some audiences watch these videos as an alternative to the entertainment content on TV or other medias. Lu et al. (2017) find that for 69% of the audiences, the primary purpose of watching the videos is to relax. For these audiences, game playing video is simply a spectator media, thus, its impact on game purchase could be trivial. However, for some other players, watching game-playing videos could increase their engagement and thus extend the shelf life of the games (The Esports Observer 2015). According to a survey on U.S. gamers, 89% of the audience watched the video game competitions because they want to get better at playing the game (Battlefy 2016), indicating that such videos could have a positive effect on game usage and sales.

In recent years, the ease of online content creation and distribution has lead to substantial changes to the landscape of game-playing videos. On the supply side, individual gamers have increasingly replaced game developers to become the backbone behind video creation (Sjöblom et al. 2018). As a result, VUGC have diversified to include all games rather than just the most popular strategy games. If one searches any game names, regardless of the genre or existing brand awareness, hours of VUGCs can be found on video-sharing sites such as YouTube or Twitch. The

content of the videos also tends to vary and include play-through, speed-run, instruction how-to, commentaries, and so on. From the viewers' perspectives, the richer VUGC content and format also bring more heterogeneous consumers with different motivations. Some perhaps are still attracted for the videos' entertainment value: they are just interested in the videos themselves and there is limited spillover effect to game sales or usage. But there could be viewers who use VUGC to learn more about new games or to improve playing skills. In a recent survey, 25% of gamers responded that they learned about new games from VUGCs (AYTM 2014). When this happens, VUGC could have important marketing implications. Some VUGC platforms now place a "buy" button alongside the VUGC content, so that viewers can purchase the game or in-game content. This indicates VUGC's possible impact on game purchase.

VUGC has also received increasing attention from game developers. However, the strategies towards the VUGC vary significantly among developers. Some encourage users to create and share content, hoping that the VUGC can work as free marketing for their games. For example, THQ Nordic, Supergiant Games, and Amanita Design [1] represent a groups of game developers with lenient policies towards VUGC. These developers would send early copies of their games to popular content creators, hoping that the content creators would create and share videos. These developers are fine with users monetizing from their created content. If user videos are blocked for some reasons (e.g., copyright claim from the background music producer), the developers may even contact the video sharing platforms on the user's behalf to help solve the dispute. Some developers, however, are more cautious toward online gaming content. They have banned gamers from uploading videos that involve their images or characters. For example, Nintendo, one of the biggest game developers, completely banned the live streaming of play-throughs on YouTube (Spencer 2017). Therefore, when it comes to user-generated gaming content, even the game developers themselves disagree on what should be the right policy to implement.

The reasons for the game developers' supporting actions can be explained by the extant literature on user-generated contents. Just like other forms of UGC such as product reviews or ratings, video gaming content has the potential to generate awareness and signal quality (Berger et al. 2010; Godes and Mayzlin 2004; Langhe et al. 2016). The scope and popularity of VUGC could help brands reach unaware consumers beyond the scope of paid marketing campaigns. Additionally,

[1] The specific policy of each developer is available at http://wholetsplay.com/.

the multi-media content of the videos can be much more engaging than the text-based user reviews to introduce new features and demonstrate how-to, potentially helping boost consumer interest and sales. There are numerous cases where a VUGC helps games earn audience and sales with little cost (New York Times 2017). For example, Crypt of the Necrodancer gained an immediate $60,000 increase in sales following a game playing video uploaded by an influencer.

Yet, unlike text-based UGC, video gaming content may backfire. Watching the play-throughs runs a risk of spoiling the fun for playing the game. Furthermore, the videos can be too entertaining that users may derive utility from consuming the VUGC instead of playing the games themselves. This was exactly what Lurking, Singapore-based game, experienced with VUGC. Some influential gamers noticed Lurking and created videos to feature their plays. Those play-throughs went viral and amassed more than 7 million views online. Lurking received a mere 8,000 downloads. Many developers share the feeling with Lurking and regard gaming VUGC to be a double-edged sword; VUGC helps with brand awareness but may run the risk of cannibalization. Thus, is VUGC a friend for video game brands, or a foe? Where is the line drawn?

Answers to this question is critical for game developers to design effective policies to either curate or to prevent the spread of VUGC. Despite its importance, the question has not yet been sufficiently studied empirically. In this research, we aim to examine the impact of VUGC on video game sales and usage. We ask two questions: (1) How do different types of VUGCs affect game sales and usage, respectively? and (2) How do the effects vary by game characteristics?

In this research, we compile a unique and comprehensive data set by collecting game sales, usage, and VUGC data from multiple sources. We specify a dynamic panel data model to examine the effect of VUGC on sales and usage, respectively, and examine the heterogeneous impacts by VUGC type and game characteristic. The results of our study indicate that VUGC affects both game sales and game engagement. When the entertainment value is high, VUGC promotes game sales. However, for high-priced games, such promotion effect seems to be offset by the substitution effect. On the contrary, when the entertainment value is low, substitution effect could dominate promotion effect on sales, especially when the game difficulty is high. Additionally, for difficult games, we could find that the highly informational VUGC positively affects game engagement. We also show that how the content characteristic —visual and audio characteristics —differences are associated with content viewers' reactions. Such findings substantiate our argument that different

types of VUGCs can lead to the heterogenous VUGC impacts on sales and/or engagement.

Our study contributes to the large body of literature on UGC. Although there has been growing interests in VUGC from both practitioners and researchers, only a few studies have examined the impacts of VUGC. Furthermore, to the best of our knowledge, this study is the first to empirically examine the two countervailing effects —promotion and cannibalization effects —of VUGC. We show that the cannibalization effect of VUGC exists and can offset the promotion effect, and the relative magnitude of each effect varies by VUGC and product characteristics. Additionally, we contribute to the literature by examining how VUGC affects not only the purchase decision but also product engagement. Lastly, we also contribute to the social media influencer research by examining the content differences between high and low influencers.

The rest of the paper is organized as follows. In section 2, we review the literature and present our theoretical framework. In Section 3, we describe the research setting and summarize data. Section 4 presents our model and discusses our identification strategy. In Section 5, we present the key findings from the analysis. In Section 6, we conduct the affective content analysis on VUGCs. Last, Section 7 concludes and discusses the future work.

2 Related Literature

2.1 Impacts of User-Generated Content

This research broadly contributes to the literature on UGC. There is a rich stream of marketing research that has examined the dynamics of UGC and its impact on consumer choices. Two potential mechanisms have been identified through which UGC could influence the performance for a brand.

First, UGC serves as a word-of-mouth channel and thus can reach a broader audience that would otherwise be unaware of the brand. (Berger et al. 2010; Trusov et al. 2009; Dhar and Chang 2009; Thompson and Malaviya 2013)

Second, UGC, especially the text-based content such as user reviews, also provides information about the brand. The content of user reviews can resolve uncertainly related to the quality of the product. In addition, users' comment may serve as a signal for the reputation. As a result, a positive effect of user reviews on sales is generally found for many product domains, including

television shows and movies (Chintagunta, Gopinath, and Venkataraman 2010; Godes and Mayzlin 2004), digital channels (Mudambi and Schuff 2010; Zhu and Zhang 2010), books (Chevalier and Mayzlin 2006; Sun 2012), and packaged goods (Moe and Trusov 2011).

2.2 User-Generated Videos in the Video Game Industry

Compared with text-based UGC, video content has its unique characteristics. Most noticeably, videos are visual and auditory so that the combination of movement and sound can attract and hold consumer attention longer than text or image contents. Thus, like conventional UGC, VUGC is also expected to create awareness and boost preferences for brands.

The base of the content contributors is so large that a VUGC may reach an audience of potential users who would otherwise be unaware of the brands. However, as far as video games are concerned, VUGC could help sales and usage through another mechanism: educating the players by doing playing-throughs or demonstrating game features. According to a survey by Google (2017), 74% of gamers report that they visit video sharing sites to learn how to get better at a game. By reducing the learning cost and introducing some hidden features, VUGC can help attract gamers with low gaming proficiency and/or increase the utility of playing, leading to increased user preference with the game. When this happens, not only new sales could be generated, but also users could engage in more repeated interactions with the game. For game developers, consumer's continuous product usage is important for another reason: revenues from in-game purchases. In-game purchases refer to the items or points that a player can buy for use within the game environment to improve a character or enhance the playing experience. For the free-to-play games, in-game purchases are where the most revenues come from. Even for the paid games, in-app purchase can make up a large portion of the firm's profits. For instance, Activision Blizzard, one of the largest game developers, made $7 billion revenue in 2017, out of which over $4 billion were from in-game purchases.

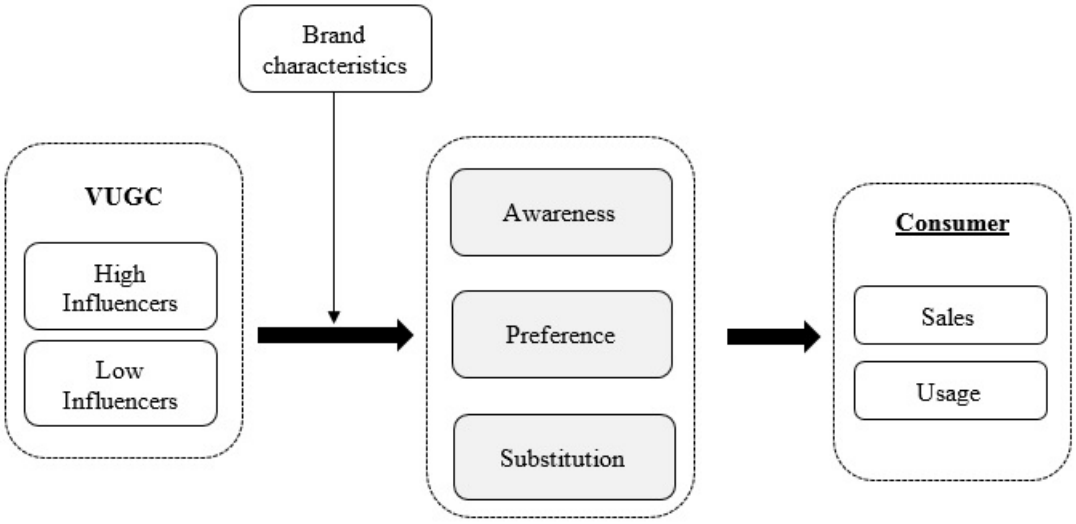
In sum, because of the potential awareness effect and preference effect, VUGC has been recognized as an important channel by practitioners, especially in the domain of video games.

However, as aforementioned, the impact of VUGC on games - or, entertainment goods in general - is less clearly defined. When used to promote tangible goods such as beauty products or electronics, VUGC is just another (perhaps more engaging) channel to boost the scope of the reach and to increase the trustworthiness of the message. However, when users create VUGC on games (typically

in the format of play-throughs or commentaries), these videos run the risk of revealing too much information about the products. If the narrative of the entertainment goods is spoiled, it may discourage users to purchase and consume the original content. Furthermore, consumers may derive utility from viewing the VUGC, as is evidenced by the millions and billions of views generated for some influential content creators. Additionally, VUGC can hurt consumer's continuous game playing. To play through video games, players need to invest time to develop diverse skills, such as problem-solving skill, manual skill, spatial and hand-eye coordination skills, as well as vision and speed skills. Then, if players feel that it requires a lot of practices or it may be almost impossible to complete missions or challenges in a game, players might derive more utility from watching a VUGC rather than from playing the game themselves. This is because users can enjoy vicarious victory without investing their time to get past a difficult section of the game. To sum up, due to the vicarious consumption and sense of accomplishment, VUGC may cannibalize sales and usage, and be detrimental to the product performance of the original content, leading to the substitution effect.

We summarize the effects - awareness, preference, and substitution - in Figure 1. While it may potentially have both the promotional effect and the cannibalization effect, the net effect of gaming VUGC perhaps ultimately depends on the characteristics of the videos and the games. Thus, a key objective of this research is to identify the conditions under which gaming VUGC would have a positive, negative, or even no impact on game sales and playtime. Next, we describe the key factors related to videos and games, respectively.

Figure 1: **Theoretical Framework on the Impacts of VUGC**



2.3 Influential Videos

In this section, we discuss how the popularity of the VUGC uploader is a key characteristic that moderates the impact of gaming VUGC on sales and engagement.

Extant research on online word-of-mouth has established that different groups of users may exert differential influences on the spread and persuasiveness of messages. Of particular interest is the so-called influencers, those who have a larger follower base and hence can reach a large number of users through content distribution (Iyengar, Van den Bulte, and Valente 2011; Nair, Manchanda, and Bhatia 2010; Trusov, Bodapati, and Bucklin 2010). In the same spirit, Kumar and Mirchandani (2012) identified influencers on social media and found that messages from them can have higher levels of reach (i.e., the number of times a message is forwarded by the recipients in general), influence (i.e., the number of times it is forwarded by the recipients to their friends), and social impact (i.e., the number of comments and replies received for each message). Because of these reasons, high-influencers (influencers with a bigger follower base) are expected to have a larger promotion impact on sales than low-influencers (influencers with a smaller follower base) and are often the targets of seeding in online marketing campaigns.

However, as far as online gaming content is concerned, videos from high-influencers differ from those from low-influencers in one substantive aspect: the entertainment value of influencer videos tends to be much higher for high-influencers. After all, there is a reason why high-influencer videos become popular in the first place. High-influencers put much effort into content creation, trying

to make their videos popular or viral. Game influencers are like celebrities in the entertainment industry. They are usually more interested in building their own brands instead of promoting the games. When gamers choose whose videos they watch and whom they follow, they consider whether their affective and tension release needs can be satisfied by watching the influencer's VUGC (Sjöblom and Hamari 2016). VUGC uploaders understand their (potential) viewers' needs, and make their videos entertaining to increase their popularity. Therefore, while videos from high-influential uploaders may reach a bigger audience and have higher content quality, they are also more likely to compete with the original game content for audience and revenue. In contrast, low-influence content creators can only reach a smaller audience and receives less attention on average, but their content may also have less cannibalization effect on sales.

The high-influencer VUGC may have a stronger positive effect on game usage than the low-influencer VUGC because a high-influencer VUGC has not only the higher entertainment value but the higher informational value than a low-influencer VUGC. According to Sjoblom and Hamari (2016), another motivation for gamers to watch VUGC is to improve game playing experience, confidence, and knowledge about games. As high-influencers are experienced video uploaders, they understand that entertainment and/or informational videos tend to go viral. Additionally, as high-influencers have a big group of followers who can provide the feedback or information on the content viewers expect to watch, high-influencers create VUGCs with the higher informational value. Thus, a high-influencer VUGC are more likely to have informational contents. Even when a high-influencer VUGC does not have higher informational value compared to a low-influencer VUGC, the VUGC viewers could feel that a high-influencer VUGC more informational, because entertaining contents are known to increase attention create positive attitude for the task itself.

On the other hand, the opposite is also possible. As previously mentioned, game players can derive a greater utility from VUGC viewing than firsthand game playing when they feel that the learning cost is high. As a high-influencer VUGC has a high entertainment and informational value, VUGC viewers feel the higher sense of accomplishment from a high-influencer VUGC than a low-influencer VUGC. Also, when a VUGC has a high entertainment value, then it is more likely that gamers watch the VUGC to the end. Then, the VUGC can spoil the fun part of playing, which discourages the repeated interaction with the game.

Thus, it remains an interesting empirical question in terms of which type of content is overall

more helpful for game developers - from the high-influencers or from low-influencers.

2.4 Game Characteristics

The extent of promotion and cannibalization effect of high-influencer and low-influencer VUGC can be determined by game brand characteristics. The first game characteristic is the size of game developer. The promotion effect of VUGC can be greater for games developed by small developers, indie games, than for games developed by big developers. The smaller the developer is, the less marketing and advertising capabilities the developer has. Accordingly, consumers are given less chance to learn about the games from paid marketing campaigns, such as trailers, advertisements or official game playthrough videos. Conversely, the substitution effect of VUGC could also be greater for indie games. To ascertain product quality, consumers rely on signals, such as brand name (Akerlof 1978) and brand advertising (Roberts 1986). This suggests that consumers could perceive indie games to be lesser quality than non-indie games, and gamers are more likely to experience the indie games from watching VUGC rather than the playing directly.

The second game characteristic is the game difficulty. VUGC can encourage gamers to purchase or play difficult games more as VUGC can make difficult games seem to be with right difficulty level and more entertaining. Also, VUGC can make difficult games actually less challenging by giving gamers specific instructions or advices on how to complete the missions. Thus, VUGC can prevent gamers, especially those less skilled, from feeling frustrated, and thus increase game engagement. In contrast, VUGC could discourage gamers from purchasing or playing difficult games. To play through difficult games, gamers need to invest their time and effort to practice their game playing skills. If gamers feel that the skill improvement requires a great deal of practice, they could prefer to substitute their first-hand game playing with VUGC watching. In video game industry, strategy games are known to have high difficulty level missions, and emphasize skillful thinking and planning to achieve victory. Thus, we examine the differential impact of VUGC on sales and product usages for strategy and non-strategy game.

The third characteristic is the price level. For high-priced games, gamers may be more likely to substitute their purchase with VUGC watching as gamers do not need to pay high prices to experience the games. However, VUGC could have the opposite effect on high-priced games for game engagement. Once consumers purchase games, the amount of money paid for the game

cannot be recovered. Thus, it seems reasonable to expect that high-priced game owners are more motivated to continue play than low-priced game owners as they invested more money. As VUGC could increase the utility of playing due to its informational value, VUGC could have a higher promotion effect on usage for high-priced games than low-priced games.

By examining the moderating impact of VUGC uploader characteristics and game brand characteristics jointly, we explore two countervailing effects, substitution and promotion effects, that VUGC might have on game sales and product usage.

3 Empirical Setting and Data

3.1 Video Game Industry

Video game industry has made a remarkable development over the past few decades and become tremendous entertainment industry. In 1999, the U.S. video game industry accounted for $7.4 billion in sales revenue, and the worldwide video game industry revenue exceeded $32 billion (Williams 2002; Chou 2003). From 2000 and onward, video game industry has become the fastest growing segment of the entire entertainment industry. The worldwide video game industry's rate of growth accounted for over 10% in 2017 with exceeding $134 billion (GamesIndustry.biz 2018). By 2018, the U.S. video game industry had matched that of the U.S. film industry on basis of revenue, with both industries having made around US$43 billion that year (VentureBeat 2019).

3.2 Data

In this research, we collect unique and comprehensive data from multiple sources, which enables us to answer the questions that have yet been empirically examined by extant research. Our data primary includes three set of variables: (1) game sales and usage, (2) user-created video content, and (3) game characteristics, which we describe in detail below.

3.2.1 Game Sales and Usage

Our primary variable of interest, game sales and usage, are collected from Steam[2], a digital platform for PC-based online games. Steam is one of the most popular online gaming portals (Business

[2]https://store.steampowered.com/

Insider 2018), where gamers can purchase and play online games. By 2017, Steam had at least 18% of global PC-based online games (PCGamesN 2018). Steam revenue in 2017 is estimated to have hit $4.3 billion, and had over 150 million registered accounts and 18.6 million concurrent users in 2018 (WePC 2019). Our variables of interest are provided by two third-party scraping websites, SteamSpy and SteamDB, which track the daily activities for steam games through an official Steam API. [3] Although Steam provides open APIs, individual user or developer cannot use each Steam game's statistics, such as sales, the number of players, or average playtime, without the right tool to scrape and aggregate the information. Game developers often rely on such information to make informed decision about whether their game could be successful, who their biggest competitors are, or when the right time to release their game. It is known that the developers of Steam games have been actively using the data on SteamSpy and SteamDB to conduct the market research (Gameinformer 2018). Especially indie game developers rely on data on these two sites because the data on these two sites are great alternative for the expensive market research report, such as the video game report by NPD group (Tech Times 2018).

We are interested in two outcome variables: sales and usage. Sales are measured as the number of units sold for each game and usage (also referred to as engagement) is measured by the daily maximum number of concurrent players for each game. A game typically attracts a different number of online players at different times of the day. Take Dota 2, the most played game on Steam as of May 30, 2019, as an example. Fewer users are playing it in the morning than in the afternoon or the evening. On May 30, 2019, at 2:00 PM, Dota 2 experiences its peak playing and attracts 849,432 users. Then, 849,423 is used to measure the engagement level for the game. The maximum number of concurrent users is a commonly adopted metric used by industry practitioners to capture the volume of usage.[4]

For this research, we randomly selected 1,208 games from all the games that were active on

[3] Initially, we studied the impact of VUGC on Steam game sales and usage using data collected from SteamSpy. However, we later found that the number of peak concurrent players from SteamSpy had data inaccuracy issue, we collected the variable from another website, SteamDB.

[4] Peak concurrent players are more frequently used than other measurement (e.g., total number of players) to measure player engagement in game industry. It is because peak concurrent players is a better indicator of game's health. In many games, it is commonly observed that players spend very short amount of time (e.g., less than 5 minutes) only to complete their daily quest then log out. Hence, the total number of players could inflate the game player engagement. On the other hand, as the peak concurrent users can reflect both the number of active players and their consistent usership, the peak concurrent players is frequently used by game developers to measure player engagement.

Table 1: **Descriptive Statistics for Consumer Response, VUGC Activity, and Game Characteristics**

	Mean	SD	Median	Min	Max
Consumer responses					
Sales units	943.09	3,158.01	234.29	0	154,651.57
Peak concurrent players	814.95	16,673.68	2.5	0	603,010.86
VUGC activities					
Number of VUGCs posted by high-influencers	0.36	7.82	0	0	309.14
Number of VUGCs posted by low-influencers	2.4	37	0	0	1,553.29
Game characteristics					
Game regular price ($)	10.12	9.70	7.99	0.99	129.99
Game price discount ($)	1.19	3.22	0.00	0.00	63.20
Number of owners	228,449.76	1348,988.63	5,187.86	0	39,197,763.4
Steam user rating	2.63	62.78	0	-1,138	8,477.14
Indie game	0.69	0.46	1	0	1
Strategy game	0.21	0.41	0	0	1
High-priced game	0.26	0.44	0	0	1
Sample size		30,627			

Note. The variables are the weekly averages. Sales unit is the total number of sales made during a week. Peak concurrent players are the average daily peak concurrent players during the week. High-influencers are the top 10,000 content contributors according to their number of followers. The remaining content contributors are referred to as the low-influencers.

Steam from June 2017 to January 2018. Our outcome variables - sales and peak play - are aggregated at the weekly basis. Table 1 presents the summary statistics. On average, the games in our sample have about 943.09 (SD = 3,158.01) weekly sales and about 814.95 (SD = 16,673.68) daily peak concurrent players. The large standard deviation for the variables indicates great heterogeneity, as is usually the case for entertainment products.

3.2.2 VUGC Activities

Our data on the VUGC is from Twitch[5], which is the market leader of video sharing platform in the domain of games.[6] [7] The primary contents on Twitch are the live-streamed videos of play-throughs from individual players, but sometimes gaming-related talk shows and eSports shorts tournaments can also be found. Viewers on Twitch can access the archives of the live streaming content through the video-on-demand service offered on the platform. The typical viewer on Twitch is a male aged

[5] https://www.twitch.tv/

[6] For video game VUGC, Twitch and Youtube are the two market leaders of the video sharing platforms. In January 2019, 63.7 thousands people posted VUGCs on Twitch, generating 1.9 million hours of live video content. 22 thousands of people posted VUGCs on YouTube, producing 460 thousand hours (Newzoo 2019).

[7] We only use Twitch VUGC activities for our analysis because Steamspy's daily number of YouTube videos is limited to 50, so if there were more videos uploaded on YouTube, the number is still 50.

between 18 and 34 years of age, which is consistent with the profile of a typical video game player. As of February 2018, Twitch had 2 million broadcasters monthly, and 15 million daily active users.

The daily VUGC postings on Twitch are tracked by third-party Twitch scarping websites, SteamSpy[8] and TwitchTracker[9], which provide the measurements of VUGC for this research. From SteamSpy, we collect the daily number of VUGC posts from all Twitch uploaders. However, SteamSpy does not distinguish content contributors according to their popularity, which is important for our research objectives. Luckily, the other scarping service, TwitchTracker, tracks the video postings from the top influencers, i.e., the most popular 10,000 Twitch VUGC uploaders.[10] By combining data from these two websites, we are able to capture the amount of the VUGCs created by contributors with high and low influence, respectively.

The summary statistics for VUGC are again presented in Table 1. On average, the games have 0.36 videos uploaded by the high-influence contributors (i.e., high-influencers) every week and 2.4 videos from the low-influence contributors (i.e., low-influencers). The number of videos from low-influencers is larger because there are much more low-influence content contributors than high influencers, but the videos from high influencers can reach a much larger audience due to their size of followers. For the analysis, we take log transformation to deal with skewness of the data (see Table 1), and mean centered VUGC activity variables to reduce the effects of multicollinearity between high-influencer and low-influencer VUGC activities.[11]

Table 2: **Correlation Matrix between Weekly Consumer Response and VUGC Activities**

	Sales	Number of player	High influencer VUGC
Number of players	0.655***	-	-
Number of high influencer VUGC	0.207***	0.485***	-
Number of low influencer VUGC	0.404***	0.632***	0.712***

Note. *: $p < 0.10, **: p < 0.05, ***: p < 0.01$

Table 2 reports the correlations between consumer responses and VUGC activities. First, the correlation between sales unit and the number of players is modestly high (0.655, p $<$ 0.01),

[8] https://steamspy.com/

[9] https://twitchtracker.com/

[10] TwitchTracker defines influencers by the number of followers for each content contributor.

[11] Before mean-centering the VUGC activity variables, the correlation between high-influencer and low-influencer VGUC variables is 0.936.

indicating that the two metrics are related but still capturing somewhat unique consumer responses. This justifies why we are interested in examining sales and user engagement separately. Second, we see that the overall correlations for VUGCs are higher with engagement than with sales, perhaps because the peak concurrent players and the VUGCs both reflect the size of active players while game sales do not necessarily capture the level of engagement. Last, the correlation between high-influencer VUGCs and low-influencer VUGCs is also high. Much of the correlation between variables is due to the heterogeneous popularity across games: popular games tend to generate sales, have a larger player base, and invite more UGC. Thus, it is critical to account for the game-level heterogeneity, which we describe in detail in the modeling section.

3.2.3 Game Characteristics

We also have game brand characteristics, such as price, the number of owners, and the user ratings on Steam (see Table 1). On Steam, game developers are allowed to change game price daily. The games on Steam have an average regular price of $10.12 and the average price discount of $1.19. We classify games into high-priced game and low-priced game based on their average retail price, depending on whether the average retail price is above or below the regular price mean. The number of owners for a game is the number of Steam users who keep the game in the user profile page on that day. The number of owners can decrease over time, as users can delete a game from the owned games list or Steam can revoke the license for the game if the user has not logged in for a while. On average, a game in our sample has an average of 228,449.76 owners each week. Apparently, this number does not reflect the size of active users; we simply use it to control for the "installation" base for each game. User ratings on Steam are on a two-point scale: positive (thumb-up) or negative (thumb-down). We use the net incremental rating, i.e., the number of thumb-ups minus the number of thumb-downs during the day. The weekly average of Steam user rating ranges from -1,138 to 8,477.14 with the mean being 62.78.

Furthermore, for each game, we know its genres. Steam categorizes a game into 22 genres and a game can belong to multiple genres. Games in our data belong to an average of 2.33 genres. From the genre information, we further create two indicator variables: (1) the indie game variable indicating whether the game was developed by an indie developer, and (2) the strategy game indicator. The number of indie and strategy games are 837 and 256, respectively.

Table 3: **Correlation Matrix between Weekly Consumer Response and VUGC Activity by Game Characteristics**

(a) Developer size

	Non-indie game		Indie game	
	Sales	Number of players	Sales	Number of players
Number of high-influencer VUGC	0.247***	0.533***	0.163***	0.419***
Number of low-influencer VUGC	0.432***	0.669***	0.375***	0.568***

Note. *: $p < 0.10, **: p < 0.05, ***: p < 0.01$.

(b) Game difficulty

	Non-strategy game		Strategy game	
	Sales	Number of players	Sales	Number of players
High-influencer VUGC	0.218***	0.506***	0.161***	0.429***
Low-influencer VUGC	0.411***	0.650***	0.385***	0.588***

Note. *: $p < 0.10, **: p < 0.05, ***: p < 0.01$.

(c) Game price level

	Low-priced game		High-priced game	
	Sales	Number of players	Sales	Number of players
High-influencer VUGC	0.146***	0.356***	0.292***	0.568***
Low-influencer VUGC	0.381***	0.484***	0.466***	0.692***

Note. *: $p < 0.10, **: p < 0.05, ***: p < 0.01$.

In Table 3, we report the correlations between consumer response and VUGC activity by VUGC content characteristic and game characteristic. Although the correlation coefficients between consumer response and VUGC activity are positive and significant across all game groups, we can observe that the size of the correlation coefficients vary by groups. For example, the correlation between sales and high-influencer VUGC is 0.292 for high-priced games and 0.146 for low-priced games, indicating a stronger association between high-influencer VUGC and sales for games with

a high price. The patterns reported in Table 3 are only descriptive and correlational, which by no means suggest causal. Thus, we formally model the impacts of high- and low- influencer VUGC activity in the next section.

4 Modeling Approach

4.1 Dynamic Panel Data Modeling

In this section we present our model to examine the impact of VUGC on game sales and usage. In particular, we model the consumer responses for game i at week t as a function of the responses from the previous time period, the amount of high- and low-influencer VUGC, and the interaction effects between high- and low-influencer VUGC and game characteristics, which goes as follows:

$$\begin{aligned} Y_{it} = \alpha_i + \rho \cdot Y_{it-1} + \gamma_1 \cdot High_{it-1} + \gamma_2 \cdot Low_{it-1} + \mu_1 \cdot High_{it-1} \cdot Indie_i + \mu_2 \cdot Low_{it-1} \cdot Indie_i \\ + \mu_3 \cdot High_{it-1} \cdot Strategy_i + \mu_4 \cdot Low_{it-1} \cdot Strategy_i \\ + \mu_5 \cdot High_{it-1} \cdot HighPriced_i + \mu_6 \cdot Low_{it-1} \cdot HighPriced_i \\ + BX_{it-1} + \Phi_t + \Lambda Z_i + e_{it}, \end{aligned} \tag{1}$$

where $Y_{it} \in \{SalesUnit_{it}, PeakCCU_{it}\}$ and α_i is the heterogeneity across games. In Equation (1), the carryover parameter ρ captures how much a shock to the past consumer responses can persist to the next period. Parameters γ_1 and γ_2 capture the main impacts of high-influencer VUGC and low-influencer VUGC on Y_{it}, respectively. We allow the impacts of VUGC to be heterogeneous across games by including the interactions between game characteristics, $Indie_i$, $Strategy_i$, $HighPriced_i$, and high-influencer VUGC and low-influencer, respectively. To address the concern for reverse causality between the outcomes and VUGC variables, we use the lagged VUGC variables instead of the VUGC from the same time period.

The outcome Y_{it} is also a function of game characteristics that change over time, denoted as X_{it}. Examples of X_{it} include price discount $Discount_{it}$, the installation base $Owner_{it}$, and the valence of user review for the game $RatingValence_{it}$. To deal with the simultaneity concern between Y_{it}

and X_{it}, We again include the lagged variables X_{it-1} in the model. Note that for the model of $SalesUnit_{it}$, $Owner_{it-1}$ is excluded from X_{it-1}, because the previous installation base is already captured in $Y_{i,t-1}$. For the model on $PeakCCU_{it}$, the price variable $Discount_{it-1}$ is excluded from X_{it-1} because price should not have a direct impact on play (the price effect should be indirect through sales). Vector Z_i includes time-invariant game characteristics such as genre, the regular price ($RegularPrice_i$), whether the game is an indie brand ($Indie_i$), whether it is a strategy game ($Strategy_i$), and whether the game has a high price ($HighPriced_i$). The definition of the variables is provided in Table 4.

Table 4: **Definition of Variables Used in the Analysis**

Variable	Definition
Outcome	
$SalesUnit_{it}$	Weekly average number of units sold
$PeakCCU_{it}$	Weekly average number of peak concurrent players
VUGC activity	
$High_{it}$	Weekly average number of videos uploaded by high Twitch influencers (top 10,000)
Low_{it}	Weekly average number of Twitch videos uploaded by Twitch influencers other than high influencers
Time varying characteristics	
$Discount_{it}$	Weekly average price discount amount
$Owner_{it}$	Weekly average number of owners
$RatingValence_{it}$	Weekly average user ratings on Steam
Time-invariant characteristics	
$RegularPrice_i$	Regular price
$Indie_i$	Indicator for an indie game
$Strategy_i$	Indicator for a strategy game
$HighPriced_i$	Indicator for a high-priced game

Lastly, in Equation (1), Φ_t is the fixed effect for week, and e_{it} is the errors that are assumed to be independent across games and time periods.

In sum, our model specification as in Equation (1), known as the dynamic linear regression modeling, allows us to assess the impacts of VUGC while controlling for both observed and unobserved heterogeneity across games. Note that we also leverage the panel data structure and include the lagged variables for VUGC activity and other game characteristics, so that the concern of simultaneity bias is addressed.

4.2 Identification

Generally, dynamic panel data models include lagged levels of the dependent variable as regressors. However, since lags of the dependent variables are necessarily correlated with the idiosyncratic error, the estimators of static panel data models, such as fixed effect model or random effect model are inconsistent. [12] To address such endogeniety concern, we choose to employ the GMM approach developed by Arellano and Bond (1991) - difference GMM. The first step in the Arellano and Bond GMM approach is taking the first difference of the model.

The first-differencing transformation can reduce the potential influence of auto-correlation and time-invariant unobservable game characteristics. However, even after first-differencing, there are still remaining sources of endogeneity problems. First, due to the presence of e_{it-1} in both terms, the first-differenced lagged dependent variable, $\Delta Y_{it-1} = Y_{it-1} - Y_{it-2}$ is correlated with the first-differenced error term $\Delta e_{it} = e_{it} - e_{it-1}$. Second, there could be unobserved time-varying factors, such as firm-level advertisement or promotion activities, that can affect both sales and usage and VUGC activities. Additionally, a shock in sales and usage for a game may lead to change in the VUGC activities, thus, VUGC activities can be correlated with the error term. Note that we already directly address the endogeneity bias due to such simultaneity or reverse causality between sales and usage and VUGC activities by estimating the effects of $High_{it-1}$ and Low_{it-1} instead of the effects of $High_{it}$ and Low_{it}. However, conservatively, we assume that $High_{it-1}$ and Low_{it-1} are endogeneous variables. This is because the unobserved the impacts of time-varying factors or a shock in sales and usage can last longer than one period. Last, the time-varying game characteristics, $Discount_{it-1}$, $Owner_{it-1}$, , and $RatingValence_{it-1}$, are also treated as endogeneous variables for the similar reasons.

Arellano and Bond (1991) proposed a solution by utilizing instrumental variables estimation. A key advantage of this method is that it does not require external instruments. Instead, the lags and lagged differences of endogenous explanatory variables are used as instruments. We used the lagged values of the endogeneous variables with two degrees and up as the instruments for their first differences. For example, $High_{i,t-p}$ $(p \geq 2)$ are used as the instrument variables for $\Delta High_{it-1}$

[12] In panel data, the fixed effects estimator is consistent when T is large with the presence of lagged dependent variable as a regressor. Although our data has a relatively large T ($T = 28$), we employed the GMM approach developed by Arellano and Bond (1991) to be conservative. Additionally, we present fixed-effect linear model estimator results in Section 5.

in the estimation.

As some games in our panel data have a two-weeks time gap due to data availability, we employ forward orthogonal deviations to preserve the sample size (Arellano and Bover 1995). This methodology has been applied in a wide variety of applications within marketing and economics (Acemoglu and Robinson 2001; Durlauf et al. 2005; Clark et al. 2009; Germann et al. 2015; Feng et al. 2015; Kang et al. 2016). Furthermore, we use Windmeijer's (2005) two-step robust estimator to correct for autocorrelation and heteroskedasticity.

The time invariant game characteristics and the week fixed effects are assumed to be exogenous in our model estimation. As aforementioned, the Arellano and Bond estimator is a GMM estimator of the first-difference. As a result, the main effects of the time-invariant game characteristics such as $RegularPrice_i$, $Indie_i$, $Strategy_i$, and $HighPriced_i$ cannot be estimated and they only enter the model in the interactions with VUGC.

5 Results

In Table 5, we present the results of Equation (1) using the fixed effect estimator and the GMM estimator proposed by Arellano and Bond (1991). The first four columns of Table 5 are the results of models including only main effects and control variables; the following four columns are the results of models including the heterogeneous effects related to game characteristics. The F-statistics for the all fixed effect models are significant at $p < 0.001$ level (Table 5). The specification tests for difference GMM suggest that the instruments used for both $Sales_{it}$ and $PeakCCU_{it}$ models are valid (Table 5). The AR(2) tests for GMM models suggest that the second-order differenced error terms $(e_{it} - e_{it-1})$ and $(e_{it-2} - e_{it-3})$ of $Sales_{it}$ and $PeakCCU_{it}$ models are uncorrelated, indicating that the assumption $E[e_{it-1}, e_{it-2}] = 0$ holds. In addition, we test for the validity of the instruments using the Hansen J test, a test that checks joint validity of GMM and IV instrument, and difference-in-Hansen C test, a test that checks whether instruments are exogenous. As Table 5 shows, we fail to reject both tests across all models. Thus, the instruments instruments used for both models are valid.

The magnitude and direction of coefficients are found to be overall consistent between the fixed effect estimates and the Arellano-Bond estimates. For example, when only main effect are

considered, $High_{it-1}$ have a positive effect on $Sales_{it}$ and $PeakCCUit$. However, there are some coefficients (e.g., Low_{it-1}) with different directions. As Arellano-Bond model can capture dynamic endogeneity that fixed effect model fails to capture, we will focus on and discuss the results using the Arellano-Bond estimator. Note that for ease of exposition, coefficients of the Φ_t are not reported here. The full list of coefficients is available from the authors, upon request.

When including only the main effects in the model(Column 3 and 4),the coefficient for high-influencer VUGC is estimated to be 0.195 ($p < 0.01$) for sales and 0.043 (n.s.) for usage. This indicates that high-influencer VUGC has a significant and positive effect on sales, but it does not significantly affect usage. On the other hand, the low-influencer VUGC estimate is -0.155 ($p < 0.01$) for sales and -0.080 ($p < 0.05$), indicating a negative co-movement between low-influencer VUGC and consumer responses. This seems to indicate that when VUGC content has a high entertainment value, it can help game developers by increasing sales. However, with the low entertainment and informational value, the average effect of videos from the large number of content generators may cannibalize both sales and usage.

Although the main effect model can shed light on the average VUGC effect, it could mask important heterogeneity across games of different characteristics. From Column 7 and 8 in Table 5, we can find some interesting heterogeneous effects of VUGC on sales and usage.

For sales, while main effect of high-influencer VUGC is positive and significant (0.482, $p < 0.01$), the interaction between high-influencer VUGC and price is estimated to be negative and significant (-0.323, $p < 0.1$). However, the main effect of low-influencer VUGC is negative and the interaction between low-influencer VUGC and price is positive although statistically insignificant at 0.10 level. These combined seem to suggest that, when a game has a high price point, gamers may substitute game purchase with watching the videos posted by popular game players, as the content tends to be of high quality and entertaining.

Table 5: **Estimation Results**

	Fixed effect		Arellano-Bond		Fixed effect		Arellano-Bond	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	$SalesUnit_{it}$	$PeakCCU_{it}$	$SalesUnit_{it}$	$PeakCCU_{it}$	$SalesUnit_{it}$	$PeakCCU_{it}$	$SalesUnit_{it}$	$PeakCCU_{it}$
$SalesUnit_{it-1}$	0.192***	-	0.067***	-	0.190***	-	0.077***	-
	(0.006)		(0.023)		(0.006)		(0.021)	
$PeakCCU_{it-1}$	-	0.707***	-	0.726***	-	0.702***	-	0.723***
		(0.005)		(0.022)		(0.005)		(0.017)
$High_{it-1}$	0.165***	0.197***	0.195***	0.043	0.337*	-0.154**	0.482**	-0.045
	(0.061)	(0.023)	(0.072)	(0.071)	(0.193)	(0.062)	(0.189)	(0.085)
Low_{it-1}	0.116***	0.047***	-0.155*	-0.080**	0.071	0.008	-0.118	-0.001
	(0.036)	(0.013)	(0.087)	(0.040)	(0.074)	(0.026)	(0.156)	(0.052)
$High_{it-1} \cdot Indie_i$					-0.191	0.157***	-0.130	0.008
					(0.131)	(0.046)	(0.129)	(0.088)
$Low_{it-1} \cdot Indie_i$					-0.024	-0.033	0.003	-0.089
					(0.076)	(0.026)	(0.169)	(0.073)
$High_{it-1} \cdot Strategy_i$					-0.028	0.327***	-0.074	0.206*
					(0.153)	(0.054)	(0.150)	(0.116)
$Low_{it-1} \cdot Strategy_i$					-0.265***	-0.002	-0.161	-0.054
					(0.086)	(0.029)	(0.143)	(0.108)
$High_{it-1} \cdot HighPriced_i$					-0.210	0.212***	-0.323*	0.076
					(0.185)	(0.059)	(0.167)	(0.101)
$Low_{it-1} \cdot HighPriced_i$					0.284***	0.115***	0.232	0.036
					(0.076)	(0.026)	(0.146)	(0.080)
$Discount_{it-1}$	0.046***	-	-0.003	-	0.046***	-	0.011	-
	(0.013)		(0.023)		(0.013)		(0.020)	
$Owner_{it-1}$	-	0.006***	-	0.007*	-	0.006***	-	0.007*
		(0.001)		(0.004)		(0.001)		(0.004)
$RatingValence_{it-1}$	0.197	0.018	0.449**	0.091	0.190	0.005	0.413**	0.086
	(0.129)	(0.053)	(0.222)	(0.066)	(0.129)	(0.053)	(0.200)	(0.078)
Specification test p-value (Arellano-Bond)								
AR(1)			0.00	0.00			0.00	0.00
AR(2)			0.52	0.10			0.70	0.10
Hansen J			1.00	1.00			1.00	1.00
Hasen C			1.00	1.00			1.00	1.00

Note. Standard deviations are shown beneath parameter estimates. *: $p < 0.10, **: p < 0.05, ***: p < 0.01$.

Next, let us discuss the results related to user engagement. First, the interactions between price and high- and low-influencer VUGC are no longer significant. This is expected as peak play captures the amount of active engagement post purchase and hence should not be directly related to the game's retail price. The insignificant interactions with price provide validation for our estimates. Second, high-influencer VUGC has a significantly positive effect for strategy games (0.206, $p < 0.10$) while the coefficient for the interaction between low-influencer videos and strategy is slightly negative yet insignificant (-0.054, $p > 0.10$). These seem to suggest that the high-quality content from high-influencer gamers may help reduce the learning cost for these difficult strategy games, thus an increase in the high-influencer videos helps boost the subsequent play from the users.

Table 6: **Total Effects of VUGC by Game Characteristics**

Game characteristics	N	Sales units		Peak concurrent players	
		High VUGC	Low VUGC	High VUGC	Low VUGC
Indie & Strategy & High-priced	52	-0.044	-0.043	0.245*	-0.109
Indie & Strategy & Low-priced	112	0.279*	-0.275*	0.169*	-0.145*
Indie & Non-strategy & High-priced	121	0.030	0.117	0.038	-0.054
Indie & Non-strategy & Low-priced	552	0.352**	-0.114	-0.038	-0.090
Non-indie & Strategy & High-priced	35	0.086	-0.047	0.237	-0.019
Non-indie & Strategy & Low-priced	57	0.409**	-0.278*	0.161	-0.055
Non-indie & Non-strategy & High-priced	112	0.160	0.114	0.031	0.035
Non-indie & Non-strategy & Low-priced	167	0.482**	-0.118	-0.045	-0.001

Note. *: $p < 0.10, **: p < 0.05, ***: p < 0.01$

To better examine the net effect of high- and low- VUGC on game sales and usage by game characteristics, we compute the significance of total effects (the linear combination of main and interaction coefficients) and report them in Table 6. The results may indicate that VUGC with high entertainment value promotes game sales by increasing awareness or preferences. However, such promotion effect seems to be offset by the substitution effects when game price is relatively high. This supports our conjecture that gamers tend to consume high-priced games vicariously with the VUGC, especially with the VUGC of high entertainment value. Additionally, the overall negative impact of low-influencer VUGC on sales may suggest that when VUGC contents are not entertaining enough, substitution effect can dominate the promotion effect, which can possibly lead to a negative net effect. In particular, we can find that the substitution effect seems to be stronger

for for strategy games than non-strategy games. As its name suggests, strategy games are known to require strategic and skillful thinking and emphasize tactical challenges, thus, they are known to have a higher difficulty level than non-strategy games.[13] Such higher difficulty of strategy games may explain the higher substitution effect of low-VUGC videos on sales. With low entertainment value, viewers could feel that vicarious play with the VUGC might be easier than firsthand play which requires certain time and effort because the content describes the game similar to its original content. However, high-influencer VUGC could decrease the substitution needs by making the firsthand game play as challenging but enjoyable at the same time with uploader's commentary and play style.

For strategy games, the result indicates that the promotion effect of VUGC on game engagement may dominate the substitution effect when VUGC has high informational value. This might suggess that VUGC positively affects game engagement when users need to advance skills, and the high informational value of high-influence VUGC helps them achieve the goal. Specifically, the promotion effect seems to be stronger for strategy games by small developers. To promote game plays, game developers publish official books or videos containing detailed gameplay information, such as complete maps of the game, cheats, advice on tactics and strategies. Due to small game developers' limited marketing capabilities and resources, such firm-created contents are less available for indie games. Accordingly, for strategy games by indie developers, VUGC with high informational value could work as a guide in a place of official books or videos. However, such promotional effect of VUGC is dominated by the substitution effect when the VUGC has a low entertainment/informational value.

We perform several tests to diagnose the robustness of our model. To estimate Equation (1), we used second lagged period or earlier lagged values of endogenous variables as instrument variables (i.e., the instruments for x_t is x_{t-2}, x_{t-3}, ...,) as all lags of endogenous variables. The results reported in Table 5 are estimated using all available instrument variables. When using instruments for GMM estimation, the number of instruments can affect estimation precision and bias. As the number of instruments increases, the estimation precision could increase, but the estimation bias

[13] We collected user-rated game difficulty level of a game from GameFAQs (https://gamefaqs.gamespot.com/), and compared the difficulty level between strategy games and non-strategy games. On a 5-point scale (1 - too easy; 5 - too difficult), the average difficulty level is 3.07 for strategy games and 2.73 for non-strategy games. The t-test (2.77, $p < 0.01$) shows that there is a significant difference in difficulty between strategy games and non-strategy games.

Table 7: **Robustness Check on Number of Instruments**

	Time length	Main effect		Indie		Strategy		High Priced	
		High VUGC	Low VUGC	High VUGC	Low VUGC	High VUGC	Low VUGC	High VUGC	Low VUGC
Sales	T=28	0.482**	-0.118	-0.130	0.003	-0.074	-0.161	-0.323*	0.232
	T=25	0.475**	-0.110	-0.115	-0.019	-0.064	-0.159	-0.322*	0.232
	T=20	0.440**	-0.108	-0.107	-0.026	-0.079	-0.147	-0.292*	0.238
	T=15	0.460**	-0.129	-0.142	-0.001	-0.119	-0.132	-0.266*	0.223
Peak CCU	T=28	-0.045	-0.001	0.008	-0.089	0.206*	-0.054	0.076	0.036
	T=25	-0.041	-0.005	0.007	-0.087	0.202*	-0.054	0.071	0.039
	T=20	-0.038	-0.002	0.016	-0.093	0.202*	-0.055	0.064	0.035
	T=15	-0.075	-0.013	0.015	-0.079	0.224*	-0.058	0.098	0.015

Note. *: $p < 0.10, **: p < 0.05, ***: p < 0.01$

could also increase (K¨ubler et al. 2018). Thus, we check the sensitivity of our results with respect to the number of instruments. Table 7 presents the estimates of VUGC main and interaction effects across diverse number of instruments. Overall, the magnitude and direction of VUGC effects are consistent across different number of instruments. As the number of instruments decreases, the estimate of the main high-infleuncer VUGC for sales decreases by 0.02 (from 0.482 to 0.460). The decrease for the interaction between high-influencer VUGC and high priced game indicator in absolute term is 0.57 (from -0.323 to 0.26). The interaction between high-influencer VUGC and strategy for usage (Peak CCU) increases by 0.018 (from 0.206 to 0.224). However, such variations do not change the substantive findings, thus, our results are robust to the variation in number of instruments.

Our main analysis is based on the classification of VUGC uploaders by their ranks on the Twitch. High-influencers are the top 10,000 content contributors according to their number of follower, and the rest of VUGC uploaders are low-influencers. Because there is uncertainty regarding the cutoff rank for high-influencers, we examine the sensitivity of the results with respect to the different cutoff ranks. As shown in Table 8, the significance and direction of the estimates are not sensitive to the change in the cutoff rank. However, there are a few coefficients whose sign or significance changes across different cutoff ranks. The coefficient for the interaction between high-influencer VUGC and strategy game for sales model is negative when the high-influencers are defined as the top 10,000 uploaders, but the coefficient becomes positive when other high-influencer cutoff ranks are used in the model. Additionally, when high-influencers are top 7,000 or 6,000 uploaders, the interaction

Table 8: **Robustness Check on Influencer Classification**

		Main effect		Indie		Strategy		High Priced	
	High-influencer cutoff	High VUGC	Low VUGC	High VUGC	Low VUGC	High VUGC	Low VUGC	High VUGC	Low VUGC
Sales	Top 10,000	0.482**	-0.118	-0.130	0.003	-0.074	-0.161	-0.323*	0.232
	Top 9,000	0.355**	-0.102	-0.063	-0.030	-0.027	-0.166	-0.231	0.229
	Top 8,000	0.332*	-0.088	-0.035	-0.035	0.036	-0.203	-0.247	0.238
	Top 7,000	0.352**	-0.105	-0.067	0.009	0.064	-0.194	-0.285	0.257*
	Top 6,000	0.328*	-0.087	-0.015	-0.038	0.109	-0.195	-0.277*	0.277**
	Top 5,000	0.259	-0.068	0.040	-0.070	0.041	-0.133	-0.188	0.208
Usage	Top 10,000	-0.045	-0.001	0.008	-0.089	0.206*	-0.054	0.076	0.036
	Top 9,000	-0.039	-0.005	0.013	-0.097	0.190*	-0.038	0.090	0.030
	Top 8,000	-0.057	-0.011	0.016	-0.097	0.190*	-0.032	0.104	0.043
	Top 7,000	-0.059	-0.022	0.027	-0.082	0.206*	-0.038	0.096	0.055
	Top 6,000	-0.079	-0.028	0.011	-0.081	0.240**	-0.037	0.139	0.057
	Top 5,000	-0.088	-0.034	-0.013	-0.081	0.258**	-0.040	0.152	0.069

Note. *: $p < 0.10, **: p < 0.05, ***: p < 0.01$

between low-influencer VUGC and high-priced game indicator becomes significant. Thus, we check whether such changes in the coefficients affect the total effect of VUGC (linear combination of main and interaction coefficient) by game characteristics. The total effect shows the consistent magnitude, sign, significance across different high-influencer VUGC cutoff ranks, therefore, we conclude that our results are robust to different classification of high- and low-influencer VUGC. The total effect of VUGC across different cutoff ranks si available from the authors, upon request.

6 Video Content Analysis

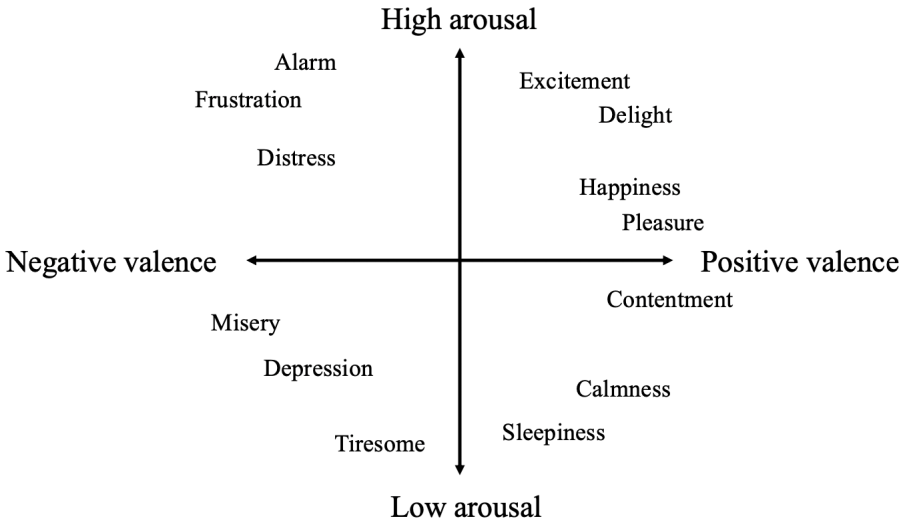
6.1 Video Affective Content Analysis

One of the key premises of this study is that VUGCs by high-influencers are different from those made by low-influencers, and thus the differences in videos lead to different levels of engagement and sales down the stream. In this section, we implement automated video affective content analysis to examine how these two types of videos differ. Conventionally, video content analysis has focused on analyzing a video at the cognitive level. The algorithms aim at extracting information that describes the "factual features" of a video, for example, the structure of the story, the composition of a scene, and the objects and people captured by the camera (Babaguchi and Nitta 2003). In this study, the video content analysis is carried out at the affective level. The purpose of the affective

content analysis is to define the amount and type of affect that the video can evoke in users during viewing (Wang and Ji 2015; Hanjalic and Xu 2005; Xu et al. 2008; Teixeria et al. 2012; Banse and Scherer 1996; Ayadi et al. 2011; Wang and Cheong 2006; Zhang et al. 2010).

Our video affective content analysis maps the type of affect on viewers into a two-dimensional space: valence and arousal (Bradley 1994; Lang et al. 1993; Russell 1980). Valence measures the degree of pleasure, with the negative values representing a "bad feeling" and the positive values representing a "good feeling". Arousal measures the activation level of the emotion. As a measure of excitement, arousal characterizes a state of heightened physiological activity, with the lower values representing a "passive feeling" and higher values an "active" or "excited" feeling. Research has shown that emotions can be plotted by these two dimensions (see Figure 2). For example, the second quadrant represents the emotions that are high in arousal and have positive valence. Examples of emotions belonging to this quadrant include "excitement" and "happiness." It is one of the objectives of this study to examine how the VUGC from high-influencers and low-influencers differs in their level of arousal and valence.

Figure 2: **Valence and arousal space of emotions by Russell (1980)**



6.2 Audio and Video features

The first step of affective content analysis is to extract visual and audio features that can capture videos' content characteristics. Wang and Ji (2015) presents a comprehensive list of visual and audio

features used for affective content analysis. For the audio features used in this study, we adopt some of the most frequently used spectral and rhythm measures (Banse and Scherer 1996; Wang and Cheong 2006; Xu et al. 2008; Teixeria et al. 2012; Ayadi et al. 2011). To capture the dynamic sound energy, we use Root-Mean-Square energy (RMSE) and zero crossing rate (ZCR). Spectral shape properties are captured with melspectogram, Mel-frequency Cepstrum Coefficients (MFCC), spectral centroid, spectral bandwidth, spectral flatness, spectral rolloff, and spectral contrast. To capture the pitch characteristics, we use Chromagram Short-Time Fourier Transform (STFT), Chromagram constant-Q transform, Chromagram Energy Normalized Statistics and tonal centroid features. To capture the rhythm characteristic, tempogram is used. Although these features were found to characterize various emotions, the features that point to the valence type and the level arousal vary across literature.

The visual features used by this study include the presence of the VUGC uploader's face in the video and emotional state of the uploader. VUGC uploaders are often observed to record their faces with a webcam and embed the recording in main VUGC game playing video. By doing so, the uploaders hope to enhance the virtual socializing experience with the viewers and establish their unique characters, so that they can increase viewers' engagement with the VUGC content. Although the size and position of the webcam section can vary across uploaders, the web cam is usually mounted on top of the streamer's primary monitor and records the uploader's upper torso and face so that viewers can focus on the emotions that the gamer feel while playing the game. Thus, we argue that the presence the uploader's face and the emotional state of facial expression can affect entertainment value of a VUGC. For example, VUGCs accompanied with recording of uploader's face with joyful emotional expressions such as laughing or smiling may have a higher entertainment value than those without the presence of face or those with negative emotional expressions such as anger.

6.3 Video Sub-Sample Summary Statistics

For video content analysis, a video is usually broken down into its most fundamental elements. For example, audio and visual features are extracted from a one-second video clip and a single frame image, respectively. Given the amount of data generated, the analysis can be computationally intensive even for a very short video. Due to this constraint, we conducted the video affective

content analysis using one game. The chosen game—the Citadel: Forged with Fire—was the most popular game among the influencers (top 10,000 Twitch VUGC uploaders) during our data collection period. The average length of VUGC in our sample is 95 minutes, which makes it still challenging to analyze all VUGCs in their full-length. Thus, we analyzed the user-generated clips of each VUGC. On Twitch, viewers of VUGC can create and share clips of that video. Twitch permits clips to last up to one minute and can be thought of as the summary of a VUGC. Across the 401 VUGCs posted on the Citadel (from 189 uploaders), we collected 1,070 clips, with the average length of the video clip being 31.73 seconds.

To measure the audio features, VUGC clips were collected in MPEG format and the accompanying audio was digitized at 22.1kHz. The initial step in acoustic feature extraction is audio type segmentation, i.e., to identify the sounds from different sources such as speech, music, and environmental sound. Compared with music and environmental sound (e.g., game play background sound), speech is more important in terms of influencing the arousal and valence of reactions from the VUGC viewers. Thus, we simplified the audio segmentation by keeping speech and removing all other types of sounds (i.e., music, noise, environmental sound). The segmentation was implemented using the audio software Audacity. It is observed that the majority of the VUGC uploaders have increased their speech sound while decreasing the volume of the game background sound. Such volume manipulation led the game background sound to be relatively constant which can be easily attenuated by the Audacity noise reduction. Thus, the proposed noise reduction procedure should allow reasonable accuracy in extracting speech features. After noise reduction, the VUGC clip's audio data was split into a series of segments with 0.1 second, and the average across all the 0.1-second segments was calculated and used as the measure of each sound feature, respectively.

To extract the visual features, each VUGC clip was split into a frame of one second and exported into the JPEG format. The 401 VUGC clips generated a total 824,571 image frames. Google Vision API was used to detect face presence and emotion for each frame[14]. Facial emotions are grouped into four states: joy, surprise, anger, and sorrow. For each image frame and each visual feature, Google Vision API returns a five-category likelihood score: very unlikely, unlikely, possible, likely, or very likely. The facial feature—face presence and emotion—is considered present in an image

[14]Google Vision API is an application program interface provided by Google that analyzes images and provides object recognition and face recognition using machine learning techniques, for example, Convolutional Networks or Linear Discriminant Analysis.

frame if the value is likely or very likely. Again, the percentage of emotion presence across all the frames of a VUGC is used as the measure for each visual feature.

Table 9: **Descriptive Statistics for Audio and Visual Features**

	Mean	SD	Median	Min	Max
Audio Features					
Sound energy: Root-Mean-Square energy	0.02	0.02	0.01	0	0.13
Sound energy: Zero crossing rate	0.12	0.04	0.11	0.02	0.54
Spectral shape: Melspectogram	0.37	0.69	0.14	0	7.96
Spectral shape: Mel-frequency Cepstrum Coefficients	-16.06	4.92	-16	-40.01	-0.81
Spectral shape: Spectral centroid	2,234.5	517.4	2,188.8	1,016.2	5,695.5
Spectral shape: Spectral bandwidth	2,211.8	293.9	2,192.7	1,219.9	3,115.1
Spectral shape: Spectral flatness	0.03	0.02	0.02	0	0.24
Spectral shape: Spectral rolloff	4,396.3	1,037.1	4,284.7	1,890.1	8,042.2
Spectral shape: Spectral contrast	19.5	1.3	19.5	16.3	24.7
Pitch: Chromagram Short-Time Fourier Transform	0.42	0.08	0.41	0.23	0.67
Pitch: Chromagram constant-Q transform	0.53	0.06	0.53	0.3	0.69
Pitch: Chromagram Energy Normalized Statistics	0.27	0.01	0.27	0.18	0.28
Pitch: Tonal centroid features	0.19	0.16	0.15	0	1.39
Rhythm: Tempogram	0.16	0.04	0.16	0.05	0.29
Visual Features					
Face presence	0.74	0.44	1	0	1
Joy	0.08	0.16	0	0	0.93
Surprise	0	0.05	0	0	1
Anger	0	0	0	0	0.06
Sorrow	0	0.01	0	0	0.3
Face size	3,857.9	7,178.9	2,090.5	0	74,824.1
Sample size			1,070		

Note. The audio and visual features are analyzed on a subset of 1,070 clips.

Table 9 presents the summary statistics of the visual and audio features for the clips on the Citadel. Overall, there are large variations in audio features among VUGCs, which indicates that selected audio features could be effective in characterizing video contents. For face presence, around 75% of the VUGCs had recorded and shown the uploader's face. Among the four emotional states, joy was the only emotional state with some variations and the other emotions had close-to zero presence and very little variation, indicating that VUGC uploaders tended to show positive valence more than negative valence.

6.4 Feature Comparisons Between High- and Low- Influencer VUGC

In section 3 and section 4, we defined the top 10,000 uploaders as the high-influencers and the rest VUGC uploaders on Twitch as low-influencers. To examine the differences in audio and

visual features between high-influencers and low-influencers, we need to collect VUGCs from both groups. Unfortunately, we do not have access to the VUGCs by low-influencers. Thus, we split the high-influencer VUGCs into upper high-influencers (i.e., the top 5,000 uploaders) and lower high-influencers (i.e., the bottom 5,000 uploaders). We then compared the video features between these two groups.

Table 10: **Feature Mean Comparison Between High- and Low-Influencers**

	Upper Influencer	Lower Influencer	t-value	p-value
Audio Features				
Sound energy: Root-Mean-Square energy	0.02	0.02	2.22	0.03
Sound energy: Zero crossing rate	0.12	0.11	1.88	0.06
Spectral shape: Melspectogram	0.38	0.34	1.08	0.28
Spectral shape: Mel-frequency Cepstrum Coefficients	-15.83	-16.47	2.09	0.04
Spectral shape: Spectral centroid	2,255.8	2,197.4	1.80	0.07
Spectral shape: Spectral bandwidth	2,216.1	2,204.3	0.63	0.53
Spectral shape: Spectral flatness	0.03	0.02	1.99	0.05
Spectral shape: Spectral rolloff	4,435.7	4,327.6	1.65	0.10
Spectral shape: Spectral contrast	19.54	19.46	0.99	0.32
Pitch: Chromagram Short-Time Fourier Transform	0.41	0.42	-0.45	0.65
Pitch: Chromagram constant-Q transform	0.53	0.54	-0.84	0.40
Pitch: Chromagram Energy Normalized Statistics	0.27	0.27	0.15	0.88
Pitch: Tonal centroid features	0.20	0.17	2.65	0.01
Rhythm: Tempogram	0.16	0.16	1.27	0.20
Visual features				
Face presence	0.76	0.71	1.72	0.09
Joy	0	0	0.14	0.89
Surprise	0.09	0.06	3.16	0.01
Anger	0	0.01	-1.13	0.26
Sorrow	0	0	1.16	0.25
Face size	4217.25	3231.24	2.35	0.02
Sample size		1,070		

Note. Within the cell is the mean of each feature.

Table 10 reports the differences in the visual and audio features between the two groups: upper high-influencers and lower high-influencers. The means are compared using an independent-sample t Test. The differences for four audio features—RMSE, spectral flatness, MFCC, and tonal centroid features—are significant at 0.05 level. Higher RMSE, spectral flatness, and MFCC indicates higher level of arousal (Xu et al. 2008; Teixeria et al. 2012), higher tonal centroid indicates more positive valence of the audio (Teixeria et al. 2012). In our data, excluding MFCC, the other three features have significantly higher mean for the VUGC from upper influencers than those from the lower influencer VUGC, indicating that upper influencer VUGCs had more positive valence and high levels of arousal. In terms of visual features, upper influencer videos were more likely to have

the recording of uploader's face. They also showed more positive facial emotion—joy—than the videos from lower influencers. Thus, we conclude that there are significant differences in content characteristics between upper influencer and lower influencer VUGC.

However, such video feature differences between upper influencer and lower influencer VUGC are based on unadjusted descriptive statistics, which do not indicate the effects of video content characteristics on gamers' responses, such as gamer's engagement or purchase intent. In the subsequent analysis, we relate the video characteristics with the number of views that each clip received on Twitch.[15] The number of views ranges from 2 to 23,132 with the mean being 197.2. Since persuasion becomes more likely with more viewing interest (Teixeira et al. 2014), the number of views can be a valid proxy for gamers' responses to the VUGC. We acknowledge that this analysis is based on observational data, which does not permit causal inference. A better study design is to examine the impact of video characteristics in a controlled lab experiment setting, which is one of our future directions.

We use Lasso regression to estimate the impact of video features on the number of views. Lasso regression minimizes the usual sum of squared errors in the linear regression model subject to constraining the sum of the absolute values of the coefficients (Tibshirani 1996). Lasso regression allows us to estimate a predictive model by simultaneously selecting only the influential features. Following the conversion, we used the log transformation and mean centered dependent variable and all independent variables during estimation.

For clip i, we regress the number of clip views on clip characteristics, audio features, and visual features, which goes as follows:

$$\begin{aligned} Views_i = \alpha + \beta_1 \cdot VUGC.Views_i + \beta_2 \cdot VUGC.Followers_i + \beta_3 \cdot VUGC.Rank_i + \beta_4 \cdot Clip.Days_i \\ + \beta_5 \cdot VUGC.Weekend_i + \beta_6 \cdot Clip.Weekend_i + + \Gamma \cdot Audio_i + \Lambda \cdot Visual_i + e_{it}, \quad (2) \end{aligned}$$

where the variable definitions can be found in Table 11. We include $VUGC.Views_i$, $Followers_i$, and $VUGC.Rank_i$ because they indicate the original VUGC content popularity which should also affect the number of clip views directly. We include the age of the clip $Days_i$ to control for the time effect. Additionally, $VUGC.Weekend_i$ and $Clip.Weekend_i$ are included because Twitch viewing

[15] The number of views is the cumulative number of views that each clip got as of June 10, 2020.

is known to be most active during the weekend.[16] Table 12 presents the summary statistics of the clip characteristics included in the Lasso regression. [17]

Table 11: **Definition of Variables Used in the LASSO Regression**

Variable	Definition
Outcome	
$Views_i$	The number of views for clip i
Clip Characteristics	
$VUGC.Views_i$	number of views for VUGC content of the clip i
$VUGC.Follower_i$	number of followers of the clip i's VUGC uploader
$VUGC.Rank_i$	rank of the clipi's VUGC among the VUGCs uploaded the sameuploader[18]
$Clip.Days_i$	date difference between clip i's upload date and the earliest uploaded clip's upload date (July 15, 2020)
$VUGC.Weekend_i$	weekend indicator of the clip i's VUGC upload date
$Clip.Weekend_i$	weekend indicator of the clip i upload date.
Audio Features ($Audio_i$)	
$Energy.RMSE_i$	Sound energy: Root-Mean-Square energy
$Energy.ZCR_i$	Sound energy: Zero crossing rate
$Spec.Mel_i$	Spectral shape: Melspectogram
$Spec.MFCC_i$	Spectral shape: Mel-frequency Cepstrum Coefficient
$Spec.Cent_i$	Spectral shape: Spectral centroid
$Spec.Band_i$	Spectral shape: Spectral bandwidth
$Spec.Flat_i$	Spectral shape: Spectral flatness
$Spec.Roll_i$	Spectral shape: Spectral rolloff
$Spec.Cont_i$	Spectral shape: Spectral contrast
$Pitch.STFT_i$	Pitch: Chromagram Short-Time Fourier Transform
$Pitch.CQT_i$	Pitch: Chromagram constant-Q transform
$Pitch.ENS_i$	Pitch: Chromagram Energy Normalized Statistics
$Pitch.Tonal_i$	Pitch: Tonal centroid features
$Rhythm.Tempo_i$	Rhythm: Tempogram
Visual Features ($Visual_i$)	
$Face_i$	Face presence
$Face.Size_i$	Face image size
Joy_i	Joy score
$Surprise_i$	Surprise score
$Anger_i$	Anger score
$Sorrow_i$	Sorrow score

By comparing the estimates of Lasso regression with and without the clip characteristics, we aim to show the importance of visual and audio features in predicting $View_i$. In Table 13, we report the linear regression and Lasso regression estimation results, which yield similar coefficients. When the clip visual and audio characteristics are not included, $VUGC.Views_i$ and $VUGC.Weekend_i$ are positive and statistically significant and $VUGC.Rank_i$ is negative and statistically significant in linear regression model.[19] Similarly, the coefficients of $VUGC.Followers_i$ and $Clip.Weekend_i$ have been shrunk to zero and the coefficient of $Clip.Days_i$ is close to zero in LASSO regression

[16] https://www.businessofapps.com/data/twitch-statistics/

[17] Since the summary statistics of audio and visual features of 829 clips are very similar to the summary statistics reported in Table 9, we only report the clip characteristic variables.

[19] If a clip has a small rank of a $VUGC.Rank_i$, it indicates that the VUGC of clip i is ranked highly among the VUGCs uploaded by the same uploader. For example, if the $VUGC.Rank_i$ is 1, the VUGC of clip i is the one with the highest views among the VUGCs, which might be seen as a counter-intuitive measure. However, considering that the total number of VUGCs uploaded varies across uploaders, it may preferable to assign smaller value to the VUGC with the highest views.

Table 12: **Descriptive Statistics for Clip Characteristics**

	Mean	SD	Median	Min	Max
$Views_i$	197.2	1271.98	11	2	23132
$VUGC.Views_i$	666.66	2317.98	160	6	28297
$VUGC.Followers_i$	127317.55	245693.28	49179	14701	2114667
$VUGC.Rank_i$	2.34	2.04	2	1	16
$Clip.Days_i$	205.41	26.16	200	0	338
$VUGC.Weekend_i$	0.36	0.48	0	0	1
$Clip.Weekend_i$	0.15	0.36	0	0	1
Sample size			829		

Note. The audio and visual features are analyzed on a subset of 829 clips because the number of views for some clips among the 1,070 clips used for the group comparison are not available.

model. Thus, three variables remain, $VUGC.Views_i$, $VUGC.Rank_i$, and $VUGC.Weekend_i$, which are the significant coefficients of LASSO regression model.

When audio and visual features are included, we find that the coefficients of some audio and visual features are statistically significant in linear regression model. The variables are $Energy.RMSE_i$, Joy_i, and $Sorrow_i$. The LASSO model estimation also show similar results. The LASSO model with the minimum Root Mean Square Error ($\lambda = 0.058$) has some non-zero video features, such as $Energy.RMSE_i$, $Spec.Cont_i$, $Pitch.CQT_i$, and Joy_i. Especially, $Eneregy.RMSE_i$ and Joy_i are commonly significant video features, and positive coefficient of $Energy.RMSE_i$ and Joy_i indicate that the VUGC with higher arousal and more positive emotion gets higher viewership. Along with findings from the video feature group comparison between upper high-influencer and lower high-influncer (Table 10, the findings may suggest that the VUGC by high-influencers may invoke more user responses than VUGC by low-influencers. Furthermore, the increase in the R-squared (from 0.354 to 0.363), although modest, suggests that including the audio and visual features that measure the affective aspects of the video can help improve predicting the viewership of VUGC.

7 Discussion

This study examines the impact of VUGC, a new form of UGC, on game product sales and usage. We utilize a unique and comprehensive panel-structure dataset complied from multiple sources. We apply the difference-GMM estimation for panel linear models to address endogeneity and un-

Table 13: **Linear Regression and Lasso Regression Estimation Results**

	(1) Linear regression without video features	(2) Lasso regression without video feature	(3) Linear regression with video features	(4) Lasso regression with video features
(Intercept)	0.060	0.072	0.122	0.051
$VUGC.Views_i$	0.614***	0.544	0.599***	0.504
$VUGC.Followers_i$	-0.054	.	-0.045	.
$VUGC.Rank_i$	-0.054***	-0.042	-0.053***	-0.027
$Clip.Days_i$	-0.025	0.000	-0.024	.
$VUGC.Weekend_i$	0.142**	0.054	0.131*	.
$Clip.Weekend_i$	0.070	.	0.055	.
$Energy.RMSE_i$			0.444**	0.103
$Energy.ZCR_i$			0.081	.
$Spec.Mel_i$			-0.093	.
$Spec.MFCC_i$			-0.024	.
$Spec.Cent_i$			-0.151	.
$Spec.Band_i$			0.073	.
$Spec.Flat_i$			-0.051	.
$Spec.Roll_i$			0.062	.
$Spec.Cont_i$			0.026	0.026
$Pitch.STFT_i$			-0.041	.
$Pitch.CQT_i$			0.015	-0.004
$Pitch.ENS_i$			-0.011	.
$Pitch.Tonal_i$			-0.243	.
$Rhythm.Tempo_i$			-0.062	.
$Face_i$			-0.093	.
$Face.Size_i$			0.030	.
Joy_i			0.064*	0.004
$Surprise_i$			-0.040	.
$Anger_i$			-0.027	.
$Sorrow_i$			0.055*	.
R-squared	0.348	0.354	0.362	0.363
RMSE	0.832	0.828	0.823	0.822

Note. *:p < 0.10, :p < 0.05, :p < 0.01.
Note. . in Column (2) and (4) indicates that the coefficients are shrunk to zero.

observed game brand-specific factors. Since VUGC can work as a word-of-mouth and a spoiler for game product at the same time, VUGC can both promote and cannibalize the sales and usage. We find that such promotion and cannibalization effects can be moderated by the VUGC type, i.e., whether the VUGC is from uploaders with high or low influencing power. Our results show that both high- and low-influencer VUGC affects game sales and usage, but the direction and magnitude of the effects vary by game product characteristics. Overall, high-influencer VUGC affect game sales positively, but such promotion effect could be offset by substitution effect when the game is more expensive. The low-influencers overall lead to substitution effect on sales and such substitution effect is more prominent for strategy games than for non-strategy games. We also find that the high-infleuncer VUGC affect game usage positively for strategy games by small

developers. With the extracted audio and visual features of VUGC, we show the differences in video content characteristics between high- and low-influencers, and estimate how VUGC audio and visual features affect viewers' responses.

This study contributes to the literature on the impact of UGC on consumer responses (Godes and Mayzlin 2004; Zhu and Zhang 2010; Chevalier and Mayzlin 2006; Moe and Trusov 2011). We empirically examine why and how VUGC affects consumers differently from the traditional textual VUGC. To the best of our knowledge, this is among the first few marketing studies that specifically model post-purchase behavior. Most of extant UGC studies have focused on UGC's impact on sales. However, for the long-term profit maximization, customer engagement with products is important in that it increases the likelihood of repeated purchasing of the same product or purchasing products by the same firm. Especially for video game developers, user engagement is critically important because it directly affects in-game purchase, one of the primary revenue sources for game developers. However, there is little empirical research on the impact of VUGC on post-purchase behavior. By examining the impact of VUGC on game usage, we provide evidence that VUGC can be used not only as a tool to affect customer awareness and attitude but also affect customer engagement.

We also contribute to the social media influencer research by examining the content differences between high and low inlfuencers. Previous studies on social media influencers mainly focus on identifying influential users (Trusov, Bodapati, and Bucklin 2010) or measuring their influence on social media platforms (Kumar and Mirchandani 2012). Instead, our study examines how video content feature characteristics differ between high- and low-influencers and show that such characteristic differences are associated with consumers' different level of reaction.

This study also has important managerial implications. Our findings suggest that game developers should consider their game's characteristics and the VUGC uploader's characteristics simultaneously when they develop and implement their policies on gaming VUGCs. For example, a game developer of a low-priced game can promote sales by providing its game for free to high-influencers or permitting them to use the game content for VUGC video upload. On the other hand, for a strategy game, the developer should implement a more strict policy on VUGC upload or mandate the gamers to get approval from the developer, so that the developers can have a control over who is allowed to upload the video. In particular, for a strategy game, our results suggest that the VUGC upload by low-influencers should be discouraged. Additionally, our video affect analysis can

help marketers to enhance their understanding in the VUGC content and to find the most effective influencers for their games.

There are several limitations to this study. First, in addition to a series of robustness checks with respect to the number of instruments and influencer cutoff points, robustness checks with an alternative model specification, for example, using the vector autoregressive model (VAR), can further validate our findings. Second, because of data unavailability, this study does not incorporate time-varying game brand-specific factors. The factors include firm's promotion level such as firm-generated video content volume, advertisement activity on TV or social media, or whether the game was featured on the front page on Steam. Note that the omission of these time-varying factors should not bias our estimates, as long as our instrumental variables are valid. That being said, future work can explicitly model these time-varying factors and shed light on their impacts on consumer responses. Third, our study can expand to include more game-level characteristics, such as sound, graphics, duration of game, use of humor, game dynamics, winning and losing features, character development. Including such game characteristics can enable us to explore richer heterogeneous effect of VUGC on game sales and usage. In the same vein, more video-content related features can be examined, such as the harmony or articulation of speech, tempo of game playing, brightness and color contrast of face, and lighting of the overall image frame. Finally, some might question the causality between VUGC and game sales and usage because viewership and game consumption do not take a place within one platform. An ideal way to show causality would be to track how many VUGC viewers click a "buy" button alongside the VUGC video window. With the access to the data, future research will be able to show the direct impact of VUGC on sales. Additionally, with the access to in-app purchase panel data, researchers can show the impact of VUGC on game usage. In the future, we also plan to conduct a lab experiment to further investigate the causal effect of video affect content features on viewers' real-time emotional responses.

8 References

Acemoglu, Daron, and James A. Robinson (2001) "A Theory of Political Transitions," American Economic Review, 91 (4), 938-963.

Akerlof, George A. (1978), "The Market for "Lemons": Quality Uncertainty and the Market Mechanism," Uncertainty in Economics, 237-251.

Albuquerque, Paulo, Polykarpos Pavlidis, Udi Chatow, Kay-Yut Chen, Zainab Jamal (2012), "Evaluating Promotional Activities in an Online Two-Sided Market of User-Generated Content," Marketing Science, 31 (3), 406-432.

Anderson, Todd W. and Cheng Hsiao (1981), "Estimation of Dynamic Models with Error Components," Journal of the American Statistical Association, 76, 598-606.

Arellano, Manuel and Stephen Bond (1991), "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," The Review of Economic Studies, 58 (2), 277-297.

Arellano, Manuel and Olympia Bover (1995), "Another Look at the Instrumental Variable Estimation of Error-components Models," Journal of Econometrics, 68 (1), 29-51.

Ayadia, Moataz El, Mohamed S. Kamel, and Fakhri Karray (2011), "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases," Pattern Recognition, 44 (3), 572-587.

AYTM (2014), "Twitch Survey: Video Game Streaming a Growing Industry" (accessed on June 10, 2020), [available at https://aytm.com/blog/twitch-survey/].

Babaguchi,Noboru and Naoko Nitta (2003), "Intermodal Collaboration: A Strategy for Semantic Content Analysis for Broadcasted Sports Video," Proceedings 2003 International Conference on Image Processing.

Bailey, Dustin (March 22, 2018). "With $4.3 billion in sales, 2017 was Steam's biggest year yet". PCGamesN. Retrieved March 22, 2018.

Banse, Rainer and Klaus R Scherer (1996), "Acoustic Profiles in Vocal Emotion Expression," Journal of personality and social psychology, 70 (3), 614-636.

Battlefy (2016), "Why Do People Watch Esports?" (accessed on June 10, 2020), [available at https://blog.battlefy.com/why-do-people-watch-esports-6ad7e8ec58b9].

Berger, Jonah, Alan T. Sorensen, and Scott J. Rasmus (2010), "Positive Effects of Negative Publicity: When Negative Reviews Increase Sales," Marketing Science, 29 (5), 815 – 827.

Blundell, Richard and Stephen Bond (1998), "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models," Journal of Econometrics, 87, 115-143.

Bradley, Margaret M. (1994), "Emotional Memory: A Dimensional Analysis," Emotions: Essays on emotion theory, Hillsdalc, NJ: Erlbaum, 97-134.

Business Insider (2018), "The World's Largest Gaming Service, Steam, is Giving Up on Regulation and Turning Over 200 Million Users into Guinea Pigs" (accessed on May 20, 2019), [available at https://www.businessinsider.com/valve-steam-no-rules-2018-6].

Chen, Pei-Yu, Shin-yi Wu, and Jongsun Yoon (2004), "The Impact of Online Recommendations and Consumer Feedback on Sales," ICIS 2004 Proceedings [available at https://pdfs.semanticscholar.org/dfae/57c8e77a85804522340d24186ccb2ee453c6.pdf].

Chevalier, Judith A. and Dina Mayzlin (2006), "The Effect of World of Mouth on Sales: Online Book Reviews," Journal of Marketing Research, 43 (August), 345-354.

Chicago Tribune (2015), "Survey Says More than Half of Shoppers Check Online Reviews," (June 2), [available at http://www.chicagotribune.com/business/ct-mintel-online-reviews-0603-biz-20150602-story.html].

Chintagunta, Pradeep K., Shyam Gopinath, and Sriram Venkataraman (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

Chou, Yuntsai (2003), "G-commerce in East Asia," Journal of Interactive Advertising, 4(1), 47-53.

Clark, C. Robert, Ulrich Doraszelski, and Michaela Draganska (2009), "The Effect of Advertising on Brand Awareness and Perceived Quality: An Empirical Investigation Using Panel Data," 7, 207-236.

CNBC (2013), "Watching other people play video games...that's esports," (accessed on June 11, 2020), [available at https://www.cnbc.com/2013/12/24/watching-other-people-play-video-gamesthats-esports.html]

Dellarocas, Chrysanthos (2006), "Strategic Manipulation of Internet Opinion Forums: Implications for Consumers and Firms," Management Science, 52 (10), 1577-15

Dhar, Vasant and Elaine A.Chang (2009), "Does Chatter Matter? The Impact of User-Generated Content on Music Sales," Journal of Interactive Marketing, 23 (4), 300 – 307.

Duan, Wenjing, Bin Gu, and Andrew B. Whinston (2008), "Do Online Review Matter? An Empirical Investigation of Panel Data," Decision Support Systems, 45 (5), 1007-1016.

Durlauf, Steven N., Paul A.Johnson, and Jonathan R.W.Temple (2005), "Chapter 8 Growth Econometrics," Handbook of Economic Growth, 1 (A), 555-677.

The Esports Observer (2015), "A profitable business: Why game publishers count on eSports", (accessed on June 11, 2020), [available at https://esportsobserver.com/a-profitable-business-why-game-publishers-and-developers-count-on-esports/].

Feng, Hui, Neil A. Morgan, and Lopo L. Rego (2015), "Marketing Department Power and Firm Performance," Journal of Marketing, 79 (3), 1-20.

Forbes (2017), "How Much Money Does 3 Billion YouTube Views Actually Generate For A Musician?," (September 18) [available at https://www.forbes.com/sites/hughmcintyre/2017/09/18/how-much-money-does-3-billion-youtube-views-bring-in/88e7da04aece].

Gamespot (2016), "That Dragon, Cancer Dev Says Let's Play Videos Hurt Sales," (March 26) [https://www.gamespot.com/articles/that-dragon-cancer-dev-says-lets-play-videos-hurt-/1100-6436034/].

Gameindustry.biz (2018), "Global Games Market Value Rising to $134.9bn in 2018," (acceesed May 29, 2019), [https://www.gamesindustry.biz/articles/2018-12-18-global-games-market-value-rose-to-usd134-9bn-in-2018].

Gameinformer (2018), "Valve Is Working On Replacing Steam Spy For Developers," (accessed on on May 29, 2019), [https://www.gameinformer.com/2018/06/30/valve-is-working-on-replacing-steam-spy-for-developers].

Germann, Frank, Peter Ebbes, and Rajdeep Grewal (2015), "The Chief Marketing Officer Matters!," Journal of Marketing, 79(3), 1-22.

Ghose, Anindya and Sang Pil Han (2011), "An Empirical Analysis of User Content Generation and Usage Behavior on the Mobile Internet," Management Science, 57 (9), 1671-1691.

Godes, David and Dina Mayzlin (2004), "Using Online Conversations to Study Word-of-Mouth Communications," Marketing Science, 23 (4), 545-560.

Google (2017), "4 Reasons People Watch Gaming Content on YouTube," (June) [https://www.thinkwithgoogle.com/consumer-insights/statistics-youtube-gaming-content/].

Gros, Daniel, Brigitta Wanner, Anna Hackenholt, Piotr Zawadzki, and Kathrin Knautz (2017), "World of Streaming. Motivation and Gratification on Twitch," Social Computing and Social Media. Human Behavior, 44-57.

Hanjalic, A. and Li-Qun Xu (2005), "Affective Video Content Representation and Modeling," IEEE Transactions on Multimedia, 7 (1), 143-154.

insivia (2018), "27 Video Stats for 2017," [available at http://www.insivia.com/27-video-stats-2017/].

Inc (2018), "Watching Video Games Is Now Bigger Than Traditional Spectator Sporting Events," [https://www.inc.com/darren-heitner/watching-video-games-is-now-bigger-traditional-spectator-sporting-events.html]

Iyengar, Raghuram, Christophe Van den Bulte, and Thomas W. Valente (2011), "Opinion Leadership and Social Contagion in New Product Diffusion," 30(2), 195-212.

Kang, Charles, Frand Germann and Rajdeep Grewal (2016), "Washing Away Your Sins? Corporate Social Responsibility, Corporate Social Irresponsibility, and Firm Performance," 80(1), 59-79.

K¨ubler, Raoul, Koen Pauwels, G¨okhan Yildirim, and Thomas Fandrich (2018), "App Popularity: Where in the World Are Consumers Most Sensitive to Price and User Ratings?," Journal of Marketing, 82 (September), 20-44.

Kumar,V. and Rohan Mirchandani (2012), "Increasing the ROI of Social Media Marketing," MIT Sloan Management Review, 54(1), 55-61.

Langhe, Bardt de, Philip M. Fernbach, and Donald R. Lichtenstein (2016), "Navigating by the Stars: Investigating the Actual and Perceived Validity of Online User Ratings," Journal of Consumer Research, 42 (6), 817-833.

Lang, Peter J., Mark K. Greenwald, Margaret M. Bradley, Alfons O. Hamn (1993), "Looking at Pictures: Affective, Facial, Visceral, and Behavioral Reactions," 30, 261-373.

Lu, Zhicong, Haijun Xia, Seongkook Heo, and Daniel Wigdor (2017), "You Watch, You Give, and You Engage: A Study of Live Streaming Practices in China," CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1-13.

Moe, Wendy W. and Michael Trusov (2011), "The Value of Social Dynamics in Online Product

Ratings Forums," Journal of Marketing Research, 48(3), 444-456.

Moe, Wendy W. and David A. Schweidel (2012), "Online Product Opinions: Incidence, Evaluation, and Evolution," Marketing Science, 31(3), 372-386.

Mudambi, Susan M. and David Schuff (2010), "Research Note: What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com," MIS Quarterly, 34(1), 185-200.

Nair, Harikesh S., Puneet Manchanda, and Tulikaa Bhatia (2010), "Asymmetric Social Interactions in Physician Prescription Behavior: The Role of Opinion Leaders," Journal of Marketing Research, 47(5), 883-895.

Narasimhan, Om, Surendra Rajiv, and Shantanu Dutta, (2006), "Absorptive Capacity in High-Technology Markets: The Competitive Advantage of the Haves," Marketing Science, 25 (5), 510-524.

New York Times (2017), "Using YouTube as an Accelerant for Video Games," (August 16) [https://www.nytimes.com/2017/08/16/technology/personaltech/using-youtube-as-an-accelerant-for-video-games.html].

Newzoo (2019), "More People Are Streaming on Twitch, But YouTube Is the Platform of Choice for Mobile-Game Streamers," (accessed on June 1, 2019), [https://newzoo.com/insights/articles/more-people-are-streaming-on-twitch-but-youtube-is-the-platform-of-choice-for-mobile-game-streamers/].

PCgamers (2017), "How much should games cost?," (accessed on June 8, 2019), [https://www.pcgamer.com/how-much-should-games-cost/].

Russell, J. A. (1980), "A Circumplex Model of Affect," Journal of Personality and Social Psychology, 39(6), 1161–1178.

Senecal, Sylvain and Jacques Nantel (2004), "The Influence of Online Product Recommendations on Consumers' Online Choices," Journal of Retailing, 80 (2), 159-169.

Sjöblom, Max and Juho Hamari (2017), "Why do people watch others play video games? An empirical study on the motivations of Twitch users," Computers in Human Behavior, 75, 985-996.

Sjöblom, Max, Maria Törhönen, Juho Hamari, and Joseph Macey (2018), "The ingredients of Twitch streaming: Affordances of game stream," Computers in Human Behavior, 20-28.

Spencer, Chloe (2017), "Nintendo Creators Program will no longer let YouTubers live-stream," Industry report, Kotaku.com.

Superdata Research (2017), "Report Launch: Gaming Video Content is bigger than streaming and sports networks," (accessed on June 13, 2019), [https://www.superdataresearch.com/report-launch-gaming-video-content-is-bigger-than-streaming-and-sports-networks].

Sun, Monic (2012), "How Does the Variance of Product Ratings Matter?," Management Science, 58(4), 696-707.

Statista (2017), "Gaming Video Content Market - Statistics Facts," (accessed on June 13, 2019), [https://www.statista.com/topics/3147/gaming-video-content-market/].

Statista (2019), "Gaming video content (GVC) revenue worldwide from 2016 to 2019 (in billion U.S. dollars)," (accessed on June 13, 2019), [https://www.statista.com/statistics/823289/gvc-revenue-worldwide/].

Tech Times (2018), "Game Devs Cry Over Steam Spy's Demise, While Others Celebrate," (accessed on May 30, 2019), [https://www.techtimes.com/articles/224977/20180412/game-devs-cry-over-steam-spys-demise-while-others-celebrate.htm].

Teixeira, René Marcelino Abritta, Toshihiko Yamasaki, and Kiyoharu Aizawa (2012), "Determination of Emotional Content of Video Clips by Low-level Audiovisual Features," Multimedia Tools and Applications, 61, 21-49,

Teixeira, Thales, Rosalind Picard, Rana el Kaliouby (2014), "Why, When, and How Much to Entertain Consumers in Advertisements? A Web-Based Facial Tracking Field Study," Marketing Science, 33 (6).

Thompson, Debora V. and Prashant Malaviya (2013), "Consumer-Generated Ads: Does Awareness of Advertising Co-Creation Help or Hurt Persuasion?," Journal of Marketing, 77(3), 33- 47.

Tibshirani, Robert (1996), "Regression Shrinkage and Selection Via the Lasso," Journal of the Royal Statistical Society: Series B (Methodological), 58 (1), 267-288.

Tuli, Kapil R., Sundar G. Bharadwaj, and Ajay K. Kohli (2010), "Ties That Bind: The Impact of Multiple Types of Ties with a Customer on Sales Growth and Sales Volatility," Journal of Marketing Research, 47 (1), 36-50.

Trusov, Michael, Randolph E. Bucklin, and Koen Pauwels (2009), "Effects of Word-of-Mouth Versus Traditional Marketing: Findings from an Internet Social Networking Site," Journal of Makreting, 73 (5), 90 – 102.

VentureBeat (2019), "NPD: U.S. game sales hit a record $43.4 billion in 2018" (accessed on

May 30, 2019), [https://venturebeat.com/2019/01/22/npd-u-s-game-sales-hit-a-record-43-4-billion-in-2018/].

Wang, Hee Lin and Loong-Fah Cheong (2006), "Affective Understanding in Film," IEEE Transactions on Circuits and Systems for Video Technology, 16 (6), 689-704.

Wang, Shangfei and Qiang Ji (2015), "Video Affective Content Analysis: A Survey of State-of-the-Art Methods," IEEE Transactions on Affective Computing, 6 (4), 410-430.

WePC (2019), "2019 Video Game Industry Statistics, Trends Data" (accessed on June 1, 2019), [https://www.wepc.com/news/video-game-statistics/].

Wikipedia (2018), "Joseph Garrett" (accessed on July 8, 2020), [https://en.wikipedia.org/wiki/Joseph$_Garrett$].

Williams, Dmitri (2002), "Structure and Competition in the US Home Video Game Industry," International Journal on Media Management, 4(1), 41-54.

Windmeijer, Frank (2005), "A Finite Sample Correction for the Variance of Linear Efficient Two-step GMM Estimators," 126 (1), 25-51.

Xiong, Guiyang and Sundar Bharadwaj (2013), "Asymmetric Roles of Advertising and Marketing Capability in Financial Returns to News: Turning Bad into Good and Good into Great," Journal of Marketing Research, 50 (6), 706-724.

Xu, Min, Jesse S. Jin, Suhuai Luo, and Lingyu Duan (2008), "Hierarchical movie affective content analysis based on arousal and valence features," Proceedings of the 16th ACM international conference on Multimedia, October, 677-680.

Xu, Min, Jinqiao Wang, Xiangjian He, Jesse S. Jin, Suhuai Luo, and Hanqing Lu (2014), "A Three-level Framework for Affective Content Analysis and its Case Studies," 70, 757-779.

Zhu, Feng and Xiaoquan Zhang (2010), "Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics," Journal of Marketing, 74 (2), 133-148.

Zhang, Xiaoquan and Chrysanthos Dellarocas (2006), "The Lord Of The Ratings: Is A Movie's Fate is Influenced by Reviews?," ICIS 2006 Proceedings [available at http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1238context=icis2006].

Zhang, Shiliang, Shuqiang Jiang, and Wen Gao (2010), "Affective Visualization and Retrieval for Music Video," IEEE Transactions on Multimedia, 12 (6), 510-522.