Please Rate After Riding: The Impact of Formal Evaluation on Peers? Feedback

Autoria

Louise Helene Goncalves Foernges - lou_gf@hotmail.com
Prog de Pós-Grad em Admin/Esc de Admin - PPGA/EA/UFRGS - Universidade Federal do Rio Grande do Sul

Cristiane Pizzutti dos Santos - cristiane.pizzutti@ufrgs.br
Prog de Pós-Grad em Admin/Esc de Admin - PPGA/EA/UFRGS - Universidade Federal do Rio Grande do Sul

Resumo

The act of sharing among peers is a growing phenomenon with many successful platforms that facilitate such interactions having emerged in the last decade. Since these new forms of sharing work mostly on the basis of sharing among strangers, reputation mechanisms have become necessary to aid peers in selecting which users can be trusted. However, often the feedback given by the users of such platforms is positively biased. To investigate this phenomenon, we conducted two scenario-based experiments using the context of on-demand transportation services. We compared feedback (in the form of rating and tip) in a formal type of evaluation to a control condition (informal). In Study 1, we confirm the existence of feedback bias in a formal evaluation system when the feedback provided is in form of ratings. Importantly, we also find that tip is a less biased form of feedback than ratings. In Study 2, we find that a high overall peer score leads to greater feedback bias in a formal type of evaluation (vs control). Additionally, in Study 2 we find anticipation of guilt to be mediator for the effect of type of type of evaluation on rating. Managerial implications and suggestions for further research are discussed.

Please Rate After Riding: The Impact of Formal Evaluation on Peers' Feedback

ABSTRACT

The act of sharing among peers is a growing phenomenon with many successful platforms that facilitate such interactions having emerged in the last decade. Since these new forms of sharing work mostly on the basis of sharing among strangers, reputation mechanisms have become necessary to aid peers in selecting which users can be trusted. However, often the feedback given by the users of such platforms is positively biased. To investigate this phenomenon, we conducted two scenario-based experiments using the context of on-demand transportation services. We compared feedback (in the form of rating and tip) in a formal type of evaluation to a control condition (informal). In Study 1, we confirm the existence of feedback bias in a formal evaluation system when the feedback provided is in form of ratings. Importantly, we also find that tip is a less biased form of feedback than ratings. In Study 2, we find that a high overall peer score leads to greater feedback bias in a formal type of evaluation (vs control). Additionally, in Study 2 we find anticipation of guilt to be mediator for the effect of type of type of evaluation on rating. Managerial implications and suggestions for further research are discussed.
Key-words: collaborative consumption; feedback; service failure; anticipation of guilt; overall peer score.

1. INTRODUCTION

In the last few years, the world has seen a growing movement towards de-ownership and sustainable use of resources (OZANNE & BALLANTINE, 2010; LINDBLOM & LINDBLOM, 2017). The term sharing economy (MALHOTRA & VAN ALSTYNE, 2014), also commonly known, with a similar meaning, as collaborative consumption (BELK, 2014b; BOTSMAN & ROGERS, 2010; MÖHLMANN, 2015; BENOIT *et al*., 2017) among other nomenclatures, essentially refers to peer-to-peer (P2P) interactions where individuals have temporary access to a good or a service without ownership transfer (BELK, 2014b; BARDHI & ECKHARDT, 2012). Usually, an online platform connects the peers who are willing to provide a service (peer-provider) or share a resource with users who are looking for that service or resource (peer-user) (BENOIT *et al*., 2017).

It is noteworthy that some collaborative service interactions are very personal (BELK, 2014a), that is, interactions that involve a high level of personal contact, such as when someone opens their house to a stranger on Airbnb or becomes a guest at a stranger's house (BRIDGES, VÁSQUEZ, 2016). Since these personal interactions occur between strangers, collaborative services platforms commonly employ reputation/feedback mechanisms (BELK, 2014b; BRIDGES, VÁSQUEZ, 2016; HOFMANN *et al.*, 2017; BENOIT *et al.*, 2017). These self-regulatory feedback systems (usually in the form of ratings and/or reviews) help minimize risks, discourage misbehavior and create trust among peers in collaborative services (BELK, 2014b; BRIDGES, VÁSQUEZ, 2016; HOFMANN, *et al.*, 2017).

According to Malhotra and Van Alstyne (2014 p. 27), "the viability of shared services hinges on the quality of review systems because people rely on them to decide whether and what to purchase (…) authenticating the validity of reviews is critical to prevent abuse". However, evidence suggests that reputation/feedback systems in collaborative services are not totally reliable and feedback bias often occurs, as peers tend to avoid giving negative ratings/reviews in this context (ZERVAS *et al*., 2015*;* FRADKIN *et al*., 2015; BRIDGES & VÁSQUEZ, 2016). Underreporting of negative experiences has, in turn, been linked to the personal nature of collaborative services, where 'social norms' are presumably being followed

(BRIDGES & VÁSQUEZ, 2016; ZERVAS *et al.*, 2015). Due to the importance of reputation/feedback mechanisms to help users decide who is trustworthy among the peers and mitigate users acting purely out of self-interest (BELK, 2014b; HAMARI *et al.*, 2016; BRIDGES & VÁSQUEZ, 2016), it seems that feedbacks are especially important in the occurrence of service failures.

We propose that anticipation of guilt, aspect connected to social harmony and empathy (MICELI, 1992), is behind feedback bias in formal evaluations in collaborative services instead of reciprocity or fear of retaliation –variables often found in literature as mediators explaining feedback bias in mutual evaluation systems (where both peer-provider and peer-user are able to evaluate each other) (DELLAROCAS & WOOD, 2008; RESNICK & ZECKHAUSER, 2002). Additional to the effect of type of evaluation (i.e. formal or informal) on feedback through anticipation of guilt, we also address the impact that different types of service failures (morality, competence and warmth) and overall driver score (high or low) have on this effect.

To investigate our hypotheses on aspects that potentially interfere in and explain feedback bias in collaborative services we conducted two cenario-based experimental studies. Using a similar approach to Zervas *et al.* (2015), we compared feedback (in the form of consumers' rating and tip) in a formal type of evaluation (i.e. the traditional in the app evaluation -or "in the system") and an informal one (i.e. to a friend -or "out of the system"), which served as a control condition, in an on-demand transportation context analogue to Uber.

According to Guyader (2018), there is a lack of research on how the peers (users and providers) integrate aspects of the market exchange and pro-social norms into their practices and interactions with one another. The author adds that further investigating collaborative consumption practices would benefit service research. Our study shows how certain aspects at play during that interaction can be determinant for the validity of feedback in collaborative services, therefore contributing to the development of theory on the subject. Also, by investigating the impact that different types of failure have on feedback we add to the literature of service failure through investigating the impact of different failures in a rather unexplored context.

2. THE IMPACT OF FORMAL EVALUATIONS ON CONSUMERS' FEEDBACK

In collaborative service markets, often feedback and reputation systems are employed in order to mitigate user's actions in self-interest and entail trust between them (RESNICK *et al.*, 2000; JØSANG, ISMAIL & BOYD, 2007; BELK, 2014b; HAMARI *et al.*, 2016). Also, the existence of such systems aims at motivating individuals to behave in a responsible manner (BOTSMAN & ROGERS, 2010; HOFMANN *et al.*, 2017). These systems allow peers to rate and/or review each other and sometimes display an overall score (average of received ratings).

There is evidence suggesting the existence of feedback bias in these feedback mechanisms (ZERVAS *et al.*, 2015*;* FRADKIN *et al.*, 2015; BRIDGES & VÁSQUEZ, 2016). According to Zervas *et al.* (2015) it is possible that 'sociological effects' lead people to be more diplomatic in their reviews in collaborative consumption services. Similarly, Bridges and Vásquez (2016) argue that when reviewing less than positive experiences, users prefer to leave neutral commentaries instead of negative ones. It is almost as if the users follow an implicit established 'norm' when leaving reviews. "Norms governing communication and interaction become established in a particular online space by the community members who interact with one another in that space" (BRIDGES & VÁSQUEZ, 2016, p.14).

> H1: In the occurrence of a given service failure, when formally (in the system) evaluating the provider (vs. informally/out of the system), users will give a more positive rating.

Another form of feedback recently incorporated to collaborative services, such as Uber, is tipping. According Azar (2009), tipping is usually a demonstration of gratitude for the quality of the service. Similarly, to Lynn and McCall (2000) argue that tipping is another way a customer can exercise quality control over the service, working, in fact, as a similar mechanism to ratings. The authors argue that, consistent with equity motivations theory, there is a positive correlation between service evaluations and tip size. However, since in a collaborative services context evaluations can potentially have long term consequences for the provider whereas the amount of tip does not directly harm the provider's livelihood, we propose that tips will be a less biased method of feedback than ratings.

H2: In the occurrence of a given service failure, tip will be a less biased form of feedback than ratings

2. ANTICIPATION OF GUILT

As Bridges and Vásquez (2016) point out, sociocultural factors such as politeness and courtesy may be one reason for the positive bias of feedbacks in Airbnb. Similarly, Zervas *et al*. (2015) argue that in collaborative consumption services (e.g. Airbnb), sociological factors often lead users to be more diplomatic in their reviews than in 'professional' services (e.g. hotels). The proximity with the provider seems to lead users to being more empathetic and avoid leaving negative reviews not to be 'unkind' (BRIDGES & VÁSQUEZ, 2016).

It is worth point out that, although extant literature points out that feedback bias can be explained by reciprocity or fear of retaliation (DELLAROCAS & WOOD, 2008; RESNICK *et al.*, 2000 RESNICK & ZECKHAUSER, 2002; BOLTON *et al.*, 2013; FRADKIN *et al.*, 2015), due to the personal nature of collaborative services, the apparent exitance of social norms in this context (which does not seem to occur in traditional services) and the fact that most evaluation systems in such services are now double-blind (i.e. peer-user and peer-provider don't have access to each other's specific evaluation or the evaluations are only made public after both parties have submitted them), we believe that these two explanatory mechanisms (reciprocity and fear of retaliation) do not apply to a collaborative service context. Instead, feelings of anticipation of guilt will take place as an underlying mechanism for the effect of the type of evaluation on feedback (in this case, in the form of rating).

For Baumeister, Stillwell and Heatherton (1994, p.243) "guilt is something that happens between people rather than just inside them. That is, guilt is an interpersonal phenomenon that is functionally and causally linked to communal relationships between people". The authors argue that guilt feelings are invoked not only for the self (such as to bolster self-control) but in a variety of human interactions (to apologize for wrongdoings or express sympathy, for example). The authors further add that the feeling of guilt comes from an anticipation -or the actual feeling- of the suffering of another. Therefore, the anticipation of guilt is responsible for an individual's performing or avoiding certain actions. In line with this, Steenhaut and Kenhove (2006) argue that the anticipation of guilt works as a mechanism to stop a certain behavior or to control action.

The anticipation of guilt may be aroused by the thought of a transgression or failure, which people tend to avoid, and a motivation to "comply with behavioral requests that will help them avoid future feelings of guilt" (LINDSEY, YUN & HILL, 2007, p.468). Therefore, we propose that anticipation of guilt is a mediator for the effect of type of evaluation on rating, instead of fear of retaliation or reciprocity.

H3: In the occurrence of a given service failure, the effect of type of evaluation on feedback will be mediated by anticipation of guilt.

3. SERVICE FAILURE

One of the main reasons why feedback systems exist in collaborative consumption services is to make sure failures are detected and reported (RESNICK *et al*., 2000). A service failure occurs when there is a problem in the delivery of a service and it fails to meet customer's expectations (HESS JR., GANESAN & KLEIN, 2003). Service failures are unavoidable, frequently elicit negative feelings and reactions in the customer (SMITH & BOLTON, 1998) and often lead to a feeling of violated trust (WANG & HUFF, 2007; BASSO & PIZZUTTI, 2016).

According to Mattila (2001), services that involve a high degree of human contact are particularly prone to failures. Three types of failures are commonly found in literature: morality, competence and warmth. Morality relates to honesty and "the relationship between integrity and trust involves the trustor's perception that the trustee adheres to a set of principles that the trustor finds acceptable" (MAYER *et al*., 1995, p.719). A competence failure -sometimes referred to as a capability failure- occurs when the seller/service provider lacks the skills and/or resources to perform a task, failing to satisfy the customer (WANG & HUFF, 2007). According to Kirmani *et at.* (2017) warmth includes traits of being sociable, playful, happy, and funny. A warmth failure occurs when these traits lack in the provider (being unfriendly, cold, unsociable etc). According to Kirmani *et al*. (2017), when choosing service providers, consumers value competence traits more than integrity -if this trait does not affect the service provided directly. The authors found that, in the context of a service failure, when choosing a service provider, knowledge and skill to perform the task is considered more important than morality or warmth traits. Since in a collaborative service's context competence and warmth failures are more likely to directly affect the service, we decided to test these two failures and not consider a morality failure.

Nierenberg and Goldstein (1976) argue that the stability of a cause determines expectancy shifts. According to the authors, if conditions of a certain situation are expected to remain same (such as the difficulty of a task or an individual's level or ability) then the outcome of past occasions is expected to reoccur. However, according to the authors, if the causal conditions are perceived as likely to change, then the present outcome may not be expected to reoccur. As a warmth failure is often depended on an individual's social skills (Kirmani *et al*., 2017) whereas a competence failure is often situation-specific, we propose that for being perceived as less stable, a competence failure will generate more positive feedback than a warmth failure.

H4: The type of failure will moderate the effect of type of evaluation on users' rating, such that a) for a competence service failure, ratings in formal and informal types of evaluation will not be significantly different, while b) for warmth failure, ratings will be higher in the formal (vs informal system).

H5: Type of failure will moderate the effect of type of evaluation on tip, such that a) for a competence service failure, tips in formal and informal types of evaluation will not be significantly different, while b) for morality and warmth failures, the amount of tip will be higher in the formal (vs informal system).

4. STUDY 1

Design and Participants

The first experiment was a factorial 2 (type of failure: competence and warmth) x 2 (type of evaluation: formal -in the system, control -informal) between-subjects design with random assignment. The aim of Study 1 was to test if formal evaluations are, in fact, positively biased, if reciprocity or fear of retaliation were mediators for the effect of type of evaluation on rating and the moderating role of type of failures. For this study a sample of 239 participants was recruited via Amazon's Mechanical Turk (MTurk) service. The questionnaire was presented in English and only US residents were eligible to respond. Most participants (65,3%) declared to be between 18-34 years old and 53,1% were females. Of the total sample (N = 239), 10 (4,2%) participants declared to be handicapped.

Procedure

The participants were presented with a 1-minute (approximately), muted and subtitled video in point-of-view format. Each vignette depicted the same scenes showing an interaction between peers in the context of on-demand transportation. However, the videos were subtitled differently, representing different interactions (dialogs) between passenger and driver, introducing the type of failure manipulation (competence or warmth). In the condition in which the driver seemed to lack competence, for example, the failure consisted in the driver telling the passenger her phone was off because her phone charger hadn't been working properly. The driver then asks if the passenger can give her directions to which the passenger replies she would try. Following the video, a text introducing the manipulation for type of evaluation was presented to the participants. In the formal (in the system) condition, participants were told they would be evaluating the driver in the service's app, in a scale from 1 to 5. In the control (out of the system) condition, participants were told they would be evaluating their experience to a friend, also in a scale from 1 to 5. After the manipulations were introduced, participants were requested to complete a questionnaire which collected data.

Measures

Rating was measured with a slider scale which users were asked to use to indicate how they would rate the driver in the video, in a scale ranging from 1 to 5. Tip was measured with a slider scale which participants were asked to use to indicate the amount of tip they would be willing to give the driver, in a scale ranging from 0% to 25% of the total price of the ride. The following variables were all measured in 7-point Likert scales.

To check the manipulations for type of failure (competence and warmth) we used the same scales provided in Kirmani et al. (2017), from Leach, Ellemers, and Barreto (2007), each consisting of 4-items. Fear of retaliation was measured with one question, adapted from Kudish, Fortunato and Smith (2006): "I fear to suffer negative consequences if I give an honest feedback to this driver". Reciprocity was also measured with one item: "I rated the driver according to how I expect the driver has rated me as a passenger". Perceived severity of the failure was measured using a 3-item scale (e.g. "If the inconvenience during the ride in the video was really happening to you, you would consider it to be"), from Weun, Beatty and Jones (2004). All measures were in 7-point scales.

Validity of Scales

An exploratory factor analysis utilizing Varimax rotation method revealed the scales exhibit satisfactory factorial structure. Cronbach's Alpha was also used to measure reliability of the scales. The statistical analysis of the scales yielded, in general, good alphas: competence ($\alpha = .89$); warmth ($\alpha = .96$) and perceived severity of the failure ($\alpha = .80$).

Main Study Results

Rating

To test for main and interactive effects between the type of failure and type of evaluation on rating we performed an ANCOVA test controlling for perceived severity of the failure. The ANCOVA analysis showed a significant direct effect of type of evaluation in rating ($F(1,234) = 19.324$, $p<.001$, $\eta^2p = 0,076$). No significant direct effect of type of failure in rating was found ($p>.05$). The initial ANCOVA test revealed that different types of evaluation methods affected the way user's rate their experience with the driver differently. In support of H1, the formal (in the system) type of evaluation yielded a significantly ($p<.001$) higher rating (M = 2,88) than the control condition (M = 2,46). However, the test showed that for rating, the interactive effect of type of failure and type of evaluation was non-significant (F (1,234) = 0.515, $p>.05$, η^2p = 0,002). When considering each failure individually, results indicate no difference in the means of rating between conditions, therefore H4a was supported but H4b was not.

Reciprocity and Fear of Retaliation

Results of mediation analysis confirmed no mediation of reciprocity or fear of retaliation exists between type of evaluation and rating. The tests revealed there is no significant path between type of evaluation and reciprocity (b = -0,09, se = 0,10, t = -0,86, $p>.05$, confidence interval (CI) between -0,30 and 0,11). The path between reciprocity and rating is significant (b = 0,09, se =0,04, t = 2,27, $p<.05$, confidence interval (CI) between 0,01 and 0,17), however the indirect effect is non-significant (confidence interval (CI) between -0,03 and 0,12). Similarly, results indicated no significant path between type of evaluation and fear of retaliation (b = -0,05, se = 0,10, t = -0,48, $p>.05$, confidence interval (CI) between -0,26 and 0,15). Nor a significant path between fear of retaliation and rating (b = 0,07, se = 0,04, t = 1,88, $p>.05$, confidence interval (CI) between -0,003 and 0,15). The indirect effect is non-significant (confidence interval (CI) between -0,02 and 0,01).

Tip

We performed an ANCOVA to test for main and interactive effects between type of service failure and type of evaluation on tip. The test indicated no main effect of type of service failure on tip ($p>.05$), nor type of evaluation on tip $F(1,234) = 0,200$, $p>.05$, $\eta^2p= 0,001$). This finding along with the finding that there is a significant difference between types of evaluation on rating lends support to H2, suggesting tips tend to be less biased than ratings. The results also showed a significant interactive effect between type of service failure and type of evaluation on tip ($F(1,234) = 4,963$, $p<.05$, $\eta^2p= 0,021$).

To further investigate the interactive effect of type of evaluation and type of failure on tip, a Spotlight analysis was conducted. When considering the competence and warmth types of failure individually, results indicate no significant difference ($p>.05$) in the means of tip between conditions, therefore H5a is supported but H5b is not. The test shows, however, that in the formal condition of type of failure, there is a significant effect in rating between the warmth (M = 4,36) and competence (M = 6,08) conditions of type of failure.

Discussion

Results of Study 1 indicated strong main effects of type of type of evaluation in rating but no interactive effect. The results revealed the mean of rating was significantly higher in the formal type of evaluation when compared to the control (informal) condition, therefore lending support to H1.

Study 1 also revealed that, contrary to rating, the mean of tip was not significantly different between type of evaluation conditions, therefore indicating a possibly less biased method of feedback and lending support to H2. In H4a, we proposed that for a competence service failure, the mean of tip in the formal type of evaluation would be significantly higher

than in the warmth condition, which was supported by our results which revealed only in formal type of evaluation, the competence type of failure condition had a significantly higher mean of tip when compared to a warmth failure.

6. OVERALL PEER SCORE

Peer scores are part of feedback/reputation mechanisms in on-demand transportation services (Uber) and room sharing (Airbnb). These mechanisms allow peers to establish a reputation based on other peer's performance evaluations (WEBER, 2014). According to Bridges and Vásquez (2016), these scores are an important tool as they serve as a cue to peer past behavior and are based in other user's personal experience. The authors point that various studies show individuals take online reviews into consideration before making decisions. Using peer scores as a clue of past behavior is in line with attribution theory. According to Weiner (1972), stability attribution is the expectancy that the cause of an event will remain stable and not fluctuant over time. The author argues that (p. 556-557) "if conditions (the presence or absence of causes) are expected to remain the same, then the outcome(s) experienced in the past will be expected to recur". In other words, the driver score helps users to attribute stability to the failure the experienced.

As literature points out, it is possible that the previous ratings/reviews serve as a cue to past behavior to users/customers of collaborative servicesp[o (BRIDGES & VÁSQUEZ, 2016). We propose that when the driver score is high, users will tend to believe the transgression is not a recurrent issue, as the score serves as cue of adequate past behavior, therefore when formally rating the driver, a driver high score will lead to more positive (i.e. biased) ratings than when informally evaluating. However, and more importantly, when the driver has a low score, pointing to inadequate past behavior, means of rating are expected to remain the same between types of evaluation conditions, i.e., consumers will evaluate the driver in a more objective way in the formal system. In other words, in this situation, passengers may think failures caused by the provider are recurrent and feel less obligated to give a positive (biased) feedback about the driver (after all, it is likely he/she has behaved badly in the past as well and does not seem to deserve to get a "false" good rating this time).

> H6: Overall driver score will moderate the effect of type of evaluation on users' rating, such that in the high driver score condition, means of rating will be significantly higher in the formal type of evaluation than in the control (informal) one, while when driver score is low means of rating will remain unaltered between types of evaluation conditions.

In the next study, besides testing the hypotheses initially proposed, we also explore H3 and H6 which concern the role of anticipatory guilty and overall peer score.

7. STUDY 2

Design and Participants

Study 2 was a 2 (competence, warmth) x 2 (overall peer -driver- score: high, low) x 2 (type of evaluation: formal -in the system, control -informal) between-subjects experimental design with random assignment. Similar to Study 1, Study 2 was also conducted online, via Mechanical Turk. The final sample included 349 individuals with mean age of 33,9 years where 53% were males. Of the total sample, 10 (2,9%) participants declared to be handicapped.

Procedure

The introduction explained to the participants the on-demand transportation service in the video was called TakeMe and was similar to services such as Uber and Lyft. The video

manipulation for type of service failure was the same as in Study 1. Following the video, an image introduced the manipulation for overall peer score. The image depicted a screenshot of a smartphone showing the app after the passenger requested the ride. In the driver high score condition a score of 4.98 was shown under the driver's name with a star next to it. In the driver low score condition, a score of 3.29 was shown. Following the type of failure and driver score manipulations, we introduced the type of rating system manipulation. This manipulation was similar to Study 1, but we included the fictitious name TakeMe for the service in Study 2 in order to bring more realism into the storytelling of the experiment. A similar questionnaire to Study 1 was then presented to participants with the additional scale for anticipation of guilt.

Measures

In Study 2 we introduced the scales for overall driver score and anticipation of guilt. Overall driver score was measured with one item: "Do you consider the current rating (score) of the driver to be", with three options answer: "high", "low" and "average". The items for anticipation of guilt were adapted from Basil *et al.* (2006). The two items were rated in 7-point scales.

Validity of Scales

An exploratory factor analysis using Varimax rotation method was conducted. The analysis revealed the scales exhibit satisfactory factorial structure. Cronbach's Alpha was also used to measure reliability of the scales. The statistical analysis of the scales yielded, in general, good alphas: competence (α = .91); warmth (α = .95) and perceived severity of the failure (α = .52). Anticipation of guilt was measured with a 2-item scale, therefore we used correlation analysis to assess reliability (r = 0,912, *N* = 522, p<.001).

Main Study Results

Rating

To test for main and interactive effects between the type of service failure and the type of evaluation on rating we performed an ANCOVA test which showed no significant main effects of type of service failure or overall driver score (p>.05), but a main effect of type of evaluation ($F(1, 340) = 5.768$, $p<.05$, $\eta^2p = 0,017$) on rating was found. In the formal condition of type of failure, ratings were, as in Study 1, significantly higher (M = 3,25) than in the informal condition (M = 3,01), therefore confirming H1. Also, the ANCOVA test revealed an interactive effect of driver score and type of evaluation on rating ($F(1, 340) = 7.035$, $p<.05$, $\eta^2p = 0,020$).

A Spotlight analyses confirmed an interactive effect occurs between overall peer score and type of evaluation. The test shows when the overall peer score is high, there is a significant difference (p<.05) in the means of rating between the formal (M = 3,40) and the control condition (M = 2,84) However, when the overall peer score is low, there is no significant difference (p>.05) in the means of rating between type of evaluation conditions, confirming H6.

Spotlight analysis indicated that when considering each type of failure individually, none of the failures showed a statistically significant difference in the means of rating (p>.05). Therefore, H4a is confirmed but H4b is not.

Anticipation of Guilt Mediation

In Study 2, we included anticipation of guilt as a possible mediator for the effect of type of evaluation on rating. Results of mediation analysis revealed a significant path from type of evaluation to anticipation of guilt (p<.05) and anticipation of guilt to rating (P<.05), while the main effect between type of evaluation and rating disappears (p>.05), revealing a full mediation.

An ANOVA test revealed that in a formal (M = 4,25) type of evaluation, means of anticipation of guilt were significantly ($F(1, 347) = 4.568$, $p<.05$) higher than in the control (informal) condition (M = 2,81).

Tip

We conducted an ANCOVA test to verify the effects of type of failure, type of evaluation and driver score on amount of tip. Results indicated a significant effect of type of failure on tip ($F(1, 340) = 4.292$, $p<.05$, $\eta^2p = 0,012$), however the main and interactive effects of type of evaluation and overall peer score were non-significant ($p>.05$). Since no difference in the means of tip was found between type of evaluation conditions in Study 2, whereas a significant difference in the means of rating between type of evaluation conditions was found, this result lends further support to H2. Test results showed the mean of rating was higher for competence failure (M = 7,69) than for warmth (M = 6,20). When considering each failure individually, no significant difference ($p>.05$) in the means of rating was found, therefore H5a was supported but H5b was not.

Discussion

Results of Study 2 confirmed most results of Study 1, revealing a significant difference in the means of rating between formal and control (informal) types of evaluation, lending further support to H1. Again, we found that tip seems to be a less biased form of feedback, since no difference in the means were found between type of evaluation conditions, supportinh H2. Anticipation of guilt was revealed to be a mediator for this effect, offering support to H3.

Contrary to our hypothesis, we found that morality and warmth failures do not have a significant difference in the means of rating between type of evaluation conditions, although competence did. Therefore, H3a and H3b were not supported. We speculate that this difference in results between Study 1 and 2 may be due to the introduction of the variable driver score in Study 2, possibly creating a cofounding effect.

Lastly, results indicated that the overall driver score moderates the effect of type of evaluation on rating. In the high driver score condition, ratings were significantly different between type of evaluation conditions. Results revealed that in the high driver score condition, means of rating were higher in the formal type of evaluation condition than in the control (informal) condition, lending support to H6. In the low driver score condition, no difference in the mean of rating was found between type of evaluation conditions. Therefore, when evaluating the driver formally, users gave a higher (more biased) rating to the driver when the driver score was high. In the informal (control) type of evaluation the mean ratings were not significantly different between low and high driver score.

8. FINAL CONSIDERATIONS

Both Study 1 and Study 2 revealed that when formally evaluating a provider, ratings were significantly higher (more biased) when compared to an informal, less ‘compromising’ type of evaluation (i.e. to a friend). In Study 1 we confirmed that fear of retaliation and reciprocity indeed are not mediators for the effect of type of evaluation on feedback. We speculate this may be due to the characteristics of the provider not being a professional and having a closer interaction with the user which may elicit empathetic concern. It is possible that knowing the rating might have negative consequences to the provider may elicit empathy from the users, when formally evaluating the peers.

In Study 2, we found anticipation of guilt to be a mediator for the effect of type of evaluation on rating. This is in line with the idea that users of collaborative services often tend, due to the nature of the service, to attenuate (bias) negative feedbacks not to harm the provider (BRIDGES & VÁSQUEZ, 2016). Furthermore, feelings of guilt have been connected to empathy and forgiveness (MICELI, 1992).

We did not find any interactive effect of type of evaluation and type of failure on rating in none of the studies. It is possible that social norms have such a strong influence in feedback in collaborative services that the type of failure has a less important 'role' than the type of evaluation (i.e. users do not want to harm a peer by giving he/she a low rating). This is made evident by some of the answers to the open question in our questionnaire (where participants were asked why they rated the driver the way they did). For example, one participant wrote: *"I didn't want to give her a bad rating especially if she could be fired"*. Revealing a strong sense of empathy and community.

In Study 2, we investigated overall driver score as another boundary condition (in addition to type of failure) for the effect of type of evaluation on rating. We found an interactive effect between type of evaluation and overall driver score on rating. Our results indicate that a high driver score leads to feedback bias when formally evaluating the provider. The results revealed means of rating were significantly higher in the formal type of evaluation than in the control (informal) condition, when respondents were submitted to high driver score condition; For respondents in the low driver score condition, the mean of rating was not significantly different between formal or informal type of evaluation conditions. Results also revealed a direct of effect of driver score on rating and tip, where in the high (vs low) score condition ratings and tips were significantly higher than in the low driver score condition. This is consistent with attribution theory which postulates that clues of past behavior (driver score) are often used as predictors of future behavior (WEINER *et al*., 1976).

Interestingly, in neither of the studies tip had a significant difference between formal and control (informal) conditions, whereas rating did in both studies. A possible explanation for this is the anticipation of guilt. It is possible that different than ratings -which users believe could have long term consequences to the driver (as demonstrated by several answers to our open question)- guilt is not elicited when tipping since this would not gravely affect the provider's livelihood. In fact, Azar (2010) found evidence that feelings of guilt and embarrassment did not have a significant impact on tip size. In line with this, Lynn (2009) found that rewarding the service was a greater motivator for tipping than avoiding guilt. Therefore, it seems that the long-established custom of tipping provides a more reliable form of feedback of service quality than other forms of service feedback. Perhaps as collaborative services and their feedback mechanisms become more permeated in our culture, we will learn how to make better use of them.

9. MANAGERIAL IMPLICATIONS

The 'sharing' economy is changing the structure of a variety of industries, and a new understanding of the consumer is needed to drive successful business models" (ECKHARDT & BARDHI, 2015). As more 'traditional' businesses expand to include collaborative services it will be extremely important for managers to understand the limitations of feedback/reputation systems in collaborative services. Problems in the self-regulating mechanism in form of feedback/reputation system may lead to incidents which in turn could damage the image of the company (BENOIT *et al*., 2017). Given that most collaborative services are self-regulated, understanding what lead users to give biased feedbacks is pivotal for the maintenance of quality in collaborative services, and consequently, company image.

Another important aspect is the Laboral activity of peer-providers. The ratings/scores of providers, especially in the case of on-demand transportation drivers, impacts directly in the provider's income. That is because companies such as Uber, give incentives to drivers with high ratings and bans those with low ratings, exerting asymmetric power over its collaborators (ROSENBLAT & STARK, 2016).

Through better comprehending these behavioral aspects that may play a role in the assessment of peers, managers might be able to create ways to go mitigate biased feedbacks

and create incentives such as training users and providers on how to make good use of evaluation tools. This is in line with Rosenblat and Stark (2016), who point that passengers' education on how to use the feedback systems is low.

10. LIMITATIONS AND SUGGESTIONS FOR FUTURE RESEARCH

This research was not without limitations. First, both studies were conducted online and in the same platform (MTurk). In order to gain greater external validity, it is suggested further studies conducted in other platforms (e.g. Prolific), or in the field. Second, we only studied one context of peer-to-peer sharing service. To allow greater generalizability of results it is recommended that future studies focus on other collaborative consumptions contexts such as hospitality (e.g. Airbnb).

As pointed by Belk (2014a), interactions in collaborative services' settings often implicate a higher level of personal contact between users and providers. It would be interesting to investigate the impact that a warmth failure that directly affects the user would have on feedback. In the open question in our survey, many participants stated that they prefer not to interact with the provider, while others stated they felt offended by the lack of interaction in the warmth failure condition. Studying the role of intimacy between peers and its impact in feedback in collaborative services and how relationship orientation (communal vs. exchange) impacts this relationship can also be interesting venues for future research.

REFERENCES

Azar, O. H. (2010). Tipping motivations and behavior in the US and Israel. *Journal of Applied Social Psychology*, *40*(2), 421-457.

Bardhi, F., & Eckhardt, G. M. (2012). Access-based consumption: The case of car sharing. *Journal of Consumer Research*, *39*(4), 881-898.

Basil, D. Z., Ridgway, N. M., & Basil, M. D. (2006). Anticipation of guilt appeals: The mediating effect of responsibility. *Psychology & Marketing*, *23*(12), 1035-1054.

Basso, K., & Pizzutti, C. (2016). Trust recovery following a double deviation. *Journal of Service Research*, *19*(2), 209-223.

Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Anticipation of guilt: an interpersonal approach. *Psychological bulletin*, *115*(2), 243.

Belk, R. (2014a). Sharing versus pseudo-sharing in Web 2.0. *The Anthropologist*, *18*(1), 7-23.

_____(2014b). You are what you can access: Sharing and collaborative consumption online. *Journal of Business Research*, *67*(8), 1595-1600.

Benoit, S., Baker, T. L., Bolton, R. N., Gruber, T., & Kandampully, J. (2017). A triadic framework for collaborative consumption (CC): Motives, activities and resources & capabilities of actors. *Journal of Business Research*, *79*, 219-227.

Bolton, G., Greiner, B., & Ockenfels, A. (2013). Engineering trust: reciprocity in the production of reputation information. *Management science*, *59*(2), 265-285.

Botsman, R., & Rogers, R. (2010). *What's mine is yours: how collaborative consumption is changing the way we live*. London: Collins.

Bridges, J., & Vásquez, C. (2018). If nearly all Airbnb reviews are positive, does that make them meaningless?. *Current Issues in Tourism*, *21*(18), 2057-2075.

Cronin Jr, J. J., & Taylor, S. A. (1992). Measuring service quality: a reexamination and extension. *The journal of marketing*, 55-68.

Dellarocas, C., & Wood, C. A. (2008). The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. *Management Science*, *54*(3), 460-476.

Eckhardt, G. M., & Bardhi, F. (2015). The sharing economy isn't about sharing at all. *Harvard business review*, *28*.

Filieri, R. (2015). What makes online reviews helpful? A diagnosticity-adoption framework to explain informational and normative influences in e-WOM. *Journal of Business Research*, *68*(6), 1261-1270.
Folkes, V. S. (1984). Consumer reactions to product failure: An attributional approach. *Journal of consumer research*, *10*(4), 398-409.
Fradkin, A., Grewal, E., Holtz, D., & Pearson, M. (2015). Bias and reciprocity in online reviews: Evidence from field experiments on Airbnb. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation* (pp. 641-641).
Guyader, H. (2018). No one rides for free! Three styles of collaborative consumption. *Journal of Services Marketing*, *32*(6), 692-714.
Hamari, J., Sjöklint, M., & Ukkonen, A. (2016). The sharing economy: Why people participate in collaborative consumption. *Journal of the Association for Information Science and Technology*, *67*(9), 2047-2059.
Hess Jr, R. L., Ganesan, S., & Klein, N. M. (2003). Service failure and recovery: the impact of relationship factors on customer satisfaction. *Journal of the Academy of Marketing Science*, *31*(2), 127-145.
Hofmann, E., Hartl, B., & Penz, E. (2017). Power versus trust–what matters more in collaborative consumption?. *Journal of Services Marketing*, *31*(6), 589-603.
Jøsang, A., Ismail, R., & Boyd, C. (2007). A survey of trust and reputation systems for online service provision. *Decision support systems*, *43*(2), 618-644.
Kirmani, A., Hamilton, R. W., Thompson, D. V., & Lantzy, S. (2017). Doing well versus doing good: The differential effect of underdog positioning on moral and competent service providers. *Journal of Marketing*, *81*(1), 103-117.
Konstam, V., Chernoff, M., & Deveney, S. (2001). Toward forgiveness: The role of shame, anticipation of guilt anger, and empathy. *Counseling and Values*, *46*(1), 26-39.
Kudisch, J. D., Fortunato, V. J., & Smith, A. F. (2006). Contextual and individual difference factors predicting individuals' desire to provide upward feedback. *Group & Organization Management*, *31*(4), 503-529.
Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group virtue: the importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of personality and social psychology*, *93*(2), 234.
Lindblom, A., & Lindblom, T. (2017). De-ownership orientation and collaborative consumption during turbulent economic times. *International Journal of Consumer Studies*, *41*(4), 431-438.
Lindsey, L. L. M., Yun, K. A., & Hill, J. B. (2007). Anticipated guilt as motivation to help unknown others: An examination of empathy as a moderator. *Communication Research*, *34*(4), 468-480.
Lynn, M. (2009). Individual differences in self-attributed motives for tipping: Antecedents, consequences, and implications. *International Journal of Hospitality Management*, *28*(3), 432-438.
Lynn, M., & McCall, M. (2000). Gratitude and gratuity: a meta-analysis of research on the service-tipping relationship. *The Journal of Socio-Economics*, *29*(2), 203-214.
Malhotra, A., & Van Alstyne, M. (2014). The dark side of the sharing economy… and how to lighten it. *Communications of the ACM*, *57*(11), 24-27.
Mattila, A. S. (2001). The impact of relationship type on customer loyalty in a context of service failures. *Journal of Service Research*, *4*(2), 91-101.
Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, *20*(3), 709-734.
Miceli, M. (1992). How to make someone feel anticipation of guilty: Strategies of anticipation of guilt inducement and their goals. *Journal for the theory of social behaviour*, *22*(1), 81-104.

Möhlmann, M. (2015). Collaborative consumption: determinants of satisfaction and the likelihood of using a sharing economy option again. *Journal of Consumer Behaviour*, *14*(3), 193-207.
Ozanne, L. K., & Ballantine, P. W. (2010). Sharing as a form of anti-consumption? An examination of toy
Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). Servqual: A multiple-item scale for measuring consumer perc. *Journal of retailing*, *64*(1), 12.
Resnick, P., & Zeckhauser, R. (2002). Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system. In *The Economics of the Internet and E-commerce* (pp. 127-157). Emerald Group Publishing Limited.
Resnick, P., Kuwabara, K., Zeckhauser, R., & Friedman, E. (2000). Reputation systems. *Communications of the ACM*, *43*(12), 45-48.
Rosenblat, A., & Stark, L. (2016). Algorithmic labor and information asymmetries: A case study of Uber's drivers.
Smith, A. K., & Bolton, R. N. (1998). An experimental investigation of customer reactions to service failure and recovery encounters paradox or peril?. *Journal of service research*, *1*(1), 65-81.
Steenhaut, S., & Van Kenhove, P. (2006). *The Mediating Role of Anticipated Guilt in Consumers' Ethical Decision-Making. Journal of Business Ethics, 69(3), 269–288.*
Tsarenko, Y., & Rooslani Tojib, D. (2011). A transactional model of forgiveness in the service failure context: a customer-driven approach. *Journal of Services Marketing*, *25*(5), 381-392.
Tsarenko, Y., & Tojib, D. (2012). The role of personality characteristics and service failure severity in consumer forgiveness and service outcomes. *Journal of Marketing Management*, *28*(9-10), 1217-1239.
Wang, S., & Huff, L. C. (2007). Explaining buyers' responses to sellers' violation of trust. *European Journal of Marketing*, *41*(9/10), 1033-1052.
Weiner, B. (1972). Attribution theory, achievement motivation, and the educational process. *Review of educational research*, *42*(2), 203-215.
Weiner, B., Nierenberg, R., & Goldstein, M. (1976). Social learning (locus of control) versus attributional (causal stability) interpretations of expectancy of success 1. *Journal of Personality*, *44*(1), 52-68.
Weun, S., Beatty, S. E., & Jones, M. A. (2004). The impact of service failure severity on service recovery evaluations and post-recovery relationships. *Journal of Services Marketing*, *18*(2), 133-146.
Worthington Jr, E. L., Hook, J. N., Utsey, S. O., Williams, J. K., & Neil, R. L. (2007). Decisional and emotional forgiveness Paper presented at the International Positive Psychology Summit. *Washington, DC*.
Zervas, G., Proserpio, D., & Byers, J. (2015). A first look at online reputation on Airbnb, where every stay is above average. SSRN Working Paper 2554500.