UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ADMINISTRAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ADMINISTRAÇÃO

Louise Helene Gonçalves Foernges

Please rate after riding: The impact of formal evaluation on consumers' feedback

Porto Alegre
2018

Please rate after riding: The impact of formal evaluation on consumers' feedback

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Administração da Universidade Federal do Rio Grande do Sul, como requisito parcial para a obtenção do título de Mestre em Administração.

Orientador: Profa. Dra. Cristiane Pizzutti dos Santos

Porto Alegre
2018

Please rate after riding: The impact of types of evaluation on consumers' feedback

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Administração da Universidade Federal do Rio Grande do Sul, como requisito para a obtenção do título de Mestre em Administração.

Aprovado em: _____de____________de______.

BANCA EXAMINADORA

____

Prof. Dr. Lélis Espartel - PUCRS

____

Prof. Dr. Leonardo Nicolao - UFRGS

____

Orientadora - Profa. Dra. Cristiane Pizzutti dos Santos - UFRGS

ACKNOWLEDGEMENTS

This research would not have been possible without the support of a few institutions and some very special people whom I would like to acknowledge.

My family without whom any achievements in my life would have been possible. Thank you for the financial support in carrying out this study and for all the support and unconditional love throughout my life. Thank you to my father Luiz, my siblings Luiz, Cristian, Carlos and Cristine. Special thanks to my mother Maria and husband Jaemoon, for their love and companionship during my studies. I would also like to thank my loyal companions of every moment, Lela, Cinza and Eros.

My advisor, Prof. Cristiane Pizzutti dos Santos, for her support, patience, teachings and guidance during my graduate studies. This research would not have been accomplished without you. Thank you for giving me the chance of being your advisee. It was an honor and a privilege, and I will always look up to you.

The other professors whose teachings and advice were essential for the accomplishment of this study: Prof. Vinicius Andrade Brei, Prof. Leonardo Nicolao, Prof. Fernando Bins Luce, Prof. Luiz Antonio Slongo and Prof. Lélis Balestrin Espartel.

My classmates and now friends, Thiago Glasenapp, Marcelo Homrich, Matheus Trucolo Conci and Maria Alice Pasdiora. I was very lucky to have met you along the way. Thank you for your academic and moral support and for making my time as master's student more enjoyable.

To my friends Lorenza, Thaísa, Nathália, Letícia, Fabiana, Débora and Woojung, thank you for always being there to help and support me with my projects.

To the Group of Studies in Services Marketing (GEMS) aka "pie group", and all its members, for the precious advice and the best coffee breaks.

I would also like to thank the Federal University of Rio Grande do Sul (UFRGS) and its staff, for making it possible for me to have high quality graduate education and CAPES – Brazilian Federal Agency for Support and Evaluation of Graduate Education within the Ministry of Education of Brazil, for granting me scholarship.

ABSTRACT

Advances in Information Technology along with changes in society have allowed for the emergence of collaborative services. The act of sharing among peers -in substitution to ownership- is a growing phenomenon with many successful companies having arisen in the last decade. Since this new economy works mostly on the basis of sharing among strangers, mechanisms for identifying good and 'bad' users have become a necessity. One popular tool is a mechanism that allows for mutual evaluation among platform users (peer-providers and peer- users) using reviews and/or ratings as forms of evaluation. However, often users will give a biased feedback or attenuate negative evaluations of their peers due to the nature of collaborative services, where interactions are more personal and social norms seem to exist. This represents a problem especially in situations where the service provided has a failure that goes unreported. Although collaborative services are growing in popularity, few studies have been carried out to investigate how pro-social norms are integrated into practices and interactions between peers. To examine factors leading to feedback bias and its boundary conditions, we conducted two scenario-based experiments online using the context of an on-demand transportation service. We compared feedback (in the form of rating and tip) in a formal type of evaluation to a control condition (i.e. informal). In Study 1, we find that feedback bias in a formal evaluation system can be explained by forgiveness. Furthermore, that the type of service failure directly impacts feedback bias with perceived quality compromised by the failure being a mediator for this effect. We also find tip to be a less biased form of feedback than ratings. In Study 2, we confirm results of Study 1, and investigate overall driver score as a boundary condition for the effect of type of evaluation on feedback. Results show that a high peer score leads to feedback bias in a formal type of evaluation. Additionally, we find anticipation of guilt to be another mediator for the effect of type of evaluation on feedback. Managerial implications and suggestions for further research are discussed.

Key-words: collaborative services; collaborative consumption; service failure; overall peer score; feedback objectivity; forgiveness; anticipation of guilt.

LIST OF FIGURES

LIST OF TABLES

SUMMARY

1. INTRODUCTION

Humans have always shared in a number of ways, a practice "as old as humankind" itself (BELK, 2014b, p.1595) In fact, since the Stone Age, the practice of sharing proved vital to our survival (BOTSMAN & ROGERS, 2010). What changed at the present time is that with modern technologies, sharing is no longer exclusively practiced among kin or communities but also among complete strangers (BELK, 2014b).

The emergence of peer-to-peer sharing services seem to have been the result of modern times. A movement towards de-ownership and more sustainable use of resources is growing (OZANNE & BALLANTINE, 2010; ALBINSSON AND PERERA, 2012; SCHAEFERS *et al.*, 2016; LINDBLOM & LINDBLOM, 2017) and the widespread usage of the internet has created a more connected world (BOTSMAN & ROGERS, 2010; BELK, 2014b). This new panorama and technologies allowed for unprecedented possibilities of interaction, including new ways of sharing. The 'sharing economy' is one of these phenomena that were made feasible by the internet era (BELK, 2014b; FIGUEIREDO & SCARABOTO 2016; HAMARI, SJÖKLINT & UKKONEN, 2016; BENOIT *et al.*, 2017, ZERVAS *et al*, 2017).

The term sharing economy (ERT, FLEISCHER & MAGEN, 2016; 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; HOFMANN, HARTL & PENZ, 2017) and sometimes called access-based consumption (BARDHI & ECKHARDT, 2012; SCHAEFERS *et al.*, 2016) or hybrid economies (SCARABOTO, 2015), essentially refers to P2P (peer-to-peer) interactions where individuals have temporary access to a good or a service without ownership transfer (BOTSMAN & ROGERS, 2010; BELK, 2014b; BARDHI & ECKHARDT, 2012). Usually, an online platform connects the users 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; CHASIN *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 intimacy, such as when someone opens their house to a stranger on Airbnb or becomes a guest at a stranger's house (BRIDGES, VÁSQUEZ, 2016). One way around the risks of such interactions and the fear of strangers is through reputation/feedback systems (BELK, 2014b; BRIDGES, VÁSQUEZ, 2016; HOFMANN *et al.*, 2017; BENOIT *et al.*, 2017). These self-regulatory feedback mechanisms (usually in the form of ratings and/or reviews) help minimize the risks, discourage misbehavior and create trust among peers[1] in collaborative consumption services (BELK, 2014b; BRIDGES, VÁSQUEZ, 2016; HOFMANN, *et al.*, 2017).

Generally, service failures take place when a customer's expectations are not met and often that comes with a feeling of broken trust (BASSO & PIZZUTTI, 2016). Since evaluation systems are an important tool for peers in collaborative consumption services to evaluate other's trustworthiness (HAMARI *et al.*, 2016; HOFMANN *et al.*, 2017), it seems that feedbacks are especially important in the occurrence of a service failure. However, despite feedback being pivotal in this context, it is when a service failure takes place that feedback bias is more likely to occur in collaborative services since peers tend to underreport negative events (BRIDGES & VÁSQUEZ, 2016). Underreporting of negative experiences has been linked to the personal nature of collaborative services, where 'social norms' are presumably being followed (BRIDGES & VÁSQUEZ, 2016; ZERVAS *et al.*, 2015). In an attempt to attenuate this bias, many collaborative services have changed the form in which feedback is displayed to users, making them 'double blind' (i.e. users are only able to see each other's assessment when both have already provided their feedbacks) (BOLTON *et al.*, 2013). One example is Airbnb, where reviews are only posted when both peer-user and peer-provider have evaluated each other, or automatically after a period of 14 days. However, even making feedbacks 'double blind' seem to not have eliminated feedback bias in collaborative services (BRIDGES & VÁSQUEZ, 2016).

We conducted two scenario-based experiments in order to investigate how certain aspects interfere in feedback bias in collaborative services. We compare two types of evaluation, from the point of view of a user of on-demand transportation evaluating a provider (i.e. driver), after a service failure has occurred. Specifically, we compare feedback (in the form of rating and tip) in a formal type of evaluation (i.e. the traditional in the app evaluation -or "in the system") and an

[1] In this study we use the terms 'user' and 'peer' interchangeably.

informal one (i.e. to a friend - out of the system), which served as a control condition. Zervas *et al.*, (2015) used a similar approach when comparing feedback on TripAdvisor and Airbnb to investigate feedback bias. The authors used feedbacks from properties listed on TripAdvisor (that only allows one-way feedbacks, from guests to properties) as a control condition to be compared with Airbnb feedbacks (that allows mutual feedback between peers, both hosts and guests).

Several authors have linked feedback bias to reciprocity and fear of retaliation (CLAYSON, 2004; DELLAROCAS & WOOD, 2008; RESNICK *et al.*, 2000 RESNICK & ZECKHAUSER, 2002; BARDHI & ECKHARDT, 2012; BOLTON *et al.*, 2013; FRADKIN *et al.*, 2015). Fradkin *et al.* (2015), for example, conducted two field experiments in the platform Airbnb. The authors found that around 70% of all reviews from guests to hosts were positive (5 stars). According to the authors, this may be evidence that fear of retaliation for negative reviews and reciprocity for positive ones lead to feedback bias. We propose that due to the nature of collaborative services, which imply closer proximity, sense of community and mutual trust between users and providers (GUYADER, 2018; BRIDGES & VÁSQUEZ, 2016; ZERVAS *et al.*, 2017) and due to most evaluation systems in this context being 'double blind', the reasons why peers give biased feedbacks in collaborative services are not due to reciprocity or retaliation. This difference in the reasons leading to feedback bias between traditional and collaborative services follows the logic of Zervas *et al.* (2015, p.2) who argues that "the social norms associated with these intimate Airbnb transactions may not be reflected in previously observed rating distributions or captured by previously proposed review generation models".

In line with this, Bridges and Vásquez (2016) pointed to sociocultural factors playing a role in collaborative services. According to the authors, users in these services tend to attenuate negative feedbacks due to the closer relationship between user and provider, which does not occur in traditional services. Therefore, we propose that anticipation of guilt and forgiveness, aspects which have been connected to social harmony and empathy (ENRIGHT, 1992; MICELI, 1992), may be behind feedback bias in formal evaluations in collaborative services instead of reciprocity or fear of retaliation.

Additionally to the effect of type of evaluation (i.e. formal or informal) on feedback through forgiveness and anticipation of guilt, we investigate the impact that different types of service failure (morality, competence and warmth) and overall driver score (high, low) have on this effect. More specifically, using rating and tip as forms of feedback, we test how these potential boundary conditions (type of service failure and overall driver score) change the effect of formal evaluation on feedback bias. Our central premise is that in the formal type of evaluation (vs informal one) feedback is more biased because social norms are at play in this situation (ZERVAS *et al.*, 2015*;* FRADKIN *et al.*, 2015; BRIDGES & VÁSQUEZ, 2016; HAMARI *et al.*, 2016). However, these two moderating variables could attenuate this biased effect (i.e. a competence failure - vs moral and warmth failure and a low driver score – vs high - will minimize feedback bias).

The different types of service failure we investigate are commonly found in literature (KIRMANI *et al*, 2017; WANG & HUFF, 2007; KIM *et al*, 2004). According to Kirmani *et al* (2017), a competence failure occurs when the provider lacks skills or abilities to effectively execute a task; a morality failure occurs when the customer has the perception the provider is being dishonest, unfair or lacks principals. Finally, for the authors, a warmth failure occurs when the provider has traits of unfriendliness and unsociability. In their research, the authors found that when people choose a service provider to perform a task (which the individual cannot perform by itself), they often value skill and knowledge to perform the task more than morality and warmth traits of the provider. Following these findings, we propose that a competence failure will lead to less biased feedbacks than morality or warmth failures, presumably having the least change on feedback between types of evaluation (since a competence failure is likely to be the one with the highest potential to harm the core delivery of the service itself, when controlling for the failure severity). Following this logic, we investigate if forgiveness is an explanatory mechanism for how these failures impact on feedback between type of evaluation conditions. Complementarily, we investigate if perceived quality compromised by the service failure explains the direct impact of service failure on feedback. Our premise is that a competence failure will be perceived as the one compromising perceived service quality the most, yielding lower ratings and amount of tip.

Moreover, most popular collaborative platforms such as Airbnb and Uber now display a "peer score". The scores are an average of all the ratings given by the peers with whom the user has interacted. Both peer-user and peer-provider are able to rate each other[2]. We included overall driver score in order to investigate the impact that cues of past behavior might have on feedback. For example, high scores may be perceived as indicators of good behavior, whereas low scores may be perceived as indicators of bad behavior, altering feedback. This follows the logic of stability attribution theory. According to Weiner *et al.* (1976), the theory postulates that future behavior is, at least in part, determined by the causes of past events. Conversely, the authors argue that if causal conditions are perceived as likely to change, then the present events may not be expected to reoccur. We propose that a high score may lead to the perception of the failure being sporadic or a one-time event, in such a way that a high provider score could increase user's rating and amount of tip (i.e. providing more biased feedback) when formally evaluating the provider (BELK, 2014b; HAMARI *et al*., 2016; GUYADER, 2018).

To summarize, Malhotra and Van Alstyne (2014 p. 27) argue that "the viability of shared services hinges on the quality of review systems because people rely on them to decide wheher 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 research shows peers avoid giving negative ratings/reviews (ZERVAS *et al*., 2015*;* FRADKIN *et al*., 2015; BRIDGES & VÁSQUEZ, 2016). Given the importance of such 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), we conducted two experimental studies aiming to answer the following questions:

1. **Does a formal evaluation system in collaborative services generate biased feedbacks?**
2. **Are forgiveness and anticipation of guilt (instead of reciprocity or fear of retaliation) underlying mechanisms that explain the effect of type of type of evaluation on feedback?**

[2] Airbnb. How do Reviews Work. Avaliable at: https://www.airbnb.com/help/article/13/how-do-reviews-work Accessed September 20th, 2018.

3. **Are the type of service failure and the overall provider (driver) score boundary conditions for the effect of type of evaluation on feedback?**

Given the fact that collaborative services are growing in terms of academic relevance and the lack of research on underlying mechanisms and boundary conditions that lead to less biased feedbacks in this context, there seems to be ample space for research into the subject.

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.

This research contributes to extend literature in the following ways:

1. Our investigation of behavioral aspects and boundary conditions interfering in feedback giving in collaborative services may serve as a starting point for further research into other factors that could possibly interfere in feedback objectivity in this rather new form of interaction.
2. 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.

Having introduced the present research in this section, the research objectives will follow next. To bring context to the studies, a review of the existing literature on collaborative services, , formal evaluation and its impact on consumers’ feedback and service failure is introduced in chapter two. In chapter three, we present results of Study 1. In chapter four, we present a review of literature on guilt and overall peer score, followed by results of Study 2 in chapter five. In chapter six we present our final discussion, while in chapter seven we present the managerial implications. Finally, we discuss limitations and suggestions for future research in chapter eight.

1.1. RESEARCH OBJECTIVES

The main goal of this study is to investigate the effect of type of evaluation (i.e. formal or informal) on feedback, in a collaborative services context. In order to better organize and develop the research and to satisfy the general objective, the following specific objectives are proposed:

1. To examine forgiveness and anticipation of guilt as mediators for the effect of type of evaluation on feedback (rating and tip), instead of reciprocity and fear of retaliation.
2. To investigate type of failure and overall provider (driver) score as boundary conditions for the effect of type of evaluation on feedback (rating and tip).

2. THEORETICAL BACKGROUND OF STUDY 1

In this chapter a review of literature will be presented in to better contextualize the research. First, the subject of collaborative services will be addressed to add to the comprehension of the scenarios where the research was developed. Then, a review of literature will follow on the subjects of feedback mechanisms and their impact on consumer's feedback and finally, service failure.

2.1 COLLABORATIVE SERVICES

> "Sharing is an alternative to the private ownership that is emphasized in both marketplace exchange and gift giving. In sharing, two or more people may enjoy the benefits (or costs) that flow from possessing a thing. Rather than distinguishing what is *mine and yours*, sharing defines something as *ours"* (BELK, 2007, p.127).

Evidence suggests that humans already practiced some form of sharing two million years ago (ISAAC, 1978). According to Belk (2009), in Medieval Europe, for example, it was natural to sit on common benches and share food. The author affirms that besides having, nowadays, a more individualistic behavior, when compared to those in Medieval Europe, we still carry many habits of sharing, especially in eating.

According to Belk (2009) not all sharing we practice nowadays are inherited habits from a more interdepend the past. The information technologies such as the Internet and the Web 2.0, have allowed us unprecedented possibilities of interaction and sharing (BELK, 2014b). O'Reilly (2005) argues that the Web 2.0 probably emerged after the dot.com collapse around the year 2001. According to the author, in contrast to the static, one-way information flow of the Web 1.0, the Web 2.0 enabled users' participation and interaction with the platforms and with one another.

The information technologies enabled the creation of online platforms where user-generated content is available and peer sharing, and collaboration is possible (HAMARI *et al.*, 2016). Examples of such platforms vary from collaborative online encyclopedias (e.g., Wikipedia), to video sharing (e.g., Youtube) to peer-to-peer file sharing (e.g., Pirate Bay) (HAMARI *et al.*, 2016). Some of these platforms were launched more than a decade ago. Wikipedia, for example, has been around since 2001 and Youtube since 2005. "In a broad sense, the Internet itself is a giant pool of shared content that can be accessed by anyone with an Internet connection, a browser, and a government that allows access to most or all web content" (BELK, 2014b p.1595).

According to Belk (2014b), the term 'sharing economy', encompasses a large number of for-profit and non-profit businesses. Among these, companies that have very different business models: such as Airbnb, Zipcar, Wikipedia, YouTube, Flickr, Facebook, Freecycle, and Twitter. Nowadays, peer-to-peer service companies connect users not only via mobile applications and websites but also public spaces -as it is often the case with car and bike sharing. There are collaborative consumption services that make it possible to share cars (Zipcar, Turo), bikes (CitiBike, Serco, Liquid), tasks (TaskRabbit, Mechanical Turk), private transportation (Uber, Cabify, Lyft), accommodations (Airbnb, HomeAway, CouchSurfing) and so on.

Another commonly used term is access-based consumption which is defined by Bardhi and Eckhardt (2012 p.881) as "transactions that can be market mediated but where no transfer of ownership takes place". In order to examine the nature of access, the authors conducted an interpretative study with users of the car sharing service Zipcar. The authors identified six dimensions "to distinguish among the range of access-based consumptionscapes" (p.881): Temporality, Anonymity, Market Mediation, Consumer Involvement, Type of Accessed Object and Political Consumerism.

It seems not only advances in technology but also a change in society and collective mindset has led to the growth of collaborative consumption services. “There are burdens to possession, as any home owner can attest. And with the increasingly rapid pace of technological change, we may see a shift toward shared ownership” (BELK, 2007 p.136). In fact, as far back as 2011, TIME magazine had already named collaborative consumption one of 10 ideas that would change the world[3].

The phenomena of collaborative consumption services possibly gained strength due to the economic crisis -from when maintaining ownership became a bigger challenge (BELK, 2014b; BARDHI & ECKHARDT, 2012; BARNES & MATTSON, 2016; BENOIT *et al*., 2017). According to Belk (2014b, p. 1599), “many of the sharing and collaborative consumption organizations that currently exist benefitted from the economic collapse that began in 2008 that caused some consumers to lose their homes, cars, and investments and made most everyone more price sensitive”.

For Sundarajan (2013) there are four factors that may have been the drivers that led to the development of sharing economy. First, the consumerization of digital technologies (the technological innovations have become mainly driven by consumers, not businesses or governments as it used to be decades ago). Second, the emergence of digital institutions (platforms that facilitate economic exchange). Third, urbanization and globalization (migrations to densely populated urban areas are increasing). Fourth, ecological and resource considerations (increasing need to use natural and other physical resources more efficiently and increasing number of people choosing to live ‘asset-light’).

The change in our values and preferences in consumption appear to have been even greater among the younger generations (BOTSMAN & ROGERS, 2010; HWANG & GRIFFITHS, 2017). Millennials are avid technology users and more conscious about the social and environmental impact of their consumption choices (LINDBLOM & LINDBLOM, 2017). Bardhi and Eckhardt (2012, p. 881), suggest that “instead of buying and owning things, consumers want access to goods

[3] Time Magazine. Avaliable at:
http://content.time.com/time/specials/packages/article/0%2C28804%2C2059521_2059717_2059710%2C00.html
Accessed Feb 23rd 2017.

and prefer to pay for the experience of temporarily accessing them". The current scenario where individuals seem to rather spend on experiences instead of material things, along with the changes in the economy, the awareness of climate change and the constant development of information networks have led to the success of peer-to-peer markets (BARNES & MATTSON, 2016; LINDBLOM & LINDBLOM, 2017).

Online peer-to-peer marketplaces where individuals can announce and buy products have been around for decades. The auction website eBay, for example, was founded in 1995, with the aim of "bringing together buyers and sellers in an honest and open marketplace"[4]. However, with the 'peer economy', the peer marketplaces now go beyond the simple trades conducted by services such as eBay (SUNDARARAJAN, 2013). Sundararajan (2013, p.2) says that "we are comfortable with the notion of commercial transactions mediated by computers or smartphones, and we've had over ten years of experience with the idea of semi-anonymous peer-to-peer exchange".

In the past few years academic papers on collaborative consumption services have started to emerge. According to Belk (2014a p.7), there has been "an explosion of studies and writings about sharing via the Internet". Belk published pioneer overviews on the subject in 2007 and 2010. Since then, a few empirical studies have also been conducted (BARDHI & ECKHARDT, 2012; HAMARI *et al*., 2016; ZERVAS, PROSERPIO & BYERS, 2015; MÖHLMANN, 2015)

Hamari *et al*. (2016), for example, studied the motivations that led users to participate in collaborative consumption. The authors surveyed 168 users registered in a collaborative consumption platform. The authors found that sustainability, enjoyment of the activity and economic gains were among the factors the motivate users to engage in peer-to-peer sharing. However, sustainability shown to be an important factor only for those users who valued ecological consumption.

Möhlmann (2015) examined the determinants of satisfaction and the likelihood of individuals using a sharing option again. The author conducted two surveys, with users of two

[4] Ebay. Our Company. Avaliable at https://www.ebayinc.com/our-company/our-history/. Accessed March 29th 2017.

distinct collaborative consumption services platforms (car2go and Airbnb). The author found that both satisfaction and the likelihood of using a sharing option again were explained by determinants that serve users' self-benefit. Furthermore, utility, cost savings, and familiarity were found to be important factors in both studies, while in the context of car2go service quality and community belonging were also essential.

While peer sharing such as in collaborative consumption may also be done face-to-face, locally, as humans did in the past, the Internet has allowed for the creation of many-to-many peer-to-peer interactions (BOTSMAN & ROGERS, 2010). In fact, technology has allowed us to connect in such new ways, that acts of cooperation are no longer bounded to kins and communities but have expanded to include unfamiliar individuals as well (BELK, 2014b).

Bardhi and Eckhardt (2012) found that often negative reciprocity (i.e. when there is an exchange but one of the users acts out of self-interest) occurs among users. Their research suggests a possible risk in engaging in this type of service, since users may be more careless and less responsible when using the shared item than they would if that item belonged to them. Obviously, some per-to-peer services implicate in more risk than others. In the context of sharing of baby products, for example, Catulli *et al.* (2013) mention the possibility of serious risks such as safety concerns due to the conduct of previous users.

In peer-to-peer markets, especially where direct face-to-face contact is necessary, the risks involved in users' interactions frequently go beyond material or financial loss. In Airbnb, for example, users often share the same physical space (i.e. peer-guest stays at peer-host's property), leaving them susceptible to violence and other forms of abuse (Ert, Fleischer and Magen, 2016). In line with that, Belk (2014b) suggests that "stranger danger" ended up leading hitchhiking out of common practice. Now, some collaborative consumption services are bringing back this old habit in new ways. One of such platforms is BlaBlaCar[5], a company which connects users who need rides to users who are willing to give rides, for a fee. Literature seem to suggest that it is to avoid that same "stranger danger" that led hitchhiking out of practice that feedback systems are so important to build trust and avoid risks in these new forms of interaction (RANZINI, 2017;

[5] For more information: BlaBlaCar. https://www.blablacar.com/. Accessed September 24th, 2018.

GUYADER, 2018). However, perhaps due to this proximity between strangers, prejudice in the form of racial discrimination has been occurring in collaborative consumption services (EDELMAN *et al*, 2017; EWENS *et al*., 2014). This is such a problem that has even led the platform Airbnb to institute anti-discrimination policies (CHENG & FOLEY, 2018).

2.2 FORMAL EVALUATION AND ITS IMPACT ON CONSUMERS' FEEDBACK

In today's peer-to-peer online marketplaces, transactions and interactions among total strangers is common practice. In collaborative consumption 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).

Online review systems have been around for a long time. Amazon.com, for example, began to offer its customers the possibility to post product comments back in 1995 (PARK, LEE & HAN, 2007). The reputation system of eBay goes back to the 90's, and it already allowed peer-to-peer evaluations, though in an exchange-based scenario (RESNICK & ZECKHAUSER, 2002). However, with the growth of the collaborative consumption phenomenon, reputation systems gained new context. Most collaborative consumption services offer review and/or rating systems that allow for mutual evaluation between users. These systems have the role of motivating individuals to behave in a responsible manner (BOTSMAN & ROGERS, 2010; HOFMANN *et al.*, 2017).

Resnick *et al.* (2000) argue that trust is naturally built in long-term relationships. According to the authors (p.46), as people interact over time, "the history of past interactions informs them about their abilities and dispositions". Also, the authors affirm that the expectation of reciprocity and fear of retaliation serves as an incentive for individuals to behave in a good manner in the future. Since internet mediated interactions often occur among strangers, this relationship lacks a past reference. According to the authors, the reputation systems serve the purpose to guide users as to what to expect from an individual in the future, based on past experiences from other users.

Botsman and Rogers (2010) argue that by enabling decentralized and transparent communities, collaborative consumption platforms allow for 'trust between strangers' to be formed.

According to the authors in the collaborative economy markets, disagreements are usually resolved among the community. The feedback/reputation systems help users to decide who among the other users they interact with is trustworthy or not, working as a sort of a self-regulatory mechanism (BELK, 2014b; HAMARI *et al.*, 2016; HOFMANN *et al.*, 2017). These systems allow peers to rate and/or review each other and form an overall score (average of received ratings) which serves as a clue to peer trustworthiness. Therefore, feedback in the form of reputation systems have an important role in helping to build trust among users (BOTSMAN & ROGERS, 2010; HOFMANN *et al.*, 2017). In fact, personal reputation appears to be, in a way, becoming an asset (BELK, 2014b) as users who sustain low scores may get banned from the platforms. In Uber, for example, peers with scores below 4.7 or 4.5 (depending on the location) may be deactivated. According to Resnick *et al*. (2000, p.46), "though few producers or consumers of the ratings know one another, these systems help people decide whom to trust, encourage trustworthy behavior, and deter participation by those who are unskilled or dishonest".

Feedback systems work either in the form of one-way feedback, where only one of the users involved in the transaction or interaction can provide feedback to the other. Or, in the case of mutual feedback systems, where both users involved can evaluate one another (ZERVAS, PROSERPIO & BYERS, 2015). While some online platforms opt for one-way feedback systems, nowadays most platforms employ mutual feedback systems.

The seller-buyer connecting platform Etsy[6], for example, allows only for one-way feedback. Buyers can rate the sellers from 1 to 5 starts. Sellers do not rate the buyers but are able to respond to evaluations of less than 3 starts. In eBay buyers are able to evaluate the sellers; however, the sellers are not able to evaluate the buyers unless it is positively. Popular services such as Uber[7] and Airbnb[8], however, use mutual feedback systems.

Despite the importance of mutual assessment in collaborative consumption, evidence suggests that the existence of a mutual evaluation system may lead to biased feedback. According to Goodrich and Kerschbaum (2011), since both users involved in the interaction have the

[6] Etsy. Seller Policy. Avaliable at: https://www.etsy.com/legal/sellers/?ref=list#reviews Accessed April 27th, 2017.
[7] Uber. Ride with Confidence. Avaliable at: https://www.uber.com/en-BR/ride/safety/ Accessed September 20th, 2018.
[8] Airbnb. How do Reviews Work. Avaliable at: https://www.airbnb.com/help/article/13/how-do-reviews-work Acessed September 20th, 2018.

possibility of evaluating each other, when a user reviews another user negatively, the user who was negatively evaluated might give a negative review back, even if not deserved, in retaliation. The opposite is also true, as evidence points to reciprocity in positive feedbacks (BOLTON *et al.,* 2013).

Dellarocas and Wood (2008) found evidence of feedback bias in eBay feedbacks. The authors conducted a study with a large dataset of eBay feedbacks. Results showed that users were more likely to post a feedback when they had a satisfactory experience than when they had a dissatisfactory one. This is in line with Bolton *et al.* (2013 p.265) idea that "reciprocity in feedback giving distorts the production and content of reputation information in a market, hampering trust and trade efficiency". In fact, one of the main challenges for peer-to-peer service platforms using a mutual feedback system is how to control for feedback bias (RESNICK & ZECKHAUSER, 2002).

Another study on the subject of reputation systems in the context of eBay and similar platforms was conducted by Bolton, Greiner and Ockenfels (2013). The authors observed feedback patterns in platforms such as eBay and then ran laboratory experiments in order to investigate how reciprocity could be better managed in reputation systems. The authors found that retaliatory feedback was a rather small phenomenon but the threat of a retaliatory negative feedback distorted feedbacks in the aggregate. According to the authors (p. 282) "reciprocity plays a major role in the leaving, timing, and content of feedback".

The problem with feedback reciprocation was such that nowadays eBay prohibits sellers from giving negative or neutral feedback to buyers. One way the collaborative consumption services platform Airbnb found to mitigate this effect, is by keeping feedbacks confidential until both the peer-provider and the peer-user have written and submitted their reviews[9] or after a period of 14 days, whichever comes first (BRIDGES & VÁSQUEZ, 2016). Only then, the reviews are made public; from when they can no longer be edited. However, the users may respond to the original feedback they were given, publicly. This type of feedback is also known as *double-blind* (Bolton *et al.,* 2013).

[9] Airbnb. Reviews. Avaliable at: https://www.Airbnb.com/help/topic/203/reviews. Accessed March 28th, 2017.

Bolton *et al*. (2013) found that making feedbacks *blind* (i.e. feedback is only made public after a period of time or after both users involved in a transaction have evaluated each other) could possibly reduce evaluation bias due to expectation of reciprocity or fear of retaliation. The authors explored the consequences of a double-feedback system through data collected from two online platforms. One platform similar to eBay, where users can sell to and buy from each other and one where software coders can bid for contracts with software buyers. The authors found evidence that a double-blind feedback system leads to more discerning feedbacks from buyers and also less correlation of feedback between trading partners.

However, other authors found evidence that even in double-blind feedback systems, such as the one in Airbnb, feedback bias also occurs. Zervas *et al.* (2015) conducted a study analyzing 226.594 properties from around the world, listed on Airbnb (which has a mutual evaluation system) and 412.223 hotels and 54.008 vacation rentals listed on TripAdvisor. The authors found that 95% of the properties had an average rating of 4.5 or 5 starts (ratings in the platform range from 1 to 5 starts). According to the authors, none of the properties in their sample had less than a 3.5 stars rating. The authors also explored the ratings of 500.000 hotels worldwide available on TripAdvisor (where only the guests are able to evaluate the properties). The authors found that the average rating for these properties was 3.8 stars –much lower than the average rating for Airbnb properties. In addition to that, the properties in TripAdvisor showed a greater variance across the reviews when compared to Airbnb reviews. The authors then compared the rating of properties listed both Airbnb and TripAdvisor. They concluded that although ratings in both platforms for the same property were similar, in Airbnb more properties received high ratings of 4.5 stars and above, which is unrealistic high. The authors argue that although the difference in these results may be due to the different tastes of each platform's users, they may also be influenced by the nature of services such as Airbnb. According to the authors it is possible that 'sociological effects' lead people to be more diplomatic in their reviews in collaborative consumption services.

Bridges and Vásquez (2016) also investigated the reciprocal feedback system of Airbnb. The authors explored Airbnb reviews using a computer-assisted approach to identify linguistic patterns. According to the authors, Airbnb reviews have a very restricted set of linguistic resources. Also, as other studies have shown, the majority of the commentaries was highly (if not unrealistic) positive, with only 7% of 400 reviews having some form of complaint. The authors argue that when reviewing less than positive experiences, users prefer to leave neutral commentaries instead of negative ones. Overall, the authors found only 2% out of 400 reviews to be entirely negative. 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).

Another form of feedback recently incorporated to collaborative services, such as Uber, is tipping. The company Uber has incorporated a function on its app that allows customers to offer a gratuity to their driver, after the ride. In fact, tips have a very similar nature to feedback/reputation mechanisms. Tipping is another way customers can exercise quality control over the service, working, in fact, as a very similar mechanism to ratings (LYNN & MCCALL, 2000). As Uber's own website states "tipping is another way to thank drivers for going the extra mile and providing a great experience".[10] Lynn and McCall (2000), conducted a meta-analysis using unpublished studies investigating the relationship between tip size and evaluations of service. The authors found that, consistent with equity motivations theory, a positive correlation between service evaluations and tip size. According to the authors, the equity motivations theory posits that (p.3) "people are socialized to feel anxiety or distress when their relationships with others are inequitable". The authors argue that a relationship becomes inequitable when the outcome for each individual is not compatible with their respective inputs. In another work from Lynn and McCall (2016), the authors found that motivations for tipping, in a restaurant setting, included social expectations, server attractiveness, server friendliness and customer mood. Based on evidence that feedback in collaborative consumption services are often biased and often unrealistically high, we propose that:

> 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) a more positive rating and b) a higher amount of tip to the provider.

[10] Riding with Uber. Tipping. Available at: https://www.uber.com/ride/how-uber-works/in-app-tipping/. Accessed 10th November 2018.

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). One mechanism that could possibly be behind this phenomenon of attenuating bad reviews in collaborative services is forgiveness. In fact, Tsarenko and Tojib (2012) found evidence that, after a service failure, emotional and decisional forgiveness have a negative impact in spreading negative word-of-mouth and intention to switch providers and, as pointed by Filieri (2014), online reviews and ratings have become one of the main sources of word-of-mouth.

Most authors agree that forgiveness is a complex emotion which lacks a clear definition (WORTHINGTON, 1998; TSARENKO & TOJIB, 2011). Enright *et al*. (1992, p.101) have defined interpersonal forgiveness as this: “one who is deeply hurt by another often fights against the other (even if only in feelings and thought toward the other); as the injured party ceases fighting against the other and gives him or her the unconditional gift of acceptance as a human being, the former is said to be forgiving”. Konstam *et al*. (2001), add that forgiveness includes offering underserved compassion, generosity and occasionally love.

For Worthington (2005), forgiveness is different for noncontinuing and close relationships. In noncontinuing relationships, such as between strangers, forgiveness may only involve reduction of the negative emotions -or giving up negative feelings. However, in close relationships, such as with romantic partners, when there is a major disappointment, forgiveness can be defined as reduction of negative emotions while replacing them with positive emotions.

Fradkin *et al.* (2015) conducted two field experiments on the platform Airbnb. The authors found that 70% of the ratings from guests to hosts were extremely positive (5 stars). Also, strong evidence of bias due to fear of retaliation for negative reviews and reciprocity for positive reviews was found in the study. Furthermore, the authors found another reason for feedback bias in evaluation: socially induced reciprocity. According to the authors, socially induced reciprocity happens when peers interact socially and therefore omit negative information from feedback or ‘inflate’ their ratings. For the authors, this may occur due to mutual empathy between peers after a social interaction and because users may feel a certain ‘obligation’ to the provider, leading to an omission of negative feedback to avoid hurting the provider. In line with this, Fradkin *et al.* (2018) found that even in double-blind feedbacks, which aim to reduce strategic reciprocity, some bias in feedback remains. The authors propose this may be due to the social nature of the interaction in collaborative services, such as Airbnb.

As pointed out by Fradkin *et al.* (2015), Bridges and Vásquez (2016) and Zervas *et al*. (2015), interactions in collaborative services have a more personal nature (when compared to more 'traditional' services). That proximity with the provider, in turn, leads users to being more empathetic and avoid leaving negative reviews not to be 'unkind' (BRIDGES & VÁSQUEZ, 2016). According to Enright (1992, p.101) "forgiveness implicates overcoming a negative affect and judgment towards the offender", specially motivated by feelings of compassion. Given the personal nature of interactions in collaborative services and the influence of sociological aspects that lead people to be more diplomatic when giving negative feedback in reviews of another individual (ZERVAS *et al*., 2015) it is possible that forgiveness in feedback giving in collaborative services follows styles of forgiveness identified by Enright (1992): Expectational Forgiveness and Forgiveness as Social Harmony. According to the authors, the former style of forgiveness occurs when one forgives due to others' pressure and expectation and the latter to keep harmony and good relations in society, avoiding conflicts. It is worth to note that extant literature points out that feedback bias can be explained by reciprocity and fear of retaliation (CLAYSON, 2004; DELLAROCAS & WOOD, 2008; RESNICK *et al.*, 2000 RESNICK & ZECKHAUSER, 2002; BARDHI & ECKHARDT, 2012; BOLTON *et al.*, 2013; FRADKIN *et al.*, 2015). However, 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, we propose that these two explanatory mechanisms (reciprocity and fear of retaliation) do not apply to a collaborative services context.

Given the more personal nature of interactions in collaborative services, the social norms at play and that often feedback in this context is double-blind, we propose the following hypothesis:

> H2: In the occurrence of a given service failure, the effect of type of evaluation on users' rating will be mediated by forgiveness.

2.3 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 (HESS JR., GANESAN & KLEIN, 2003) and the it fails to meet customer's expectations (HOLLOWAY & BEATTY, 2003). Service failures are unavoidable and frequently elicit negative feelings and reactions in the customer (SMITH & BOLTON, 1998; GOODWIN & ROSS, 1992). Furthermore, often a service failure will lead to a feeling of violated trust (WANG & HUFF, 2007).

According to Mattila (2001), services that involve a high degree of human contact are particularly prone to failures. The author (p. 93) argues that because of the human factor "customers realize that being loyal to a particular service provider is no guarantee against occasional service failures". Since peer-to-peer services that involve face-to-face interaction demand a high degree of human interaction, it is possible to conclude that it creates an environment where service failures are prone to occur.

Plenty of research on service/seller failure and recovery has been done in the last few years. Researchers investigated aspects such as handling of customer's complaints following a service failure (TAX, BROWN & CHADRASHEKARAN, 1998), customer's reaction to service failure and recovery encounters (SMITH & BOLTON, 1998), equity and repurchase intention after a failure (PALMER, BEGGS & MC-MULLAN, 2000), customer's satisfaction following service failure and recovery (MCCOLLOUGH, BERRY & YADAV 2000), the impact of type of relationship on customer's loyalty following a service failure (MATILLA, 2001) and service failure in the context of online retailing (HOLLOWAY & BEATTY, 2003; PIZZUTTI & FERNANDES, 2010), to cite some.

When it is perceived that the seller/service provider has some control over the failure, this type of trust violation can generally be distinguished between morality-based and competence-based (WANG & HUFF, 2007). 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).

A third type of failure explored in literature is warmth. According to Kirmani *et at.* (2017) the lack of more extensive literature related to warmth failures can be explained by the fact that morality (i.e. integrity) and warmth traits were often considered together. According to the authors, despite sharing common traits such as gratitude and kindness, morality and warmth are conceptually and empirically distinct. For the authors, 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 hiring a service, the consumer expects to accomplish a goal with the service provider’s help, one which a person may not be able to accomplish on its own. Accordingly, the authors found that choosing a service provider who is knowledgeable and skilled to perform a certain task is considered more important than that provider’s morality or warmth traits.

Martijn *et al.* (1992) conducted a study in order to investigate negatively and positively effects in trait inferences and impression formation. The authors found that negative behavioral information leads to more certain inferences concerning morality and positive behavioral information leads to more certain inferences concerning ability. The authors also found that information regarding morality is more influential in forming an evaluative impression than equivalent information related to ability. Wang and Huff (2007) argue that customers react more negatively when they perceive a lack of integrity from the seller than when they perceive lack of competence. According to the authors, when it is perceived that the seller lacks capability, the customer may attribute this to factors ou t of the seller’s control. However, when it is perceived that the seller lacks morality, it is likely the customer will assume the seller intentionally behaved in a harmful way.

Contrary to prior research, Kirmani *et al*. (2017) found that when choosing service providers, consumers value competence traits more than integrity -if this trait does not affect the service provided (e.g. when a person knows a provider is highly skilled but acts immorally in their personal life). The authors conducted a research in order to investigate the role of competence, integrity and warmth failures when choosing a service provider and how underdog positioning affects that choice. The authors found that, in the context of service failure, when choosing a service provider, knowledge and skill to perform the task is considered more important than morality or warmth traits. That means, according to the authors, that individuals tend to value the ability to accomplish the service more than ethics, given that the moral or warmth failures do not harm the service. Results also revealed that when a moral service provider is positioned as underdog, consumers tend to feel empathy towards him/her and it attenuates the importance of competence (or lack of) traits. However, underdog positioning had no effect for the competent provider to overcome a deficit in morality or a warm provider to overcome a deficit in competence.

Following the logic of Kirmani *et al*. (2017), that users tend to value competence more than morality and warmth traits in a service provider and given that the morality failure (in the way the it was manipulated by authors and by us in this study) does not directly harm the user (therefore not compromising the quality of service), we propose that: controlling for failure severity, a competence failure will be perceived as the one compromising the quality of the core service the most, therefore leading to less willingness to forgive than other types of failure (i.e. morality or warmth). Thus, a competence failure will present no significant difference in rating and amount of tip between types of evaluation (will be less biased); while morality or warmth failures will have higher means of rating and amount of tip in the formal type of evaluation (vs. control - informal), that is, following the logic of our hypothesis 1.

Therefore:

H3: 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 morality and warmth failures, ratings will be higher in the formal (vs informal system).

H4: 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).

Studies have shown perceived quality to be an antecedent to customer satisfaction (CRONIN & TAYLOR, 1992). In line with this, a study conducted by Mohlmann (2015) with Airbnb users, revealed perceived service quality to have a positive effect on the satisfaction with collaborative services and likelihood to use that service again. For Parasuraman (1988) service quality perceptions are the result of customer's expectations versus service performance. Furthermore, according to the authors, customers evaluate quality not only based on service outcome but also the process of delivery of the service. The authors identified 5 determinants of service quality: tangibles, reliability, responsiveness, assurance and empathy. Among these, empathy involves warmth traits such as respect, consideration and friendliness. Assurance involves morality traits such as trustworthiness and honesty, and reliability traits such as skills and knowledge to perform a service. Therefore, about the main effect of type of failures on feedback, we expect that a morality failure will yield higher ratings and amount of tip since it will be perceived as the one compromising the outcome of the service the least (when compared to competence or warmth failures).

H5: In the occurrence of a given service failure, perceived quality of the service will explain the effect of type of failure on users' feedback.

Evidence points to the existence of a sense of community and social norms due to the proximity between user and provider in collaborative services. This, in turn, could bias evaluations in collaborative services, where these mechanisms are pivotal for the maintenance of service quality and even safety. Given the importance of accurate service assessment in this context, we conducted a pre-test to test whether we could successfully manipulate variables related to the phenomena under investigation, followed by an experimental study to test our hypotheses.

3. STUDY 1

Design and Participants

The first study was a factorial 3 (type of failure: competence, warmth, integrity) x (type of evaluation: formal -in the system, control -informal) between-subjects experimental design with random assignment.

For this study a sample of 373 participants was recruited via Amazon's Mechanical Turk (MTurk) service.

In Study 1, we presented a questionnaire to participants which took around 12 minutes to be completed. Thus, we compensated the workers with $1,20 per HIT (fair amount suggested by workers). The qualifications workers had to meet in order to be eligible to participate in the study were: a minimum of 95% approval rate in previous HITs, having more than 100 completed and approved HITs, being a US resident and not having participated in the pretests.

Sample size was determined using G*Power 3.1.9.2 software (FAUL et al., 2007), a tool to compute statistical power analyses[11]. The software considers the type of statistical test that will be used in the study, the expected effect size, confidence level, degrees of freedom, number of groups and number of covariates. Since the pretest showed a significant difference in perceived severity of the failure across conditions, we determined it was necessary to control for this variable. Therefore, we set analysis of covariance (ANCOVA) with fixed effects, main effects and interactions as the statistical test of choice. Expected effect size was set to medium (0.25). According to Hair et al. (2009), 95% is the suggested confidence level for the social sciences. Therefore, we set the confidence level at 95% for the sample size. The number of degrees of freedom was set to 2 (3-1)*(2-1), the number of groups was set to 6 (3x2 factorial design study) and the number of covariates to 1 (perceived severity of the failure). Within these parameters, the software determined a total sample size of a minimum of 251 subjects.

[11] For more information see http://www.gpower.hhu.de/fileadmin/redaktion/Fakultaeten/Mathematisch-Naturwissenschaftliche_Fakultaet/Psychologie/AAP/gpower/GPowerManual.pdf for a tutorial. Accessed March 10th 2018.

In order to maintain our sample as true to reality as possible, we decided to take a conservative approach and to not exclude any outlier from the sample. From our sample of 373 individuals, 125 were assigned to the competence condition, 114 were assigned to the warmth condition and 134 were assigned to the morality condition. The formal evaluation (in the system) condition had 190 participants assigned to it while the control (informal – out of the system) had 183 participants assigned to it. Most participants (49,6%) declared to be between 25-34 years old and 50,1% were females. Of the total sample (N = 373), 15 (4%) participants declared to be handicapped.

Procedure

In Study 1, the data collection instrument was created on Qualtrics software which generated a link to the questionnaire made available to MTurk participants. Participants were first presented with a short introduction to the study which included a generic description of what the research was about. The first manipulation (type of failure) was in the form of a 1-minute (approximately), muted and subtitled video in point-of-view format (dialogs used in the videos are available in appendix A). Our first manipulation was challenging as it demanded a somewhat complex role playing from the participants. The purpose of using a video for the first manipulation was to give the participant a more realistic and accurate perception of the scenario. A study conducted by Bateson and Hui (1992) showed photographic slides and videotapes to have ecological validity when used as environmental simulation of a service setting in a context of crowding. Hughes and Huby (2002), further argue that videotaped vignettes provide a more solid basis when attempting to simulate elements of reality and are superior to written vignettes since observed behavior is more easily retained and remembered.

We decided to use the same mute video for all conditions only changing the subtitle depicting a dialog between driver and passenger. We chose to use the same mute video and subtitle it according to the manipulation instead of use different videos and spoken dialogs. That was to avoid cofounds such as the tone of voice of the driver, speed of the car, the car wobble or external noises and therefore guarantee higher internal validity. Also, the subtitles made it possible for us to easily change any details in the manipulations that could be necessary after the pretest. Each vignette depicted the same scenes but a different interaction (dialog) between passenger and driver.

In all conditions a screen was shown after the passenger embarks the car and has the first exchanges with the driver informing that 15 minutes passed until reaching the destination. After which the destination is reached, there is a final interaction and the passenger disembarks the vehicle.

In the condition in which the driver seemed to lack competence, the failure was manipulated by the driver telling the passenger her phone was off because her cable to charge the phone hadn't been working properly. The driver then asks if the passenger can give her directions to which the passenger replies she would try.

In another condition the driver seemed to lack morality, bragging to the passenger she had just bought a sticker that would allow her to park in handicapped parking spots. In this condition a screen showed a text explaining the driver bought the sticker illegally and was not handicapped. We included this screen after the pretest as some participants were not sure if the driver was in fact handicapped. Also, after the pretest we included a friendlier dialog from the driver in the morality and competence conditions, to avoid overlap with the warmth condition. This included greetings and asking if the passenger was comfortable with the temperature inside the vehicle.

In the third condition, lack of warmth, the driver was unfriendly to the passenger. The driver didn't respond to greetings and only spoke once during the entire ride, replying rudely to a question asked by the passenger. The videos were randomized in all conditions.

Following the video, a text which introduced the manipulation for type of evaluation was presented to the participants. In the formal (in the system) condition, the text read: *"You can now rate your driver in the service's app. The feedback system used by the transportation service in the video is a two-way type. That means you can rate the driver and the driver can also rate you as a passenger. Drivers and passengers with a low average rating may get banned from the service"*. In the formal condition participants were told they would be evaluating the driver in the service's app, in a scale from 1 to 5. The text in the control (informal) condition read: "*You arrived at your destination and met a friend. Your friend tells you he has never used an on-demand transportation service before (like the one you just used). Your friend asks you how your experience with the service was today"*. Then, participants were asked to informally evaluate the driver to their friend as a way of telling that friend more about their experience with the service. In both conditions participants were told to rate the driver in a scale of 1 to 5. This manipulation was also randomized.

After the manipulations were introduced, participants were requested to complete a questionnaire (available in appendix B) which collected data and included measurements for the dependent variables, covariables, manipulation checks, perception of realism and demographic variables.

Measures

Dependent Variables (Feedback)

Rating

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, allowing for decimals.

Tip

We measured tip 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.

Manipulation and Attention Checks

Perception of Type of Failure – Manipulation Check

To determine if each of the failures presented to participants was perceived as intended and did not overlap, we included a scale to measure the perception of type of failure. The measures were the same used by Kirmani et al. (2017). Morality and competence were measured according to Leach, Ellemers, and Barreto (2007), with 4 items each, rated in a 7-point scales. Items for morality measured the extent to which participants perceived the provider as dishonest/honest, insincere/sincere, manipulative/not manipulative and not trustworthy/trustworthy. Competence items measured the extent to which participants perceived the provider to be incompetent/competent, not clever/clever, not knowledgeable/knowledgeable and unskilled/skilled. Warmth was measured according to Kirmani et al. (2017) also with 4 items on a 7-point scale (unfriendly/friendly, cold/warm, unsociable/sociable and not nice/nice). The three scales have been previously tested and showed to be reliable for measuring perception of type of failure (LEACH, ELLEMERS & BARRETO, 2007; KIRMANI, 2017).

Type of Evaluation Manipulation Check

The manipulation check for type of evaluation consisted of one question: “you evaluated the driver:”. Possible answers were: “to your friend”, “in the service’s app” and “do not remember”.

Attention Check

The attention check was measured with one item: “What happened during the trip shown in the video?”. Possible answers were: “driver was bragging about buying a handicap parking permit sticker”, “driver couldn’t charge her phone and did not know how to get to the destination”, “driver was cold/unfriendly to the passenger” and “none of the above”.

Mediators

Forgiveness

We adapted the EFS (Emotional Forgiveness Scale) to our study’s context, in order to measure the degree to which participants would forgive the service failure depicted to them. The EFS was created by Worthington, Hook, Utsey, Williams, and Neil (2007) and consists of eight items that measure the degree to which an individual has experienced emotional forgiveness in face of a transgression. The EFS consists of two measures: Presence of Positive Emotion and Reduction of Negative Emotion. For this study we used only the items (except item 2 which was not appropriate for this study) measuring Reduction of Negative Emotion, since this aspect of forgiveness fits better the context of our study. Participants rated each item on a seven-point rating scale from 1 = strongly disagree to 7 = strongly agree. The scale has shown evidence of internal consistency and construct validity (WORTHINGTON ET AL., 2007).

Fear of Retaliation

Fear of retaliation was measured with one question, in a 7-point Likert scale, adapted from Kudish , Fortunato and Smith (2006): "I fear to suffer negative consequences if I give an honest feedback to this driver". The scale ranged from "Strongly Disagree" to "Strongly Agree".

Reciprocity

Reciprocity was also measured with one item, in a 7-point Likert scale: "I rated the driver according to how I expect the driver has rated me as a passenger". The scale ranged from "Strongly Disagree" to "Strongly Agree".

Perceived Quality (compromised)

Perceived quality was measured with a single-item: "To which extent you believe that what happened during the trip has compromised the quality of the service provided (transportation service to the destination)? – The items were rated with a 7-point Likert-type scale ranging from "Not at All" to "Very Much".

Control Variables

Perceived Severity of the Failure

Perceived severity of the failure was measured using a 3-item scale from Weun, Beatty and Jones (2004). One example is "If the inconvenience during the ride in the video was really happening to you, you would consider it to be: 1 = Not Severe at All 7 = Very Severe". All measures were in seven-point scales. The scale was previously tested for convergent and discriminant validity and exhibited satisfactory results (WEUN, BEATTY & JONES, 2004).

"*Used Similar Services Before"*

We measured if participants had used similar services before with a one-item scale: "Have you ever used transportation services such as Uber, Cabify, Lyft or similar before?". Possible answers were "yes" and "no".

Frequency of Use

Frequency of use was also measured with one item: "Considering the past 6 months, with which (average) frequency have you used on-demand transportation services, such as the one shown in the video?" Possible answers were: "Less than one ride a month", "2-3 times a month", "once a week", "2-3 times a week", "4-5 times a week", "5-6 times a week" or "7+ times a week".

Handicapped

We measured if participants were handicapped with a one-item question: "Are you handicapped?". Possible answers were "yes" and "no".

Gender

Gender was measure with the question: "What is your gender?". Possible answers were: "male", "female" or "do not wish to answer".

Age

Age was measured with one question: "what is your age?". Possible answers were: "18-24", "25-34", "35-44", "45-54" and 55+.

Perception of Realism

Perception of realism as measured with one item, in a 7-point Likert scale: "I believe that the situation presented in the video could happen during a ride with on-demand transportation services". The scale ranged from "Strongly Disagree" to "Strongly Agree".

Validity of Scales

An exploratory factor analysis utilizing Varimax rotation method was conducted with the 4 items measuring competence, 4 items measuring morality, 4 items measuring warmth, 3 items measuring perceived severity of the failure, and 3 items measuring forgiveness. The analysis revealed the scales exhibit *satisfactory factorial* structure. The items measuring the competence showed factor loadings between .822 and .663. The items that measured morality showed factor loadings between .837 and .759. The items measuring the warmth factor had factorial loadings between .946 and .866. Perceived severity of the failure had loadings ranging from .744 to .536. Forgiveness had loadings between .857 and .828. The Kaiser-Meyer-Olkin was .879 which is above the threshold of .6 (Kaiser, 1974) and the Bartlett's Test of Sphericity reached statistical significance (p<.001).

Cronbach's Alpha was also used to measure reliability of the scales. The statistical analysis of the scales yielded, in general, good alphas[12]. The scale measuring competence consisted of 4 items (α = .89); the scale for morality also consisted of 4 items (α = .88); the scale measuring warmth (α = .96) was composed of 4 items as well. Perceived severity of the failure and forgiveness were measured with 3-item scales each (α = .80 and α = .93, respectively).

Pretests

The main objective of the pretest was to test whether the manipulations of the independent variables were effective and if the scenarios were sufficiently distinct, realistic and well understood by the participants. For this study we ran two pretests. The first pretest had a sample of 151 participants recruited on Amazon's Mechanical Turk. Through a frequency distribution we identified missing values. From the 151 participants in the sample, we excluded seven who did not complete the survey, reducing the sample to 144 subjects.

In the pretest the manipulation for type of failure (competence, morality and warmth) was tested using a scale developed by Kirmani et al. (2017), available in the appendix. An analyses of variance (One-way ANOVA) with Post Hoc test Tukey HSD showed an effect of type of failure manipulation in warmth (F(2, 141) = 47.206, p<.001) and morality (F(2,141) = 25.568, p<.001) perceptions but there was no significant effect in competence (F(2, 141) = 1.858, p = .160). Post Hoc comparisons using the Tukey HSD test indicated that in the competence condition the mean score for competence (M = 4.27, SD = 1.58) was not significantly different from the mean score for warmth (M = 3.81, SD = 1.59) and morality (M = 4.41, SD = 1.58).

[12] All scales are 7-point Bipolar Likert type.

We also included an attention check for type of failure which consisted in the following question: “what happened during the trip shown in the video?”. Possible answers were “Driver was bragging about buying a handicap parking permit sticker”, “driver couldn’t charge her phone and didn’t know how to get to the destination”, “driver was cold/unfriendly to the passenger” and “none of the above”. A Crosstabs test showed that most respondents answered the manipulation check correctly: competence (83,3%), morality (89,5%) and warmth (83,3%). A chi-square test showed a statistically significant association between the type of failure and the attention check ($x^2(6, N = 144) = 200.86$, $p<.001$)

Perception of realism (M = 5.82, SD= 1.35) was tested using the affirmation "I believe that the situation presented in the video could happen during a ride with on-demand transportation services" to which participants were required to answer with a 7-point Likert scale ranging from “Strongly disagree” to “Strongly agree”. The mean value for perception of realism was higher than the mean of the scale (4>), and type of failure had a significant effect on perception of realism with mean values varying between the competence (M = 5,6, SD = 1,49), morality (M= 5,69, SD = 1,32) and warmth (M = 6,17, SD = 1,19) conditions. Perceived severity of the failure was tested using ANOVA ($F(2,141) = 0.261$, $p>.05$). The test results indicated perception of severity of the failure was similar between the competence (M = 3,4, SD = 1,01), morality (M = 3,56, SD = 1,23) and warmth (M = 3,43, SD = 1,15) conditions.

After the first pretest we performed adjustments in the manipulation for type of failure. We made the scenarios clearer and more distinguishable to participants by increasing warmth in the competence and morality conditions. In the new videos the driver was friendlier, greeting the passenger in a warmer manner and asking about the car temperature We also made the competence failure more evident by adding another dialog in which the driver shows ignorance regarding the use of the service’s app.

We ran a second pretest in order to verify the effectiveness of the manipulations after the changes performed in the scenarios. Through a frequency distribution we identified missing values. From the 181 participants in the sample, we excluded 16 who did not complete the survey, reducing the sample to 165 subjects. Participants were also recruited on Amazon’s Mechanical Turk.

In the second pretest, we included a control group. The control condition was deemed necessary in this case because we needed a group where the treatment is withheld in order to have a base for comparison with the treatment group (GOODWIN & GOODWIN, 2003). An analyses of variance (One-way ANOVA) for type of failure manipulation showed a significant difference between morality (F(2, 162) = 32.820, p<.001), competence (F(2, 162) = 16.978, p<.001), and warmth (F(2, 162) = 66.498, p<.001) scenarios. Post Hoc tests were also performed. The comparisons within conditions revealed a significant difference in perception of type of failure (p<.01) in most cases. However, the test indicated that within the competence condition (M = 2.92), warmth (M = 4.19) and morality (M = 4.16) types of failure did not have a significant difference (p>.05) and within the warmth condition (M = 2.14), competence (M = 4.89) and morality (M = 4.83) also did not have a significant difference (p>.05). In the morality condition the mean for morality (M = 3.04) was lower than the mean for warmth (M = 4.35) and competence (M = 5.09), indicating a significant difference between all groups (p<.05). Therefore, the results allow us to conclude that the failure manipulations were effective in all conditions.

The Crosstabs test for the attention check for type of failure (participants were asked what happened in the video) showed that most respondents answered the manipulation check correctly: competence (98,2%), morality (92,4%) and warmth (90,9%). A chi-square test showed a statistically significant association between type of failure and the attention check (x^2(6, *N* = 165) = 295.138, p<.001).

The manipulation check for type of evaluation was included. The sentence “You evaluated the driver” was presented to which participants had three choices for an answer: “in the service’s app”, “to a friend” and “do not remember”. A crosstabs test showed most participants answered the manipulation check correctly: “in the service’s app” (96%) and “to a friend” (92,4%). A chi-square test showed a statistically significant association between the type of evaluation and the manipulation check (x^2(4, *N* = 165) = 139.788, p<.001).

The mean of perception of realism was also high in the second pretest (M = 5.78, SD = 1.33, 7-point Likert scale), above the mean of the scale (<4) and similar across conditions, indicating participants believed the situation presented to them in the manipulations was realistic in all conditions.

Data Analysis and Assumptions for Statistical Tests

For the data analysis we used IBM SPSS Statistics software as a tool to operationalize statistical testing. To analyze manipulation and attention checks Crosstabs and Pearson's Chi Square tests were used. To test the hypothesis analysis of covariance (ANCOVA) was used. The ANCOVA differs from the ANOVA by allowing covariables to be controlled in the model and have its influence adjusted before the ANOVA procedure (Hair et al., 2009). Since some of the variables under study were categorical variables, we were interested in an interaction effect and there were covariables, this method was chosen for the analysis of hypothesis. In addition to the ANOVA, Post Hoc test Turkey HSD was also used to test for differences between conditions when the ANOVA indicated significance in the effects. According to Sprinthall and Fisk (1990), Turkey HSD is the most commonly used Post Hoc choice.

We also measured the effect sizes using the eta partial squared (η^2p) which, according to Cohen (1988), is the proportion of the effect added to the error of variance which is attributed to the error. For the author, an eta partial squared (η^2p) around 0,01 is considered small, 0,06 is considered a medium effect size and from 0,13 the effect is considered large.

For mediation and moderation effects, procedures suggested by Preacher and Hayes (2004) were used and operationalized through the PROCESS macro for SPSS. The PROCESS macro is a modeling tool which performs mediation, moderation and conditional process analysis and bootstrapping procedures (resampling technique) (Hayes, 2013).

The PROCESS macro also performs procedures known as Spotlight analysis and Floodlight analysis, techniques that were used for statistical assessment in this study. According to Spiller et al. (2013, p.277), "the Spotlight analysis provides an estimate and statistical test of the simple effect of one variable at specified values of another continuous variable". Still according to the authors, the Floodlight analysis from Johnson-Neyman in the other hand, identifies the simple effects of a variable Z in every possible value of X. It identifies the regions in X where the effect of Z is significant and the regions where it is not. The authors suggest that unless the researcher is interested in a particular value of X, the Floodlight analysis (Johnson-Neyman technique) should substitute the Spotlight analysis (Spiller et al., 2013).

According to Hair *et al.* (2009), there are three assumptions for the use of analysis of variance (ANOVA): independency of observations, equality of variance-covariance matrices, and normality of distribution of dependent variables. The author also suggests that before running ANOVA and/or ANCOVA tests, a verification of missing values and atypical values (Outliers) should be performed.

Independency of observations was achieved through the random distribution and a between-subjects design (each participant was allocated to only one experimental condition). Through an analysis of frequencies, it was possible to determine there were no missing values in the database of Study 1. Atypical values (standardize values >|3|), which according to Hair et al. (2009) could denote the sample contains Outliers, were verified using a Z-test. Outliers were only detected in the duration variable (the amount of time respondents took to complete the study), but we decided to keep these cases in the sample to preserve the authenticity and validity of the database. Therefore, statistical analysis was conducted using all 373 observations from the original sample.

The equality of variance-covariance matrices or homoscedasticity was verified using Levene test. According to Hair *et al.* (2009), homoscedasticity of data indicates that the dependent variables exhibit equal variations across different levels of the predictor variable. The Levene test verifies if the variances of a variable are equal or not between groups. An ANOVA was conducted which showed that the dependent variable rating varied across different levels of the independent variables (type of failure and type of evaluation) under study. The Levene test ($F(5,367) = 2,791$, $p<.05$) indicated the heteroscedasticity of the depended variable rating. Another ANOVA was conducted to verify the homoscedasticity of the dependent variable tip. The Levene test showed that this dependent variable is also heterogeneous (F(5, 367) = 5,032, p<.001), therefore indicating a difference in the variance of the variable across different levels of the independent variables.

The normality of distribution of the dependent variables was verified using kurtosis values. For the rating variable, skewness (0,98) and kurtosis (-1.037) values were acceptable for normal distribution, as were the skewness (0,84) and kurtosis (-0,322) values for tip[13]. In addition to the

[13] Hair *et al.* (2006)

univariate test, the Shapiro-Wilk test was also performed. The results of the Shapiro-Wilk test for rating indicate that the null hypothesis (H0) could not be rejected, since the probability was lower than .05 (S-W = 0,959, p<.001). The tip variable the test yielded a similar result (S-W = 0,859, p<.001), indicating that the null hypothesis could not be rejected for this variable as well. However, due to the large sample size, parametric tests were used despite the data not meeting all the assumptions for parametric tests[14].

Main Study Results

The data collected for Study 1 was analyzed according to the procedures described previously. Results for manipulation checks and tests for main and interactive effects are presented next.

Manipulations and Attention Checks

Type of Failure

The manipulation check for type of failure was conducted using a One-way ANOVA. The analysis showed a statistically significant difference between groups: morality (F(2,370) = 58.316, p<.001), competence (F(2,370) = 65.202, p<.001), and warmth (F(2,370) = 244.202, p<.001).

Post Hoc test Tukey HSD showed that within the competence failure condition there was a significant difference for perceptions of type of service failure between competence (M = 2.71) and warmth (M = 4.20) (p<.001) and competence and morality (M = 4.26) (p<.001). Within the warmth failure condition there was a significant difference for perception of type of failure between warmth (M = 1.91) and competence (M = 5.06) (p<.001) and warmth and morality (M = 4.87) (p<.001). Within the morality condition there was a significant difference in perception of type of failure between morality (M = 2.94) and competence (M = 4.78) (p<.001) and morality and warmth

[14] According to Ghasemi and Zahediasl (2012, p. 486) "with large enough sample sizes (> 30 or 40), the violation of the normality assumption should not cause major problems; this implies that we can use parametric procedures even when the data are not normally distributed"

(M = 4.13) (p<.01). Therefore, allowing us to conclude that the participants perceived the conditions correctly and the manipulation for type of failure was successful.

As in the pretest, we also included an attention check for type of failure. For the attention manipulation check, the question "what happened during the trip shown in the video?" was presented to participants. The participants had to choose one of the following answers: "Driver was bragging about buying something illegal", "Driver couldn't charge her phone and didn't know how to get to the destination", "Driver was cold/unfriendly" or "none of the above". The data test for the attention check for type of failure showed that most respondents answered it correctly: competence (97,6%), morality (93,2%) and warmth (96,4%). A Pearson Chi-Square test showed a statistically significant association between the type of failure and the attention check ($x^2(6, N = 373) = 677.092, p<.001$).

Type of Evaluation

In the manipulation check for type of evaluation participants were presented with the same question as in the pretest: the sentence "You evaluated the driver" was shown to which participants had three choices for an answer: "in the service's app", "to a friend" and "do not remember". A Crosstabs test showed the manipulation check for informal type of evaluation was successful for the control (informal) condition (93,4%)[15] and the formal condition (96,8%) with most participants answering it correctly. The Pearson Chi-square test yielded a statistically significant relationship between the manipulated variable and the manipulation check ($x^2(2, N = 373) = 315.472, p<.001$). The results of the statistical analyses confirm the manipulation for type of evaluation was effective.

Control Variables

We conducted an ANCOVAs in order to test for potential covariables in our model. The test revealed perceived severity of the failure to have a significant effect on the variable rating (F(1,

[15] We also tested as a control condition another informal type of evaluation, or "out of the system", where the respondent would directly tell the researcher (instead of a friend) how he or she would rate the driver. We dropped this control as there was no significant difference between this condition and the condition where the respondent would tell a friend about his or her assessment of the driver. We decided to maintain only the latter as we perceived it to be more realistic.

292) = 189.467, p<.001). To identify the effects that different types of service failures had on perceived severity of the failure we conducted a second ANOVA (F(2, 370) = 4.627, p<.05). Post Hoc test Tukey HSD indicated a significant difference (p<.05) in the means of perceived severity between the competence service failure condition (M = 4.65, SD = 1.27) and the warmth service failure condition (M = 4.14, SD = 1.38) forming two subsets; the mean of perceived severity in the morality service failure condition (M = 4.28, SD = 1.35) was not significantly different from competence (p>.05) or warmth (p>.05), therefore, perceived severity of the failure was included in the model as a covariable. Other potential covariates for the dependent variable rating were tested. These included if participants had used similar services before (p>.05), frequency of use (p>.05), age (p>.05), gender (p>.05), and being handicapped (p>.05). Therefore, no other control variables were included in the model.

These potential covariates were also tested in the analysis model for the dependent variable tip. Severity of the failure was shown to have a significant effect on this variable (F(1, 292) = 55.529, p<.001). We also tested if participants had used similar services before (p>.05), frequency of use (p>.05), gender (p>.05), if participants had had issues with similar services in the past (p>.05) and being handicapped (p>.05). The variable age (F(1, 292) = 7.410, p<.05) had a significant effect on the dependent variable tip. However, a Pearson Chi-Square test revealed no significant difference in the variable age between conditions in all age groups (p>.05). Therefore, we decided not to include this variable in the model.

In addition to testing control variables, we also tested perception of realism. A one-item question was included in the instrument of data collection to assess this variable. The mean of perception of realism was high (M = 5,79, SD = 1,36), above the mean of the scale (<4) and similar across conditions, with an ANOVA test pointing to no statistically significant difference between conditions (F(2, 367) = 0,99, p>.05), indicating participants believed the situation presented to them was realistic in all conditions.

Hypotheses Tests

To test the hypothesis, we used the ANCOVA method for statistical analysis. The ANCOVA assess the interaction between type of failure and type of evaluation as independent variables, rating and tip as dependent variables, and perceived severity of the failure as a covariable. We included perceived severity of the failure as a covariable since statistical analysis showed a significant difference in perceived severity across type of failure conditions.

Rating

To test for main and interactive effects between the type of failure and type of evaluation on rating we performed an ANCOVA test (Table 1) with perceived severity as a covariable in the model. The test showed that for rating, the interactive effect of type of failure and type of evaluation was non-significant ($F(2,366) = 0.206$, $p>.05$, $\eta^2 p = 0{,}001$)[16]. However, the ANCOVA analysis showed a significant direct effect of type of failure in rating ($F(2,366) = 18.420$, $p<.001$, $\eta^2 p = 0{,}091$) and type of evaluation in rating ($F(1,366) = 24.511$, $p<.001$, $\eta^2 p = 0{,}063$). Perceived severity of the failure was confirmed as having a significant effect on rating ($F(1,366) = 281.783$, $p<.001$, $\eta^2 p = 0{,}435$), therefore it was kept in the model. This allows us to conclude that different types of failure and types of evaluation in on-demand transportation services affect user's ratings of drivers differently (Figure 1).

[16] The eta partial squared ($\eta^2 p$) measures effect size. According to Cohen (1988), eta partial squared values of .01 are considered small, 0,06 are considered medium and 0,13 and above are considered strong effects.

Table 1 – Results of ANCOVA – Main and Interactive Effects of Type of Failure and Type of Evaluation on Rating

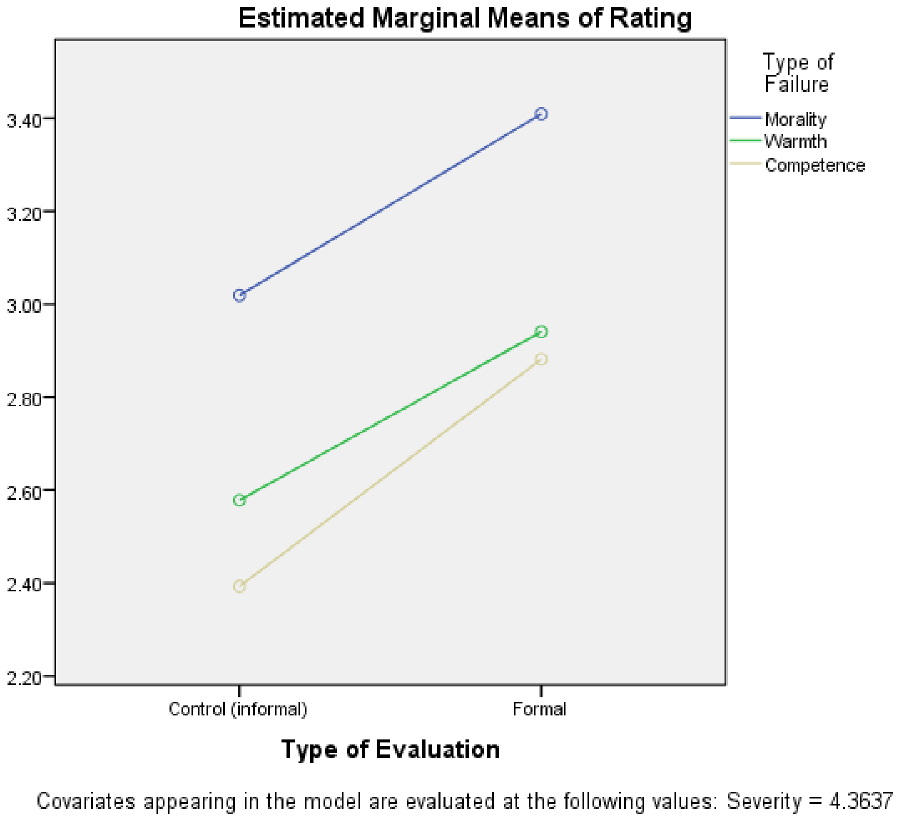
Tests of Between-Subjects Effects

Dependent Variable: Rating

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	236.908[a]	6	39.485	61.216	.000	.501
Intercept	846.996	1	846.996	1313.170	.000	.782
Severity_failure	181.750	1	181.750	281.783	.000	.435
Type_failure	23.761	2	11.881	18.420	.000	.091
Type_eval	15.810	1	15.810	24.511	.000	.063
Type_failure * Type_eval	.265	2	.133	.206	.814	.001
Error	236.070	366	.645			
Total	3582.130	373				
Corrected Total	472.978	372				

Source: Research data (2018)

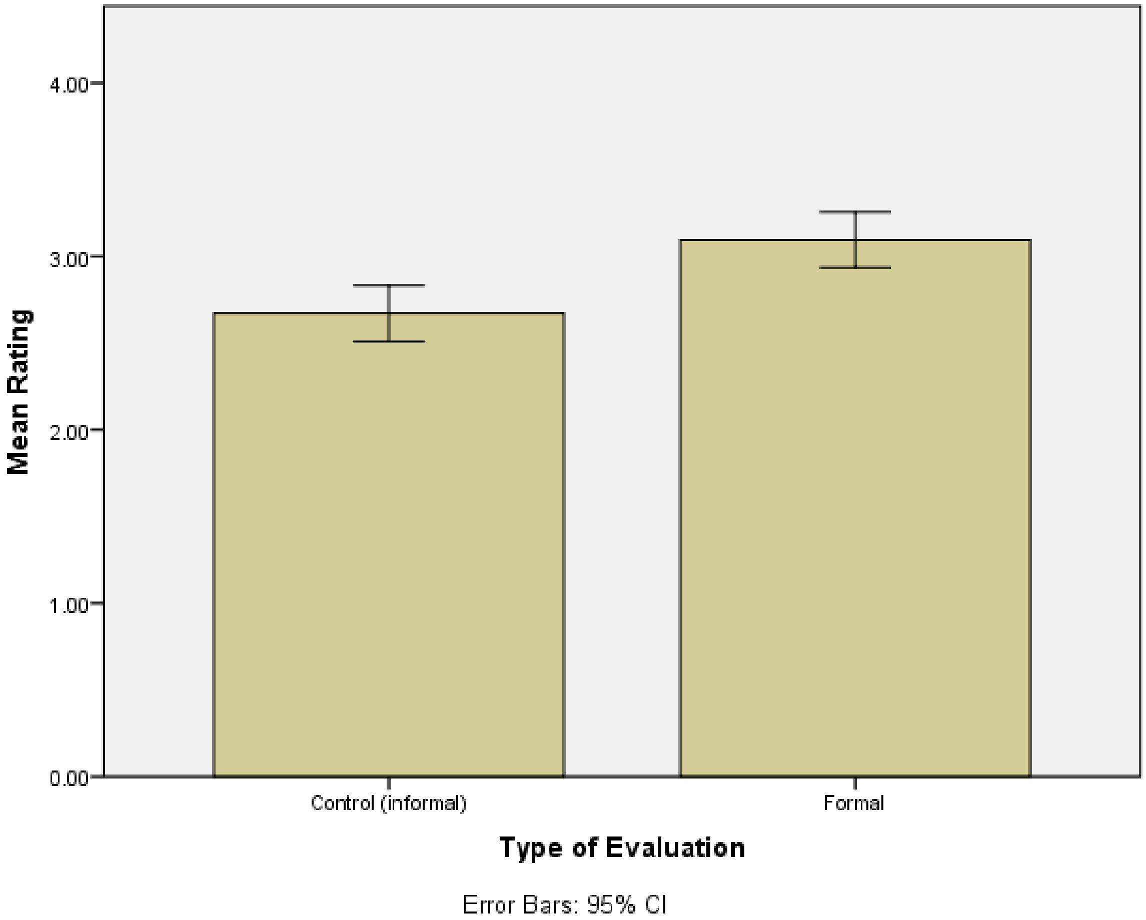
Figure 1 – Graphic Representation of the Means of Rating Between Conditions of Type of Failure and Type of Evaluation



Source: Research data (2018)

The initial ANCOVA test revealed different types of evaluation methods affect the way user's rate their experience with the driver (Table 1) differently. In support of H1a, the formal (in the system) type of evaluation yielded a significantly higher total average rating (M = 3,09, SD = 1,11) than the control condition (M = 2,67, SD = 1,10) (Figure 2).

Figure 2 – Means of Rating Between Type of Evaluation Conditions

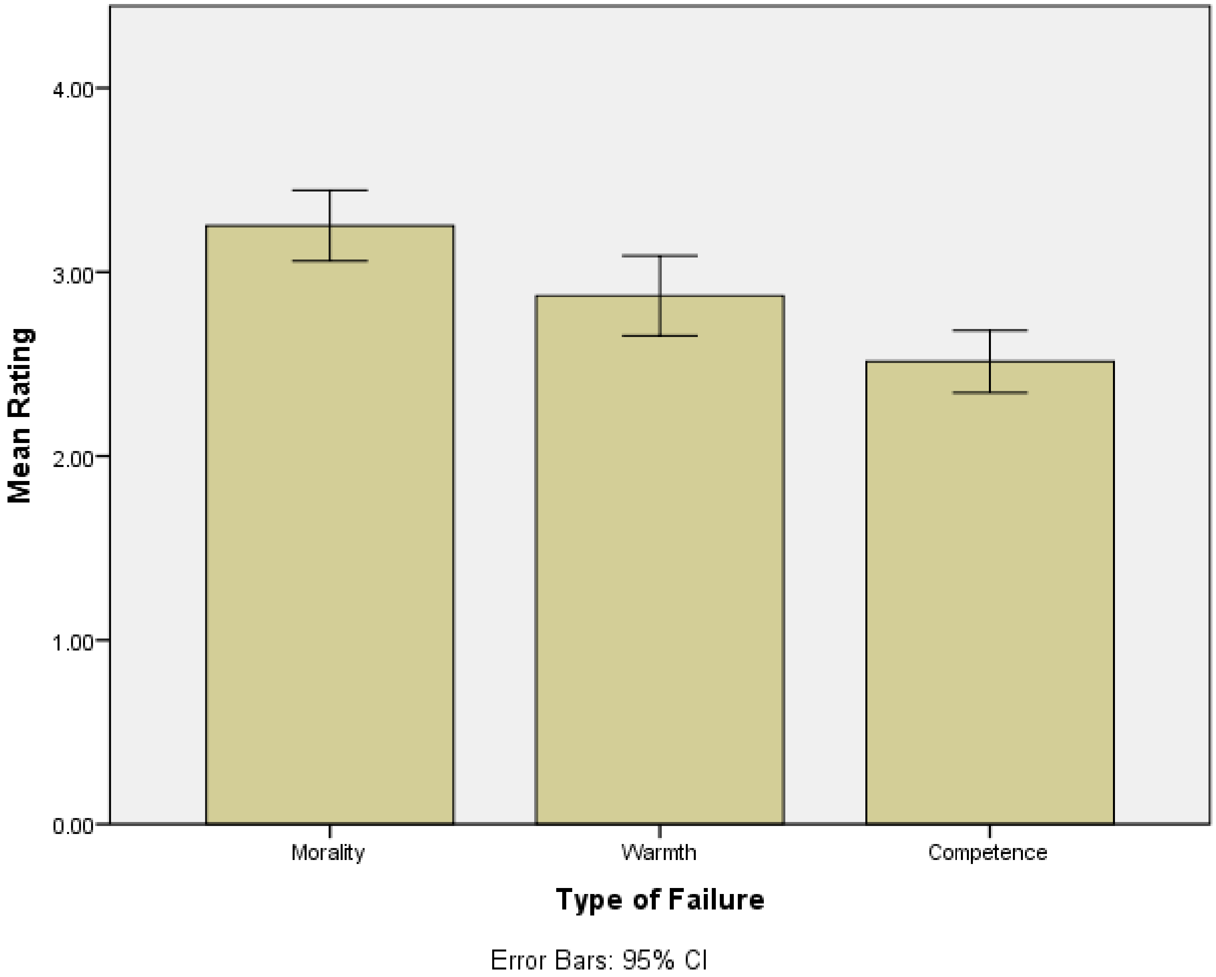


Source: Research data (2018)

When considering each type of failure individually, Spotlight test indicated no significant difference (p>.05) in the means of rating between type of evaluation conditions. The means of rating in the competence condition were M = 2,80 in the control condition and M = 3,34 in the formal condition. For the morality type of service failure, the means of rating were M = 2,94 in the control condition of type of evaluation and M = 3,13 in the formal one. For the warmth type of service failure, the means of rating between the were M = 3,46 in the control condition and M = 3,71 in the formal one. Therefore, H3a which posited that there would be no statistically significant in the mean of rating in the competence condition of type of failure between conditions of type of evaluation was supported but H3b was rejected.

We conducted an ANCOVA test, controlling for perceived severity, to compare the means of rating between different types of service failure. When considering both types of evaluation (Figure 3), Pairwise Comparisons indicated a significant difference in the means of rating between the morality (M = 3,25) and warmth (M = 2,87) types of service failure (p<.05) and the morality and competence (M = 2,49) types of service failure (p<.05). However, no difference was found in the means of rating between the competence and warmth types of service failure (p>.05).

Figure 3 – Means of Rating Between Type of Failure Conditions



Source: Research data (2018)

Mediators

In literature (RESNICK *et al*., 2000; BOLTON *et al.*, 2013; FRADKIN, 2015) fear of retaliation and reciprocity are often mentioned in connection to feedback bias in mutual feedback systems. In Study 1, we included two questions to explore these possible mediators "I fear to suffer negative consequences if I provide an honest feedback to this driver" and "I rated the driver according to how I expect the driver has rated me as a passenger" respectively. In line with our proposition, that in collaborative services this mediation does not occur (and instead forgiveness and anticipation of guilt mediate this effect), results indicated the type of evaluation did not influence reciprocity (F(1,209) = 1.038, p>.05), nor fear of retaliation (F(1,209) = 1.061, p>.05). This was investigated further using the macro PROCESS developed by Hayes (2013) for SPSS software. The model 4 for mediation by Hayes (2013), which considers the effect of an X variable (type of failure) on a Y variable (rating), mediated by a M variable (quality compromised) was used in the test. As recommended by Hayes (2013) the number of bootstrap samples was set to 5000. Bias corrected was the chosen method to generate the confidence intervals (CI) via bootstrapping. Perceived severity of the failure was included as a covariable in the model.

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* = -0,20, *se* = 0,20, *t* = 1,02, *p*>.05, confidence interval (CI) between -0,60 and 0,19), the path between reciprocity and rating is significant (*b* = 0,15, *se* =0,03, *t* = 4,86, *p*<.05, confidence interval (CI) between 0,09 and 0,22), however the indirect effect is non-significant (*b* = -0,03, *bootse* = 0,03, confidence interval (CI) between -0,10 and 0,02).

Similarly, results indicated no significant path between type of evaluation and fear of retaliation (*b* = -0,19, *se* = 0,19, *t* = -1,02, *p*>.05, confidence interval (CI) between -0,58 and 0,18). The test indicated a significant path between fear of retaliation and rating (*b* = 0,12, *se* = 0,03, *t* = 3,59, *p*<.05, confidence interval (CI) between 0,05 and 0,19). The indirect effect was confirmed as non-significant (*b* = 0,05, *bootse* = 0,02, confidence interval (CI) between -0,08 and 0,02).

Perceived Quality Mediation

An analysis of covariance (ANCOVA), including perceived severity as a covariable, showed that the type of failure had a significant effect in the perception of how the quality of the service had been compromised by the failure (F(2, 369) = 23.235, p<.001). This result suggests that the perception of quality could be a mediator for the effect of type of service failure in rating, as proposed in H5. Pairwise Comparisons results (Table 2) revealed a difference in the means of perceived quality (compromised) between the three types of service failure conditions (p<.05). The morality failure condition yielded the lowest perceptions that the quality of the service had been compromised by the failure (M = 3,64), followed by the warmth failure (M = 4,17). The competence failure, however, was perceived as the one which compromised the quality of the service the most (M = 4,90). The means of perceived quality (compromised) are depicted in Figure 4. Thus, the effect of type of failure on rating (with respondents indicating higher rating in the morality failure condition), could be explained by the fact that this type of failure is perceived as compromising the outcome of the core service less than warmth or competence failures, as proposed in H5.

Table 2 – Mean comparisons of perceived compromised quality between different types of failure

Pairwise Comparisons

Dependent Variable: Perceived Compromised Quality

		Mean			95% Confidence Interval for Difference[b]	
(I) Type of Failure	(J) Type of Failure	Difference (I-J)	Std. Error	Sig.[b]	Lower Bound	Upper Bound
Morality	Warmth	-.530*	.188	.005	-.899	-.160
	Competence	-1.256*	.184	.000	-1.618	-.893
Warmth	Morality	.530*	.188	.005	.160	.899
	Competence	-.726*	.193	.000	-1.105	-.346
Competence	Morality	1.256*	.184	.000	.893	1.618
	Warmth	.726*	.193	.000	.346	1.105

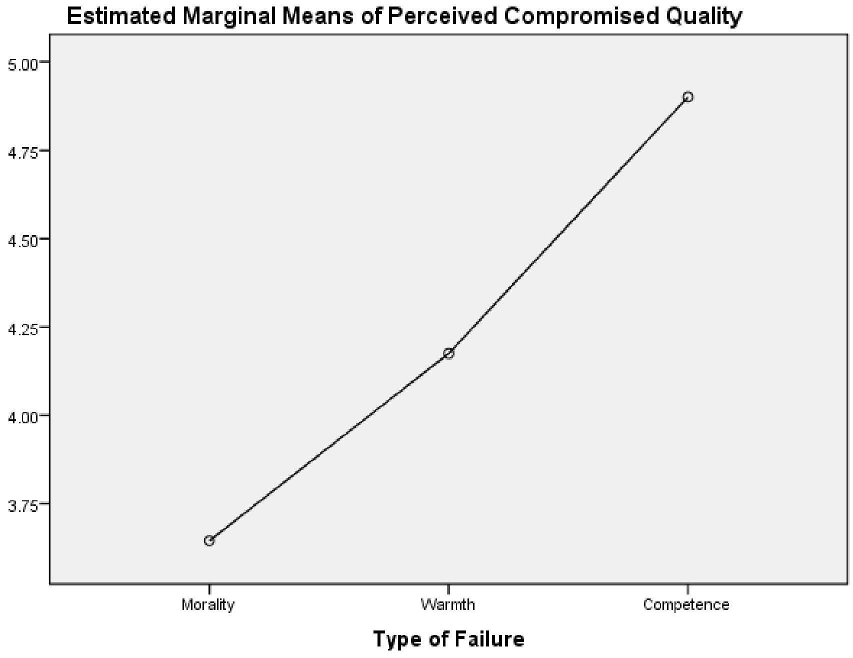
Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Source: Research data (2018)

Figure 4 – Graphic Representation of Perceived Compromised Quality of the Service Between Type of Failure Conditions



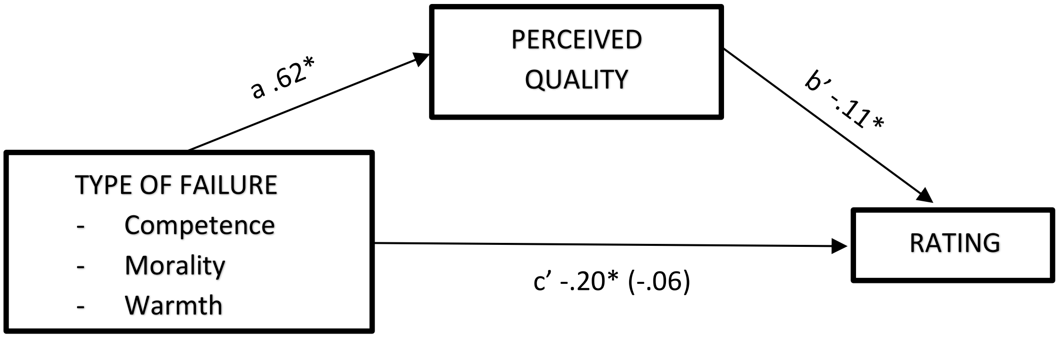
Covariates appearing in the model are evaluated at the following values: Severity = 4.3637

Source: Research data (2018)

This possible mediation was further investigated using the PROCESS macro. Perceived severity of the failure was included in the model as a covariable.

The test results indicated that type of service failure has an effect on perceived compromised quality (i.e., the more the perception that the quality was compromised the lower the rating). However, a direct effect remains between the predictor and the outcome variables with the mediator in the model (*b* = -0,20, *se* = 0,05, *t* = -3,80, *p*<.05, confidence interval (CI) between -0,31 and -0,09). Therefore, indicating perception of quality compromised as a partial mediator for the effect of type of service failure on rating (Figure 5), confirming H5. The explained variance of the model is 0,46.

Figure 5 – Theoretical Model of Mediation Between Type of Failure and Rating[17]



Source: Research data (2018)

Forgiveness Mediation

In H2 we proposed that the effect of type of evaluation on feedback would be mediated by forgiveness. In order to verify this hypothesis, we first conducted an ANOVA test. Results (Table 3) revealed a significant effect of type of evaluation on forgiveness (F(1,371) = 8.502, p<.05). The

[17] Figure 1. Standardized regression coefficients for the relationship between type of service failure and rating mediated by perceived quality. The standardize regression coefficient between type of service failure and rating, controlling for perceived quality, is in the parenthesis. **p*>.05

mean of forgiveness was significantly higher (p<.05) in the formal (in the system) type of evaluation (M = 4,69, SD = 1,73) when compared with the control for type of evaluation (M = 4,16, SD = 1,83). The means are depicted in Figure 6.

Table 3 – ANOVA Results – Main Effect of Type of Evaluation on Forgiveness

Tests of Between-Subjects Effects

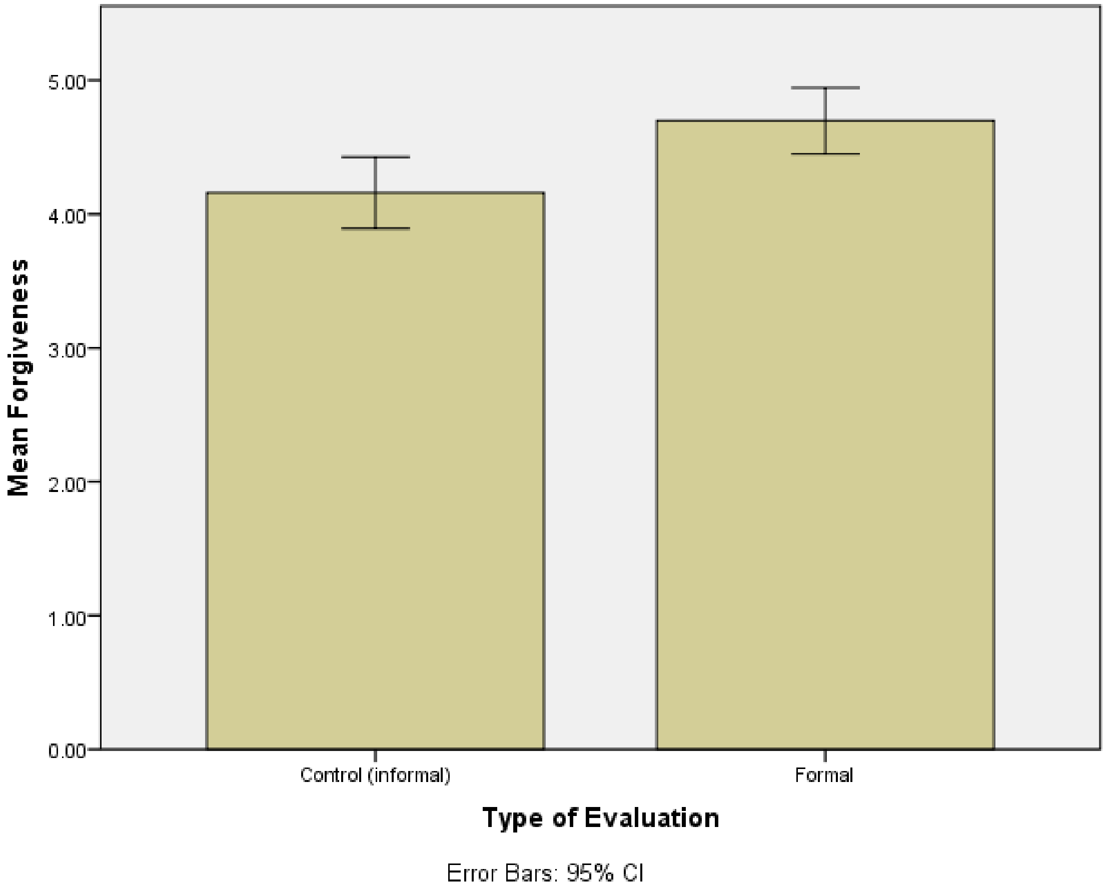
Dependent Variable: Forgiveness

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	26.977[a]	1	26.977	8.502	.004
Intercept	7315.093	1	7315.093	2305.564	.000
Type_evaluation	26.977	1	26.977	8.502	.004
Error	1177.109	371	3.173		
Total	8538.444	373			
Corrected Total	1204.085	372			

a. R Squared = .022 (Adjusted R Squared = .020)

Source: Research data (2018)

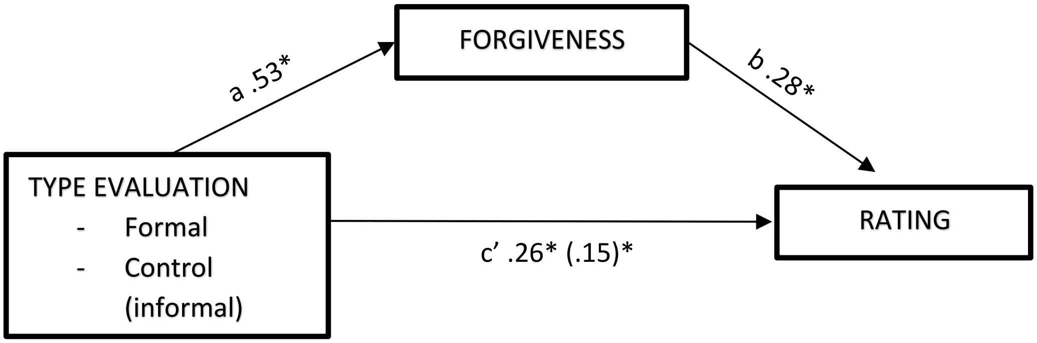
Figure 6 – Graphic Representation of Forgiveness Between Different Types of Evaluation



Source: Research data (2018)

After finding a significant result in the ANOVA test, we conducted a mediation analysis using the PROCESS macro. Results showed that forgiveness is indeed a mediator for the effect of type of evaluation on rating. The analysis revealed a significant effect of type of evaluation on forgiveness (*b* = 0,53, *se* = 0,18, *t* = 2,91, *p*<.05, confidence interval (CI) between 0,17 and 0,90) and a significant effect of forgiveness on rating (*b* = 0,28, *se* = 0,02, *t* = 10,11, *p*<.001, confidence interval (CI) between 0,23 and 0,34). However, a direct effect between type of evaluation and rating remained (*b* = 0,26, *se* = 0,10 *t* = 2,60, *p*<.05, confidence interval (CI) between 0,06 and 0,47), reveling a partial mediation (Figure 7). Therefore, forgiveness was found to be a mediator for the effect of type of evaluation on rating, offering support to H2. The explained variance of the model is 0,03. In Study 2 we investigate anticipation of guilt as another possible mediator for this effect.

Figure 7 – Theoretical Model of Mediation Between Type of Evaluation and Rating[18]



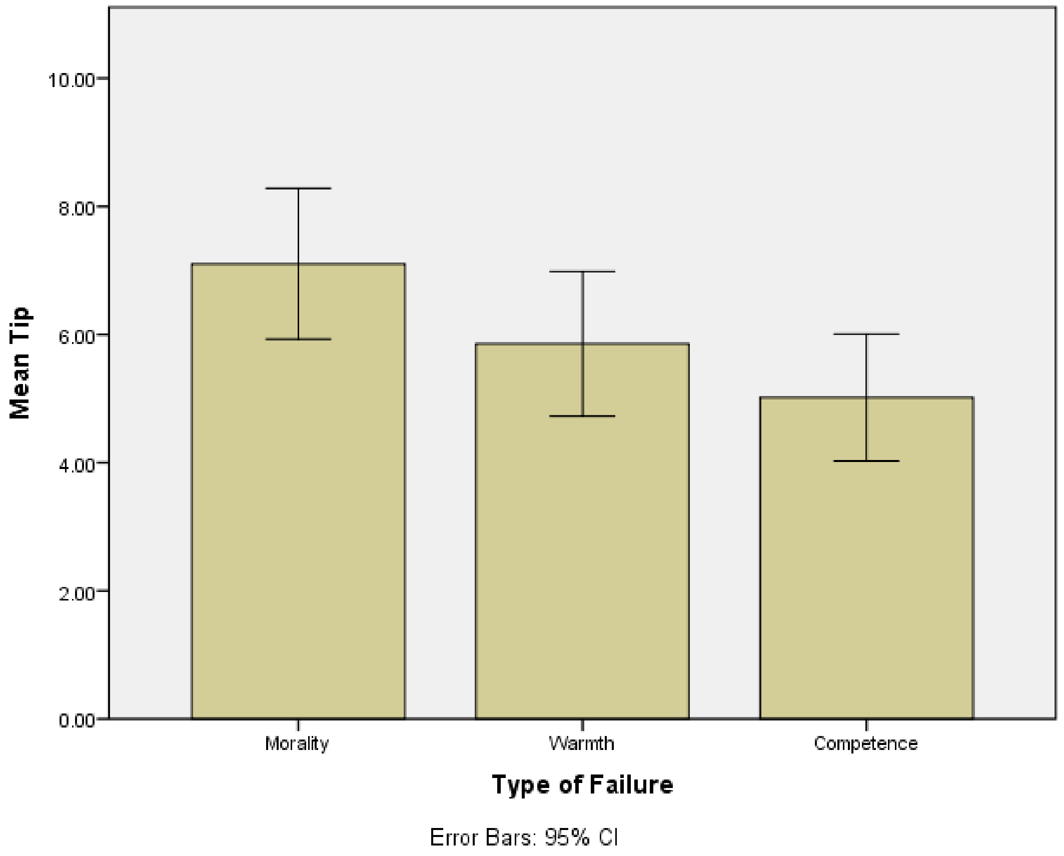
Source: Research data (2018)

Tip

We performed an ANCOVA to test for main and interactive effects between type of service failure and type of evaluation on tip (Table 4). Again, perceived severity of the failure was included as a covariable in the model (F(1,366) = 82,428, p<.001, η²p = 0,184). The test indicated a main effect of type of service failure on tip (F(2,366) = 3,037, p<.05, η²p = 0,016), as seen in Figure 8 and a significant interactive effect between type of service failure and type of evaluation on tip (F(2,366) = 3,000, p=.051, η²p= 0,016). Type of evaluation did not have a significant direct effect on tip (*F*(1,366) = 0,403, p>.05, η²p = 0,001), therefore H1b was not supported. When considering both types of evaluation, Pairwise Comparisons indicated a significant difference (p<.05) in the means of tip between the morality (M = 6,96) failure condition and competence (M = 5,52) failure condition. Results also showed a significant difference (p<.05) between the morality failure condition and the warmth (M = 5,40) failure condition. No difference in the means of tip (p>.05) was found between the warmth failure condition and competence failure condition.

[18] Figure 1. Standardized regression coefficients for the relationship between type of evaluation and rating mediated by forgiveness. The standardize regression coefficient for type of evaluation and rating, controlling for forgiveness, is in the parenthesis. **p*<.05

Figure 8 – Means of Tip Between Type of Failure Conditions



Source: Research data (2018)

Table 4 – Results of ANCOVA – Main and Interactive Effects Between Type of Failure and Type of Evaluation on Tip

Tests of Between-Subjects Effects

Dependent Variable: Tip

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	3079.468[a]	6	513.245	16.230	.000	.210
Intercept	6817.097	1	6817.097	215.567	.000	.371
Severity_failure	2606.711	1	2606.711	82.428	.000	.184
Type_failure	192.076	2	96.038	3.037	.049	.016
Type_eval	12.745	1	12.745	.403	.526	.001
Type_failure * Type_eval	189.739	2	94.870	3.000	.051	.016
Error	11574.394	366	31.624			
Total	28185.260	373				
Corrected Total	14653.862	372				

a. R Squared = .210 (Adjusted R Squared = .197)

Source: Research data (2018)

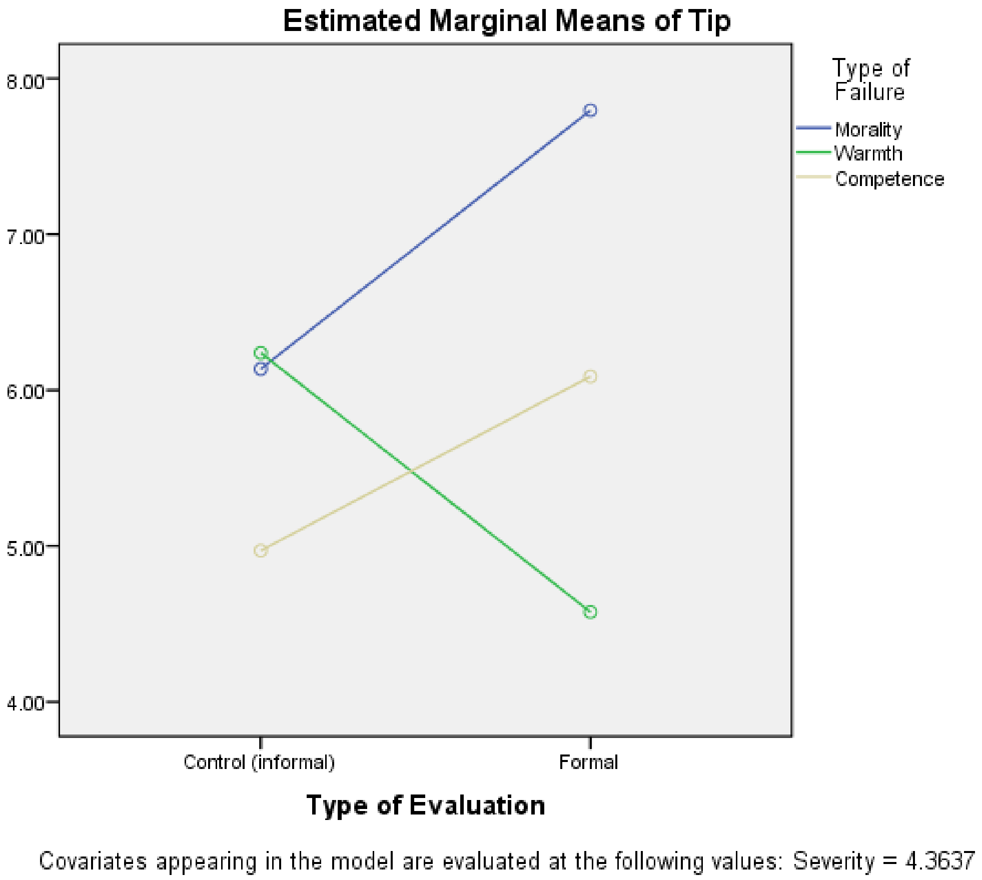
Spotlight Analysis of interactive effect

To further investigate the interactive effect of type of evaluation and type of failure on tip, a Spotlight analysis was conducted using the PROCESS macro. Type of service failure conditions were coded -1 = Morality, 0 = Warmth and 1 = Competence and type of evaluation -1 = control and 1 = formal. When considering only the control (informal) condition of type of evaluation, results indicated no significant difference (p>.05) in the means of tip between morality (M = 6,13), competence (M = 4,96) and warmth (M = 6,23) types of service failure. When considering only the formal condition of type of evaluation, Spotlight analysis indicated a significant difference (*b* = -3,21 *se* = 1,02, *t* = -3,13, *p* < .05, with confidence intervals (CI) between -5,23 and -1,20) between morality (M = 7,79) and warmth (M = 4,57) types of failure. No difference in the means of tip was found between morality and competence (M = 6,08) types of service failure (p>.05) and competence and warmth types of failure (p>.05).

The interaction can be better visualized in Figure 9. Spotlight also indicated no significant differences in the means of tip in any type of service failure between control and formal types of evaluation (p>.05).

When considering each type of failure individually, no difference in the means of tip was found between the two types of evaluation (p>.05), offering support to H4a but not H4b. The mean of tip in the competence failure condition, in the formal type of evaluation condition was M = 5,38, and M = 4,55 in the control (informal) type of evaluation. In the morality failure condition, the mean of tip was M = 8,11 in the formal condition, and M = 6,15 in the control (informal) condition. Finally, in the warmth condition the mean of tip was, in the formal type of evaluation M = 5,27 and M = 6,41 in the control (informal) condition. The means can be visualized in Table 5.

Figure 9 – Graphic Representation of the Means of Tip Between Conditions



Source: Research data (2018)

Table 5 – Means of Tip Between Type of Evaluation and Type of Failure Conditions

Type of Failure	Type of Evaluation	Mean of Tip
Morality	Informal	6,13
Warmth	Informal	6,23
Competence	Informal	4,96
Morality	Formal	7,79
Warmth	Formal	4,57
Competence	Formal	6,08

Source: Research data (2018)

Discussion

Results of Study 1 indicated strong main effects of type of service failure and type of evaluation in rating but no interactive effect. The interactive effect did not occur since the difference in the means of rating between types of evaluation was significant for all conditions of type of failure and it followed the same direction. The means of rating in the competence failure condition indeed were not significantly different between type of evaluation conditions as hypothesized in H3a. However, nor were the means different in the morality and warmth conditions of type of failure between types of evaluation, therefore H3b is rejected. 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 H1a. The study also revealed that the mean of rating was significantly higher for the morality condition when compared to the competence and warmth conditions. We found perception of quality compromised to be a mediator for this effect, consistent with H5. Participants perceived the morality failure as affecting the quality of the service significantly less than the competence and warmth failures. In study 1 the morality failure affected the passenger less directly than the other two failures, which could possibly explain the difference in perception of quality of the service compromised in this condition. This is consistent with Kirmani et al. (2017), who found evidence that when the immoral behavior does not harm the customer directly, competence has a bigger influence on customer's assessment of the service provider.

Forgiveness was found to be a mediator for the effect of type of evaluation on rating, consistent with H2. People were more forgiving of the failure of the service when formally evaluating the driver in the app than when informally evaluating to a friend, lending support to H2. This is consistent with a study by Tsarenko and Tojib (2011), which revealed consumer's emotional and decisional forgiveness has a negative impact in willingness to spread negative word of mouth about the provider, after a service failure. Studies on forgiveness frequently associate other emotions to the act of forgiving, such as shame and guilt (KONSTAM *et al.*, 2001). According to Bridges & Vásquez (2016), the nature of personal experience in services such as Airbnb, leads individuals to be less critic when compared to 'professional service'. Since it is common to experience closer relationships in collaborative services, we propose that individuals may feel anticipation of guilt (i.e. not wanting to harm the provider) for giving a negative feedback when formally evaluating the provider. In Study 2, we investigate anticipation of guilt as another mediator for the effect of type of evaluation on feedback.

Study 1 also revealed a direct effect of type of failure on tip and an interactive effect between type of service failure and type of evaluation on tip. The mean of tip was not significantly different between type of evaluation conditions, therefore H1b was not supported. In H4a, we proposed that for a competence service failure, the mean of tip in formal and informal types of evaluation would not be significantly different, while in H4b we proposed that for morality and warmth failures, the amount of tip would be higher in the formal (vs informal system). Our study revealed support to H4a but not H4b, since the means of tip did not vary between types of evaluation condition for any type of failure. When formally evaluating the provider, when a morality service failure occurs, means of tip would higher when compared to competence or warmth failures. Our results indicated that, only in formal type of evaluation, the morality type of failure condition had a significantly higher mean of tip when compared to a warmth failure but no significant difference in the mean of tip was found between the morality failure condition and the warmth failure condition. Results point to a significant difference in the means of tip between morality and warmth service failures only when the type of evaluation used is formal.

4. THEORETICAL BACKGROUND FOR STUDY 2

Now, we present the theoretical background for Study 2, a review of literature on anticipation of guilt and overall drive score will follow.

4.1 ANTICIPATION OF GUILT

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. According to the authors, consumers are likely to let their behavior be guided by these feelings of anticipatory guilt, leading to an avoidance to engage in unethical behavior. For the authors, guilt is a moral emotion, linked to the welfare of others and the society in general.

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). Vangelisti, Daly and Rudnick (1991) conducted four studies in order to examine how people would elicit guilt in conversations. The research revealed a link between close relationships and anticipation of guilt. Their study revealed guilt is more likely to be elicited the more intimate is the relationship. Miceli (1992) also conducted a study on guilt inducing. The study pointed out that guilt emerges when individuals feel somewhat responsible for an event or for failing to avoid something to occur. The author (p. 81) points that "the sense of guilt plays a crucial role in developing the sense of individual and social responsibility and of moral behavior in general". In the context of charitable donations, Basil *et al.* (2008) found guilt to be a mediator for the effect of empathy on donation intentions, linking feelings of guilt to generosity. In line with this, Konstam *et al.* (2001) found a strong positive relationship between proneness to feel guilty and total forgiveness.

We propose that anticipation of guilt and forgiveness are both mediators for the effect of type of evaluation in rating. This follows the logic of Enright *et al.* (1992), who identified several styles of forgiveness. One of such styles is what the authors called Conditional or Restitution Forgiveness. According to the authors, Conditional or Restitution Forgiveness means that if one feels guilty for withholding forgiveness, then the individual can forgive to be relieved of one's guilt.

Therefore, due to the nature of collaborative services, which involve more personal interactions and because users are aware of the power their rating has to harm the provider, in the occurrence of a service failure, users will show higher levels of anticipation of guilt when evaluating the provider formally (vs. informally).

> H6: In the occurrence of a given service failure, the effect of type of evaluation on feedback will be mediated by anticipation of guilt.

4.2 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. The logic is that people interpret behavior in terms of its causes and that these interpretations play an important role in determining their reactions to the behavior (Kelley and Michela 1980). Weiner (1972) explains stability as the expectancy that the cause of an event will remain stable and not fluctuant over time. According to the author (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.

Folkes (1984) conducted a study relating consumer reactions to product failures to attributional theory. The study confirmed stability to be linked to future expectations. Furthermore, the more stable the perception of the cause the more customers were certain that the product would be “bad” again. Weiner *et al*. (1976) argue that the stability of a cause determines expectancy shifts. “Attribution theorists postulate that future behavior is in part determined by the perceived causes of past events” (WEINER *et al*., 1976 p.55). 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. For the authors, success should generate anticipation of future success while failure would likely lead to the belief that subsequent failures will occur in the future. 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 literature points out, it seems that the previous ratings/reviews, serve as a cue to past behavior to users/customers (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 the failure action is recurrent and stable and feel less obligate 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).

In line with this, we propose the following hypothesis:

H7: 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.

5. STUDY 2

In Study 2 we investigate overall driver score as another possible boundary condition for the effect of type of evaluation on feedback. Further, we test anticipation of guilt as another possible mediator for the effect of type of evaluation on feedback.

Design and Participants

Study 2 was a 3 (type of failure: morality, 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 total sample included 543 individuals.

In Study 2, we presented the same videos for type of failure manipulation as in Study 1, followed by images showing the current score of the driver and finally a text describing the type of evaluation as used in Study 1. A questionnaire which took around 12 minutes to be completed followed. The qualifications workers had to meet in order to be eligible to participate in the study were the same as in Study 1.

Sample size was, as in study 1, determined using G*Power software (FAUL et al., 2007). We set analysis of covariance (ANCOVA) with fixed effects, main effects and interactions as the statistical test of choice. Expected effect size was set to medium (0.25) and confidence level at 95% for the sample size. The number of degrees of freedom was set to 2, the number of groups was set to 12 and the number of covariates to 1. Within these parameters, the software determined a total sample size of a minimum of 251 subjects.

From our initial sample of 543 individuals, 20 were excluded from analysis due to incomplete answers or missing data. Our final sample consisted of 522 participants of which 256 were assigned to the formal (in the system) evaluation condition and 266 were assigned to the control (informal) evaluation condition. The competence condition of type of failure had 178 participants assigned to it, the morality condition had 160 and 184 were assigned to the warmth condition. The driver high score condition had 259 participants assigned to it and the driver low score condition had 263 participants assigned to it. The mean age of participants was 34,5 years and 51,9% were males. Of the total sample, 17 (3,3%) participants declared to be handicapped.

Procedure

In Study 2, the data collection instrument was also created on Qualtrics software and made available to MTurk participants. Participants were first presented with a consent form to which they had to agree in order to continue with the research. A short introduction to the study was then presented. 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 same as in Study 1.

Following the video, an image introduced the manipulation for driver score. Participants were asked to pay close attention to the image. The image depicted a screenshot of a smartphone showing the app after the passenger requested the ride (available in appendix D). In the image there was a white square containing information about the driver (photo, name and score) and the car (manufacturer, model, color), similar to real life on-demand ride apps. In the driver high score condition a score of 4.98 was shown under the driver's name with a star next to it (most on-demand transportation apps have some symbol of achievement next to high scores of drivers, such as stars or trophies). In the driver low score condition, a score of 3.29 was shown and in the control condition there was no information about the driver score. Only one of the images was randomly presented to the participants.

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. Also, we made clear that drivers and passengers who sustained a score lower than 4 over a certain period, could get banned from using the service. This manipulation was randomized.

After the manipulations were introduced, participants were requested to complete a questionnaire (Appendix C) which collected data and included measurements for the dependent variables, manipulation checks, covariables and demographic variables.

Measures

The measures for the dependent variables, manipulation and attention checks (except for overall driver score), control variables (except age), forgiveness and perception of quality were the same used in Study 1.

Manipulation and Attention Checks

Overall Driver Score

The manipulation for 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”

We also included another attention check: “What was the current rating (score) of the driver?”, the possible answers were: “above 4.7”, “below 4.7” and “do not remember”.

Mediators

Anticipation of guilt

The items for anticipation of guilt were adapted from Basil, Ridgway and Basil (2006). The two items were rated in 7-point scales. One example is: “I gave the driver a rating different than I thought she really deserved because I would feel anticipation of guilty if the driver suffered negative consequences due to my rating – 1 = Strongly Disagree, 7 = Strongly Agree”.

Control

Age

In Study 2, we changed the form in which age was measured, in order to have a more accurate measure. We used the same question as in Study 1, however changed the scale to a continuous one ranging from 18 to 100.

Validity of Scales

An exploratory factor analysis using Varimax rotation method was conducted with the 4 items measuring competence, 4 items measuring morality, 4 items measuring warmth, 3 items measuring severity of the failure and 3 items measuring forgiveness. The analysis revealed the scales exhibit satisfactory factorial structure. The items measuring competence showed factor loadings between .914 and .796. The items measuring morality showed factor loadings between .906 and .751. The items measuring the warmth factor had factorial loadings between .943 and .842. Perceived severity of the failure had loadings ranging from -.826 to -.331. Forgiveness had loadings between .916 and .863. The Kaiser-Meyer-Olkin was .863 which is above the threshold of .6 (Kaiser, 1974) and the Bartlett's Test of Sphericity reached statistical significance (p<.001).

Cronbach's Alpha was also used to measure reliability of the scales. The statistical analysis of the scales yielded, in general, good alphas[19]. The scale measuring competence consisted of 4 items ($\alpha = .91$); the scale for morality also consisted of 4 items ($\alpha = .89$); the scale measuring warmth ($\alpha = .95$) was composed of 4 items as well. Perceived severity of the failure was measured with a 3-item scale ($\alpha = .52$). Forgiveness was also measured with a 3-item scale ($\alpha = .93$). Anticipation of guilt was measured with a 2-item scale, therefore we used correlation analysis to assess reliability. Results demonstrated a positive correlation between the two items (r = 0,912, *N* = 522, p<.001).

Pretests

We conducted three pretests to verify if the manipulation for driver score was effective and identify which aspects of that manipulation needed refinement.

In the first pretest, we had a total sample of 122 participants, 64,8% of them were male with a mean age of 33,5 years. In pretest 1, we were mainly interested in testing the manipulation of driver score. We used screenshot of a smartphone showing the app after the passenger requested

[19] All scales are 7-point Bipolar Likert type.

the ride. In the image there was a white square with information about the driver (photo, name, score, number of trips completed) and the car (manufacturer, model, color). Initially, we tested a driver score of 4.98 for the high score condition and 4.51 for the low score condition. In the high score condition, there was a star next to the score similar to real on-demand transportation services which put a star next to high scores (average close to 5).

The manipulation check for driver score consisted of the question "Do you consider the current rating (score) of the driver to be", with three options answer: "high", "low" and "average". A crosstabulation test demonstrated the manipulation in the first pretest was not effective. In the low driver score condition, only 8% of respondents considered the score low (27,8% considered it average) while in the high driver score condition, 65% considered the score high (29,5% considered it average).

We also included an attention check, which consisted of the question "what was the current rating (score) of the driver?", the possible answers were: "above 4.7", "below 4.7" and "do not remember". In the high driver score condition, 68% of participants answered the attention check correctly and in the low driver score condition only 47,5% of participants answered correctly. Perception of realism was tested as in Study 1 and had a mean of 5,84, higher than the mean of the scale (<4), and similar across conditions (p>.05).

In the second pretest, we had a sample of 161 valid observations. Among the participants, 62,1% were male and the mean age was 33,1 years. In the second pretest, we attempted to make the high and low driver score conditions more distinct. We maintained the manipulation of high driver score as used in the first pretest but lowered the driver score to 3,81. We used the same manipulation check for driver score as we used in the first pretest. However, since we lowered the score in the low driver score condition, we changed the possible answers for the attention check to: "above 4.5", "below 4.5" and "do not remember".

A crosstabulation test showed the manipulation in the second pretest was also not effective. In the low driver score condition, 30,8% of respondents considered the score low and 29,6 considered it to be average. In the high driver score condition, 75% considered the score high and 22,5% considered it to be average. In the attention check, a crosstabulation test showed that 76,25%

of participants answered the attention check for high driver score correctly and in the low driver score condition 56,79% of participants answered correctly. Perception of realism had a mean of 5,67, higher than the mean of the scale (<4), and similar across conditions (p>.05).

In the second pretest we also included a manipulation check for type of evaluation. This check was the same as in Study 1 and crosstabulation was used to test its effectiveness. The formal (in the system) condition had 69,7% participants answering the manipulation check correctly and the control (informal) condition 73,9%.

In the third and final pretest, we had a valid sample of 101 participants, among which 54,5% were male and the mean age was 34 years. In this pretest, we maintained the high driver score condition unaltered and made further adjustments to the low driver score condition. We attempted to make the manipulation of low driver score more effective by lowering the driver score in that condition even further, to 3.34. In this pretest we included a name for the fictious on-demand transportation service we were presenting to participants. Since we lowered the score once more in the low driver score condition, we made new changes to the possible answers for the attention check: “above 4.2”, “below 4.2” and “do not remember”.

A crosstabulation test revealed that in low driver score condition, 61,22% of respondents considered the score low and 24,4% considered it average. In the high driver score condition 80% considered the score high and 13% considered it average. Therefore, we conclude manipulations were effective. In the attention check, a crosstabulation test showed that 84% of participants answered the attention check for high driver score correctly and in the low driver score condition 80,3% of participants answered it correctly.

In the third pretest, the type of evaluation check was also tested using crosstabulation. The test revealed 68,6% of participants in the control (informal) condition answered the manipulation check for type of failure correctly and 94% in the formal (in the system) answered it correctly.

Data Analysis and Assumptions for Statistical Tests

The data analysis procedures used were the same as in study 1. Independency of observations was achieved through the random distribution and a between-subjects design (each participant was allocated to only one experimental condition). Through an analysis of frequencies, it was possible to determine there were 15 missing values in the database of Study 2.

Similar to study 1, the Z-test only detected outliers in the duration variable (amount of time respondents took to complete the study). However, as in study 1, we decided not to exclude from the sample outliers due to extreme duration values. Therefore, statistical analysis was conducted using all 567 valid observations from the original sample.

The equality of variance-covariance matrices or homoscedasticity was verified using Levene test. An ANOVA test revealed that the dependent variable rating did not vary across different levels of the independent variables (type of failure, type of evaluation and driver score) under study. The Levene test (*F*(11, 510) = 1.254, *p*>.05) indicated the homoscedasticity of the depended variable rating. Another ANOVA was conducted to verify the homoscedasticity of the dependent variable tip. The Levene test showed that this dependent variable is homogeneous (F(11, 510) = 1.472, p>.05), therefore indicating there is no difference in the variance of the variable across different levels of the independent variables.

The normality of distribution of the dependent variables was verified using skewness and kurtosis values. For the rating variable skewness value (-0,31) was acceptable for normal distribution. The kurtosis value (-0,96) for rating was also acceptable for normal distribution. The results for the Shapiro-Wilk test for rating indicated that the null hypothesis (H0) could not be rejected, since the probability was lower than .05 (S-W = 0,951, p<.001). For the tip variable, skewness (0,54) and kurtosis (-0,88) values were acceptable for normal distribution. The results for the Shapiro-Wilk test for rating indicated that the null hypothesis (H0) could not be rejected, since the probability was lower than .05 (S-W = 0,889, p<.001).

Main Study Results

The data collected for Study 2 was analyzed according to the procedures described previously. Results for manipulation checks and tests for main and interactive effects are presented next.

Manipulations and Attention Checks

Type of Service Failure

The first manipulation check analysis was conducted via One-way ANOVA. The analysis showed a statistically significant difference between groups: competence (F(2,519) = 44.878, p<.001), morality (F(2,519) = 64.376, p<.001), and warmth (F(2,519) = 305.483, p<.001).

Post Hoc test Tukey HSD showed that within the competence condition there was a significant difference for perception of type of failure between competence (M = 3,49) and warmth (M = 4,39) (p<.001) and competence and morality (M = 4,88) (p<.05). Within the morality condition there was a significant difference in perception of type of failure between morality (M = 3,60) and competence (M = 5,23) (p<.001), and morality and warmth (M = 5,21) (p<.001) failures. Within the warmth condition there was a significant difference for perception of type of failure between warmth (M = 2,45) and competence (M = 5,41) (p<.001) and warmth and morality (M = 5,33) (p<.001). Therefore, allowing us to conclude that the participants perceived the conditions correctly and the manipulation for type of failure was successful.

The attention check for type of service failure showed that most respondents answered the manipulation check correctly: competence (93,2%), morality (79,3%) and warmth (87,5%). A Pearson Chi-Square test showed a statistically significant association between the type of failure and the attention check (x^2(6, *N* = 522) = 720.556, p<.001).

Type of Evaluation

In the manipulation check for type of failure participants were presented with the same question as in study 1: the sentence “You evaluated the driver” was shown to which participants had three choices for an answer: “in the service’s app”, “to a friend” and “do not remember”. A Crosstabs test showed the manipulation check for informal vs formal type of evaluation was effective for the informal (control) condition (70,3% correct answers) and the formal condition (82,8% correct answers). The Pearson Chi-square test yielded a statistically significant relationship between the manipulated variable and the manipulation check (x^2(2, *N* = 522) = 182.637, p<.001).

Driver Score

The manipulation check for driver score consisted of the question “Do you consider the current rating (score) of the driver to be”, with three options for an answer: “high”, “low” and “average”. A crosstabs test showed that in the low driver score condition 46,2% of participants considered the driver score to be low and 36,5% considered it average. In the high driver score condition 79,8% of participants considered the driver score high and 13,5% considered it average. A Pearson Chi-Square test showed a statistically significant association between the type of driver score and the manipulation check ($x^2(2, N = 481) = 194.009$, p<.001). The attention check for driver score consisted in the question “The current rating (score) of the driver was” to which there were three alternatives for answer presented to participants: “above 4.2”, “below 4.2”, and “there was no information”. A Crosstabs test showed the manipulation for high (86,1% correct answers) vs low (75,2% correct answers) driver score was successful. A Pearson Chi-Square test showed a statistically significant association between the driver score and the attention check ($x^2(2, N = 522) = 274.635$, p<.001).

Control Variables

We tested the same potential covariables as in Study 1, to control for intervenient effects that could distort the results of the study. These were included as covariables in the ANCOVA test.

In the model where rating was the dependent variable, as in Study 1, the analysis of variance showed severity to have a significant cofounding effect ($F(1, 377) = 146.186$, p<.001) on rating. Other variables potentially affecting the dependent variable rating were tested using ANOVA. Having used similar services before (p>.05), being handicapped (p>.05), age (p>.05), gender (p>.05) and frequency of use (p>.05) were found to not have intervenient effects with the variable under study.

These potential covariates were also tested for the dependent variable tip. Again, severity of the failure was shown to have a significant effect in the model (F(1, 377) = 29.073, p<.001). Among the other covariables tested, gender (p>.05), having used similar services before (p>.05), frequency of use of similar services (p>.05), being handicapped (p>.05) and age (p>.05) had no effect on tip.

In addition to testing control variables, we also tested perception of realism. The same one-item question as in Study 1 was included in the instrument of data collection to assess this variable. The mean for perception of realism was (M = 5,87, SD = 1,16), above the mean of the scale (>4) and not statistically different across conditions (F(2, 510) = 0,361, p>.05), therefore revealing participants believed the situation presented to them was realistic in all conditions.

Hypothesis Tests

To test the hypothesis, again we used the ANCOVA method for statistical analysis. The ANCOVA assess the interaction between type of failure, type of evaluation and driver score as independent variables, rating and tip as dependent variables, and severity of the failure as a covariable.

Rating

Similar to Study 1, ANCOVA was the method of statistical analysis of choice for testing main and interactive effects. We included severity as covariable since statistical analysis showed a significant effect of this variable on rating between conditions. In Study 2 we tested type of failure, type of evaluation and driver score main effects on rating and tip and its interactive effects on these two dependent variables (rating and tip).

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 (Table 6) which showed significant main effects of type of service failure (*F*(2, 509) = 18.191, p<.001, η^2p = 0,067), type of evaluation (*F*(1, 509) = 15.692, p<.001, η^2p = 0,030) and driver score (*F*(1, 509) = 10.487, p<.05, η^2p = 0,020) on rating. Also, the ANCOVA test revealed an interactive effect (Figure 10) of driver score and type of evaluation on rating (F(1, 509) = 4.743, p<.05, η^2p = 0,009).

Table 6 – Results of ANCOVA – Main and Interactive Effects of Type of Failure, Type of Evaluation and Driver Score on Rating

Tests of Between-Subjects Effects

Dependent Variable: Rating

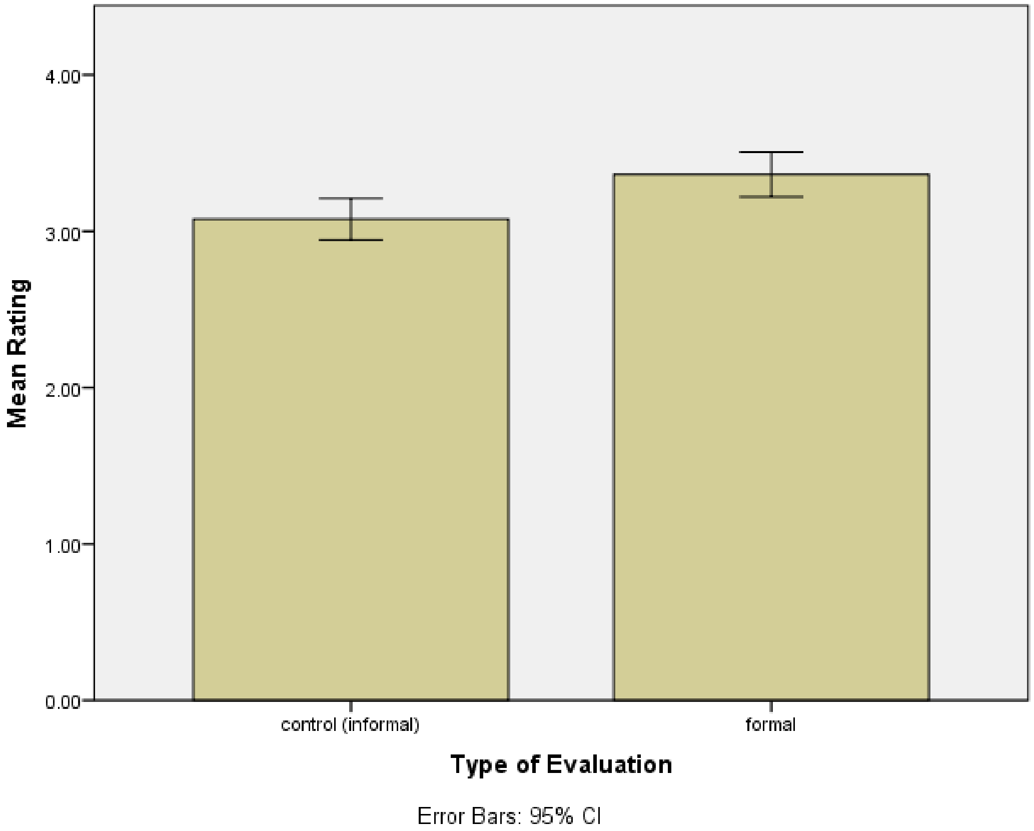
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	222.170[a]	12	18.514	20.730	.000	.328
Intercept	1040.295	1	1040.295	1164.779	.000	.696
Severity	163.998	1	163.998	183.622	.000	.265
Driver_Rating	9.366	1	9.366	10.487	.001	.020
Type_of_Failure	32.493	2	16.247	18.191	.000	.067
Type_of_Evaluation	14.015	1	14.015	15.692	.000	.030
Driver_Rating * Type_of_Failure	.093	2	.046	.052	.949	.000
Driver_Rating * Type_of_Evaluation	4.236	1	4.236	4.743	.030	.009
Type_of_Failure * Type_of_Evaluation	2.917	2	1.458	1.633	.196	.006
Driver_Rating * Type_of_Failure * Type_of_Evaluation	.800	2	.400	.448	.639	.002
Error	454.601	509	.893			
Total	6080.450	522				
Corrected Total	676.771	521				

a. R Squared = .328 (Adjusted R Squared = .312)

Source: Research data (2018)

The ANCOVA results showed that the means of rating were significantly different higher in the formal (M = 3,40) condition of type of evaluation than in the control (informal) (M = 3,07) condition (Figure 10), therefore lending support to H1a.

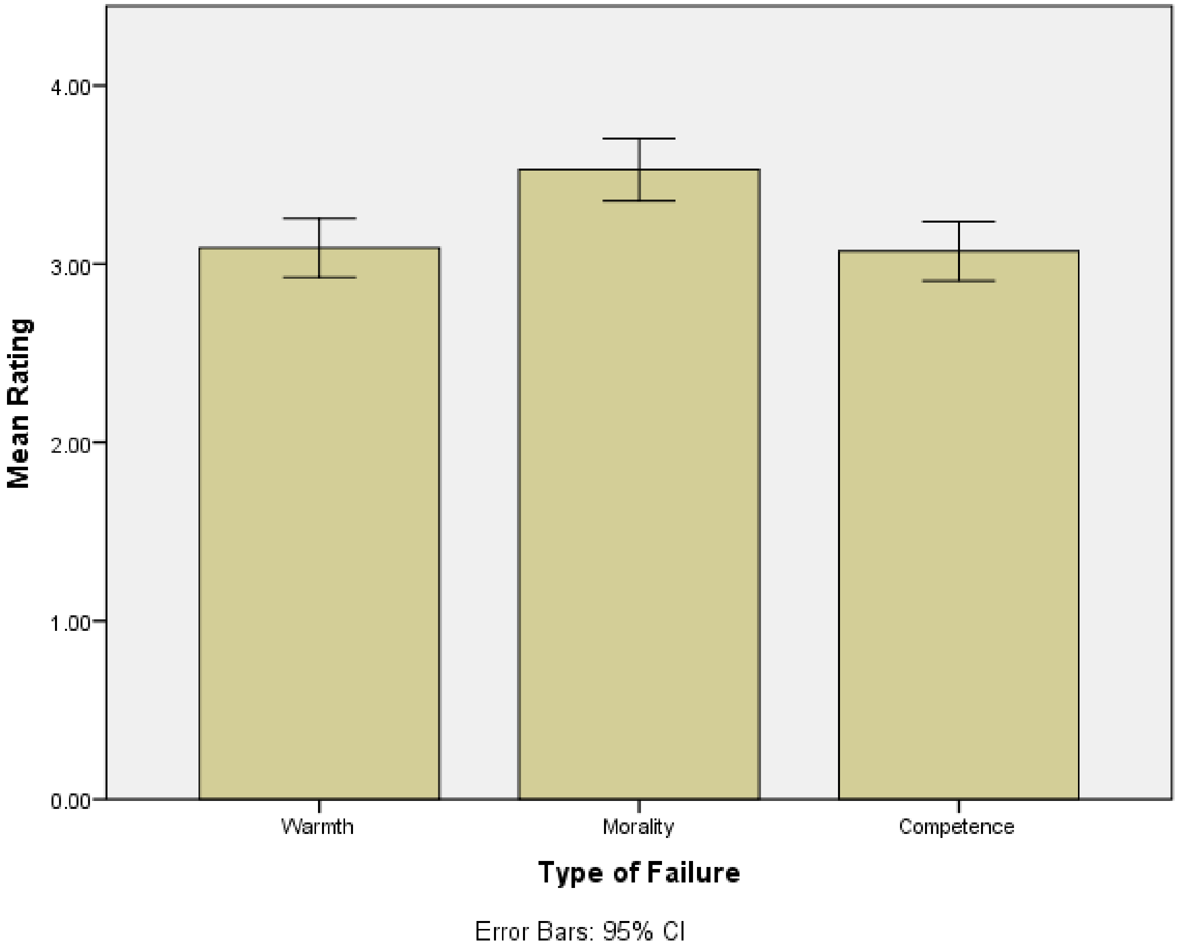
Figure 10 – Means of Rating Between Type of Evaluation Conditions



Source: Research data (2018)

The ANCOVA test revealed that when considering both types of evaluation and driver score conditions, there is a significant difference (p<.001) in the means of rating between the morality (M = 3,60) and warmth (M = 3,03) service failures. Similarly, morality and competence (M = 3,08) failures had a significant difference (p<.001) in the means of rating between conditions. Results showed no significant difference (p>.05) in the means of rating between the competence and warmth conditions. Results are depicted in Figure 11.

Figure 11 – Means of Rating Between Type of Failure Conditions

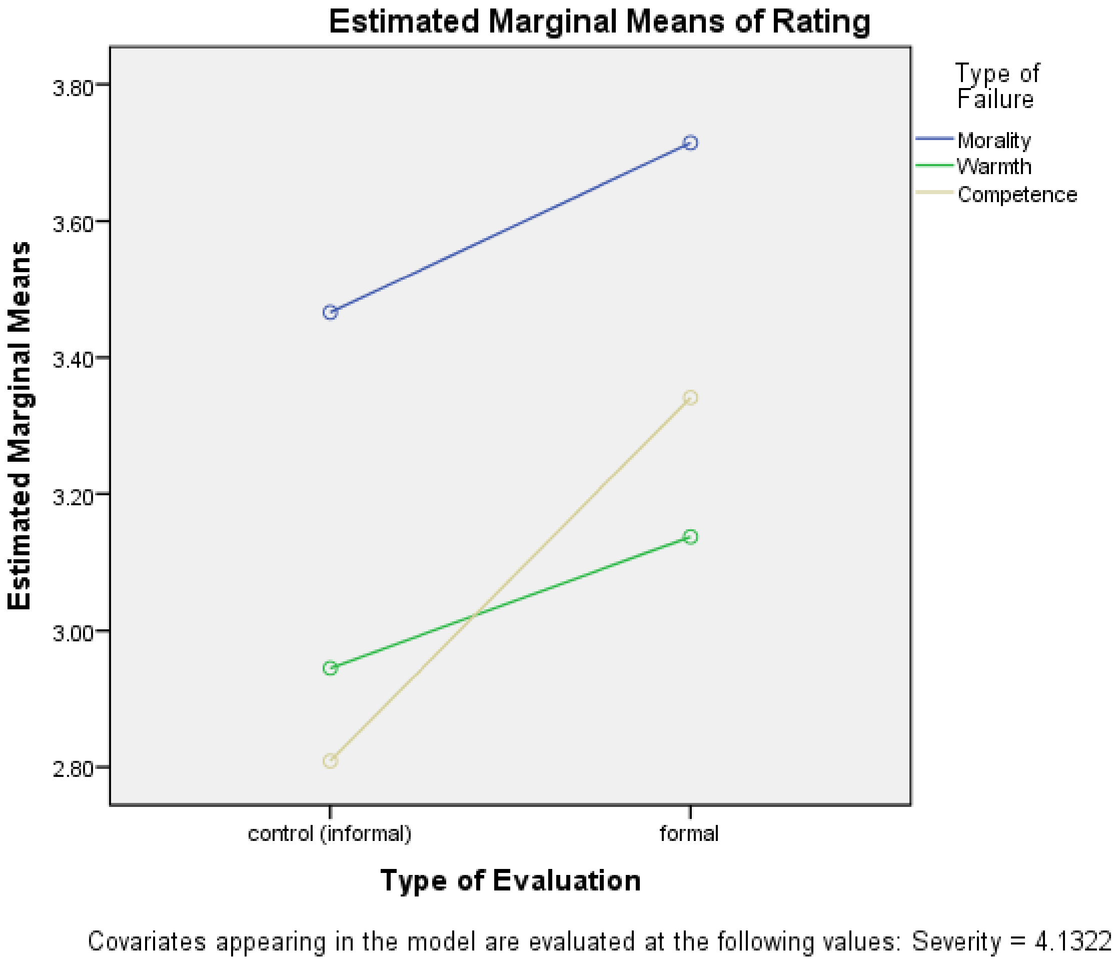


Source: Research data (2018)

Spotlight analysis indicated that when considering each type of failure individually, only in the competence failure condition a statistically significant difference in the means of rating was found (*b* = 0,27 *se* = 0,08, *t* = 3,29, *p*<.05, with confidence intervals (CI) between 0,11 and 0,43) between the formal (M = 3,35) and the control (M = 2,80) conditions of type of evaluation. No difference was found (*b* = 0,12 *se* = 0,08, *t* = 1,44, *p*>.05, with confidence intervals (CI) between -0,04 and 0,30) in the morality failure condition between the formal (M = 3,65) and the control (M = 3,39) type of evaluation conditions. When considering only the warmth failure, results also showed no statistically significant difference (*b* = 0,01 *se* = 0,08, *t* = 0,14, *p*>.05, with confidence intervals (CI) between -0,14 and 0,17) in the means of rating between the formal (M =

3,10) and the control (M = 3,07) type of evaluation conditions. Therefore, H3a and H3b were not supported.

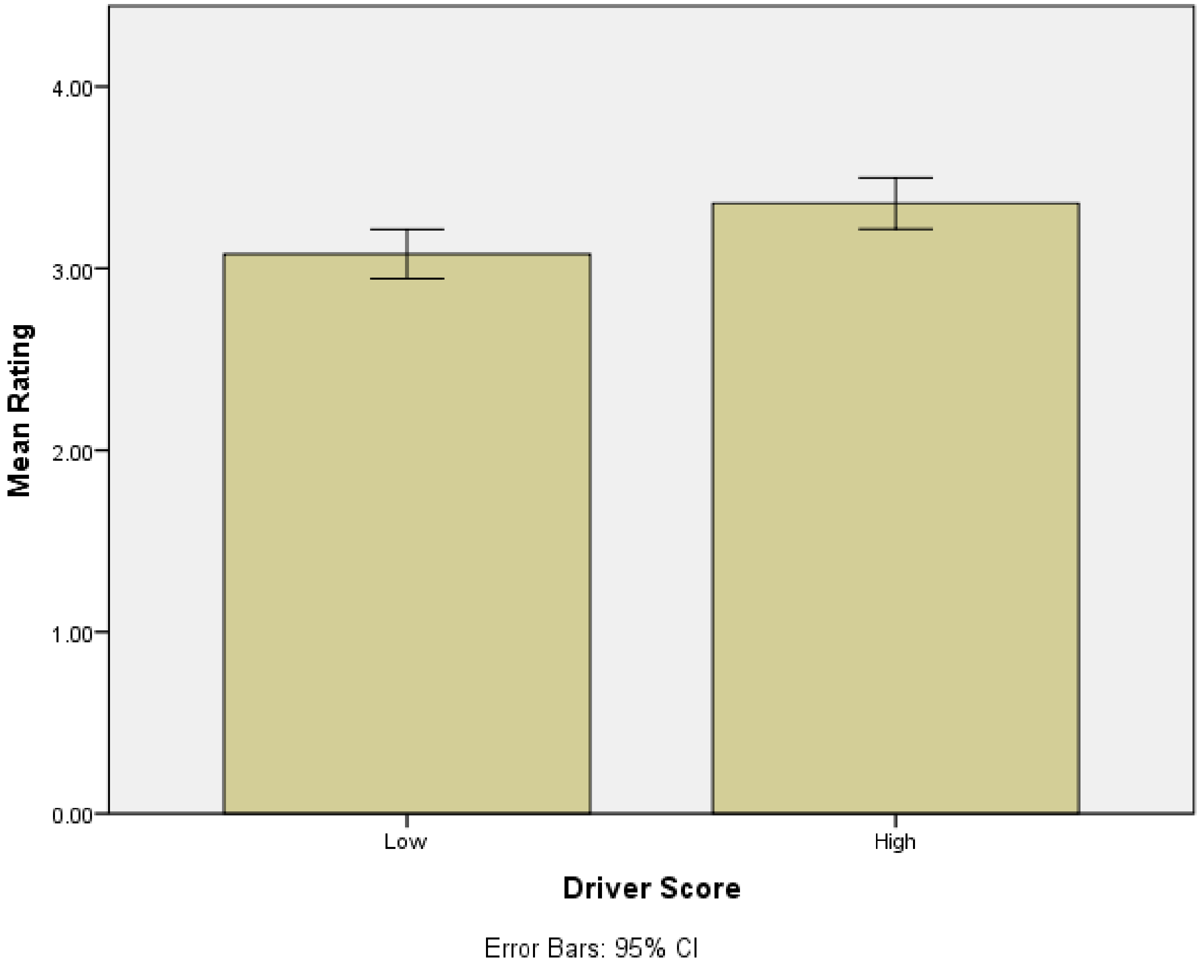
Figure 12 – Graphic Representation of the Means of Rating Between Type of Failure and Type of Evaluation Conditions



Source: Research data (2018)

The ANCOVA results also revealed that driver score had a main effect on rating. In the high driver score condition (M = 3,37), means of rating were considerably higher than in the low driver score condition (M = 3,10), as depicted in Figure 13.

Figure 13 – Means of Rating Between Driver Score Conditions



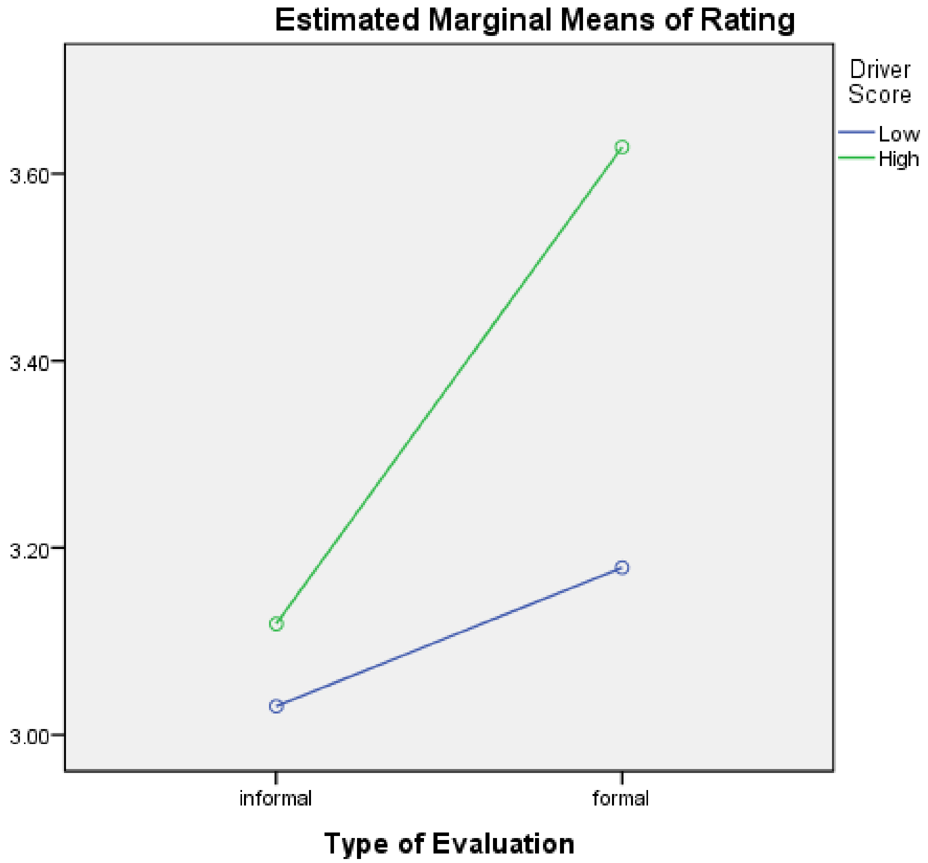
Source: Research data (2018)

Spotlight analysis of interactive effect

We conducted Spotlight analyses to investigate the interactive of type of evaluation and driver score on rating. Driver score was coded -1 = low, 1 = high and type of evaluation was coded 1 = formal, -1 = control (informal). We used perceived severity of the failure as a covariable in the model. The interaction is depicted in Figure 14.

Spotlight analyses revealed an interactive effect occurs when the driver score is high, between the formal (M = 3,60) and control (M = 3,10) type of evaluation conditions (*b* = 0,24 *se* = 0,06, *t* = 4,09, *p* <.001, with confidence intervals (CI) between 0,12 and 0,36). Also, the analysis revealed that a significant effect exists between high (M =3,60) and low (M = 3,18) driver score conditions in the formal type of evaluation condition (*b* = 0,21 *se* = 0,06, *t* = 3,48, *p*<.05, with confidence intervals (CI) between 0,09 and 0,33). The means for each condition can be better visualized in Table 7. This result is consistent with H7, which proposes 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 in the low driver score condition means of rating would remain unaltered between types of evaluation conditions.

Figure 14 – Graphic Representation of the Means of Rating Between Driver Score and Type of Evaluation Conditions



Source: Research data (2018)

Table 7 – Means of Rating Between Type of Evaluation and Driver Score Conditions

Type of Evaluation	Driver Score	Mean of Rating
Informal	Low	2,93
Formal	Low	3,08
Informal	High	3,03
Formal	High	3,51

Source: Research data (2018)

Mediators and Moderators

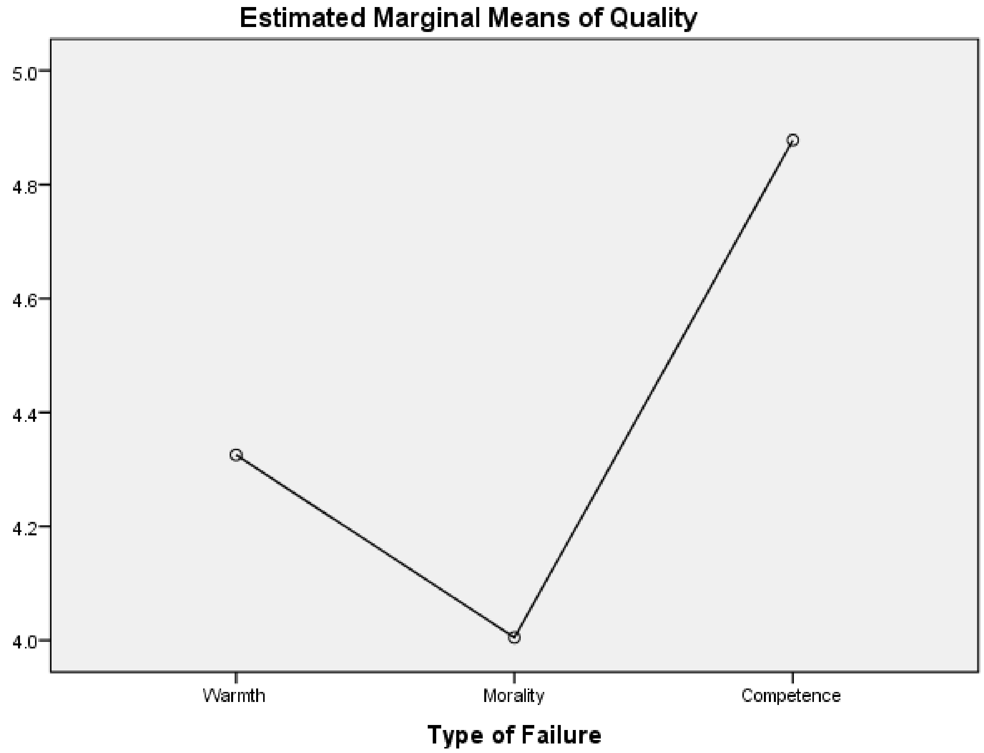
Perceived Quality Mediation

Confirming the results of study 1, regression analysis using the PROCESS macro revealed perceived compromised quality is a mediator for the effect of type of service failure on rating. We included perceived severity of the failure as covariable in the model.

A significant effect of type of service failure on perceived compromised quality was found (*b* = 0,27 *se* = 0,07, *t* = 3,46, *p* < .05, with confidence intervals (CI) between 0,11 and 0,42). Results also revealed a significant effect of perceived compromised quality on rating (*b* = -0,06 *se* = 0,02, *t* = -2,17, *p* < .05, with confidence intervals (CI) between -0,12 and -0,006). This result offers support to H5, which predicted perceived quality would mediate the effect of type of failure on rating. Whereas in study 1 perceived compromised quality was found to be only a partial mediator for this effect, in study 2 the mediation was full, with the direct effect of type of service failure on rating becoming non-significant once the mediator was introduced (*b* = 0,03 *se* = 0,05, *t* = 0,63, *p* > .05, with confidence intervals (CI) between -0,07 and 0,13). The explained variance of the model is 0,24.

An analysis of variance, controlling for perceived severity of the failure, revealed a significant difference in the means of perceived compromised quality between all type of failure conditions (p<.05). Morality failure was perceived as the failure compromising the quality of the service the least (M = 4,00), followed by warmth (M = 4,32) and competence (M = 4,87), which was perceived as the failure affecting the quality of the service the most. The means of perceived compromised quality are displayed in Figure 15.

Figure 15 – Graphic Representation of the Means of Perceived Quality Between Type of Failure Conditions



Source: Research data (2018)

Anticipation of Guilt and Forgiveness Mediation

In Study 2, in addition to forgiveness, we included anticipation of guilt as a possible mediator for the effect of type of evaluation on rating. We conducted a mediation analyses controlling for perceived severity of the failure to investigate this effect.

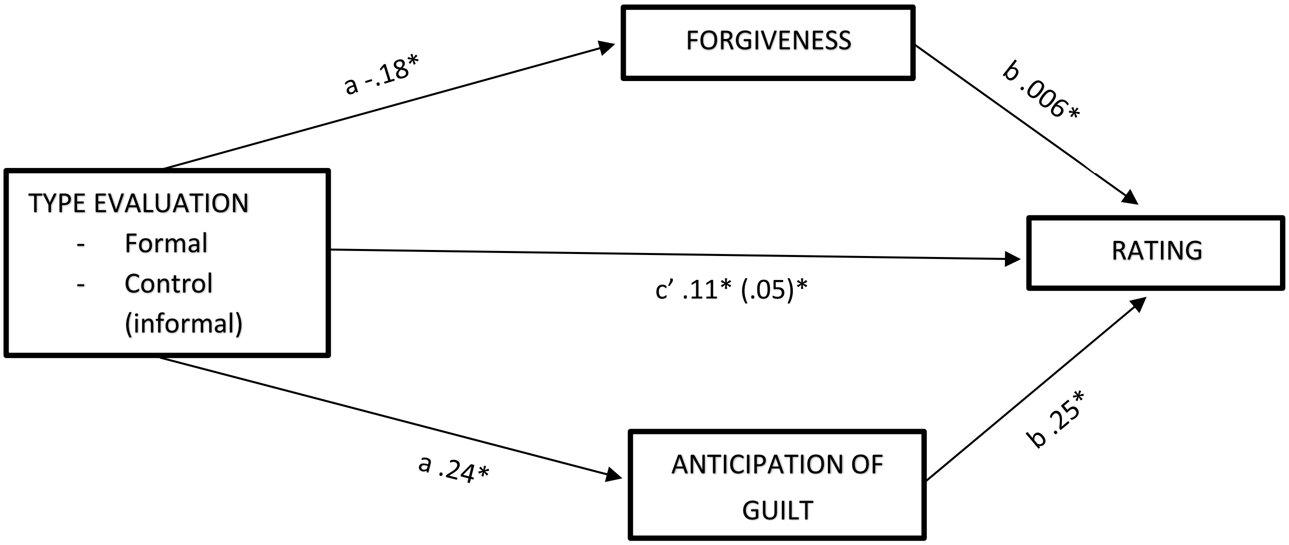
In this model, we tested if forgiveness and anticipation of guilt are parallel mediators for the effect of type of evaluation on rating. Results show a significant path from type of evaluation to forgiveness (*b* = -0,18, *se* = 0,07, *t* = -2,56, *p* < .05, with confidence intervals (CI) between -0,32 and -0,04) and forgiveness to rating (*b* = 0,06, *se* = 0,02, *t* = 2,23, *p* < .05, with confidence intervals (CI) between 0,007 and 0,11). The test also revealed a significant path from type of evaluation to anticipation of guilt (*b* = 0,24, *se* = 0,08, *t* = 2,79, *p* < .05, with confidence intervals (CI) between

0,07 and 0,41) and anticipation of guilt to rating (*b* = 0,25, *se* = 0,01, *t* = 13,30 *p* < .001, with confidence intervals (CI) between 0,22 and 0,29).

However, a direct effect of type of evaluation on rating remained after forgiveness and anticipation of guilt were included as mediators (*b* = 0,11, *se* = 0,03, *t* = 3,13, *p* < .05, with confidence intervals (CI) between 0,04 and 0,19). The mediation analysis revealed a partial parallel mediation of forgiveness and anticipation of guilt in the effect of type of evaluation on rating, consistent with H2 and H6. When including both mediators, the explained variance of the model is 0,44. The relationship between these variables is depicted in the theoretical model in Figure 16.

An ANOVA test revealed that in the formal condition (M = 4,66) of type of evaluation, means of forgiveness were significantly (F(1, 519) = 400.392, *p*<.05) higher than the means in the control (informal) condition (M = 4,17). Similarly, in the formal type of evaluation condition (M = 3,37) the means of anticipation of guilt were significantly (F(1, 519) = 7.801, *p*<.05) higher than in the control (informal) condition (M = 2,91).

Figure 16 – Theoretical Model of Mediation Between Type of Evaluation and Rating[20]



Source: Research data (2018)

[20] Figure 1. Standardized regression coefficients for the relationship between type of evaluation and rating mediated by forgiveness and guilt. The standardize regression coefficient for type of evaluation and rating, controlling for forgiveness and guilt, is in the parenthesis. **p*>.05

Tip

We conducted an ANCOVA test (Table 8) 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*(2, 509) = 26.088, p<.001, η^2p = 0,093) and driver score on tip (*F*(1, 509) = 4.264, p<.05, η^2p = 0,008). No significant effect between type of evaluation conditions and tip was found (*F*(1, 509) = 0.227, p>.05), therefore H1b was not supported.

Table 8 – Results of ANCOVA – Main and Interactive Effects of Type of Failure, Type of Evaluation and Driver Score on Tip

Tests of Between-Subjects Effects

Dependent Variable: Tip

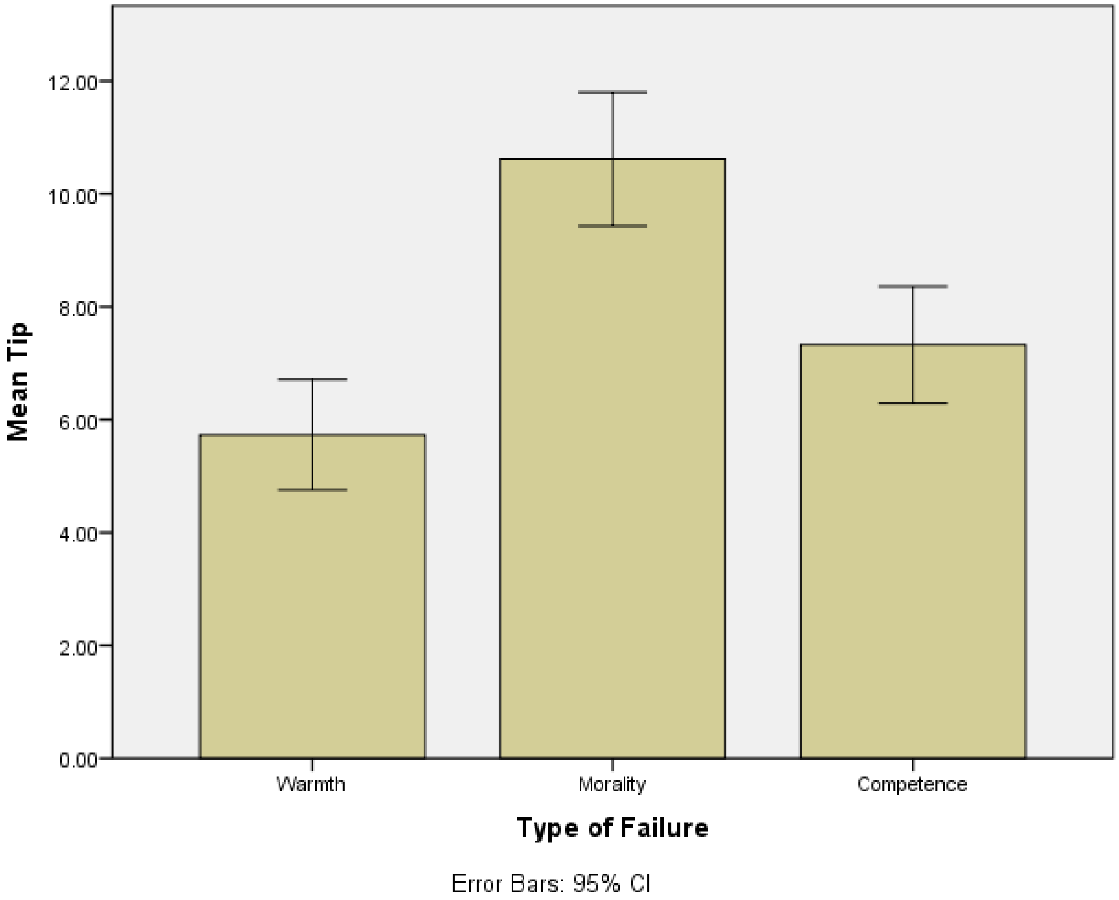
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	4654.983[a]	12	387.915	8.387	.000	.165
Intercept	8253.933	1	8253.933	178.466	.000	.260
Severity	1925.508	1	1925.508	41.633	.000	.076
Type_of_Evaluation	10.493	1	10.493	.227	.634	.000
Driver_Rating	197.225	1	197.225	4.264	.039	.008
Type_of_Failure	2413.134	2	1206.567	26.088	.000	.093
Type_of_Evaluation * Driver_Rating	13.685	1	13.685	.296	.587	.001
Type_of_Evaluation * Type_of_Failure	134.131	2	67.066	1.450	.236	.006
Driver_Rating * Type_of_Failure	10.973	2	5.486	.119	.888	.000
Type_of_Evaluation * Driver_Rating * Type_of_Failure	126.012	2	63.006	1.362	.257	.005
Error	23540.879	509	46.249			
Total	59730.100	522				
Corrected Total	28195.863	521				

a. R Squared = .165 (Adjusted R Squared = .145)

Source: Research data (2018)

Pairwise comparisons revealed a significant difference (p<.001) in the means of tip between morality (M = 10,82) and warmth (M = 5,53) conditions of type of failure. Also, between morality and competence (M = 7,35) conditions of type of failure (p<.001) and between competence and warmth conditions of type of failure (p<.05). The difference in the means of tip between type of service failure conditions is represented in Figure 17.

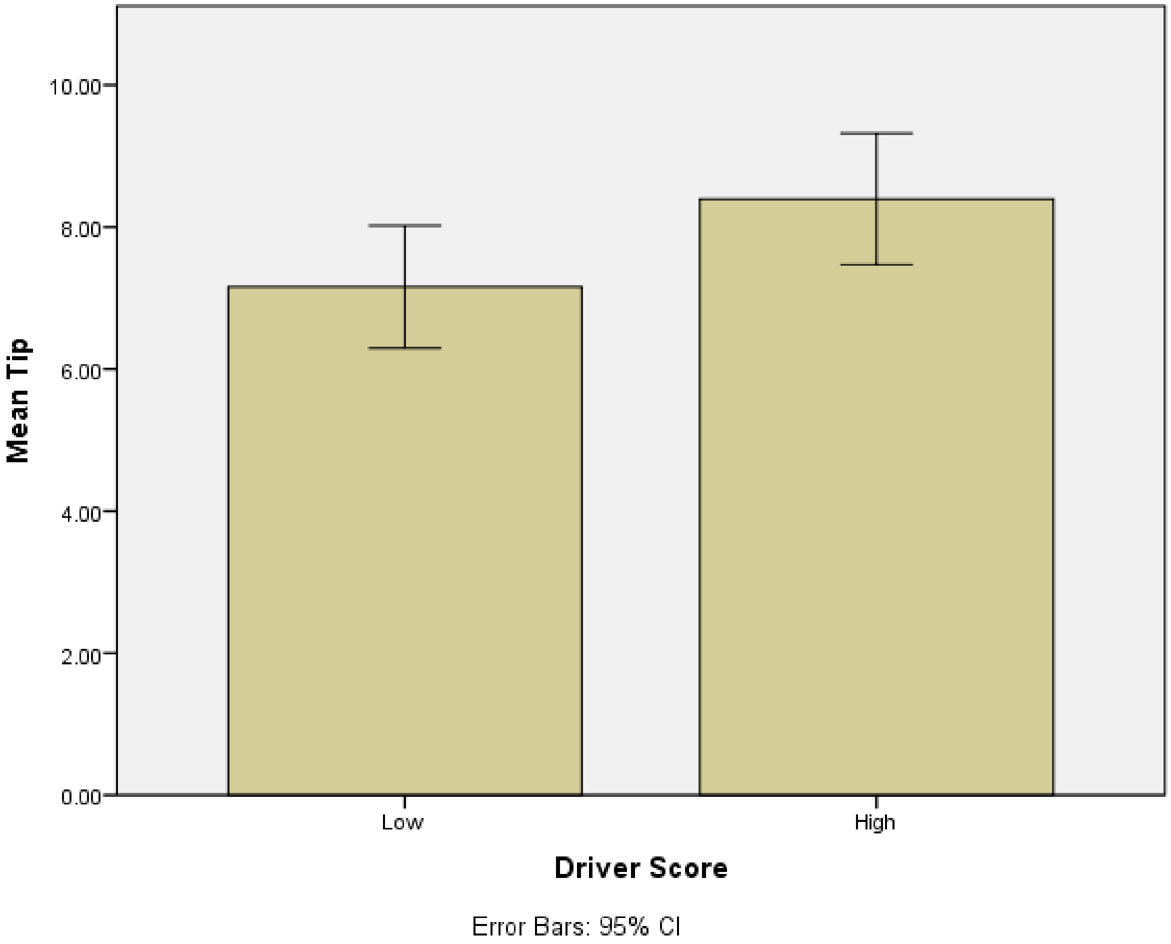
Figure 17 – Means of Tip Between Type of Failure Conditions



Source: Research data (2018)

Pairwise comparisons revealed a significant difference (p<.05) in the means of tip between the high driver score condition (M = 8,52) and the low driver score condition (M = 7,29), as seen in Figure 18.

Figure 18 – Means of Tip Between Driver Score Conditions



Source: Research data (2018)

We conducted Spotlight tests to verify if any difference in the means of tip existed in type of failure conditions between types of evaluation. When considering each type of failure individually, no statistically significant difference was found between type of evaluation conditions (p>.05), therefore H4a was supported but not H4b. When considering only the morality type of failure, the mean of tip was (M = 10,05) in the formal condition of type of evaluation and (M =

11,21) in the control (informal) condition. When considering only the warmth type of failure, the mean of tip was (M = 6,06) in the formal condition of type of evaluation and (M = 5,42) in the control (informal) condition of type of evaluation. When considering only the competence type of failure, the mean of tip was (M = 8,02) in the formal condition of type of evaluation and (M = 6,66) in the control (informal) condition.

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 H1a. Also, results showed that morality failures lead to higher ratings when compared to competence and warmth failures. Perceived compromised quality was found to be a mediator of this effect, lending further support to H5. 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, however 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.

Results of Study 2 confirmed forgiveness as a partial mediator for the effect of type of evaluation in rating, providing further support for H2. Results indicate that anticipation of guilt is also a partial mediator for this effect, in a parallel mediation model, confirming H6.

Similar to Study 1, in Study 2 it was found that morality failures lead to a higher amount of tip than warmth or competence failures. This is in line with Kirmani *et al*. (2017) who found that when hiring a service, customers value ability more than other traits. In the case of the morality service failure, the service outcome was not compromised as in the competence failure, indicating feedback would be the more biased in the morality type of service failure. As in Study 1, the biggest gap in the means of tip was between morality and warmth type of failure conditions. In H4a we proposed that in the occurrence of a competence failure the means of tip would not be significantly different between type of evaluation conditions, this hypothesis was supported. However, the other two failures (morality and warmth) also showed no difference in the means of tip between type of evaluation conditions, inconsistent with H4b, which posited that in the occurrence of these two failures, tips would be higher in the formal condition of type of evaluation than in the control (informal) condition.

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 H7. 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.

Table 9 – Summary of Findings

Hypotheses	Study 1	Study 2
H1a: 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.	**Confirmed**	**Confirmed**
H1b: 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 higher amount of tip.	**Rejected**	**Rejected**
H2: In the occurrence of a given service failure, the effect of type of evaluation on users' rating will be mediated by forgiveness (and not by reciprocity or fear of retaliation).	**Confirmed**	**Confirmed**
H3a: The type of failure will moderate the effect of type of evaluation on users' rating, such that for a competence service failure, ratings in the formal and informal types of evaluation will not be significantly different.	**Confirmed**	**Rejected**
H3b: The type of failure will moderate the effect of type of evaluation on users' rating, such that for morality and warmth failures, ratings will be higher in the formal (vs informal system).	**Rejected**	**Rejected**
H4a Type of failure will moderate the effect of type of evaluation on tip, such that for a competence service failure, tips in formal and informal types of evaluation will not be significantly different.	**Confirmed**	**Confirmed**

H4b: Type of failure will moderate the effect of type of evaluation on tip, such that for morality and warmth failures, the amount of tip will be higher in the formal (vs informal system).	**Rejected**	**Rejected**
H5: In the occurrence of a given service failure, perceived quality of the service will explain the effect of type of failure on users' feedback.	**Confirmed**	**Confirmed**
H6: In the occurrence of a given service failure, the effect of type of evaluation on feedback will be mediated by anticipation of guilt.	**N/A**	**Confirmed**
H7: 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.	**N/A**	**Confirmed**

6. FINAL CONSIDERATIONS

Our studies 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). We proposed that given the nature of collaborative services (more personal interactions than in traditional services) where social norms seem to lead to an attenuation of negative feedback, forgiveness and anticipation of guilt would mediate the effect of type of evaluation on rating instead of previously postulated mediators for this effect, namely reciprocity and fear of retaliation. 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 by knowing a low average rate will have drastically less consequences to them than to the provider, users are less concerned about the consequences that a low score will have for them and more about how it could affect the provider.

We did find forgiveness is a partial mediator for this effect. Forgiveness has been connected to feelings of empathy (ENRIGHT *et al.*, 1992) which in turn seems to play a role in the context of collaborative services (FRADKIN *et al*., 2015) where sociological norms are likely to exist. In Study 2, we included anticipation of guilt as another possible mediator for this effect, as 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). Anticipation of guilt was indeed found to be another partial mediator for the effect of type of evaluation on feedback along with forgiveness. Perhaps another mediator for this effect is an aspect more closely related to sense of community and social norms, such as empathy -which in turn is linked to anticipation of guilt and forgiveness. 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. This is in line with Fradkin (2015), who argue that there is socially induced reciprocity when peers interact socially, leading them to bias feedback. For the authors, this may occur due to mutual empathy between peers after a social interaction and because users may feel a certain 'obligation' to the provider, leading to an omission of negative feedback to avoid hurting the provider.

In Study 1 and 2, we investigated feedback bias after a service failure. In Study 1, we found that type of evaluation and type of failure directly affect the rating given by users to providers. In line with Kirmani *et al*. (2017), our research revealed that competence failures are perceived as affecting the quality of services outcome more than warmth and morality failures (when the latter does not directly harm the user), leading to lower ratings. Hence, perceived quality compromised mediated the effect of type of failure on rating. This mediation was found in both studies.

We did not find any interactive effect of type of evaluation and type of failure on rating in none of the studies. This could possibly be due to the ratings for all failures varying in the same direction. 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). One participant wrote "*A phone dying or not being completely familiar with an area is not a big deal. She seemed pleasant so why be harsh?"*. Another participant wrote *"I didn't want to give her a bad rating especially if she could be fired".* Another one *"I rated the driver more favorably than I probably should have because I didn't want her to get a bad rating that might affect her ability to work"*. Revealing a strong sense of empathy and community.

However, an interactive effect of type of evaluation and type of failure occurred on tip, in Study 1. Specifically, when the type of evaluation was formal, means of tip were significantly higher in the morality condition of type of failure than in the warmth condition of type of failure. This result is consistent with the idea that a morality failure compromises the quality of the service less than other failures. Interestingly, in the control (informal evaluation) condition no difference in the means of tip was found, pointing to bias when formally evaluating the provider. However, this result was not confirmed in Study 2. Two reasons seem to lead to this result, the first one is that the significant interaction found in study 1 could be spurious or the addition of overall driver score in Study 2 may have created a confounding effect on the interaction of type of evaluation and type of failure on tip. Note that, in Study 2, a direct effect of overall driver score on tip was revealed.

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, where in the high score condition ratings were significantly higher than in the low driver score condition. A direct effect of overall driver score on tip was also found. Results show the mean tip given by users to the provider was significantly higher when the driver score was high (vs low). This is consistent with stability attribution theory which postulates that clues of past behavior are often used as predictors of future behavior (WEINER *et al*., 1976). We argue that when the driver score is high, users might be more forgiving of the failure, as a high score can be an indicator of 'good' behavior, therefore leading users to believe the failure is not a recurrent issue. Conversely, a low score may lead users to believe the provider has failed or behaved badly before, and in turn compel users to provide less biased feedbacks.

In Study 2, the main effects of type of evaluation and type of failure on rating were also found. Similar to Study 1, morality which was perceived as the failure affecting the quality of the service the least had the highest means of rating (when compared to competence and warmth). However, no difference in the means of rating were found between competence and warmth failures. Also similar to Study 1, the means of rating were significantly higher in the formal type of evaluation than in the control (informal) condition. In Study 2 the main effect of type of failure on tip was found, as in Study 1. In both studies the mean of tip was significantly higher in the morality condition of type of failure when compared to competence and warmth failures and in both studies the biggest gap in the mean of tip was between morality and warmth conditions of type of failure.

Interestingly, in neither of the studies tip had a significant difference between formal and control (informal) conditions, whereas rating did in both studies. This result seems to indicate tip is a less biased form of feedback in collaborative services than ratings. Since tips implicate a tangible 'loss' and a cost (whereas leaving a rating/review does not) for the user, it is possible that the user would be less willing to have a generous attitude when tipping after a failure. A possible explanation for this is the anticipation of guilt. Our results revealed that anticipation of guilt is a mediator for the effect of type of evaluation on rating. 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. Furthermore, tipping is usually a demonstration of gratitude for the quality of the service (AZAR, 2009) which is not necessarily true for ratings. This is evident since literature seems to suggest ratings are often biased and unrealistically high. In fact, Azar (2009) 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 feedback of service quality and satisfaction than the new forms of service evaluation.

Perhaps as collaborative services and its feedback mechanisms become more permeated in our culture, we will learn how to make better use of them.

7. 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 (e.g. car-rental company Avis bought car-sharing service Zipcar in 2013, and Fox Rent A Car acquired the car-sharing service JustShareIt in 2017), 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 to 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[21]. 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). Our results showed that a competence failure, which may be a one-time event, leads to a lower rating than a morality failure, which is likely a recurring issue, because it affects the perceived quality of the service more than a morality failure. Just one low rating, from a one-time occurrence, may translate into a decrease in the average rating of the driver and harm their livelihood. As our study points out, the current score of the provider also impacts user's assessment, which could also be an issue. That makes it even harder for providers to increase their average rating, simply because feedback was biased by a cue of past, not current behavior. This provides further support to the notion that the automated algorithms used in

[21] An Analysis of the Entrepreneurial Aspects of Uber's Driver-Partner Platform. Available at: https://www.brown.edu/academics/engineering/sites/brown.edu.academics.engineering/files/uploads/UberCaseBrownUniversityMcQuown.pdf. Accessed November 9th 2018.

feedback/reputation systems may unfairly harm providers by not considering cases individually (ROSENBLAT & STARK, 2016).

Results of our study also revealed that behavioral aspects such as forgiveness and anticipation of guilt impact in feedback objectivity. 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.

Another important managerial contribution of our studies is the finding that tip seems to be a less biased form of feedback than ratings. This result could indicate that maybe other forms of feedback, such as tips, are more reliable than ratings or reviews.

8. 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 a lab. Also, other types of methods could be employed such as surveys and nethnographies for greater generalizability. 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 (Airbnb).

In line with Zervas *et al.* (2015), who compared feedbacks in traditional services with their collaborative 'equivalent', a good venue for future research would be to explore the impact of type of failure, type of evaluation and driver score in feedback, comparing traditional and collaborative services. Another topic worth of attention is how other passenger's attitude may compromise the perceived quality of the service, since, nowadays, most on-demand transportation services offer a "pool" option. The pool option is cheaper, and it implies sharing the car with other (unknown) passengers along the selected route, similar to a car picking up hitchhikers. Airbnb also offers the option of renting a room in a house, which may be shared with other guests. Understanding how the behavior of other users in a collaborative service impacts the experience is an interesting topic for research.

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

According to Bhattacharjee *et al*. (2013), moral decoupling is a moral reasoning process that results in judgments of performance being separated from judgements of morality. It allows people to support another's performance while also condemning their transgressions. The more relevant the transgression, the harder it is to decouple. As in Study 1, in Study 2 the morality failure had a less direct impact to the passenger (i.e, the lack of morality shown by the driver was towards others, not directly affecting the passenger -nor the service) than the other failures (warmth and competence). For this reason, it is possible that decoupling from the moral transgression was easier, affecting less the perception of quality being compromised. Further studies into this process in collaborative services settings could be another interesting topic for research, perhaps investigating the consequences of a morality failure which directly affects the service outcome.

REFERENCES

Albinsson, P. A., & Yasanthi Perera, B. (2012). Alternative marketplaces in the 21st century: Building community through sharing events. *Journal of consumer Behaviour*, *11*(4), 303-315.

Ashton, M. C., Paunonen, S. V., Helmes, E., & Jackson, D. N. (1998). Kin altruism, reciprocal altruism, and the Big Five personality factors. *Evolution and Human Behavior*, *19*(4), 243-255.

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.

Barnes, S. J., & Mattsson, J. (2016). Understanding current and future issues in collaborative consumption: A four-stage Delphi study. *Technological Forecasting and Social Change*, *104*, 200-211.

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.

Basil, D. Z., Ridgway, N. M., & Basil, M. D. (2008). Guilt and giving: A process model of empathy and efficacy. *Psychology & Marketing*, *25*(1), 1-23.

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. (2007). Why not share rather than own? *The Annals of the American Academy of Political and Social Science*, *611*(1), 126-140.

_______. (2009). Sharing. *Journal of consumer research*, *36*(5), 715-734.

_______. (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.

Bhattacharjee, A., Berman, J. Z., & Reed, A. (2012). Tip of the hat, wag of the finger: How moral decoupling enables consumers to admire and admonish. *Journal of Consumer Research*, *39*(6), 1167-1184.

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.

Catulli, M., Lindley, J. K., Reed, N. B., Green, A., Hyseni, H., & Kiri, S. (2013). What is Mine is NOT Yours: Further insight on what access-based consumption says about consumers. In *Consumer Culture Theory* (pp. 185-208). Emerald Group Publishing Limited.

Chasin, F., von Hoffen, M., Cramer, M., & Matzner, M. (2017). Peer-to-peer sharing and collaborative consumption platforms: a taxonomy and a reproducible analysis. *Information Systems and e-Business Management*, 1-33.

Cheng, M., & Foley, C. (2018). The sharing economy and digital discrimination: The case of Airbnb. *International Journal of Hospitality Management, 70*, 95-98.

Chung, E., & Beverland, M. (2006). An exploration of consumer forgiveness following marketer transgressions. *ACR North American Advances*.

Clark, M. S., & Mills, J. (1994). Communal and exchange relationships: Controversies and research. *Theoretical frameworks for personal relationships*.

Clayson, D. E. (2004). A test of the reciprocity effect in the student evaluation of instructors in marketing classes. *Marketing Education Review*, *14*(2), 11-21.

Cohen, J. (1988). Statistical power analysis for the behavioural sciences.

Costa, P., de Carvalho-Filho, M. A., Schweller, M., Thiemann, P., Salgueira, A., Benson, J., ... & Quince, T. (2017). Measuring medical students' empathy: exploring the underlying constructs of and associations between two widely used self-report instruments in five countries. *Academic Medicine*, *92*(6), 860-867.

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

Cronin Jr, J. J., Brady, M. K., & Hult, G. T. M. (2000). Assessing the effects of quality, value, and customer satisfaction on consumer behavioral intentions in service environments. *Journal of retailing*, *76*(2), 193-218.

De Corte, K., Buysse, A., Verhofstadt, L. L., Roeyers, H., Ponnet, K., & Davis, M. H. (2007). Measuring empathic tendencies: Reliability and validity of the Dutch version of the Interpersonal Reactivity Index. *Psychologica Belgica*, *47*(4), 235-260.

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.

Derrida, J. (2003). *On cosmopolitanism and forgiveness*. Routledge.

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

Edelman, B., Luca, M., & Svirsky, D. (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics*, *9*(2), 1-22.

Enright, R. D., Gassin, E. A., & Wu, C. (1992). Forgiveness: a developmental view. Journal of Moral Education, 21(2), 99–114.

Ert, E., Fleischer, A., & Magen, N. (2016). Trust and reputation in the sharing economy: The role of personal photos in Airbnb. *Tourism Management*, *55*, 62-73.

Ert, E., Fleischer, A., & Magen, N. (2016). Trust and reputation in the sharing economy: The role of personal photos in Airbnb. *Tourism Management*, *55*, 62-73.

Ewens, M., Tomlin, B., & Wang, L. C. (2014). Statistical discrimination or prejudice? A large sample field experiment. *Review of Economics and Statistics*, *96*(1), 119-134.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*(2), 175-191.

Figueiredo, B., & Scaraboto, D. (2016). The Systemic Creation of Value Through Circulation in Collaborative Consumer Networks. *Journal of Consumer Research*, *43*(4), 509-533.

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. (2018). The determinants of online review informativeness: Evidence from field experiments on Airbnb.

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).

Freedman, S. R., & Enright, R. D. (1996). Forgiveness as an intervention goal with incest survivors. *Journal of consulting and clinical psychology*, *64*(5), 983.

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism*, *10*(2), 486.

Goodrich, M. T., & Kerschbaum, F. (2011, February). Privacy-enhanced reputation-feedback methods to reduce feedback extortion in online auctions. In *Proceedings of the first ACM conference on Data and application security and privacy* (pp. 273-282). ACM.

Goodwin, C. J., & Goodwin, K. (2003). Research in psychology . Hoboken.

Goodwin, C., & Ross, I. (1992). Consumer responses to service failures: influence of procedural and interactional fairness perceptions. *Journal of Business research*, *25*(2), 149-163.

Goodwin, C., & Ross, I. (1992). Consumer responses to service failures: Influence of procedural and interactional fairness perceptions. *Journal of Business research*, *25*(2), 149-163.

Guyader, H. (2018). No one rides for free! Three styles of collaborative consumption. *Journal of Services Marketing*, *32*(6), 692-714.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). *Multivariate Data Analysis*. Bookman.

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.

Hardin, G. (1968). The tragedy of the commons. *science*, *162*(3859), 1243-1248.

Hayes, Andrew F.(2013). Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach. New York, NY: The Guilford Press. *Journal of Educational Measurement*, *51*(3), 335-337.

Heider, F. (1958|2013). *The psychology of interpersonal relations*. Psychology Press.

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.

Holloway, B. B., & Beatty, S. E. (2003). Service failure in online retailing: A recovery opportunity. *Journal of service research*, *6*(1), 92-105.

Hope, D. (1987). The healing paradox of forgiveness. *Psychotherapy: Theory, Research, Practice, Training*, *24*(2), 240.

Hughes, R., & Huby, M. (2002). The application of vignettes in social and nursing research. *Journal of advanced nursing*, *37*(4), 382-386.

Hwang, J., & Griffiths, M. A. (2017). Share more, drive less: Millennials value perception and behavioral intent in using collaborative consumption services. *Journal of Consumer Marketing*, *34*(2), 132-146.

Isaac, G. L. (1978). The Harvey Lecture series, 1977-1978. Food sharing and human evolution: archaeological evidence from the Plio-Pleistocene of east Africa. *Journal of Anthropological Research*, *34*(3), 311-325.

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions the attribution process in person perception. *Advances in experimental social psychology*, *2*, 219-266.

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.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31-36.

Kelley, Harold H. and John L. Michela (1980), ‘‘Attribution Theory and Research,’’ Annual Review of Psychology, 31 (1), 457-501.

Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the shadow of suspicion: the effects of apology versus denial for repairing competence-versus morality-based trust violations. *Journal of applied psychology*, *89*(1), 104.

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.

Lynn, M., & McCall, M. (2016). Beyond gratitude and gratuity: A meta-analytic review of the predictors of restaurant tipping [Electronic version]. Retrieved November 18th 2018, from Cornell University, SHA School site: http://scholarship.sha.cornell.edu/workingpapers/21

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.

Martijn, C., Spears, R., Van der Pligt, J., & Jakobs, E. (1992). Negativity and positivity effects in person perception and inference: Ability versus morality. *European Journal of Social Psychology*, *22*(5), 453-463.

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.

McCollough, M. A., Berry, L. L., & Yadav, M. S. (2000). An empirical investigation of customer satisfaction after service failure and recovery. *Journal of service research*, *3*(2), 121-137.

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.

O'reilly, T. (2005). What is web 2.0.

Ozanne, L. K., & Ballantine, P. W. (2010). Sharing as a form of anti-consumption? An examination of toy

Palmer, A., Beggs, R., & Keown-McMullan, C. (2000). Equity and repurchase intention following service failure. *Journal of Services Marketing*, *14*(6), 513-528.

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). Servqual: A multiple-item scale for measuring consumer perc. *Journal of retailing*, *64*(1), 12.

Park, D. H., Lee, J., & Han, I. (2007). The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement. *International journal of electronic commerce*, *11*(4), 125-148.

Pizzutti, C., & Fernandes, D. (2010). Effect of recovery efforts on consumer trust and loyalty in e-tail: a contingency model. *International Journal of Electronic Commerce*, *14*(4), 127-160.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior research methods, instruments, & computers*, *36*(4), 717-731.

Ranzini, G., Newlands, G., Anselmi, G., Andreotti, A., Eichhorn, T., Etter, M., & Lutz, C. (2017). Millennials and the Sharing Economy: European Perspectives.

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.

Riaz, Z., & Khan, M. I. (2016). Impact of service failure severity and agreeableness on consumer switchover intention: Mediating role of consumer forgiveness. *Asia Pacific Journal of Marketing and Logistics*, *28*(3), 420-434.

Rosenblat, A., & Stark, L. (2016). Algorithmic labor and information asymmetries: A case study of Uber’s drivers.

Scaraboto, D. (2015). Selling, sharing, and everything in between: The hybrid economies of collaborative networks. *Journal of Consumer Research*, *42*(1), 152-176.

Schaefers, T., Lawson, S. J., & Kukar-Kinney, M. (2016). How the burdens of ownership promote consumer usage of access-based services. *Marketing Letters*, *27*(3), 569-577.

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.

Spiller, S. A., Fitzsimons, G. J., Lynch Jr, J. G., & McClelland, G. H. (2013). Spotlights, floodlights, and the magic number zero: Simple effects tests in moderated regression. *Journal of marketing research*, *50*(2), 277-288.

Sprinthall, R. C., & Fisk, S. T. (1990). *Basic statistical analysis*. Englewood Cliffs, NJ: Prentice Hall.

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.*

Sundararajan, A. (2013). From Zipcar to the sharing economy. *Harvard Business Review*, *1*.

Tax, S. S., Brown, S. W., & Chandrashekaran, M. (1998). Customer evaluations of service complaint experiences: implications for relationship marketing. *The journal of marketing*, 60-76.

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.

Vangelisti, A. L., Daly, J. A., & Rudnick, J. R. (1991). Making People Feel Anticipation of guilty in Conversations:" Techniques and Correlates". *Human Communication Research*, *18*(1), 3.

Vasquez, C. (2014). *The Discourse of Online Consumer Reviews*. Bloomsbury Publishing.

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.

__________ (2009). *Forgiving and reconciling: Bridges to wholeness and hope*. InterVarsity Press.

__________ (2013). *Forgiveness and reconciliation: Theory and application*. Routledge.

___________ (Ed.). (2005). Handbook of forgiveness. New York, NY: Brunner-Routledge.

Worthington Jr, E. L. (1998). The pyramid model of forgiveness: Some interdisciplinary speculations about unforgiveness and the promotion of forgiveness. *Dimensions of forgiveness: Psychological research and theological perspectives*, 107-137.

Worthington Jr, E. L., Berry, J. W., & Parrott III, L. (2001). Unforgiveness, forgiveness, religion, and health.

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*.

Worthington, E. L., & Scherer, M. (2004). Forgiveness is an emotion-focused coping strategy that can reduce health risks and promote health resilience: Theory, review, and hypotheses. *Psychology & Health*, *19*(3), 385-405.

Worthington, E. L., Witvliet, C. V. O., Pietrini, P., & Miller, A. J. (2007). Forgiveness, health, and well-being: A review of evidence for emotional versus decisional forgiveness, dispositional forgivingness, and reduced unforgiveness. *Journal of behavioral medicine*, *30*(4), 291-302.

Zervas, G., Proserpio, D., & Byers, J. (2015). A first look at online reputation on Airbnb, where every stay is above average.

__________ (2017). The rise of the sharing economy: Estimating the impact of Airbnb on the hotel industry. *Journal of Marketing Research*, *54*(5), 687-705.

APPENDIX A

Video Dialogs[22]

MORALITY

Passenger: Hi.

Driver: Hello. How are you?

Passenger: I'm good.

Driver: Is the temperature inside the car ok?

Passenger: Yes

Driver: You know… I just bought a sticker that will allow me to park in handicapped parking spots. That way I can waste less time looking for parking spots.

Subtitle The passenger realizes that the driver isn't handicapped and is just bragging about illegally buying a handicap parking sticker.

Screen [15 minutes later]

Driver: Thank you. Have a great day.

Passenger: Thank you, bye.

COMPETENCE

Passenger: Hi.

Driver: Hello. How are you?

Passenger: I'm good.

Driver: Is the temperature inside the car ok?

[22] Videos available at:

Competence Failure https://youtu.be/LDeK7IkffnU
Morality Failure https://youtu.be/VHdHg-y3ORs
Warmth Failure https://youtu.be/JVjz-afILjM

Passenger: Yes

Driver: The cable of my charger hasn’t been working well lately and I think I’ll run out of battery soon. Do you think you can guide me to your destination?

Passenger: I’m not sure but I will try

Subtitle The passenger realizes the driver doesn’t know any of the main streets and would have gone in the opposite direction if the passenger didn’t know which way to go.

Passenger: How can I reserve a ride for later today in the app?

Driver: Oh, I don’t know how to use the app except for accepting and ending rides.

Screen [15 minutes later]

Driver: Thank you. Have a great day.

Passenger: Thank you, bye.

WARMTH

Passenger: Hi

Subtitle (driver does no reply)

Passenger: Do you know how to get to the address?

Driver: Yep

Screen [15 minutes later]

Passenger: Thank you, bye.

Subtitle (driver does not reply again)

APPENDIX B

Questionnaire Study 1

TCLE Welcome!

Thank you for showing interest in this study.

Please read the following before starting:

This is a short study on consumer behavior conducted by researchers from the Federal University of Rio Grande do Sul. Please complete this survey in one go. In other words, please only participate if you have about 12 minutes that you can dedicate to it, although you might need less that.

Informed Consent Form

What will I do if I choose to be in this study?
If you agree to take part in this study, you will be asked to complete a short online questionnaire which will involve watching a short video and seeing an image and then answering a few questions.

What will be the benefits of participating in this study?
There is no direct benefit in participating in this research. However, your participation will help us to better understand consumer behavior in collaborative consumption services.

What are the possible risks or discomforts of participating in this study?
We believe that there are no risks associated with this study. However, as with any online related activity, the risk of a breach of confidentiality is always possible. Your answers in this study will remain confidential and will only be used for academic purposes, including publication of results. Your personal information will not be stored with the data collected from your responses.

What are my rights if I take part in this study?
Your participation in this study is voluntary. As researchers we are not qualified to provide counselling services and we will not be following up with you after this study. However, if you have questions about this project or if you have research-related problems, you may contact the researchers by e-mail at louise.foernges@gmail.com.

By clicking "I agree" I declare that I am 18 years of age or over, and agree to participate in this research. I declare that I was informed that my participation in this study is voluntary, that I can leave this survey at any time without penalty, and that all data is confidential. I understood that this study does not offer serious risks.

○ I agree

Dear respondent,

You will watch a video that presents a situation in which a passenger is using an on-demand transportation service called TakeMe (similar to Uber, Lyft etc.) which works through an app installed in smartphones.
The video is mute.

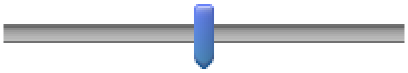
[VIDEO]

FORMAL CONDITION

[Info] You can now rate your driver in the service’s app. The feedback system used by the transportation service in the video is a two-way type. That means you can rate the driver and the driver can also rate you as a passenger. Drivers and passengers with a low average rating may get banned from the service.

[Rating_app]
In a scale of 1 to 5 how would you rate the driver service’s app?

(You don't necessarily have to use whole numbers)

	1 5
1 = Very Bad 5 = Excellent ()	

INFORMAL CONDITION

[Info] You arrived at your destination and met a friend. Your friend tells you he has never used an on-demand transportation service before (like the one you just used). Your friend asks you how your experience with the service was today.

[Rating_friend] In a scale of 1 to 5, how would you rate the driver to your friend? Remember you are not rating in the app, but only informally telling your friend more about your experience.

(You don't necessarily have to use whole numbers)

	1 5
1 = Very Bad 5 = Excellent ()	

Please tell us in your own words why you rated the driver the way you did:

[Tip] You can tip the driver directly through the app. The amount of tip (in a scale of 0% to 25% of the total amount paid for the trip) you would be willing to give the driver is

	0 3 5 8 10 13 15 18 20 23 25
In % ()	

[Info] Now, please answer a few questions

[Reciprocity] I rated the driver according to how I expect the driver has rated me as a passenger

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

[Affect_job] When rating the driver, I took into consideration how that rating can affect her job

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

[Fear of Negative Consequences] I fear to suffer negative consequences if I give an honest feedback to this driver

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

[Forgiveness] I'm bitter about what the driver did

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

[Forgiveness] I'm mad about what happened during the ride

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

[Forgiveness] I resent what she did during the ride

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

[Info] For the next five questions, please indicate to which extent you agree or disagree with the following remarks:

[Recommend] If possible to do so, I would recommend this driver

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

[Perceived Severity] If the inconvenience during the ride in the video was really happening to you, you would consider it to be:

	1	2	3	4	5	6	7	
Not Severe at All	○	○	○	○	○	○	○	Very Severe

[Perceived Severity] If the situation shown in the video was really happening to you, it would make you feel:

	1	2	3	4	5	6	7	
Not Angry at All	○	○	○	○	○	○	○	Very Angry

[Perceived Severity] If the situation shown in the video was really happening to you, you would consider it to be:

	1	2	3	4	5	6	7	
Not Pleasant at All	○	○	○	○	○	○	○	Very Pleasant

[Quality] To which extent you believe that what happened during the trip has compromised the quality of the service provided? (transportation service to the destination)

	1	2	3	4	5	6	7	
Not at All	○	○	○	○	○	○	○	Very Much

[Realism] "I believe that the situation presented in the video could happen during a ride with on-demand transportation services"

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

[Warmth] In a 7-point scale, to which degree you consider the driver in the video to be:

	1	2	3	4	5	6	7	
Unfriendly	○	○	○	○	○	○	○	Friendly
Cold	○	○	○	○	○	○	○	Warm
Unsociable	○	○	○	○	○	○	○	Sociable
Not nice	○	○	○	○	○	○	○	Nice

[Morality] In a 7-point scale, to which degree you consider the driver in the video to be:

	1	2	3	4	5	6	7	
Dishonest	○	○	○	○	○	○	○	Honest
Insincere	○	○	○	○	○	○	○	Sincere
Manipulative	○	○	○	○	○	○	○	Not Manipulative
Not trustworthy	○	○	○	○	○	○	○	Trustworthy

[Competence] In a 7-point scale, to which degree you consider the driver in the video to be:

	1	2	3	4	5	6	7	
Incompetent	○	○	○	○	○	○	○	Competent
Not Clever	○	○	○	○	○	○	○	Clever
Not knowledgeable	○	○	○	○	○	○	○	Knowledgeable
Unskilled	○	○	○	○	○	○	○	Skilled

[Type of Evaluation Manipulation Check] You evaluated the driver

○ To your friend

○ In the service's app

○ Do not remember

[Manipulation Check One-way, Two-way] The evaluation system used in the situation presented to you is one-way (only you are rating the driver) or two-way (you are rating the driver and the driver is also rating you as a passenger)?

○ One-way

○ Two-way

○ Do not remember

[Attention Check] What happened during the trip shown in the video?

○ Driver was bragging about buying something illegal

○ Driver couldn't charge her phone and didn't know how to get to the destination

○ Driver was cold/unfriendly

○ None of the above

[Used before] Have you ever used transportation services such as Uber, Cabify, Lyft or similar before?

○ Yes

○ No

[Frequency of use] Considering the past 6 months, with which (average) frequency have you used on-demand transportation services, such as the one shown in the video?

(Please consider each individual ride. For example, if you used the service to go and later come back from a location, consider it 2 times)

○ Less than one ride a month

○ 2-3 times a month

○ Once a week

○ 2-3 times a week

○ 4-5 times a week

○ 5-6 times a week

○ 7+ times a week

[Gender] What is your gender?

○ Male

○ Female

○ Do not wish to answer

[Age] What is your age?

○ 18-24

○ 25-34

○ 35-44

○ 45-54

○ 55+

[Handicapped] Are you handicapped?

○ Yes

○ No

Please write in a few words what do you think this research is about

Suggestions Any suggestions regarding this survey? (optional)

APPENDIX C

Questionnaire Study 2

TCLE Welcome!

Thank you for showing interest in this study.

Please read the following before starting:

This is a short study on consumer behavior conducted by researchers from the Federal University of Rio Grande do Sul. Please complete this survey in one go. In other words, please only participate if you have about 12 minutes that you can dedicate to it, although you might need less that.

Informed Consent Form

What will I do if I choose to be in this study?
If you agree to take part in this study, you will be asked to complete a short online questionnaire which will involve watching a short video and seeing an image and then answering a few questions.

What will be the benefits of participating in this study?
There is no direct benefit in participating in this research. However, your participation will help us to better understand consumer behavior in collaborative consumption services.

What are the possible risks or discomforts of participating in this study?
We believe that there are no risks associated with this study. However, as with any online related activity, the risk of a breach of confidentiality is always possible. Your answers in this study will remain confidential and will only be used for academic purposes, including publication of results. Your personal information will not be stored with the data collected from your responses.

What are my rights if I take part in this study?
Your participation in this study is voluntary. As researchers we are not qualified to provide counselling services and we will not be following up with you after this study. However, if you have questions about this project or if you have research-related problems, you may contact the researchers by e-mail at louise.foernges@gmail.com.

By clicking "I agree" I declare that I am 18 years of age or over, and agree to participate in this research. I declare that I was informed that my participation in this study is voluntary, that I can leave this survey at any time without penalty, and that all data is confidential. I understood that this study does not offer serious risks.

○ I agree

Dear respondent,

You will watch a video that presents a situation in which a passenger is using an on-demand transportation service called TakeMe (similar to Uber, Lyft etc.) which works through an app installed in smartphones.
The video is mute.

[VIDEO]

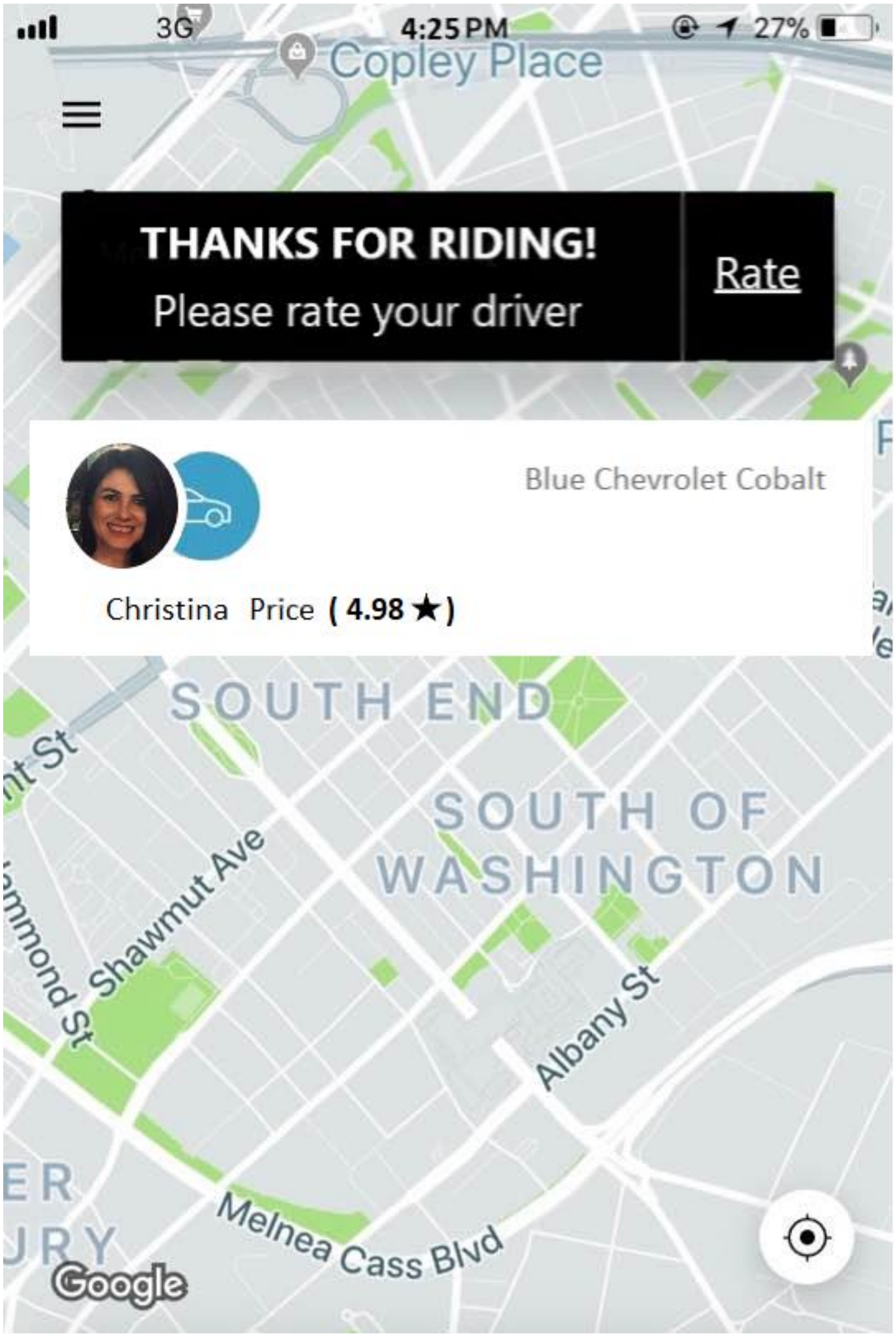
Info In this service, drivers and passengers are able to rate each other in a scale that goes from 1 to 5.

Drivers and passengers who sustain a rating lower than 4 for more than a month may get banned from the service.

As soon as you leave the car, you grab your phone and see the following screen

(Please pay closer attention to the white square, you will be asked about this image later)

HIGH DRIVER SCORE



THANKS FOR RIDING!

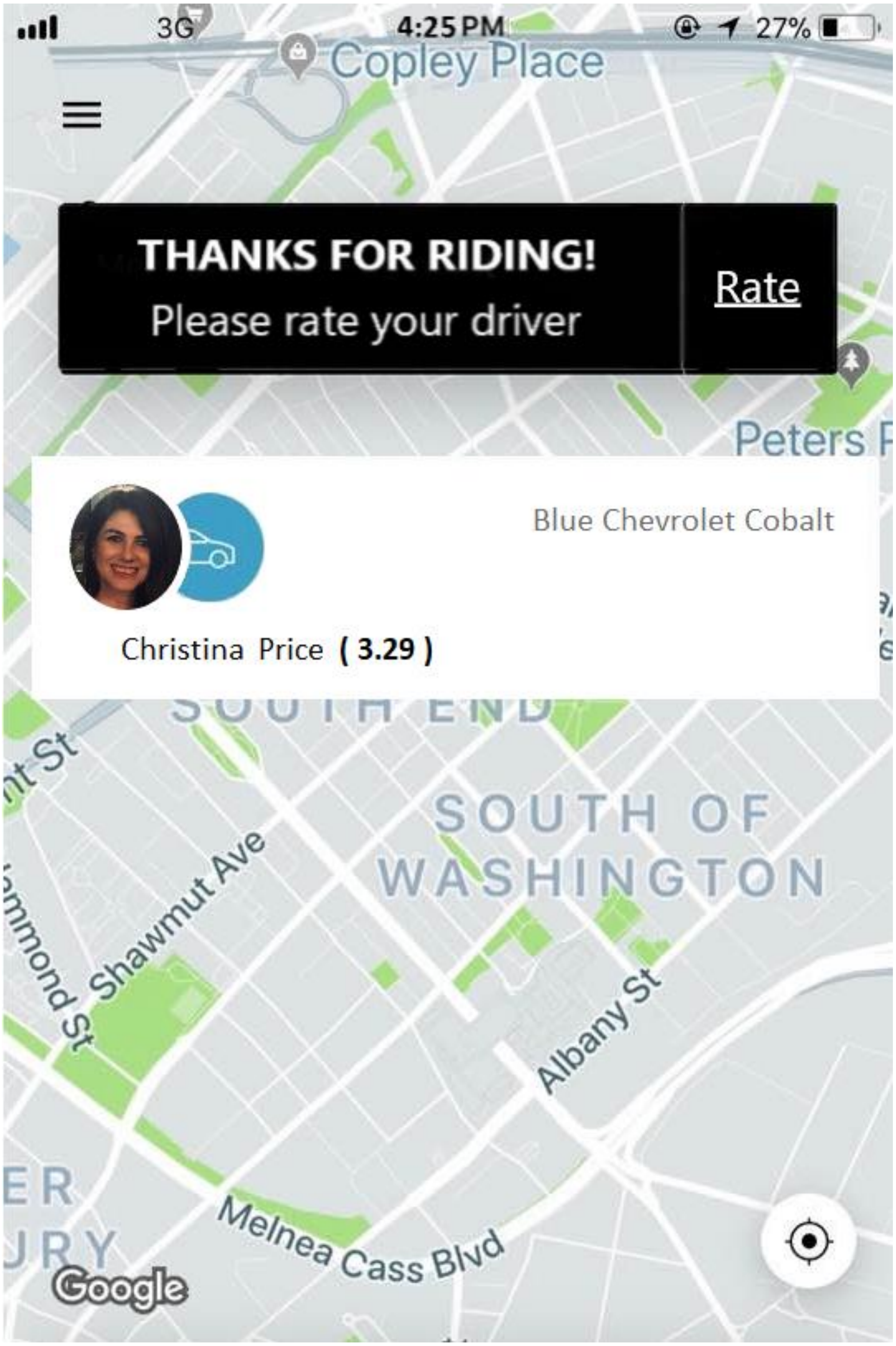
Please rate your driver

Rate

Blue Chevrolet Cobalt

Christina Price (4.98 ★)

LOW DRIVER SCORE



THANKS FOR RIDING!

Please rate your driver

Rate

Blue Chevrolet Cobalt

Christina Price **(3.29)**

FORMAL CONDITION

[Info] You can now formally rate your driver in the TakeMe service app.

[Rating_app]
In a scale of 1 to 5 how would you rate the driver in the TakeMe service app?

(You don't necessarily have to use whole numbers)

	1 5
1 = Very Bad 5 = Excellent ()	

INFORMAL CONDITION

[Info] You arrived at your destination and met a friend. Your friend tells you he has never used an on-demand transportation service before (like the one you just used). Your friend asks you how your experience with the TakeMe service was today.

[Rating_friend] In a scale of 1 to 5, how would you rate the driver to your friend? Remember you are not rating in the app, but only informally telling your friend more about your experience.

(You don't necessarily have to use whole numbers)

	1 5
1 = Very Bad 5 = Excellent ()	

Please tell us in your own words why you rated the driver the way you did:

____

[Tip] You can tip the driver <u>directly through the TakeMe app</u>. The amount of tip (in a scale of 0% to 25% of the total amount paid for the trip) you would be willing to give the driver is

	0	25
In % ()		

[Info] Please answer a few questions about yourself

[Forgiveness] I'm bitter about what the driver did

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

[Forgiveness] I'm mad about what happened during the ride

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

[Forgiveness] I resent what she did during the ride

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

[Anticipation of guilt] I gave the driver a rating different than I thought she really deserved because I would feel anticipation of guilty if the driver suffered negative consequences due to my rating

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

[Anticipation of guilt] I gave the driver a rating different than I thought she really deserved because I would feel bad if the driver suffered negative consequences due to my rating

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

[Stability] Please select to which degree you believe that the cause of the problem/inconvenience during the ride is something that is

	1	2	3	4	5	6	7	
Temporary	○	○	○	○	○	○	○	Permanent
Variable over time	○	○	○	○	○	○	○	Stable over time
Changeable	○	○	○	○	○	○	○	Unchangeable

[Perceived Severity] If the inconvenience during the ride in the video was really happening to you, you would consider it to be:

	1	2	3	4	5	6	7	
Not Severe at All	○	○	○	○	○	○	○	Very Severe

[Perceived Severity] If the situation shown in the video was really happening to you, it would make you feel:

	1	2	3	4	5	6	7	
Not Angry at All	○	○	○	○	○	○	○	Very Angry

[Perceived Severity] If the situation shown in the video was really happening to you, you would consider it to be:

	1	2	3	4	5	6	7	
Not Pleasant at All	○	○	○	○	○	○	○	Very Pleasant

[Recommend] If possible to do so, I would recommend this driver

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

In a scale that goes from strongly disagree to strongly agree, please select to which extent you agree with the following statement:

[Realism] "I believe that the situation presented in the video could happen during a ride with on-demand transportation services"

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

[Quality] To which extent you believe that what happened during the trip has compromised the quality of the service provided? (transportation service to the destination)

	1	2	3	4	5	6	7	
Not at All	○	○	○	○	○	○	○	Very Much

[Warmth] In a 7-point scale, to which degree you consider the driver in the video to be:

	1	2	3	4	5	6	7	
Unfriendly	○	○	○	○	○	○	○	Friendly
Cold	○	○	○	○	○	○	○	Warm
Unsociable	○	○	○	○	○	○	○	Sociable
Not nice	○	○	○	○	○	○	○	Nice

[Morality] In a 7-point scale, to which degree you consider the driver in the video to be:

	1	2	3	4	5	6	7	
Dishonest	○	○	○	○	○	○	○	Honest
Insincere	○	○	○	○	○	○	○	Sincere
Manipulative	○	○	○	○	○	○	○	Not Manipulative
Not trustworthy	○	○	○	○	○	○	○	Trustworthy

[Competence] In a 7-point scale, to which degree you consider the driver in the video to be:

	1	2	3	4	5	6	7	
Incompetent	○	○	○	○	○	○	○	Competent
Not Clever	○	○	○	○	○	○	○	Clever
Not knowledgeable	○	○	○	○	○	○	○	Knowledgeable
Unskilled	○	○	○	○	○	○	○	Skilled

[Attention Check] The current rating (score) of the driver was

○ Above 4.2

○ Below 4.2

○ There was no information

[Driver Score Manipulation Check] You consider the current rating (score) of the driver to be

○ High

○ Low

○ Avarage

[Attention Check] What happened during the trip shown in the video?

○ Driver was bragging about buying something illegal

○ Driver couldn't charge her phone and didn't know how to get to the destination

○ Driver was cold/unfriendly

○ None of the above

[Type of Evaluation Manipulation Check] You evaluated the driver

○ To your friend

○ In the service's app

○ Do not remember

[Used before] Have you ever used transportation services such as Uber, Cabify, Lyft or similar before?

○ Yes

○ No

[Frequency of use] Considering the past 6 months, with which (average) frequency have you used on-demand transportation services, such as the one shown in the video?

(Please consider each individual ride. For example, if you used the service to go and later come back from a location, consider it 2 times)

○ Less than one ride a month

○ 2-3 times a month

○ Once a week

○ 2-3 times a week

○ 4-5 times a week

○ 5-6 times a week

○ 7+ times a week

[Gender] What is your gender?

○ Male

○ Female

○ Do not wish to answer

[Age] What is your age?

▼ 18 ... 100

[Handicapped] Are you handicapped?

○ Yes

○ No

Please write in a few words what do you think this research is about

[Suggestions] Any suggestions regarding this survey? (optional)