UNIVERSITA' COMMERCIALE "LUIGI BOCCONI" PhD SCHOOL

PhD program in Business, Administration and Management

Cycle: 32nd

Disciplinary Field (code): SECS-P/08

Divergent Effects of Technological Solutions in Services

Advisor: Andrea ORDANINI

PhD Thesis by

Anastasia NANNI

ID number: 1441112

Year 2021

Technology Detour in Personal Service Settings: When Promoting Convenience Inadvertently Reduces Customer/FLE interactions and Backfires the Shopping Experience

Target Journal: Journal of Marketing Research

Anastasia Nanni

Doctoral Student in Marketing, PhD in Business Administration & Management
Bocconi University, Via Rontgen 1, 20136 Milan, Italy
Phone +39 02 58363510
anastasia.nanni@unibocconi.it

In joint with:

Andrea Ordanini

BNP Paribas Professor of Marketing & Service Analytics
Bocconi University, Via Rontgen 1, 20136 Milan, Italy,
Phone +39 02 58363623
andrea.ordanini@unibocconi.it

ABSTRACT

Providing service convenience is pivotal to both customer satisfaction and retention. Service providers and retailers increasingly rely on high convenience technologies that allow customers to save time and effort, to make their customers' experiences smooth and frictionless. A high convenience technology is narrowcasting which enables the real-time dissemination of information in the service setting, using artificial intelligence (AI) and digital displays. In the present research, we investigate the effect of high convenience technologies, in particular narrowcasting, in personal service settings, in which customers enter into close, collaborative relationships with frontline employees (FLEs). We advance propositions drawn from services marketing, system theories, and organizational behavior literature, which we test using a multimethod approach based on controlled field experiments, intercept surveys, and online studies. We secured the collaboration of a fashion retail chain, which represents a typical personal service setting, that recently introduced a narrowcasting technology in one of its stores in an effort to streamline and improve the efficiency of customers' shopping experience. The results reveal that narrowcasting technology can exert negative effects on customer shopping behavior and experiences in personal service settings, due to its unintended effects on FLEs' extra-role behaviors. Specifically, the technology increases customers' service convenience, but it limits FLEs' efforts to adapt and respond effectively to customer needs, resulting in an overall negative effect on spending and perceptions of the service experience. We confirm the causal mechanism associated with FLEs' extra-role behavior by noting that the narrowcasting technology exerts a positive effect on service outcomes if the interaction between FLEs and customers is peripheral to the experience (e.g., when customers shop under time pressure).

Keywords: High Covenience Technology; PersonalService Settings; Field Experiments; Customer Role Behavior; Frontline Employee Role Behavior; Customer-Employee Interaction

INTRODUCTION

Service convenience, defined as customers' perceptions of the time and effort associated with the use of a service, is critical to both customer satisfaction and retention (Seiders et al. 2005). Service providers and retailers rely increasingly on high convenience technologies to help make their customers' experiences smooth and frictionless (Grewal et al. 2019); ultimately, these technologies aim to change the shopping process and thereby allow customers to save time and effort (Larivière et al. 2017). In particular, *narrowcasting* technologies (Floor 2006; Metzger 2017) enable the real-time dissemination of information in the service setting, using artificial intelligence (AI) and digital displays (Grewal et al. 2019). Uses of narrowcasting technology are spreading to various service domains, including hotels (e.g., Bastion chain), banks (e.g., Holifax), restaurants (e.g., McDonald's), and specialty stores (e.g., Pandora jewelry, Expresso fashion). Yet tests of these technologies mainly have involved relatively straightforward encounters in which customers mostly interact and the technology (e.g., self-service, online) (Heller et al. 2019; Johnson, Bardhi, and Dunn 2008), rather than personal settings, in which customers enter into close, collaborative relationships with frontline employees (FLEs) (Frohele and Roth 2004; Gutek 1999).

The lack of evidence in personal service settings is problematic, because the outcomes of high convenience technologies, including narrowcasting, might vary in personal interactions. In particular, despite their promise to streamline the process, high convenience technologies tend to increase (cf. reduce) complexity and interactions among actors. That is, even if the technologies focus on the customer, they affect both the customer's shopping behavior and the FLE's service behavior, through their core relationship, for which the content also has changed (Cadwallader et al. 2010). Accordingly, it is timely and relevant to consider whether the introduction of a high convenience technology such as narrowcasting

might (inadvertently) alter FLEs' behaviors in personal interactions with customers, in what ways, and what effects this change might have on customers' experience and behaviors.

To investigate these questions, we advance propositions drawn from services marketing, system theories, and organizational behavior literature, which we test using a multimethod approach based on controlled field experiments, intercept surveys, and online studies. We secured the collaboration of a fashion retail chain, which represents a typical personal service setting, that recently introduced a narrowcasting technology in one of its stores in an effort to streamline and improve the efficiency of customers' shopping experience. The results reveal that narrowcasting technology can exert negative effects on customer shopping behavior and experiences in personal service settings, due to its unintended effects on FLEs' extra-role behaviors. Specifically, the technology increases customers' service convenience, but it limits FLEs' efforts to adapt and respond effectively to customer needs, resulting in an overall negative effect on spending and perceptions of the service experience. We confirm the causal mechanism associated with FLEs' extra-role behavior by noting that the narrowcasting technology exerts a positive effect on service outcomes if the interaction between FLEs and customers is peripheral to the experience (e.g., when customers shop under time pressure).

As its contributions, our research thus responds to calls for more research that acknowledges and addresses the complex and contingent nature of technology transformations in retail and service settings (Grewal et al. 2019). Failing to consider the systemic nature of service settings can lead to unexpected consequences of technology introductions. For example, if service managers focus solely on the immediate, desired outcomes of a promising technology, without paying attention to its indirect effects on other elements of the system (e.g., personal interactions), the overall effects can deviate from expectations, markedly and negatively. Our findings reveal specifically that narrowcasting

technologies, with their great promise for streamlining and facilitating customer shopping experiences, actually increase the complexity of these experiences in personal service settings. Various technological options are becoming increasingly available, especially those driven by AI solutions, so managers must be vigilant to their risks, avoid potential mismatches with service settings, and carefully address the level of resulting complexity.

We also contribute to the organizational frontline research stream by offering novel insights into technology use and its impacts on customers and FLEs (Marinova et al. 2017; Rafaeli et al. 2017). As our results show, even if a technology does not targets FLEs directly, its ultimate effects may deviate from expectations, due to the unplanned changes it produces in FLEs' behavior. This evidence helps address some gaps in services marketing literature, which tends to overlook the "multidimensional nature" of service employees' behavior (Hogreve et al. 2017). We document a case in which a technology (inadvertently) makes employees more prone to engage in self-oriented behavior, at the expense of customer-oriented behavior. Our analysis also reveals the importance of FLEs' extra-role efforts to engage in customer-oriented behaviors, which are appreciated by service customers.

Finally, our research provides recommendations for service managers who operate in personal service settings, as well as in more general situations where technology is not expected to substitute fully for FLEs. Before investing in high convenience technologies, managers and firms must carefully scrutinize and examine the extent to which each technology might change the interaction between employees and customers. Without such careful assessments and appropriate management, the changes may result in poor customer service experiences and lower levels of spending.

BACKGROUND LITERATURE

Service convenience arises when consumers can save the time and effort they devote to various phases of the shopping activity; it also is a critical goal for retailers and service

providers (Berry, Seiders, and Grewal 2002). The time and effort that consumers devote to their shopping experiences significantly affect their overall evaluations, such that empirical evidence strongly confirms the positive effect of service convenience on customers' evaluations of the service, satisfaction, and retention (Rust, Lemon, and Zeithaml 2004; Seiders et al. 2005). To increase service convenience, many companies are starting to implement high convenience technologies, which include tools explicitly designed to make store experiences more comfortable, easy, or frictionless (Grewal et al. 2019). Typical examples include self-checkout solutions, which help consumers avoid time-consuming checkout lines; interactive kiosks that allow consumers to automate their ordering activities and information searches; and augmented reality tools (e.g., virtual fitting rooms) that eliminate certain tasks (e.g., choosing a matching accessory). Narrowcasting technologies offer another promising example; they provide specific information on digital screens located in convenient locations, to facilitate customer shopping experiences. All these example technologies transform the service provision process, in that they alter the role and activities required of consumers. Arguably, giving consumers more control over their shopping time and effort enables them to be more efficient in allocating their effort to various phases during their service experience (Larivière et al. 2017).

Previous literature suggests positive effects of such high convenience technologies, especially in *encounter* service settings, in which customers have few connections or interactions with FLEs, and technology largely shapes the customer experience (Froehle and Roth 2004; Gutek 1999). The beneficial effects of high convenience technologies—in terms of satisfaction, evaluations, and willingness to pay—appear prominent in self-service modes (Bitner, Ostrom, and Meuter 2002; Johnson, Bardhi, and Dunn 2008) and online service settings (Heller et al. 2019; van Beuningen et al. 2009). But in other service settings, the customer experience depends on a *personal relationship* with an FLE, such that human

interactions are fundamental for the service to succeed (Gremler and Gwinner 2008), but technology is not necessarily required (Froehle and Roth 2004; Gutek 1999). For example, in physical retail stores selling jewelry, customers generally review a few products, then request help from a FLE who interacts with and focuses closely on them, by displaying other product options selected from showcases that otherwise would be inaccessible to the consumer. In gyms, personal trainers might assist customers throughout their workout sessions.

Even as high convenience and narrowcasting technologies are entering such personal service settings in practice though (Grewal et al. 2019), we lack empirical evidence of their effects. In particular, we posit that even if the technologies modify customers' shopping behaviors as intended, they also are likely to affect FLEs' behaviors and, indirectly, the scope of the personal interactions that define the customer experience (Cadwallader et al. 2010). A particular concern is that, instead of streamlining the consumption experience, high convenience technologies could make personal services settings more complex and interactive (Mende and Noble 2019; Vargo and Lusch 2018). In situations characterized by such *interactive complexity*, a change in any one element can generate unexpected interactions with and changes to other elements (Perrow 1988), with unintended or unpredictable consequences at the system level (Quint 2007; Sargut and McGrath 2011).

Traditionally, FLEs perform in-role and extra-role behaviors while interacting with customers (Netemeyer and Maxham 2007). Their in-role behaviors usually reflect basic, "scripted rules" and specified job descriptions; extra-role behaviors imply going beyond the actions prescribed by these job requirements (Humphrey and Ashforth 1994). In personal service settings, extra-role behaviors are particularly important, because they can define the customer experience, such as when an FLE takes the initiative to communicate in detail with customers or works to adapt her or his responses to emerging customer needs (Bettencourt, Gwinner, and Meuter 2001). When a high convenience technology enters such a personal

service setting though, it alters these relationships, and we predict it might indirectly constrain extra-role behavior. Companies introduce new technologies mainly to attain greater efficiency (Marinova et al. 2017) and take over some tasks from FLEs, yet reducing the scope of FLEs' actions may reduce their commitment and, consequently, the effort they devote to responding to customers' needs (DiMascio 2010). The technologies also enable customers to reallocate their time and effort, which may lead them to violate the social norms for close relationships with FLEs, with potentially negative consequences for the FLEs (Williams and Aaker 2002). For example, if customers unexpectedly reject offers of service and assistance, FLEs might experience a negative sense of being rebuffed and become less willing to devote extra effort to satisfying customers' needs (Giebelhausen et al. 2014).

Therefore, we predict that the overall effect of high convenience technologies on customer experiences in personal service settings involves a trade-off: The high convenience technologies may provide benefits by allowing customers to reallocate their time and effort and perceive more control over the service process (Botti, McGill, and Iyengar 2003), but they also might limit FLEs to their in-role duties and discourage them from taking on extra-role tasks. Bitner, Brown, and Meuter (2000) cite a lack of unprompted and unsolicited actions by FLEs as a main cause of service customer dissatisfaction; Wels-Lips, van der Ven, and Pieters (1998) also show that extra-role behaviors increase customers' reports of positive experiences. Accordingly, we predict that introducing narrowcasting technology in personal service settings exerts a weaker (and even potentially negative) effect on the overall customer experience, compared with encounter-based service settings. We also anticipate that this effect is mediated by the (unintended) spillover effect of reduced FLE extra-role behaviors.

To test these predictions, we conduct a series of studies, as presented in Figure 1. In Study 1, we test the effect of introducing a high convenience technology (narrowcasting) in a personal service setting (apparel shop) on observed customer shopping behaviors (customer

spending and number of purchased items). We randomly manipulate the presence and absence of the technology in the store for different time windows, during a two-week period. The introduction of the technology did not provide tangible benefits; rather, it reduced customer spending and the number of purchased items. To obtain a potential explanation for these results, in Studies 2a and 2b we check for any unintended changes in FLEs' behavior. With Study 2a, we establish that in the personal service setting, the technology tends to reduce perceived interactions of customers and employees, providing initial evidence of this unintended effect. Study 2b replicates this evidence but also reveals that the reduction of extra-role behavior mediates the effect of the technology on both the customer experience and shopping behavior (including non-purchase).

Finally, in Study 3 we confirm the central role of an unintended reduction in FLEs' extra-role behavior, by determining that the narrowcasting technology results in the expected benefits when the interaction between FLEs and customers is not central to the service experience. That is, an online experiment provides evidence, consistent with prior literature (Marinova et al. 2017; Menon and Dubé 2007), that shopping during lunch breaks increases customers' time pressure and lowers their expectation that FLEs exhibit extra-role behaviors. Then we manipulate again the presence or absence of the technology across lunch break periods (vs. rest of the day), which provides confirmatory process-by-moderation evidence (Vancouver and Carlson 2015): During lunch breaks, when shoppers confront time pressures, FLEs' extra-role behavior is no longer important, and the narrowcasting technology is beneficial for the shopping experience.

-Insert Figure 1 about here-

EMPIRICAL SETTING

High Convenience Technology in Personal Service Settings

To investigate our research questions empirically, we sought the collaboration of a chain of fashion stores located in a large European city. Customers spend significant time in these stores (on average, 1.5 hours), and multiple FLEs (ten in each work shift) are available to serve them, for as much time as they need, and offer expert advice, creative ideas, or simple assistance with item selection and matching. Furthermore, the stores in this chain aim to maximize the available assortment, by displaying only one size of each clothing item for sale. Customers choose items, visit the service counter, and ask for the size they need. At the counter, a FLE sends the request to the warehouse and gives the customer a number, to hold while waiting in line next to the counter to receive the requested item. Similar display options appear in various retail stores, such as Argos (apparel and sport equipment) or IKEA; they require customers to interact with employees to customize and complete the shopping experience.

The focal store recently decided to modify the routine by introducing narrowcasting technology, which it hoped would provide customers with more freedom and autonomy. Specifically, the warehouse is located in a separate section of the store. When customers ask for their preferred clothing items, the counter employee still sends a request to the warehouse, and warehouse assistants pick the clothes to be sent back to the counter. However, special sensors embedded in the counter read the labels on the clothing items, and linked software matches the items to the order. It registers that the order has arrived at the service counter and automatically displays the order number on digital screens, installed throughout the shop. Thus, customers receive alerts that their orders are ready, but they can continue to browse, instead of having to stand in a queue near the counter waiting for the order to be delivered.

This technology thus represents a typical high convenience technology (Grewal et al. 2019), in that it enables customers to reallocate their time as they prefer; instead of being required to devote time to waiting for the FLE, they can engage in other activities, such as

shopping in other departments. According to the store manager, the retailer anticipated that the technology would enhance the customer experience and spending, by freeing customers to walk around, spot more clothes, and shop more.

For our studies, we randomly manipulate the presence or absence of the narrowcasting technology in the store across time periods, such that in one period, customers' orders do not appear on the screens (instead, the numbers are called by the FLEs, and customers must stand near the counter to hear them), whereas in the other period, the order numbers appear on screens around the shop, so customers can see them wherever they are. For simplicity, we call the first condition "technology absent" and the second "technology present." Appendix A displays the digital screens in both conditions. We started the empirical analyses about 8 months after the new technology had been introduced, which helped us obtain more accurate effects and reduced potential biases associated with excited responses, as tend to arise in the early stages of technology introduction. In particular, we avoid the so-called honeymoon effect, which occurs when adopters get excited by the novel option (Wells et al. 2010), as well as skeptical responses that might reflect potential adopters' mental rigidities or resistance to change (Mani and Chouk, 2017).

Preliminary Study: Effect of Narrowcasting Technology on Service Convenience

Although both service convenience literature and experienced store managers predicted that the narrowcasting technology would increase service convenience, we decided to test this assumption before initiating the field studies. To this purpose, 229 participants (48.03% women; average age: 28.74 years) from the Prolific Academic panel participated in an online study for $0.44. The study presents a mockup of the focal fashion store; we use this setting for our other studies too. Participants were randomly exposed to two technology scenarios before responding to some questions. In the technology absent scenario, customers ask for their size and wait to be served by the employee at the corner; in the technology

present scenario, they ask for their size, then may walk around the store and check digital screens in the store to learn when their order is ready (see Appendix A). We measure service convenience with three items adapted from the SERVICON scale (Seiders et al. 2007) and three general items (α = .94). Next, we measure their involvement in the fashion category with three items from Goldsmith and Emmert (1991) (α = .82). Appendix B contains additional details about the measurement instruments.

With an analysis of variance (ANOVA), we test whether the presence of the narrowcasting technology affects the level of service convenience. As expected, we find a significant difference in the means of service convenience between conditions ($F_{(1,227)}$ = 5.35; p = .021),[1] such that the presence of the narrowcasting technology increases the perceived level of service convenience ($M_{Presence}$ = 3.68; $M_{Absence}$ = 3.23). This effect is general, in the sense that it does not depend on the level of involvement of the respondent in the fashion category ($F_{(1,225)}$ = .25; p = .62). Thus, the narrowcasting technology that we manipulate in our field experiments represents a high convenience technology, and we use the manipulation in all the studies that follow.

STUDY 1: HIGH CONVENIENCE TECHNOLOGY AND CUSTOMER SHOPPING BEHAVIOR

Procedure and Design

This field experiment aims to test the effect of the narrowcasting technology on customers' shopping behavior. We use two observed measures of shopping behavior: the amount of money each customer spends (i.e., customer spending) and the number of purchased items. We manipulated the presence or absence of the narrowcasting technology as noted previously and ran the experiment over two weeks (cf. Mondays, when the store is

[1] The results also hold when we use three items from Seiders et al.'s (2007) scale of benefit convenience ($F_{(1,227)}$ = 3.79; p = .05) or general measures of service convenience ($F_{(1,227)}$ = 6.89; p = .01) as the dependent variable.

closed). To avoid potential confounds, we selected two weeks in May 2018 in which there were no new promotions, changes in assortment, or special events in the shop. We established 2-hour windows, from 11:00 am to 1:00 pm and then from 3:00 pm to 5:00 pm, to implement the manipulation. During these timeslots, customer order numbers were not displayed on the screens, and the employee at the counter called the numbers of ready orders (high convenience technology absent) *or* the numbers remained regularly displayed on the screens (high convenience technology present). We rotated the manipulation in the timeslots to avoid any confounds related to the time of the day, as detailed in Table 1.

-Insert Table 1 about here-

From the retailer, we collected the observed dependent variables for each customer: the amount spent and the number of purchased items during the manipulation. The descriptive statistics for this and our subsequent studies are in Table 2.

Results

At first glance, the customer spending variable appears heavily right-skewed (sk = 3.6), with several outliers in the upper tail of the distribution (mean = 155; 75th percentile = 1.014). Therefore, we start by predicting the log of customer spending. The results of an ordinary least squares (OLS) analysis with technology present/absent as the independent variable and day and hour fixed effects reveal that the narrowcasting technology had a negative effect on customer spending (b = -.18; p = .082). A log transformation can adjust skewness, but it is less powerful for correcting the bias produced by influential outliers (Cook and Wang 1983). Noting the presence of influential cases even after our log transformation, we therefore turn to an estimation procedure based on robust regression, which has a long tradition in marketing (Mahajan, Sharma, and Wind 1984). The robust regression does not transform the original variable but rather inversely weights each case, according to its level of influence on the estimates, such that it minimizes the influence of outliers (Berk 1990). The

results of this robust regression reveal that, when customer order numbers are displayed on screens (technology present), customer average spending is significantly lower than in the technology absent condition (b = -22, p = .036). With a marginal analysis, we also find that estimated average spending drops from 126€ to 104€ in the presence of the narrowcasting technology. The effect holds when we exclude the day and hour fixed effects too (b = -16, p = .016).

For the number of purchased items, we first use a Poisson regression; the estimation results indicate that the narrowcasting technology reduces the number of products bought (b = -.12, p = .080). To address the extreme skewed and kurtotic distribution of the number of purchased items, we define the original ordinal variable according to a scale with three levels: 1 = 1 item bought (41%), 2 = 2 or 3 items bought (38%), and 3 = more than 3 items bought (21%). With an ordinal Probit model, we replicate the results; the presence of the high convenience technology reduces the number of purchased items (b = -.25, p = .042). A marginal analysis suggests that the technology increases the chance people buy only one product by 9.6% (p = .041) and lowers the chances that they buy two or three products, by 2.5% (p = .048), or more than three products, by 7.1% (p = .042).

On weekends, more customers visit the shop (464 vs. 315 in our study period), so the effect of the narrowcasting technology could differ on those days. We checked this possibility by replicating our analysis after adding a dummy variable for weekend days (Fridays, Saturdays, Sundays). However, it does not interact with the presence of the high convenience technology to predict customer spending (b = -.77, p > .1) or the number of purchased items (b = -.07, p > .1). The negative effect of the narrowcasting technology on customer shopping behavior thus appears relatively homogeneous throughout the week. When we consider only Saturdays and Sundays as the weekend, it still does not interact with the presence of the technology to predict customer spending (b = -9.8, p > .1), though there is

a negative interaction effect (b = -.35; p = .033) on number of purchased items. This notable result shows that even when customers seemingly should benefit more from the high convenience technologies (i.e., crowded store), a negative effect persists and even may be stronger.

Overall, the Study 1 findings provide empirical evidence that a narrowcasting technology does not enhance purchase behaviors in a personal service setting. When randomly exposed to shopping conditions involving the presence or absence of such a technology, consumers in our study spend more and buy more products when the technology is absent. We designed and executed Study 1 with great care, but it still is limited in terms of the evidence it can provide. The data only include customers who make a purchase; it was not possible to track the behavior of customers who did not purchase. Thus, we cannot predict whether the manipulation affects customers' probability of buying. In addition, this study relies on observed measures provided by the service provider, which do not reveal how the narrowcasting technology might influence customer–employee interactions, FLEs' extra-role behaviors, or customers' perceptions of their experience. With Studies 2a and 2b, we aim to address these limitations.

STUDY 2A: HIGH CONVENIENCE TECHNOLOGY AND CUSTOMER–EMPLOYEE INTERACTIONS

Procedure and Design

To investigate our intuition that the narrowcasting technology might have (unintended) effects on how FLEs interact with customers, we replicated the Study 1 manipulation during a weekend in June 2018, but for this study, a researcher stood near the service counter the entire time the store was open and asked each customer who approached how much he or she interacted with employees while waiting for their orders. She directly observed customers' shopping behavior too. To measure perceived interaction levels, we used a single-item scale

(1 = "I didn't interact with employees" to 7 = "I interacted with employee the whole time"). We obtained usable information from 103 customers; the descriptive statistics are in Table 2.

Results and Discussion

An ANOVA with the interaction level as the dependent variable and the presence of the narrowcasting technology as the independent variable reveals a significant effect ($F_{(1,101)}$ = 16.9; p = .000). The presence of the technology reduces the level of perceived interaction from 3.22 to 1.81. According to this initial evidence, the narrowcasting technology modifies the way customers and employees interact, such that when it is present, it acts like a sort of barrier to the interaction of customers and FLEs. If the numbers are not displayed on digital screens (technology absent), the FLEs interact with customers waiting near the service counter. The direct observation findings indicate that they also perform extra-role behaviors during these interactions, by entertaining and assisting customers, including helping customers find additional clothes and accessories. In the technology present conditions though, these interactions dramatically decrease, because customers walk around the store while they wait for their orders, and FLEs do not have the opportunity or motive to display their extra-role behaviors.

STUDY 2B: EFFECTS OF HIGH CONVENIENCE TECHNOLOGY AND MEDIATING ROLE OF FLEs' EXTRA-ROLE BEHAVIOR

Procedure and Design

In Study 2b, we directly investigate a possible mediating role of FLEs' extra-role behavior; we also seek to replicate the effect of the narrowcasting technology on customer shopping behavior, by investigating customers who do not buy. That is, we estimate the probability that customers spend some non-zero amount of money and determine if that probability differs across conditions. Finally, we gather customer perceptions of FLEs' extra-

role behavior to test for its potential mediating role in explaining the effect of the narrowcasting technology on both customers' shopping behavior and their experience.

The process mimics the previous manipulation, conducted in the same store during two weekends in June 2018. To measure customer experience and FLEs' extra-role behavior, we intercepted and surveyed customers at the end of their shopping activity. Thus we could collect information about the shopping behaviors of a sample of non-buyers, as well as gather perceptual measures of FLEs' extra-role behavior and customer experience. To exclude mere visitors who might not have noticed the technology, we surveyed only those shoppers who indicated they had requested at least one item from the warehouse. We obtained usable information from 186 customers, 97 in the technology present condition and 89 in the technology absent condition. Among these respondents, 144 (77%) made purchases and 42 (23%) did not. Thus, our survey effort captured a sizable and relevant number of potential customers; considering that the total number of purchasers was 335 over two weekends in Study 1, we estimate that our survey effort intercepted about 43% (144/335) of the total purchasers in the two weekends during which we executed Study 2b.

The post-purchase survey design cannot guarantee complete randomization of the participants, but we also are relatively unconcerned with the potential biases associated with this situation. First, the proportion of customers who refused to participate in the survey was equivalent across conditions (37.8% vs. 36.2%, $z = .41$, $p = .679$). Second, the survey participants did not differ in their average age ($t = -1.51$, $p = .132$) or education level ($\chi^2_{(3)} = 1.39$, $p = .707$) across conditions. Third, a formal non-response bias test (Armstrong and Overton 1977) reveals no clear differences between early (first quartile) and late (fourth quartile) respondents on the core constructs of customer spending ($\chi^2_{(2)} = 4.80$, $p = .091$), customer experience ($t = -.21$, $p = .902$), or FLEs' extra-role behavior ($\chi^2_{(1)} = .15$, $p = .70$).

Measures

The key dependent variable is the amount of money spent and number of purchases. Because these data were self-reported, at the end of the shopping experience, we also sought to minimize the chance of non-response due to privacy issues, so we use an ordinal scale with three levels for the amount spent: 0 = 0€; 1 = less than 100€ (which is approximately the median value spent in Study 1); and 2 = more than 100€. We use a similar procedure to collect self-reports of the number of items purchased, for which the ordinal scale includes 0 = 0 items (32%), 1 = 1 item (23%), and 2 = more than 1 item (45%).

For the perceived customer experience dependent variable, we considered several approaches that marketing scholars have developed to evaluate this concept (Lemon and Verhoef 2016). Because our purpose is to measure expectations of the experience in a specific context, we adapt a subset of eight items from the scale developed by Schouten, McAlexander, and Koenig (2007), who investigate customer experience in the specific context of a brand community. A principal component analysis reveals that the subscale is unidimensional and offers good reliability ($\alpha = .89$).

To measure FLEs' extra-role behavior, as the core mediator, we asked customers to provide their perceptions of the extent to which FLEs performed extra-role behaviors during their shopping experience. To capture these perceptions, we use a four-item scale, adapted from Tax and Brown (1998) and Schneider et al. (1998), which measures customers' perceptions of FLEs' readiness to address special requests or issues, high level of attention to the interaction, and empathy. A principal component analysis reveals that the scale is unidimensional (eigenvalue of the first factor = 3.41; 85% of variance extracted) and exhibits high levels of reliability ($\alpha = .94$). However, the measure is left-skewed, with a median of 6, a standard deviation of 1.13, and a kurtosis index greater than 5. This floor effect might be explained by the importance of FLE–customer interactions in a personal service setting. For empirical reasons, we thus dichotomize the original scale with a median split, such that we

account for the presence of FLEs' extra-role behavior only if customer perceptions are strictly above the median level (37%). Despite the loss of information associated with such dichotomizing, the practice is justified in the presence of extreme skewness in the distribution of the original metric, which can lead to significant bias (MacCallum et al. 2002). As noted, Appendix B contains all the measurement instruments.

Finally, we collected customers' demographic information (age and education) and also leveraged an opportunity to gather further evidence about the personal nature of the service setting. That is, we asked respondents to rate the relevance of their interaction with FLEs in that store, using a single-item scale from Parasuraman (2000; 1 = "very low" to 7 = "very high"). The overall mean is 6.02, with no difference across conditions (t = 1.08, p = .28), suggesting that the store represent a personal service setting, in which FLE–customer interactions are central to the shopping experience. The descriptive statistics are in Table 2.

-Insert Table 2 about here-

As in Study 1, we included *day* and *hour* dummies as control variables for the estimation, but the results do not change if we remove them.

Results: Single Effects

Considering the ordered nature of our dependent variable, we began by testing the main effect of the narrowcasting technology on customer spending using an ordinal Probit model with robust standard errors. The results replicate the evidence from Study 1, though in this case, we include the possibility of non-purchase (b = -.48, p = .005). With a marginal analysis, we find that the presence of the technology increases the chance that people do not buy by 14% (p = .004) and the chance that they spend less than the median value by 4% (p = .012). Conversely, it lowers the probability that customers spend more money by 18% (p = .003). A similar pattern of results emerges for the number of purchased items; the presence

of the technology has a significant negative effect on consumption volume (b = -.44, *p* = .019).

An OLS regression with customer experience as the dependent variable and the presence of the narrowcasting technology as the predictor reveals a significant negative effect (b = -.49, *p* = .008); the high convenience technology worsens customers' perceptions of the experience. Then an ordinal Probit model with the ordinal measure of customer spending as the dependent variable and customer experience as the key predictor reveals the expected positive effect (b = .21, *p* = .005). We also replicate this analysis with the number of products bought and find the same positive effect (b = .21, *p* = .01).

Robustness Checks: Possible Confounders

With further analyses, we address some potential confounds. First, we implemented the manipulation 8 months after the introduction of the narrowcasting technology, so we checked whether customers exposed to the technology absent condition perceived it as a service failure. In that case, the differential effects of the technology present/absent conditions might have stemmed from failure-related reasons. Specifically, we asked the participants about the extent to which they felt they had suffered a service failure during their shopping experience, on a scale from 1 to 7. The average value did not differ across groups (2.10 vs. 2.05; t = .25, *p* = .80), so the behavioral differences across groups are unlikely to depend on different evaluations of a service failure related to the technology.

Second, narrowcasting technology primarily is designed to help customers save time and reallocate their time to other activities, so we checked how much time customers perceived that they waited for their orders (1 to 7 scale). Again, we find no difference in perceived waiting time across the two conditions (3.45 vs. 3.23; t = 1.05, *p* = .30). Even if customers might save time in the technology present condition, they do not assign any particular penalty to the experience if they must spend time queuing at the counter.

Third, we investigate whether the effect of the narrowcasting technology on the customer experience and shopping behavior depends on visit frequency. We asked respondents how often they visit the store, on a scale from 1 ("this is the first time") to 7 ("everyday"). Visit frequency does not moderate the effect of the technology though, whether on customer spending (b = .14, $p > .1$), number of purchased items (b = .089, $p > .1$), or the customer experience (b = .074, $p > .1$). The results thus appear independent of the regularity with which customers shop at this store.

FLEs' Extra-Role Behavior as Mediator

We test a mediation model, with the prediction that customers' perceptions of FLEs' extra-role behavior can help explain the effect of the narrowcasting technology on customers' shopping behavior. Assessing a mediation model in this case requires some care, because conventional methods to test indirect effects cannot accommodate our ordinal dependent variable. Therefore, we turn to the KHB routine (Kohler, Karlson, and Holm 2011), which can decompose the total effect in its direct and indirect components for various nonlinear models. This mediation analysis with bias-corrected bootstrapping (1,000 replications) reveals a significant indirect effect of FLEs' extra-role behavior on the relationship between the presence of the narrowcasting technology and our ordinal measure of customer spending (b = -.09, $p = .07$). In terms of effect size, it captures about 18% of the total (negative) effect of the technology. With regard to the number of purchased items, FLEs' extra-role behavior also has a significant indirect effect (b = -.12, $p = .04$) explaining about 28% of the effect. Finally, the mediating role of FLEs' extra-role behavior in the relationship between the presence of the narrowcasting technology and customer experience reveals a significant indirect effect (b = -.12, $p = .05$), explaining about 27% of the effect on customer experience.

These results are consistent with our theoretical reasoning; high convenience technologies, probably unintentionally, lead FLEs not to perform their extra-role behaviors,

which worsens both the customer experience and their shopping behaviors, beyond any potential effect on convenience. However, we measure FLEs' extra-role behaviors using customer perceptions, and Study 2b does not constitute a purely randomized design, which means we cannot assess the causality of our mediation model. We address this limitation in Study 3.

STUDY 3: FLEs' EXTRA-ROLE BEHAVIOR, MODERATION OF PROCESS

Methodology

In Study 2b, we measured FLEs' extra-role behavior and used traditional mediation analyses to obtain empirical evidence that the variable represents a carryover mechanism, from the negative influence of the narrowcasting technology to customer shopping behavior. While informative, this approach cannot establish causal effects conclusively (Stone-Romero and Rosopa 2008). Therefore, in Study 3, we adopt a process-by-moderation approach (Vancouver and Carlson 2015) and assess the mediation by testing the interaction effect of a moderating variable that strongly affects the mediator and the independent variable. This process-by-moderation approach can be used in cases in which the mediator cannot be manipulated. Furthermore, once the relationship between the moderator and the mediator has been theoretically and practically established, we no longer need to measure the mediator. These traits are crucial for our study, because manipulating the mediating variable (i.e., preventing FLEs from performing extra-role behavior) would be impractical and ineffective. Moreover, generating enough variance with just self-generated measures is complicated.

We include time pressure as moderator variable to assess our mediation model, which we operationalize as specific timeslots, namely, whether customers are shopping during their lunch breaks or not. During their lunch breaks, consumers generally have limited time to shop and thus behave as if they are under time pressure. According to prior service marketing literature, consumers under time pressure tend to focus on products and are less responsive to

environmental stimuli (Inman, Winer, and Ferraro 2009). They also tend to prefer that the FLEs limit their behavior to instrumental actions (i.e., in-role behavior) rather than perform extra-role behaviors (Marinova et al. 2017 Menon and Dubé 2007). Accordingly, we expect the time pressure (shopping during lunch break) to nullify the positive effect of FLEs' extra-role behavior on customer shopping behavior, because this outcome instead should be dominated by efficiency and convenience goals (as in encounter-like service settings). We thus expect an interaction effect, between the presence of the narrowcasting technology and time pressure, such that during the lunch break, its presence has a positive effect on customer shopping behavior, but during other periods, the effect is negative. This pattern of findings would help establish FLEs' extra-role behavior as central explanation for the effects we have reported thus far. However, before proceeding with the manipulation, we test to confirm whether customers who shop during their lunch breaks are under time pressure and want FLEs to limit their extra-role behavior.

Pretest: Effect of Shopping During Lunch Breaks on Time Pressure

Three hundred ninety-nine participants (48.37% women; average age: 30.56 years) drawn from the Prolific panel participated in exchange for $0.44. They were randomly assigned to the "shopping during lunch break" or "shopping during other time of the day" conditions and asked to read a scenario carefully and respond to some questions. In the lunch break condition, they read: "Imagine you want to shop for some clothes and decide to stop at a fashion store during your lunch break"; otherwise, they read: "Imagine you want to shop for some clothes and decide to stop at a fashion store either in the morning or in the afternoon."

We collected measures of (1) time pressure, with three items adapted from Vermeir and Van Kenhove's (2005) scale of time pressure; (2) expectations of FLEs' extra-role behaviors, with the scale we used in Study 2a; and (3) fashion category involvement, using three items from Goldsmith and Emmert (1991) (see Appendix B). The principal component

analysis indicates that the scales are unidimensional, with high levels of reliability (time pressure $\alpha = .92$; extra-role behavior $\alpha = .80$; fashion category involvement $\alpha = .80$).

The related ANOVA reveals, as expected, a statistically significant difference in consumer time pressure between conditions ($F_{(1, 397)} = 180.87$; $p = .001$), so customers who shop during their lunch break feel more time pressure than customers shopping at other times ($M_{Lunch} = 5.77$; $M_{Other} = 4.02$). Regarding FLEs' extra-role behavior, we confirm that customers shopping during their lunch break do not expect FLEs to perform these behaviors ($M_{Lunch} = 4.59$; $M_{Other} = 5.01$; $F(1, 397) = 14.74$; $p = .001$). These effects of shopping during a lunch break on FLEs' extra-role behaviors do not depend on customers' level of involvement in fashion ($F_{(1,395)} = .18$; $p > .1$), though we find a negative interaction effect on time pressure ($F_{(1,395)} = 8.36$; $p = .004$). That is, when consumers are highly involved in fashion, implying their greater knowledge of fashion products, the positive effect of shopping during their lunch break on perceived time pressures is weaker. Therefore, shopping during a lunch break generally causes customers to sense greater time pressure, and it should act as a suppressing moderator of our core mediator. We consider it a good candidate to test the causal mechanism associated with FLEs' extra-role behavior.

Main Study: Procedure and Design

In this study, we implement the process-by-moderation approach to test the mediating role of FLEs' extra-role behavior. We manipulate the presence of the technology for 8 days across three time slots (morning: 11:00 am–1:00 pm; lunch break: 1:00–3:00 pm; afternoon: 3:00–5:00 pm). During the experimental period, the overall number of customers did not vary substantially across the three timeslots: morning, 270 customers; lunch break, 234 customers; and afternoon, 223 customers. As in Study 1 we use two observed measures of shopping behavior: the amount of money each customer spent and the number of purchased items. The descriptive statistics are in Table 2.

Results

As in Study 1, customer spending is right-skewed (p = .001). Using robust regressions, we uncover a significant interaction effect between the presence of the narrowcasting technology and shopping during lunch breaks on customer spending (b = 76, p = .004). When customers are under time pressure and expect less extra-role behavior from FLEs, the presence of the technology increases customer spending by 37€ (p = .033); if they are not under time pressures, it reduces customer spending by 39€ (p = .003).[2] We graph these results in Figure 2.

-Insert Figure 2 about here-

Regarding the effect on the number of purchased items, we use an ordinal Probit regression and find a significant interaction between the technology and shopping during lunch breaks (b = .96, p = .002). In the marginal analysis, we note that when customers shop in the morning or the afternoon, the presence of the technology increases their probability of buying only one product by 15% (p = .008) but reduces the probability that they buy 2 or 3 products by 4.27% (p = .007) and the probability that they buy more than 3 products by 10.63% (p = .01). Among customers shopping during their lunch breaks though, the technology reduces the chances of buying only one product by 19.33% (p = .004), has no effect on the probability of buying 2 or 3 products (b = .02, p > .1), and increases the chances of buying more than 3 products by 17.86% (p = .005) (see Figure 3).

-Insert Figure 3 about here-

These results provide strong support for the mediation model we tested in Study 2b. In particular, when customers are under time pressure, they prefer FLEs to not perform extra-role behavior, nullifying the positive effect of such behaviors on customer shopping behavior.

[2] The interaction effect between the narrowcasting technology and shopping during lunch breaks is significant when we use the log-transformation of customer spending as the dependent variable too (b = .61, p = .019).

The combination of the presence or absence of the narrowcasting technology with different time pressure conditions produces a significant interaction effect, and because time pressure correlates with the salience of FLEs extra-role behavior, this interaction affirms our proposed causal mechanism. When customers face high (low) time pressure, high convenience technologies exert positive (negative) overall effects on their shopping behaviors, because the customers regard FLEs' extra-role behavior as less (more) salient for their experience.

GENERAL DISCUSSION AND IMPLICATIONS

With this research, we provide to the best of our knowledge, the first investigation of the effect of high-convenience technologies on customer shopping behavior in a personal service setting, in which the FLE–customer relationship is foundational to the service experience. Our research propositions anticipate the possible lack of benefits of such technologies, because their introduction can inadvertently weaken those critical relationships (by reducing FLEs' extra-role behavior). To investigate these propositions, we conducted three field studies (experiments and customer intercept surveys) in the personal service setting of a fashion retail store. In Study 1, with a controlled field experiment, we identify a non-positive effect of the high convenience technology (narrowcasting) on the volume and value of customer purchases. In Studies 2a and 2b, we also find evidence that the technology inadvertently reduces FLE–customer interactions, inhibiting employees from performing their valuable extra-role behaviors, reducing customer shopping behaviors, and harming their experiences. In Study 3, we establish the role of FLEs' extra-role behavior as a carryover mechanism: When the technology operates in a context in which FLEs' extra-role behaviors are not desirable (e.g., when the customer faces time pressures), the overall effect of the technology on customer shopping transforms back, from negative to positive.

These results inform marketing literature pertaining to in-store technology infusions and address recent calls for more attention to the complex, contingent nature of such

technology transformations in retail settings (Grewal et al. 2019). Failing to consider the sociotechnical, systemic nature of service settings can lead to unexpected consequences of technology introductions. The main cause of such unplanned effects is the exclusion of considerations of how "investments in or resources in technology, marketing, and facilities, also impacts employees" (Schneider and Bowen 2019, p. 4). Employees' roles vary widely across service settings, so introducing technologies solely because they have proven successful in other contexts is not sufficient. If service managers focus on immediate, specific, and desired outcomes of a technology, without attending to potential indirect effects on other elements of the service system (e.g., other actors), the overall effect may deviate from expectations and even be detrimental (Baxter and Sommerville 2011). The increasing range of technological options available for service transformation, such as those driven by AI solutions, means the risk of these potential mismatches is not trivial. As our study reveals, high convenience technologies that aim to streamline and facilitate customer shopping experiences actually may end up increasing their complexity in personal service settings.

Our study also contributes to the emerging organizational frontline research stream (Singh et al. 2019), at the interface of marketing and organizational behavior literature. Our findings support the view that complex sets of interactions among technology, employees, and customers, occurring at the service frontline, ultimately determine service outputs. Assuming that a technology can efficiently substitute for human relations in service delivery is naïve at best. This point deserves particular attention, because in various service settings, customers seek "curated" personal experiences, not necessarily efficient ones (Ozkok et al. 2019). Adding AI solutions to such settings likely will increase (instead of reduce) the complexity of the service frontline, thereby demanding more coordination efforts from a managerial perspective. This prediction is exactly what we document in our analysis; the focal technology is not addressed to FLEs, but its final effect conflicted with expectations

precisely because it involved an unplanned change in FLEs' behavior. A meta-analysis (Hogreve et al. 2017) has indicated that marketing literature signals an incomplete understanding of the "multidimensional nature" of service employee behavior, suggesting the need for simultaneous considerations of both customer-oriented and self- or company-oriented behaviors by these employees (Bowen and Schneider 2014). Our analysis sheds light on this question, by illustrating a case in which a technology (inadvertently) makes employees more prone to self-oriented behavior, to the detriment of their far more important customer-oriented behavior.

Our findings thus suggest some practical insights for managers investing in technological solutions designed to optimize their service frontlines. These managers must realize that even if such technologies are designed and claim to modify a specific element of the frontline (e.g., customer behavior), they are likely to alter other elements of the service system too (e.g., FLEs' behavior). Recognition and understanding of this interactive complexity should be at the top of managerial agendas when considering a new technology introduction, to avoid unplanned and unnecessarily adverse effects.

LIMITATIONS AND FURTHER RESEARCH

We sought to ensure good consistency in the design and execution of our studies, though our research, as a first effort in a relatively unstudied and novel field, inevitably has limitations. We briefly discuss two of them and how they might be addressed with continued research. First, our studies refer to a single, specific consumption category (fashion apparel), whereas in other categories, additional variables might determine the impact of high convenience technologies on customers' shopping behavior. For example, it would be interesting to investigate the effect of narrowcasting technologies in more utilitarian settings, such as investment banking, to determine if customers evaluate FLEs' extra-role behavior

similarly, as well as in settings in which FLEs provide the service content directly, such as teachers in an education setting.

Second, we tested the effect of one type of high convenience technology; it is not interactive. Yet we recognize that customers' active participation with a technology might attenuate the negative effects we find; such investigations likely need to include other explanatory variables. Service marketing scholars have offered some theoretical propositions about technologies that might provide a sense of social presence (Grewal et al. 2019; van Doorn et al. 2017). Testing them would constitute a useful extension to existing knowledge about the multifaceted potential impacts of future technology on service encounters.

REFERENCES

Armstrong, J. Scott, and Terry S. Overton (1977), "Estimating Nonresponse Bias in Mail Surveys," *Journal of Marketing Research,* 14(3), 396-402.

Baxter, Gordon, and Ian Sommerville (2011), "Socio-Technical Systems: From Design Methods to Systems Engineering," *Interacting with Computers,* 23 (1), 4-17.

Berk, Richard (1990), "A Primer on Robust Regression," in *Modern Methods of Data Analysis,* Scott J. Long and John Fox, eds. Newbury Park, CA: Sage, 291-324.

Berry, Leonard L., Kathleen Seiders, and Dhruv Grewal (2002), "Understanding Service Convenience," *Journal of Marketing,* 66 (3), 1-17.

Bettencourt, Lance A., Kevin P. Gwinner, and Matthew L. Meuter (2001), "A Comparison of Attitude, Personality, and Knowledge Predictors of Service-Oriented Organizational Citizenship Behaviors," *Journal of Applied Psychology,* 86 (1), 29-41.

Bitner, Mary Jo, Amy L. Ostrom, and Matthew L. Meuter (2002), "Implementing Successful Self-Service Technologies," *Academy of Management Perspectives,* 16 (4), 96-108.

Bitner, Mary Jo, Stephen W. Brown, and Matthew L. Meuter (2000), "Technology Infusion in Service Encounters," *Journal of the Academy of Marketing Science,* 28(1), 138-149.

Botti, Simona, Ann L. McGill, and Sheena S. Iyengar (2003), "Preference for Control and Its Effect on the Evaluation of Consumption Experiences," in *Advances in Consumer Research,* Vol. 30, Punam Anand Keller and Dennis W. Rook, eds. Valdosta, GA: Association for Consumer Research, 127-28.

Bowen, David E., and Benjamin Schneider (2014), "A Service Climate Synthesis and Future Research Agenda," *Journal of Service Research,* 17 (1), 5-22.

Cadwallader, Susan, Cheryl B. Jarvis, Mary Jo Bitner, and Amy L. Ostrom (2010), "Frontline Employee Motivation to Participate in Service Innovation Implementation," *Journal of the Academy of Marketing Science,* 38 (2), 219-239.

Cook, R. Dennis, and P. C. Wang (1983), "Transformations and Influential Cases in Regression," *Technometrics,* 25 (4), 337-343.

Di Mascio, Rita (2010), "The Service Models of Frontline Employees," *Journal of Marketing,* 74 (4), 63-80.

Farquhar, Jillian Dawes, and Jennifer Rowley (2009), "Convenience: A Services Perspective," *Marketing Theory*, 9 (4), 425-438.

Floor, Ko (2006), *Branding a Store: How to Build Successful Retail Brands in a Changing Marketplace*. London: Kogan Page Publishers.

Froehle, Craig M., and Aleda V. Roth (2004), "New Measurement Scales for Evaluating Perceptions of the Technology-Mediated Customer Service Experience," *Journal of Operations Management*, 22 (1), 1-21.

Giebelhausen, Michael, Stacey G. Robinson, Nancy J. Sirianni, and Michael K. Brady (2014), "Touch Versus Tech: When Technology Functions as a Barrier or a Benefit to Service Encounters," *Journal of Marketing*, 78 (4), 113-124.

Goldsmith, Ronald E., and Janelle Emmert (1991), "Measuring Product Category Involvement: A Multitrait-Multimethod Study," *Journal of Business Research*, 23 (4), 363-37.

Gremler, Dwayne D., and Kevin P. Gwinner (2008), "Rapport-Building Behaviors Used by Retail Employees," *Journal of Retailing*, 84 (3), 308-324.

Grewal, Dhruv, Stephanie M. Noble, Anne L. Roggeveen, and Jens Nordfalt (2019), "The Future of In-Store Technology," *Journal of the Academy of Marketing Science*, 48 (1), 96-113.

Gutek, Barbara A. (1999), "The Social Psychology of Service Interactions," *Journal of Social Issues*, 55 (3), 603-617.

Heller, Jonas, Mathew Chylinski, Ko de Ruyter, Dominik Mahr, and Debbie I. Keeling (2019), "Touching the Untouchable: Exploring Multi-Sensory Augmented Reality in the Context of Online Retailing," *Journal of Retailing*, 95 (4), 219-234.

Hogreve, Jens, Anja Iseke, Klaus Derfuss, and Tonnjes Eller (2017), "The Service–Profit Chain: A Meta-Analytic Test of a Comprehensive Theoretical Framework," *Journal of Marketing*, 81(3), 41-61.

Humphrey, Ronald H., and Blake Ashforth (1994), "Cognitive Scripts and Prototypes in Service Encounters," *Advances in Services Marketing and Management*, 3(C),175-199.

Inman, J. Jeffrey, Russell S. Winer, and Rosellina Ferraro (2009), "The Interplay Among Category Characteristics, Customer Characteristics, and Customer Activities on In-Store Decision Making," *Journal of Marketing* 73 (5), 19-29.

Johnson, Devon S., Fleura Bardhi, and Dan T. Dunn (2008), "Understanding How Technology Paradoxes Affect Customer Satisfaction with Self-Service Technology: The Role of Performance Ambiguity and Trust in Technology," *Psychology & Marketing*, 25 (5), 416-443.

Kohler, Ulrich, Kristian Bernt Karlson, and Anders Holm (2011), "Comparing Coefficients of Nested Nonlinear Probability Models," *The Stata Journal,* 11 (3), 420-438.

Larivière, Bart, David Bowen, Tor W. Andreassen, Werner Kunz, Nancy J. Sirianni, Chris Voss, Nancy V. Wünderlich, and Arne De Keyser (2017), "'Service Encounter 2.0': An Investigation Into the Roles of Technology, Employees and Customers," *Journal of Business Research*, 79 (October), 238-246.

Lemon, Katherine N., and Peter C. Verhoef (2016), "Understanding Customer Experience Throughout the Customer Journey," *Journal of Marketing*, 80(6), 69-96.

MacCallum, Robert C, Shaobo Zhang, Kristopher J. Preacher, and Derek D. Rucker (2002), "On the Practice of Dichotomization of Quantitative Variables," *Psychological Methods*, 7 (1), 19-40.

Mahajan, Vijay, Subhash Sharma, and Yoram Wind (1984), "Parameter Estimation in Marketing Models in the Presence of Influential Response Data: Robust Regression and Applications," *Journal of Marketing Research,* 21 (3), 268-277.

Mani, Zied, and Inès Chouk (2017), "Drivers of Consumers' Resistance to Smart Products," *Journal of Marketing Management* 33 (1/2), 76-97.

Marinova, Detelina, Ko de Ruyter, Ming-Hui Huang, Matthew L. Meuter, and Goutam Challagalla (2017), "Getting Smart: Learning from Technology-Empowered Frontline Interactions," *Journal of Service Research*, 20 (1), 29-42.

Mende, Martin and Stephanie M. Noble (2019), "Retail Apocalypse or Golden Opportunity for Retail Frontline Management?" *Journal of Retailing,* 95 (2), 84-89.

Menon, Kalyani, and Laurette Dubé (2007), "The Effect of Emotional Provider Support on Angry Versus Anxious Consumers," *International Journal of Research in Marketing*, 24 (3), 268-275.

Metzger, Miriam J. (2017), "Broadcasting Versus Narrowcasting," *The Oxford Handbook of Political Communication*, https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199793471.001.0001/oxfordhb-9780199793471-e-62

Netemeyer, Richard G., and James G. Maxham III (2007), "Employee Versus Supervisor Ratings of Performance in the Retail Customer Service Sector: Differences in Predictive Validity for Customer Outcomes," *Journal of Retailing*, 83 (1), 131-145.

Ozkok, Ozlem, Jagdip Singh, Kwanghui Lim, and Simon J. Bell (2019), "Service Innovation from the Frontlines in Customer-Centric Organizations," in *Handbook on Customer Centricity*. Northampton, MA: Edward Elgar Publishing, 79-107.

Parasuraman, Ananthanarayanan (2000), "Technology Readiness Index (TRI) a Multiple-Item Scale to Measure Readiness to Embrace New Technologies," *Journal of Service Research*, 2 (4), 307-320.

Perrow, Charles (1982), "The President's Commission and the Normal Accident," in *Accident at Three Mile Island: The Human Dimensions*, David Sils, C. P. Wolf, and Vivien B. Shelanski, eds. Boulder CO: Westview Press, 173–184.

Quint, Susan (2017), "The SEAM Four-Leaf Clover, Revisited," *The Theory and Practice of Socio-Economic Management*, 2 (1), 30-41.

Rafaeli, Anat, Daniel Altman, Dwayne D. Gremler, Ming-Hui Huang, Dhruv Grewal, Bala Iyer, Ananthanarayanan Parasuraman, and Ko de Ruyter (2017), "The Future of Frontline Research: Invited Commentaries," *Journal of Service Research*, 20 (1), 91-99.

Rust, Roland T., Katherine N. Lemon, and Valarie A. Zeithaml (2004), "Return on Marketing: Using Customer Equity to Focus Marketing Strategy," *Journal of Marketing*, 68 (1), 109-127.

Sargut, Gökçe, and Rita Gunther McGrath (2011), "Learning to Live with Complexity," *Harvard Business Review*, 89 (9), 68-76.

Schneider, Benjamin, and David E. Bowen (2019), "Perspectives on the Organizational Context of Frontlines: A Commentary," *Journal of Service Research*, 22 (1), 3-7.

Schneider, Benjamin, Susan S. White, and Michelle C. Paul (1998), "Linking Service Climate and Customer Perceptions of Service Quality: Tests of a Causal Model," *Journal of Applied Psychology*, 83(2), 150-163.

Schouten, John W., James H. McAlexander, and Harold F. Koenig (2007) "Transcendent Customer Experience and Brand Community," *Journal of the Academy of Marketing Science*, 35 (3), 357-368.

Seiders, Kathleen, Glenn B. Voss, Andrea L. Godfrey, and Dhruv Grewal (2007), "SERVCON: Development and Validation of a Multidimensional Service Convenience Scale," *Journal of the Academy of Marketing Science*, 35 (1), 144-156.

Seiders, Kathleen, Glenn B. Voss, Dhruv Grewal, and Andrea L. Godfrey (2005), "Do Satisfied Customers Buy More? Examining Moderating Influences in a Retailing Context," *Journal of Marketing*, 69 (4), 26-43.

Singh, Jandip, Karen Flaherty, Ravipreet S. Sohi, Dawn Deeter-Schmelz, Johannes Habel, Kenneth Le Meunier-FitzHugh, Avinash Malshe, Ryan Mullins, and Vincent Onyemah (2019), "Sales Profession and Professionals in the Age of Digitization and Artificial Intelligence Technologies: Concepts, Priorities, and Questions," *Journal of Personal Selling & Sales Management*, 39 (1), 2-22.

Stone-Romero, Eugene F., and Patrick J. Rosopa (2008), "The Relative Validity of Inferences about Mediation as a Function of Research Design Characteristics," *Organizational Research Methods*, 11 (2): 326-352.

Tax, Stephen S., and Stephen W. Brown (1998), "Recovering and Learning from Service Failure," *MIT Sloan Management Review*, 40 (1), 75-88.

Van Beuningen, Jacqueline, Ko De Ruyter, Martin Wetzels, and Sandra Streukens (2009), "Customer Self-efficacy in Technology-based Self-service: Assessing Between-and Within-Person Differences," *Journal of Service Research*, 11(4), 407-428.

Vancouver, Jeffrey B., and Bruce W. Carlson (2015), "All Things in Moderation, Including Tests of Mediation (at least some of the time)," *Organizational Research Methods*, 18 (1), 70-91.

Van Doorn, Jenny, Martin Mende, Stephanie M. Noble, John Hulland, Amy L. Ostrom, Dhruv Grewal, and Andrew J. Petersen (2017), "Domo Arigato Mr. Roboto:

Emergence of Automated Social Presence in Organizational Frontlines and Customers' Service Experiences," *Journal of Service Research*, 20 (1), 43-58.

Vargo, Stephen L., and Robert F. Lusch (2018) *The SAGE Handbook of Service-Dominant Logic*. Los Angeles: Sage.

Vermeir, Iris, and Patrick Van Kenhove (2005), "The Influence of Need for Closure and Perceived Time Pressure on Search Effort for Price and Promotional Information in a Grocery Shopping Context," *Psychology & Marketing*, 22 (1), 71-95.

Wells, John D., Damon E. Campbell, Joseph S. Valacich, and Mauricio Featherman, M. (2010), "The Effect of Perceived Novelty on the Adoption of Information Technology Innovations: A Risk/Reward Perspective," *Decision Sciences*, 41 (4), 813-843.

Wels-Lips, Inge, Marleen van der Ven, and Rik Pieters (1998), "Critical Services Dimensions: An Empirical Investigation Across Six Industries," *International Journal of Service Industry Management*, 9 (3), 286-309.

Williams, Patti, and Jennifer L. Aaker (2002), "Can Mixed Emotions Peacefully Coexist?" *Journal of Consumer Research*, 28 (4), 636-649.

Table 1. Timeslots in Study 1

TIMESLOTS	*FIRST WEEKEND*	*SECOND WEEKEND*
Tuesday (11 am- 1 pm)	Technology ABSENT	Technology PRESENT
Tuesday (3pm- 5pm)	Technology PRESENT	Technology ABSENT
Wednesday (11 am- 1 pm)	Technology PRESENT	Technology ABSENT
Wednesday (3pm- 5pm)	Technology ABSENT	Technology PRESENT
Thursday (11 am- 1 pm)	Technology ABSENT	Technology PRESENT
Thursday (3pm- 5pm)	Technology PRESENT	Technology ABSENT
Friday (11 am- 1 pm)	Technology PRESENT	Technology ABSENT
Friday (3pm- 5pm)	Technology ABSENT	Technology PRESENT
Saturday (11 am- 1 pm)	Technology ABSENT	Technology PRESENT
Saturday (3pm- 5pm)	Technology PRESENT	Technology ABSENT
Sunday (11 am- 1 pm)	Technology PRESENT	Technology ABSENT
Sunday (3pm- 5pm)	High convenience technology ABSENT	High convenience technology PRESENT

Table 2. Descriptive statistics of the studies

STUDY 1					
Technology Condition	Customers	Customer Spending (€)		Number of Purchased Items	
		Mean	SD	Mean	SD
ABSENT	841	159.30	176.42	2.51	2.38
PRESENT	1,037	148.88	166.51	2.42	2.16

STUDY 2A			
Technology Condition	Customers	Customer–Employee Interaction	
		Mean	SD
ABSENT	46	3.22	2.14
PRESENT	57	1.81	1.32

STUDY 2B								
Technology Condition	Customers	Customer Spending (% of customers in condition)			Customer Experience		FLEs' Extra-Role Behavior	
		0€	<100€	>100€	Mean	SD	Mean	SD
ABSENT	89	13.95%	31.40%	54.65%	4.60	1.27	5.90	1.25
PRESENT	97	30.93%	29.90%	38.18%	4.23	1.18	5.86	1.00

STUDY 3						
Technology Condition	Lunch Break	Other Time of Day	Customer Spending (€)		Number of Purchased Items	
			Mean	SD	Mean	SD
ABSENT	122	239	165.05	169.18	2.459	1.847
PRESENT	112	254	149.48	154.05	2.472	1.947

Figure 1. Overview of the studies

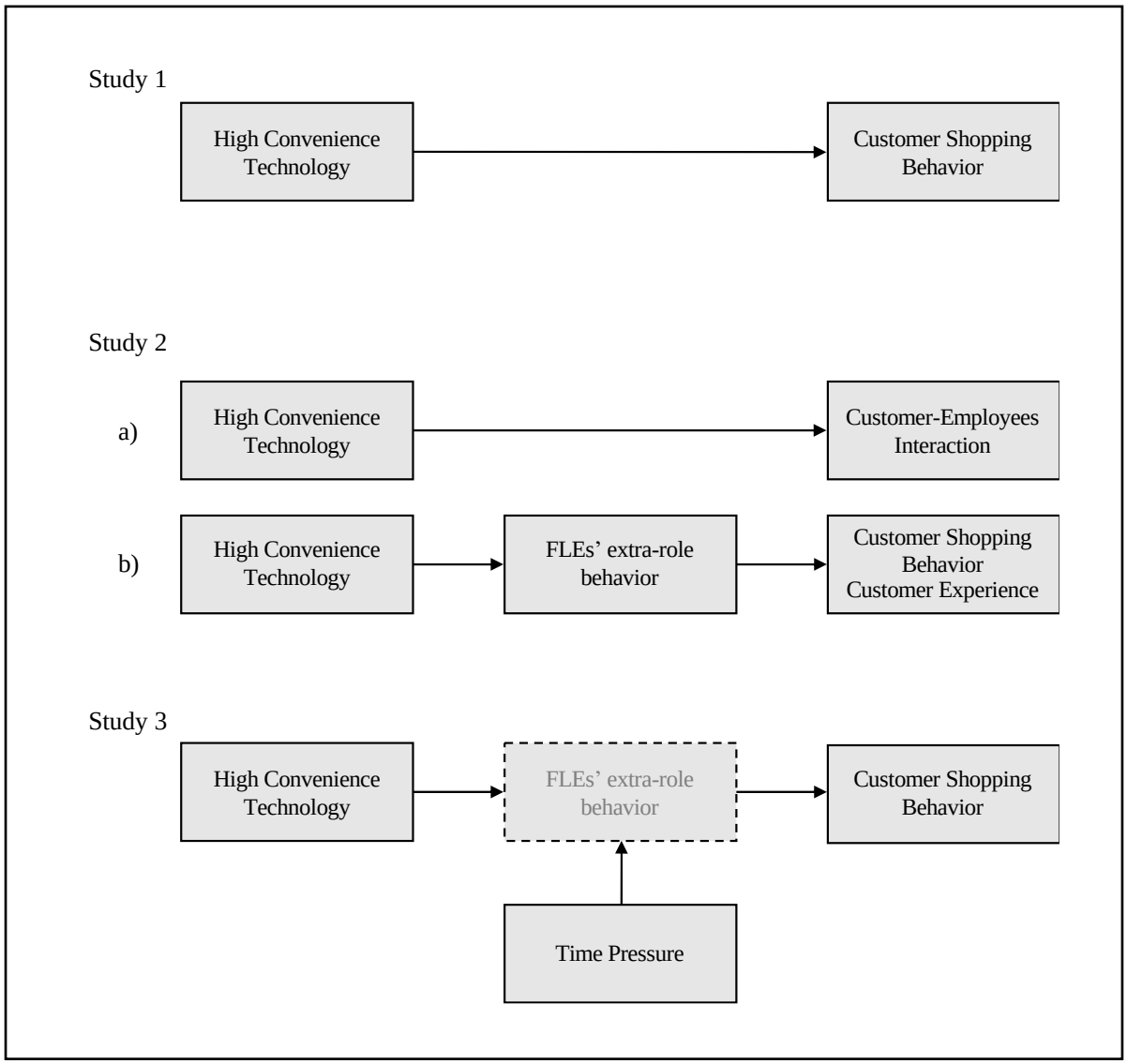


Figure 2. Interaction effect of shopping during lunch break and high convenience technology on customer spending (95% confidence intervals)

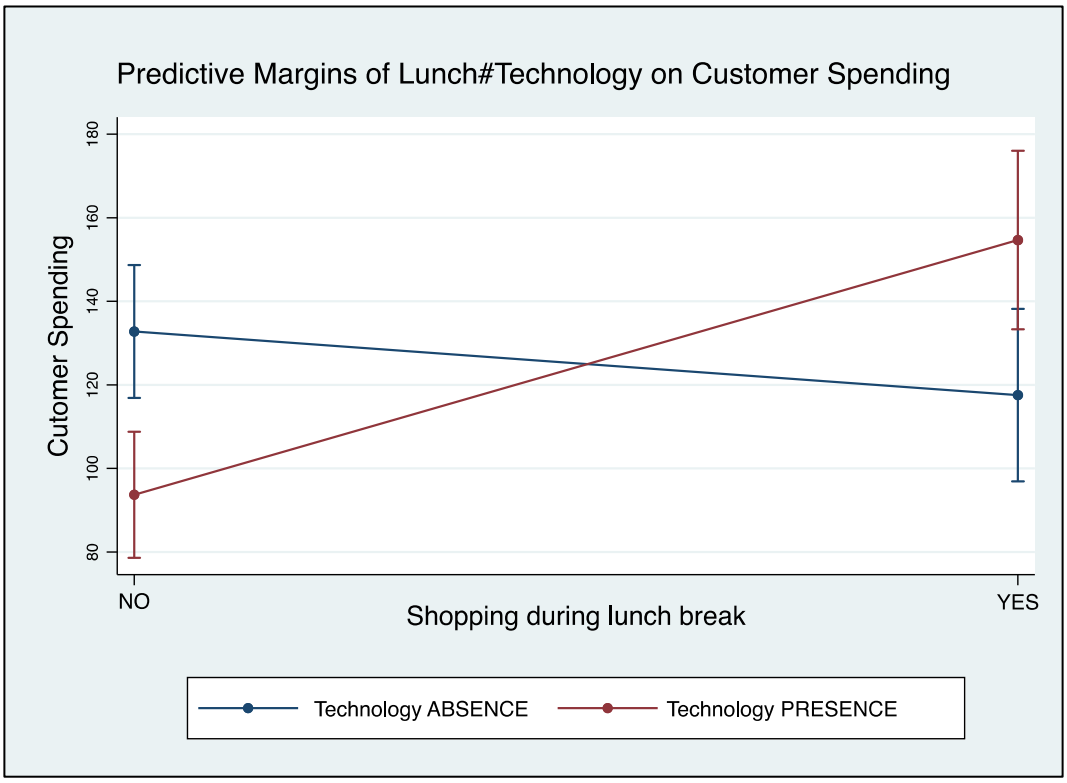
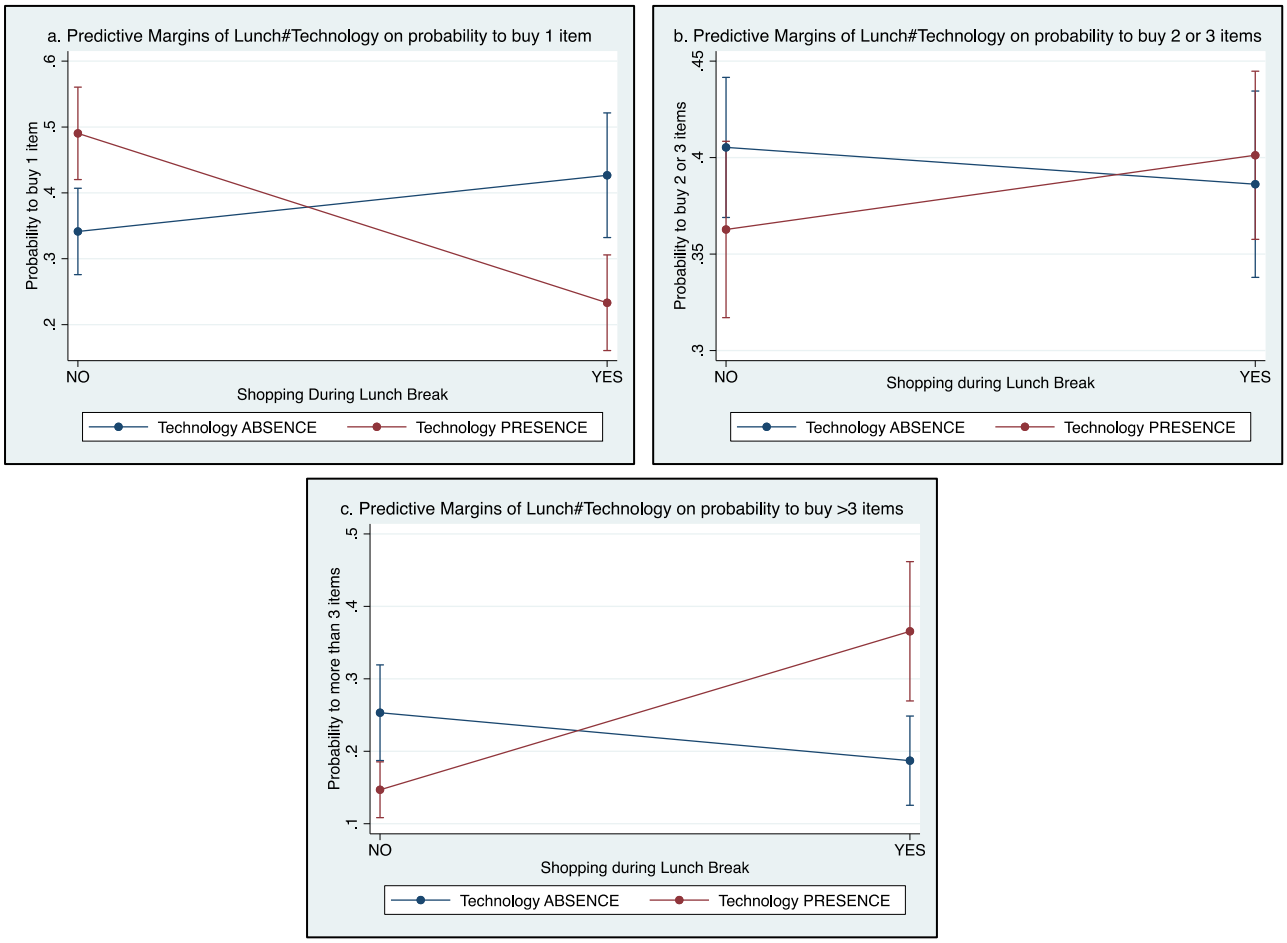


Figure 3. Interaction effect of shopping during lunch break and high convenience technology on probability to buy (a) only 1 item, (b) 2 or 3 items, and (c) more than 3 items (95% confidence intervals)



APPENDIX A

Digital screens in the two conditions

b. Digital screen in technology absent condition

a. Digital screen in technology present condition

Preliminary Study: The Effect of Narrowcasting Technology on Service Convenience.

1. Technology absent condition:

"At Fashion Z, a large fashion clothing store, only one size for each product is displayed. When you find interesting clothes, to get them in your size, you have to go to a counter and ask an employee for the size you need. The employee sends your request to the warehouse and gives you a number. At this point: You stand in line near the counter while you are waiting your turn to get your clothes. Once your order is ready, the employee calls your number and gives you the clothes so that you can continue with the shopping experience."

2. Technology present condition:

"At Fashion Z, a large fashion clothing store, only one size for each product is displayed. When you find interesting clothes, to get them in your size, you have to go to a counter and ask an employee for the size you need. The employee sends your request to the warehouse and gives you a number. At this point: You are free to walk around the store and continue shopping. Once your order is ready, your number appears on several digital screens around the shop, and then you can go back to the counter to get the clothes so that you can continue with the shopping experience."

APPENDIX B

Measurement instruments

Preliminary Study: Effect of Narrowcasting Technology on Service Convenience

Service convenience (Seiders et al. 2007) (eigenvalue of the first factor = 2.53, 84% of variance extracted; α = . 88)

- "It is easy to evaluate the merchandise at Fashion Z."
- "The merchandise I want at Fashion Z can be obtained quickly."
- "It is easy to obtain the products I am looking for at Fashion Z."

(Farquhar and Rowley 2009) (eigenvalue of the first factor = 2.56, 85.38% of variance extracted; α = .91)

- "The shopping experience at Fashion Z is smooth."
- "The shopping experience at Fashion Z is effortless."
- "The shopping experience at Fashion Z is quick."

Fashion category involvement (Goldsmith and Emmert 1991)

- "If I heard that a new designer fashion was available in the store, I would be interested enough to buy it."
- "I know the names of new designer fashion before other people do."
- "I buy clothes very often."

Study 2b

Relevance of the interaction with frontline employees in high contact setting (Parasuraman 2000)

- "While shopping in this store, the contact with employees is very important."

Customer experience (Schouten, McAlexander, and Koenig 2007)

- "Shopping in this store made me feel different (it changed my perception of me)."
- "I felt like I was having an ideal shopping experience."
- "My buying routine was new."
- "I enjoyed the shopping experience fully."
- "I learned new things as a result of this experience."
- "I'd like to have a similar shopping experience in the future."
- "The shopping experience was emotionally intense."
- "After shopping, I feel more positive about myself."

FLEs' extra-role behavior (Schneider, White, and Paul 1998; Tax and Brown 1998)

- "Frontline employees are able to handle special requests and any problems."
- "Frontline employees are ready to do anything to solve customers' problems."
- "Frontline employees are able to relate with the customer, giving them full attention."
- "Frontline employees are friendly and empathetic."

Service failure

- "I perceived my experience of shopping in this store as characterized by malfunctions or problems."

Perceived waiting time

- "How much time do you think you waited for your order?"
    - less than 1 minute
    - from 1 to 3 minutes
    - from 3 to 5 minutes
    - from 5 to 7 minutes
    - from 7 to 10 minutes
    - from 10 to 15 minutes
    - from 15 to 20 minutes
    - more than 20 minutes

Visit frequency

- "How often do you visit this store?"
    - this is my first visit
    - once or twice per year
    - bimonthly
    - monthly
    - twice per month
    - weekly
    - daily

Pretest: Effect of Shopping During Lunch Breaks on Time Pressure

Time pressure (Vermeir and Van Kenhove 2005)
"In the shopping experience described above:

- "I find myself pressed for time."
- "I am in a hurry when I shop for the clothes."
- "I have only a limited amount of time available to shop for my shirt."

FLEs' extra-role behavior expectations (Schneider, White, and Paul 1998; Tax and Brown 1998)
"In the shopping experience described above, I expect that frontline employees:

- "are able to handle special requests and any problems."
- "are ready to do anything to solve customers' problems."
- "are able to relate with the customer, giving them full attention."
- "are friendly and empathetic."

Fashion category involvement (Goldsmith and Emmert 1991)

- "If I heard that a new designer fashion was available in the store, I would be interested enough to buy it."
- "I know the names of new designer fashion before other people do."
- "I buy clothes very often."

Using Text Analytics Tools to Improve Online Reputation: Evidence from a Field Experiment and a Lab Study

Reject& Resubmit at Management Science

Anastasia Nanni

Doctoral Student in Marketing, PhD in Business Administration & Management
Bocconi University, Via Rontgen 1, 20136 Milan, Italy
Phone +39 02 58363510
anastasia.nanni@unibocconi.it

In joint with:

Andrea Ordanini

BNP Paribas Professor of Marketing & Service Analytics
Bocconi University, Via Rontgen 1, 20136 Milan, Italy,
Phone +39 02 58363623
andrea.ordanini@unibocconi.it

Raji Srinivasan

Sam Barshop Professor of Marketing Administration
Red McCombs School of Business, University of Texas at Austin
raji.srinivasan@mccombs.utexas.edu

Abstract

With the advent of online review platforms, online reputation management has become an important issue for marketing managers. The extant literature has recently examined the *response* process of online reputation management, by which managers decide whether and how to respond to online reviews. However, the *integration* process, by which managers leverage feedback from online reviews to improve their product offerings has been largely overlooked.

The integration process is crucial because it provides management with key inputs for changing the offer according to the most salient customer feedback. However, it is also challenging, as the most informative part of online reviews is unstructured text, which is not straightforward to process. To alleviate this challenge, managers can rely on text analytics tools that identify salient topics and their valence in online reviews. However, there is no evidence about the effectiveness of text analytic tools in improving firms' online reputation.

Accordingly, we investigate whether and how text analytics tools can serve as *cognitive repairs* (i.e., techniques that mitigate the limitations of managerial decisions) and test the effects of managerial use of text analytics tools on firms' online reputation. Empirically, we use a mixed-method approach with two studies: a randomized controlled trial of 201 hotel managers and a protocol analysis of 100 marketing managers' responses in a lab study. The findings indicate that: i) the use (vs. non-use) of text analytics tools significantly improves the hotel's online reputation, ii) it results in a more objective and less uncertain managerial cognitive process that effectively identifies weak performance areas, but iii) it does so only when the hotel's current performance is unsatisfactory.

*Keywords***:** reputation management, text analytics, cognitive repair, field experiment, protocol analysis

INTRODUCTION AND RELATED LITERATURE

Recent years have witnessed a drastic change in the customer feedback landscape from post-experience customer surveys initiated by companies to the rapid diffusion of third-party online review platforms, which has resulted in a key role for online customer reviews on consumers' purchase decisions (You, Vadakkepatt, and Joshi 2015). Not surprisingly, online reputation management is a key strategic challenge for managers as a "consumer is motivated to write reviews not only because reviews may impact other consumers, but because reviews may impact the management and the quality of the service" (Chevalier, Dover, and Mayzlin 2018, p. 688).

Based on the systematic collection of online customer feedback, online reputation management includes two key outbound processes: *responding* in a customized manner to online reviews and *integrating* online customer feedback into the firm's business processes to improve the quality of the firm's offerings (Niu and Fan 2018). Recent literature on online reputation management has largely focused on the response process by which managers decide whether and how to respond to online customer reviews. The empirical evidence in this area is mixed with some studies suggesting that managerial responses improve online reputation (Proserpio and Zervas 2017), others stating that the response effect depends on reviews' valence (Wang and Chaudhry 2018), and still others finding that managerial response stimulates the generation of further negative reviews (Chevalier, Dover, and Mayzlin 2018). A possible reason for this mixed evidence is that online reputation outcomes not only depend on managerial responses but also depend on the extent to which managers are able to *integrate* the most salient customer feedback in updating their product offerings (Büschken and Allenby 2016). Unfortunately, despite its managerial relevance, the integration process of online reputation management has been overlooked in the literature. This study represents a first effort in this area.

While the response process entails a volitional activity focused on whether or not to respond to each online review, the integration process includes *cognitive* tasks (understand and generate insights in the online reviews) and managerial *motivation* (accordingly changing their product offerings) (Niu and Fan 2018). The cognitive tasks are not straightforward because the most informative component of online customer reviews—where there is likely to be insightful feedback—is unstructured text (Ludwig et al. 2013). Getting insights from review text is complex and

challenging because the disembodied nature of the online customer review environment makes it hard for managers to interpret the reviews (Dellarocas 2003). Given the subjectivity of the online review content, managers may not be able to identify all relevant inputs for their decisions and may take shortcuts in their decision making (Van Bruggen and Wierenga 2010).

To address the cognitive challenges of the reputation management integration process, companies are investing in *text analytics tools*, i.e., algorithms that capture the main topics in the text, the valence of these topics, and any potential interdependencies among them (Blei 2012). Text analytics tools' big promise is to provide managers with a very diagnostic set of visual information to prioritize hypotheses and actions regarding strengths and weaknesses of their offerings (Seifert et al. 2014). In line with the literature in cognitive psychology (Trope and Liberman 1996) and organizational behavior (Heath, Larrick, and Klayman 1998), text analytics tools can act as *cognitive repairs*, practices that mitigate limitations and shortcomings of managerial decisions by "improving the mental procedures individuals use to decide which task to pursue and how to pursue it" (Heath, Larrick, and Klayman 1998, p. 5). Yet this literature suggests that text analytics tools also provide vivid additional information (e.g., tables, maps) that need to be processed by managers beyond the review text. This can result in locally rational decision making when one (say salient) piece of information is prioritized over others or a decisional impasse may occur when two pieces of information are incongruent (Lurie and Mason 2007).

Developments in the extant literature on the managerial use of text analytics tools indicate that text analysis of online reviews, by identifying latent dimensions of customer satisfaction and quality (Tirunillai and Tellis 2014), can help managers handle segmentation and mapping tasks. Other developments indicate that text analytics tools can help predict online reputation companies' scores using latent dimensions of customer satisfaction (Büschken and Allenby 2016). But a causal evidence of whether text analytics tools are effective cognitive repairs in the online reputation management integration process is still lacking. In other words, we do not know *whether* and *how* text analytics tools improve managerial ability to take appropriate remedial actions on product offerings and, in turn, increase online reputation. This is our first set of research questions.

Beyond the cognitive side, the integration of customer feedback in online reputation management is also influenced by the motivational features at the heart of the managerial action (Powell, Lovallo, and Fox 2011). Motivations are especially important when managerial decisions entail change (Miller and Chen 1994) as is the case in the integration process of online reputation management where product offerings may have to be changed based on online customer feedback. According to managerial behavioral theory (Cyert and March 1963; Greve 1998), an attainment discrepancy (i.e., a level of performance judged as being unsatisfactory) is a key driver for organizational change. When managers perceive that their firms' current performance is below a satisfactory threshold, they are likely to engage in "problemistic" search and risk-taking. In contrast, when current performance is considered satisfactory, managers are risk-averse and prioritize status quo over change (March and Shapira 1987). If managers do not experience an attainment discrepancy in their firms' actual performance, they may not be motivated to act based on online customer feedback. Hence, we do not know the extent to which attainment discrepancy may act as a motivational boundary to the effectiveness of text analytics tools. This is our second set of research questions.

We address our research questions using a multi-method approach with two studies investigating managerial subjects. The main study (Study 1) is a randomized controlled trial in the hotel industry aimed at identifying the causal effect of managerial use of text analytics tools and the potential boundary conditions of this effect. These findings then inform a lab experiment (Study 2) where we use protocol analysis to generate insights on the cognitive process employed by managers when using text analytics tools for online reputation management.

The findings from Study 1 indicate that, in a six-month time period, the managerial use (vs. non-use) of text analytics tools increases the treated hotels' online reputation in terms of their average TripAdvisor scores in a range of +4.4% to +6.2%. These findings are robust to alternative assumptions about sample attrition, measurement, and potential confounders. The findings also indicate that the managerial use of text analytics tools leads to different outcomes because managers in the treatment condition are better than those in the control condition in handling the weakest spot of their offering. Other evidence from the field experiment indicates that attainment discrepancy

effectively acts as a motivational constraint. Specifically, for managers in 45% of the hotels in our sample who indicated that they were satisfied with their hotel's current performance, there is no evidence of a positive effect of managerial use of text analytics tools on the hotel's online reputation.

The findings of Study 2 extend the insights of Study 1 by showing that the use of text analytics tools affects the cognitive process of managers involved in the reputation management integration process. Specifically, managers using text analytics tools exert greater cognitive effort, perceive lower uncertainty, and have a more insightful diagnosis of their offerings' weak spots. Moreover, they place lower emphasis on subjective criteria in weighting the various attributes of the product offering.

Taken together, these findings suggest that text analytics tools act as effective cognitive repairs and can be beneficial to a firm's online reputation by making customer feedback more diagnostic and improving the quality of the cognitive processes of managers. Yet the findings indicate that such benefits are bounded by motivational constraints of managers, specifically performance attainment, which erases the benefits of these tools.

THE MIXED-METHOD FRAMEWORK

Given the breadth of our research questions and to obtain more robust evidence, we use a multi-method approach in our empirical strategy. In particular, following the taxonomy proposed by Davis, Golicic, and Boerstler (2011), we implemented the *interpretation* design. This typically consists of two studies with unequal weight: a primary study that is used to robustly detect an effect and a secondary study that is used to interpret and better understand some of the results obtained from the primary study.

The primary study in our case is a randomized controlled trial (RCT) in the hotel industry. In the RCT, we randomly assigned 201 hotel managers in the real-world setting to either the treatment or control conditions. In the treatment (control) condition, they received (did not receive) online review text analytics as a stimulus in addition to the texts of online reviews. At the end of a six-month period, the difference in the hotels' average online reputation score between the treatment and control conditions represents the average treatment effect of the managerial use of text analytics tools. In

addition to identifying the causal effect of managerial use of text analytics tools on the hotel's online reputation, the RCT allows us to examine the presence of a heterogeneous effect of performance attainment discrepancy.

However, RCTs are not well-suited to investigate the mechanism underlying the causal effect because: (i) in the field context, it is difficult to completely control the conditions and (ii) collecting process measures during the RCT may interfere with the treatment effect (Banerjee and Duflo, 2017). For this reason, we complement the RCT with a second study, collecting qualitative data to help interpret the results of the primary study and generate additional insights into the managers' cognitive process in the reputation management process. Specifically, we conduct a lab study in which marketing managers in the treatment and control conditions (i.e., with and without text analytics tools) are asked to take decisions to improve a specific hotel's offering on the basis of online feedback. After managers performed the reputation management task, we asked them to report their thoughts and the steps that they undertook in making their decisions. We then used protocol analysis to analyze the managers' narratives (Bell and O'Keefe 1995).

To sum up, our mixed-method approach first includes a primary study (RCT) to investigate the role of text analytics tools as cognitive repairs, testing their causal effect on online reputation. The RCT also tests the moderation effect played by performance attainment discrepancy as a motivational boundary. This is then followed by a secondary study (lab experiment) to generate insights on the cognitive process underlying the causal effect of text analytics tools detected in the primary study. Figure 1 displays our conceptual framework.

---- Insert Figure 1 about here ----

STUDY 1—METHOD

Context and Sampling

We first investigate the role of text analytics tools as effective cognitive repairs for online reputation management using a RCT in the hotel industry. We chose the context of hotels as online reviews are central to consumers' pre-purchase decision making and therefore to hotels' online reputations (Litvin, Goldsmith, and Pan 2008; Proserpio and Zervas 2017). We enlisted the

cooperation of Federalberghi, the major Italian hotel industry association, to conduct a RCT on a sample of Italian hotels. Federalberghi fields a quarterly survey of the economic situation of its member hotels. In its January 2016 survey, we included a section introducing our study with a request for participation in the RCT. In this section, we informed managers that we sought their participation in a study investigating hotels' responses to online consumer feedback. We indicated that we would send them consumers' online reviews of their hotels posted on Tripadvisor, Italy at the beginning of the treatment period (March 1, 2016) related to the previous one-year period followed by another similar report after three months (June 1, 2016) as a stimulus reinforcement. As an incentive for their participation, we informed participants that at the end of the study participating hotels would receive a summary report of their hotel's online reputation synthesized from consumer reviews of their hotels. To ensure that the RCT would be handled by managers accountable for online reputation management, we contacted either the hotel's general manager or the manager responsible for managing the hotel's online consumer reviews.

598 hotel managers responded to this initial survey and 215 hotel managers (34%) expressed willingness to participate in the study. As the study focuses on text analytics tools, we excluded from the RCT, 12 hotels who reported using consultants and/or software to process online consumer reviews. We also excluded two hotels with no TripAdvisor reviews in the period under investigation. This resulted in an initial sample of 201 hotels. On average, the participating hotels had 42 rooms (range: 3-389) and most of them (59%) were moderately priced (i.e., class three hotels) with 18% being budget hotels (i.e., class one or class two) and 23% being boutique hotels (i.e., class four or class five). All hotels in our study were independent hotels (vs. chain). Further, the hotels were equally split between city (44%) and tourist locations (56%). The 201 participating hotels did not differ from the non-participating hotels (n=397) on type (segment, size, location) and on characteristics of managers (education, age, work experience, and their attitudes toward online reviews) (see Table A in the Appendix).

Randomization and Manipulation Procedure

In March 2016, we randomly assigned each of the 201 hotels willing and qualified to participate in the RCT to either the control or treatment condition. We stratified the random assignment of hotels to the two conditions to achieve an unbiased assignment of hotels based on size, type, and location (see Table B in the Appendix). In the control condition (n =101), we emailed managers a report that included their hotel's online reviews and ratings from TripAdvisor for the previous twelve months (March 2015-February 2016) (see Table C in the Appendix). The information that we provided to managers in the control condition mimics their hotel's TripAdvisor, Italy's online reviews webpages and represents the "baseline" condition in which hotel managers process online reviews (i.e., without text analytics tools).

We emailed a similar report to managers in the treatment condition (n=100). However, in addition to the online reviews, we also included the output of a text analytics tool (the treatment). To ensure that the manipulation was realistic and had external validity, an online reputation management consultant designed the text analytics output based on templates in use in the Italian hotel industry at that time (see Appendix, Table D). This output consists of two components. The first component is a table summarizing the net sentiment score (ranging from 0 to 100) of a fixed set of attributes common across all hotels: location, rooms, service, food, welcome, cleaning, convenience, and Internet connection. The second component is a set of hotel-specific tag clouds capturing the most influential underlying concepts with their associated valence extracted from the raw text. Concepts with positive (negative) valence are represented with green (red) tag clouds. The size of the tag cloud reflects the frequency of occurrence, recency, and strength of the sentiment associated with that concept.

As the stimulus for each hotel's treatment is customized based on that hotel's online customer reviews, the outcome of a hotel cannot be affected by treatments received by other hotels, ensuring stability of the unit treatment values (Athey and Imbens 2017). To reinforce the stimuli and reduce potential non-compliance effects (Glennerster 2017), at the end of May 2016, we contacted all managers and emailed them a second report with the same information as in the first report, updated with online reviews between March 2016 and May 2016. At the end of the treatment period (August 2016), we fielded a brief survey of the hotel managers to collect additional information and to check if

they had been active during the RCT. As might be expected in longitudinal field experiments, there was some sample attrition. Fifty-six hotel managers (28%) did not reply to our survey and might be considered as having dropped out of the study. We subsequently address threats to validity from sample attrition.

Our criterion variable is the hotel's online reputation, which we measure through its online ratings on TripAdvisor, Italy (Proserpio and Zervas 2017). The pre-treatment online reputation of the hotel is the monthly weighted (by the number of reviews) average online rating of the hotel on TripAdvisor between June and August 2015. We use the time between June and August as the period for the RCT as it is the peak season for Italian hotels and there are a higher number and more diverse set of customers. A t-test reveals no difference in the pre-treatment online reputations of hotels in the control and treatment conditions ($\mu_{treat} = 4.07$; $\mu_{ctrl} = 4.12$; $t = 0.56$; $p = .58$). The monthly weighted average online ratings of the hotel between June and August 2016 is the measure of its post-treatment online reputation.

Analytical Set Up

Our design reflects a typical RCT with a repeated (pre-post) measure. Each participant hotel is randomly assigned to a control or a treatment condition, an observed baseline outcome measure is obtained for each hotel before the treatment period, the treatment protocol is administered, and a follow-up outcome is measured at the end of the treatment period. This design is also known as mixed design as the average treatment effect depends on both a fixed factor that varies across hotels (the manipulation) and a random factor (the hotel) by which the pre-post observations are clustered. In short, our aim is to see if, on average, the within hotels pre-post treatment difference in the TripAdvisor scores differs between hotels across the control and the treatment conditions. The mixed design is well-suited for our focus on the change (and not the absolute value) of online reputation following the treatment. In addition, the repeated-measures structure of this design ensures more power and allows each subject (the hotel) to serve as its own control. At the same time, this design attenuates carryover (outcome measures are observed, not perceived) and demand effects (as both conditions receive a stimulus).

We estimate a linear mixed model with a random intercept, which is more efficient and flexible than traditional approaches (e.g., Repeated Anova, Ancova) in accommodating data with a nested and crossed structure. This method is also superior in handling unbalanced designs and missing data (Searle, Casella, and McCulloch 1992). Formally, with the *j* subscript representing the hotel level (between) and *i* the baseline vs. follow-up observation level (within), the following equation, adapted from Laird and Ware (1982), represents the model that we estimate:

$$[1] \quad Y_{ij} = a + b_1(Treat)_j + b_2(Follow\text{-}up)_i + b_3(Treat^* Follow\text{-}up)_{ij} + U_j + eij.$$

Coefficients a, b_1, b_2, and b_3 represent the typical fixed portion of the model, i.e., the intercept and the regression parameters. The treatment effect of interest is represented by the coefficient of the interaction term (b_3) of the hotel in the treatment (vs. control) condition and in the follow-up (vs. baseline) condition. This coefficient captures the mixed (between and within) average effect on the TripAdvisor scores (Y) associated with the treatment in the RCT. The random part of the model is represented by coefficients U_i, a vector of estimated random effects at the hotel level that accommodates the clustered nature of the data and reflects the variance of the hotel-specific intercepts and e_{ij} that captures the residual random error of the estimation.

STUDY 1—RESULTS

The Intention-to-Treat Model

Given the design of the experiment and the procedure to administrate the stimuli, our first analysis is based on an Intention-To-Treat (ITT) model. A straightforward approach to conduct an ITT analysis is to include all the initially randomized hotels, regardless of their compliance in the experiment. While such analysis may rely on unrealistic assumptions, an ITT model estimation is useful as: i) it does not suffer from identification bias since the causal effect is achieved through randomization and ii) it provides a conservative (i.e., lower bound) estimate of the average treatment effect (White et al. 2011).

For the ITT analysis, we estimate the linear mixed model on the full sample of 201 hotels. We use the Maximum-Likelihood approach for this model (and all other models in this study) and assess the significance of the treatment effect using a bias-corrected bootstrap with 2,000 replications. We

provide the results of this model estimation in Model 1 in Table 1, which shows a positive treatment effect (b_3 = .145; se = .077; p =.059; $bs_{95\%}$ = .006 / .307). Marginal effects analysis reveals that hotels in the treatment condition increased their online reputations (4.25 vs. 4.07; $\chi^2_{(1)}$ = 11.14; p = .001) significantly more than did hotels in the control condition for whom the increase was not statistically significant (4.16 vs. 4.13; $\chi^2_{(1)}$ =.48; p = .491). In Figure 2 panel A, the plot of pre-post differences in online reputation scores reveal that for hotels in the control group such differences are spread roughly around the horizontal axis, while for hotels in the treatment group, most of the points are above the horizontal axis, revealing the prevalence of positive differences in reputation scores of the treated hotels.

The elasticity of the effect on the performance in the treatment condition is 4.4%. In terms of effect size, the f^2 index (Selya et al. 2012) provides a statistic of .070, i.e., an effect whose size is between small and medium (Cohen 1998). Considering the conservative assumptions of the ITT model, these results provide preliminary evidence of a positive effect of managerial use of text analytics tools on their hotel's online reputation.

---- Insert Table 1 about here ----

---- Insert Figure 2 about here ----

Non-Compliance "Completely at Random"

Non-compliance is an endemic problem in longitudinal RCTs because researchers do not have full control over the treatment. In other words, although we randomize the assignment of subjects to the treatment and control conditions, we cannot ensure that they fully comply with the treatment in the RCT (Fisher et al. 1990). Non-compliance is problematic as it can threaten causal identification if non-compliers are systematically different from compliers (Winkels and Withers 2000).

We do have some evidence of non-compliance in the treatment period in our RCT. As noted above, 56 hotel managers (28%) did not provide any feedback when we sent them the second report in May 2016 and also did not respond to our end-of-treatment survey. Hence, these hotels might be considered to be non-compliant with the RCT procedure. In the presence of systematic attrition, the results of the ITT model might be biased, and the analysis should reveal the extent to which results

change across the different assumptions underlying sample attrition (Carpenter and Kenward 2008), which we do next.

The simplest way to handle the non-compliance problem in the RCTs is to assume that the 56 non-compliant hotels dropped out completely at random from the experiment. That is, the non-complier hotels are not different from compliers on any observed or unobserved characteristics that may affect the outcome (Rubin 1976). If this is the case, a linear mixed model using the sample of compliers (n=145) provides unbiased estimates (Carpenter, Kenward, and Vansteelandt 2006). We provide the results of this estimation in Model 2 in Table 1. The results indicate a stronger positive effect (b_3 = .193; se = .097; p = .046; $bs_{95\%}$ = .015 / .381) of managerial use of text analytics tools on online reputation than in the above ITT model. The elasticity of the effect on the performance in the treatment condition rises to 6%, and the effect size in terms of f^2 statistics increases to .091.

Non-Compliance "At Random"

However, we note that the assumption of missingness completely at random may be optimistic (Schafer and Graham 2002). Hence, we examine whether non-compliance in our RCT is associated with some observed characteristics of the hotels. We first run a probit model to see if non-compliance is predicted by hotels' characteristics including type, location, size, treatment status, and pre-treatment performance. We provide the estimates of this probit model in Table 2. The goodness-of-fit of the probit model is good ($\chi^2_{(8)}$ =3.98; p = .859) and the Wald test for the hotel's size is significant ($\chi^2_{(1)}$ =8.52; p = .004), suggesting that the data fit the model well. The results indicate that non-compliance in our RCT is balanced across treatment and control conditions (b_{treat} = .058; se = .192; p = .763) and is not affected by pre-treatment outcome level (b_{perf} = .049; se = .139; p = .725). Moreover, it is also not associated with structural variables like type or location. However, the likelihood of non-compliance increases as the size of the hotel increases (b_{size} = .491; se = .238; p=.039); a hotel larger than the median size in our sample had a 15.8% greater likelihood of not complying with the experiment protocol.

---- Insert Table 2 about here ----

These results suggest that our assumption of non-compliance in our RCT being completely at random is not valid. Using Rubin's terminology, non-compliance could be simply at random (AR), meaning that non-compliers and compliers can be considered to be similar only after the predictor of non-compliance (i.e., the size of the hotel) is taken into account. We do this by estimating the linear mixed regression with an inverse probability weighting associated with the size of the hotel (Seaman and White 2011). We provide these results in Model 3 in Table 1. These results support the positive treatment effect (b_3 = .213; se = .098; p = .030; $bs_{95\%}$ = .020 / .395) of the managerial use of text analytics tools on the hotel's online reputation. Additional marginal effects analysis reveals that hotels in the treatment condition significantly increased their online reputations (4.30 vs. 4.04; $\chi^2_{(1)}$ = 10.80; p = .001) more than hotels in the control condition (4.19 vs. 4.15; $\chi^2_{(1)}$ =.47; p = .491). Compared to the results of Model 2, the elasticity of the effect of managerial use of text analytics tools on the online reputation of hotels in the treatment condition increases to 6.2% and the effect size in terms of f^2 statistics increases to .098.

Non-Compliance "Not at Random"

Although non-compliance in the sample is correlated with an observed variable—the size of the hotel—we cannot conclude that it is not correlated with unobserved variables. If this is the case, non-compliance would be not at random and the correction proposed in Model 3 above based on the inverse probability weighting may not be enough to detect a true causal effect (Wooldridge 2002). To explore this possibility, we use a parametric approach to model non-compliance using a Heckman two-step selection correction. We first calculate the Inverse Mills Ratio (IMR) from the probit selection equation and include it as a covariate in the mixed model on the unweighted sample of complete cases (Grilli and Rampichini 2010). The analysis reveals the coefficient of IMR is not statistically significant (b_{Mills} = -.198; se = .356; p = .579) and the treatment effect of managerial use of text analytics tools on online reputation is still positive and significant (b_3 = .199; se = .097; p = .041; $bs_{95\%}$ = .008 / .404). We also replicate the Heckman model estimating the outcome and the selection equations simultaneously, using generalized structural equation modeling (Skrondal and Rabe-Hesketh 2004). In this multi-level setting, selection bias is modeled using a correlated latent

variable that enters both the selection and the outcome equation. Results in the outcome equation replicate those of the ITT model while the latent variable reflecting selection bias results is non-significant (b = -.046; p = .431). These analyses suggest that the treatment effect holds even when unobserved variables are considered for non-compliance.

While the previous analyses suggest that the parameter estimates in Model 3 may be valid, the Heckman selection approach relies on assumptions related to the parametrization employed for the selection model (Puhani 2000). To address this limitation, we examine any potential effect of non-compliance caused by unobserved variables using a non-parametric approach (Tauchmann 2013), which does not provide point estimates but a confidence interval for them. Specifically, we estimate this model adopting Lee's (2009) approach using non-compliance as the selection dummy variable, the coefficient b_3 in equation [1] as the predictor, and the size of the hotel as a tightening variable. The analysis estimates a lower bound of .160 and an upper bound of .219 (both with $p < .10$) for the treatment effect of managerial use of text analytics tools on the hotel's online reputation. We note that this confidence interval includes all the estimates in the various options of non-compliance above described. This analysis also provides a 90% confidence interval for the "true effect" [.020 / .405], which is generally similar to those identified by the bootstrapping efforts reported in Table 1 and also includes the ITT model estimates.

To summarize, this first set of results in Study 1 indicate that: i) the manipulation results in a positive treatment effect of the managerial use of text analytics tools on the hotel's online reputation, when ignoring non-compliance (i.e., ITT); ii) the positive effect of text analytics tools on the hotel's online reputation is replicated under alternative scenarios of different sources of non-compliance; and iii) a non-parametric analysis provides a confidence interval of the parameter estimates that is consistent with the range of the treatment effect estimates obtained across the different models. Figure 3 summarizes the different estimates of the treatment effects, which provide evidence that the managerial use of text analytics tools indeed improved the online reputations of treated hotels.

---- Insert Figure 3 about here ----

Robustness Checks

We next report additional analyses that examine the robustness of the treatment effect to alternative assumptions and biases beyond those arising from non-compliance. A first concern that we rule out is the possibility that a few influential cases may be driving the results. Inspection of Figure 2 of the plot of the single hotels' pre-post differences in the online reputation score suggests that it is unlikely results are driven by outliers. However, using the procedure for detecting outliers in the case of linear mixed models (Moehring and Schmidt-Catran 2013), we first identify influential cases (n = 4) as those with a DFBETA diagnostic greater than the threshold value. We then re-estimate Model 3 excluding these four influential cases. These results indicate that the treatment effect is supported and is even stronger (b_3 = .257; se = .096; p = .007).

The second concern that we address is the effect of exogenous events during the treatment period. If such exogenous events affect hotels differently across conditions, the RCT's results may be biased (Glennerster 2017). To address a possible confounder, when we contacted hotel managers for stimulus reinforcement in May 2016, we asked them to report on any extraordinary events (e.g., natural phenomena, renovation, refurbishing) that may have occurred during the treatment period. Twelve (8.3%) of 145 hotels reported such events; the proportion of these hotels was equal across the two conditions: 7 in the control condition and 5 in the treatment condition ($\chi^2_{(1)}$ =.43; p = .51). We then estimate Model 3 with the inclusion of a covariate to account for exogenous events. Again, the treatment effect is positive and statistically significant (b_3 = .213; se = .098; p = .030) and does not interact with the covariate for extraordinary events ($\chi^2_{(1)}$ =.11; p = .74).

The third concern that we address is whether the treatment effect varies based on the hotel's volume of online reviews. While the average number of reviews included in the reports used for the stimuli in the RCT is 46 and does not vary across treatment and control conditions (t=.82; p = .41), its distribution is dispersed (σ= 55) and shows a strong left-skewness (skew=3.2). Therefore, we investigate if the positive effect of managerial use of text analytic tools depends on the number of online reviews processed by the manager. We thus re-estimate Model 3 including the (log) number of the hotel's online reviews as a covariate in the estimation. Again, the treatment effect is positive and statistically significant (b_3 = .213; se = .099; p = .032); hotels that had more reviews increased their

online reputation in the treatment period (b = .114; se = .050; p = .023), but this was similar across both the treatment and control conditions ($\chi^2_{(1)}$ =1.31; p = .25).

The fourth concern that we examine is whether the treatment effect varies according to the experience of managers. In other words, is the beneficial effect of text analytics tools moderated by managers' tenure? Hence, we use information about managers' job tenure obtained in our initial survey. We first note that managerial tenure does not vary across treatment and control conditions either within the hotel industry (μ_{treat}=2.6 vs. μ_{ctrl}=2.9; t =1.2; p =.23) or outside the hotel industry (μ_{treat}=8.0 vs. μ_{ctrl}=5.8; t =1.43; p =.16). We then re-estimate Model 3 including the years of managerial tenure as a covariate. The treatment effect is positive and statistically significant (b_3 = .201; se = .098; p = .041). Hotels with managers with lower tenure in the treatment condition increased their online reputation in the treatment period (b = -.081; se = .040; p = .044), but this effect was similar in the treatment and control conditions ($\chi^2_{(1)}$ =0.07; p = .80).

The fifth concern that we address is whether the positive effect of the managerial use of text analytics tools in the integration process of online reputation management is not confounded with the effect that may have been generated earlier in the response process. We do this as some developments in the literature (e.g., Proserpio and Zervas 2017; Chevalier, Dover, and Mayzlin 2018; Wang and Chaudhry 2018) have identified benefits associated with managerial responses to online reviews. We thus obtained information about how many hotel managers in our sample responded to online reviews (in 2016) and found that this proportion does not vary across treatment and control conditions (54% vs 59%; $\chi^2_{(1)}$ =0.38; p =.54). We then re-estimate Model 3 including a dummy variable measuring whether the managers did or did not respond to online reviews. The treatment effect remains positive and significant (b_3 = .213; se = .098; p = .030). Managers' responses to online reviews did not affect their hotels' online reputations in the treatment period (b = .043; se = .102; p = .671), and the treatment effect was independent of managers' response activity ($\chi^2_{(1)}$ =0.06; p = .80).

The final concern that we address is whether the treatment effects were specific to the treatment period. We obtained the online reputation scores of hotels in a time window (March-May 2015) before the baseline period (June-Aug 2015) and find no difference across treated and untreated

hotels (μ_{treat}=4.12 vs. μ_{ctrl}=4.09; t =-.27; p =.79). These results suggest that in a three-month period just before the baseline (and about one year before the treatment), the treated and untreated hotels had similar levels of online reputation.

To summarize, the various robustness checks indicate that in our RCT: i) the treatment effect is not driven by the few influential outliers, ii) potential confounders in the treatment period do not bias the treatment effect, iii) the treatment effect does not depend on either the volume of customer feedback received by participants or the managers' tenure, iv) the treatment effect is independent of whether or not managers responded to online reviews, and v) the treatment effect is specific to the treatment period as hotels' online reputations are similar across control and treatment conditions in a time window before the RCT.

Preliminary Evidence of the Causal Mechanism

In this section, we provide preliminary evidence on *how* text analytics tools effectively act as cognitive repairs. As the integration process of online reputation management implies an organizational change driven by customer feedback (Niu and Fan 2018), we expect that the benefits of managerial use of text analytics tools would result in managers being better able to fix the weak points of their product offerings. To test this assumption, we used the sentiment scores in the treatment stimulus table to identify the weakest attribute of the hotel's offering. We also secured the sentiment scores for the hotels in the control condition so that we were able to identify, for each hotel in our sample, the attribute with the lowest sentiment score (from 0 to 100), i.e., the weakest attribute according to the online reviews. We excluded the attributes of hotel's location, restaurant, and room elements from this analysis as they cannot be reasonably modified by the hotel during the short time period of our RCT. We also excluded wi-fi issues in the hotel because few observations in our sample reported problems with them. Therefore, the set of attributes we considered included: service (worst attribute in 48% of the cases), price (30%), cleaning (12%), and welcoming (10%). The distribution of these worst attributes is similar across the treatment and control conditions ($\chi^2_{(3)}$ = 0.98; p = .99).

We then built the criterion variable for all the hotels with detailed sentiment data (n=114) as the difference between the level of sentiment on the worst attribute before and after the treatment (μ =

11.3; $\sigma = 31.7$) and predicted this difference using our treatment variable. An OLS regression controlling for hotel type, location, and size reveals that the average increase of the worst attribute's score for hotels in the treatment condition is different from zero (Δ=16.5; $p < .01$) and substantially higher than in the control condition (Δ=5.3; $p = .23$) (b = 11.26; se = 5.90; p = .059). This preliminary evidence suggests that managerial use of text analytics tools in the treatment condition enables managers to identify and remedy the weakest attribute of their hotels better than managers in the control condition. This finding provides preliminary insights on how text analytics can effectively act as cognitive repairs for hotel managers who use online reviews for their online reputation management. We later extend this preliminary evidence in Study 2 by further exploring the cognitive process associated with the managerial use of text analytics tools.

Heterogeneous Effect: Performance Discrepancy as a Motivational Boundary

We next address our second research question, namely the presence of a possible motivational boundary to the positive effect of managerial use of text analytics tools on online reputation. Behavioral theory (March and Shapira 1987; Greve 1998) suggests that when decisions involve change, as is the case in the integration process of reputation management, managers tend to be risk-averse. They may be prone to inertia, thus eliminating any potential benefits of cognitive repairs through the use of text analytics tools. But when managers perceive their organization's performance as being unsatisfactory, they become less prone to inertia and start search activities to improve their condition. Therefore, the positive effect of text analytics tools may be contingent on managerial perceptions about their organization's current performance. If managers perceive (do not perceive) an attainment discrepancy, they will (will not) have an incentive to leverage the potential of text analytics tools.

To investigate this research question, in our pre-treatment survey, we collected perceptions of managers about their hotels' current performance. We specifically asked them about their hotel's occupancy rate, a key performance metric in the three months prior to the treatment period on a scale from 1 (below a minimum threshold) to 5 (completely satisfactory). The average level of satisfaction with their organization's performance ($\mu = 3.10$, $\sigma = 1.15$) did not vary across the control and

treatment conditions (t = 0.24; p = 0.81). We then test for a possible heterogeneous effect by re-estimating Model 3 with the inclusion of a three-way interaction term (treatment × time × performance discrepancy). This parameter estimate is negative and statistically significant (b = -.188; p = .025), indicating that the treatment effect is weaker as the hotel's performance attainment discrepancy decreases (i.e., satisfaction with current performance increases). In Figure 4, we plot the pre-post change in the TripAdvisor reputation score (Y axis) over the treatment and control conditions at different levels of the hotel's performance attainment discrepancy (X axis).

---- Insert Figure 4 about here ----

Figure 3 reveals that hotels in the treatment condition have higher online reputations than hotels in the control condition only when their performance attainment discrepancy is high and satisfaction with their current performance is low, i.e., below the mean of 3.1. This evidence suggests that the boundary effect provided by the motivational factor is tangible; in about 45% of our sample, managers did not offer reasons to act to change and improve their offer. To summarize, our investigation of the second research question indicates that the effect of managerial use of text analytics tools is heterogeneous and the (lack of) attainment discrepancy perceived by the manager emerges as a boundary condition for this effect.

STUDY 2— METHOD

Context

Study 1 provides preliminary evidence on the role of managerial use of text analytics tools as cognitive repairs for marketing managers dealing with online customer feedback. When available, such tools thus help managers improve their organizations' online reputations by addressing the weak spots of their offerings. While interesting, the evidence on fixing the weaknesses of their product offerings can only offer a partial account of the mechanism underlying the differential effect of text analytics tools. First, as weak spots in product offerings are hotel-specific, it may be that those were simply easier to address or more impactful on their online reputations for hotels in the treatment condition than for hotels in the control condition. Second, in general, RCTs face limitations in

investigating the effects underlying mechanisms so that additional studies in more controlled environments may be useful to generate insights on the underlying mechanism (Simester 2017).

Against this backdrop and in line with our mixed-method approach, we set up a lab experiment in which a sample of 100 marketing managers across treatment and control conditions (with or without the text analytics tools) were exposed to the same context of product offering. We asked the managers to identify and remedy weaknesses in a specific hotel offering that participated in our Study 1. To focus only on the potential differences in the cognitive process, we selected a hotel in which participants were exposed to a limited number of online reviews (n=22) and across both conditions, they were asked to reach the same decisional outcome (i.e., the identification of the weaknesses). Hence, our approach represents a conservative test of managerial use of text analytics tools as cognitive repairs such that managers in the treatment condition may reveal a different cognitive process even when their task outputs (i.e., their decisions) produce results similar to those of managers in the treatment condition. To increase the level of control in the experimental conditions and ensure the proper selection of participants, we used the services of an independent marketing research company to conduct this study.

Sampling

One hundred Italian managers from different industries who work with online reviews as part of their job duties participated in the study. The sample consisted of 49 females and 51 males whose mean age was 44 years. As participants were managers with experience in decision-making with high opportunity cost for their time, we offered a monetary incentive to increase their motivation to participate to the study. Mimicking the manipulation in the RCT, we randomly assigned participants to two conditions. Participants in the control condition received only the raw text of the online reviews of an Italian hotel while participants in the treatment condition received the same text analysis outputs (tag clouds and sentiment table) used in Study 1 about the same hotel in addition to the online reviews. In both the conditions, we replaced the actual name of the hotel with a fictional one to avoid potential bias from any prior knowledge.

We asked managers in both conditions to read the material and identify the two most problematic aspects of the hotel's offering (which signal the presence of weak spots) and list the actions that they would implement to solve the two identified problems. After managers performed the task, we asked them to recall and report all their thoughts and the steps that they undertook in making their decisions. In particular, we asked participants to tell us (in an open format) their experience while working on the task, asking them to report on the rules followed, problems faced, and feelings experienced. We did not give participants any time limit for the decision. We then conducted protocol analysis with the transcripts provided by the managers to develop measures of their decision processes as we describe in the next section.

Protocol Analysis

Protocol analysis is a process tracing method that is used in investigating decision-making processes (Bell and O'Keefe 1995). We collected protocol data in a verbal form, which we then transcribed and used a neutral (vs. structured) probing protocol in which the subjects were simply instructed to describe their problem-solving and decision-making processes without imposing a priori structure (Hsieh and Shannon 2005) or theoretical perspective (Laureiro-Martínez and Brusoni 2018). To attenuate potential interference from the problem-solving process, we used a retrospective (vs. concurrent) protocol approach in which subjects recalled their processes after the task (Todd and Benbasat 1987). In short, we used protocol analysis to examine the activities of decision makers between the onset of a stimulus (in our case, the text analytics tool) and the eventual response (the experimental task of analyzing the hotel's situation).

Two researchers coded the protocols, tabulating the frequencies of the concepts in the protocols (Todd and Benbasat 1987). To ensure the reliability of the scores, two researchers external to the author team and blind with respect to the study's purpose evaluated the coding. We then calculated inter-rater agreement for qualitative data using the proportional reduction in loss (PRL) approach (Rust and Cooil 1994). For three raters and two categories, the proportion of inter-judge agreement is 90%, which corresponds to a PRL level of 98, suggesting a good degree of confidence in our coding output. To get a further internal validation of the scores generated by the researchers and

to obtain additional evidence, we also analyzed raw data using the Linguistic Inquiry & Word Count (LIWC) automated text analysis software (Pennebaker et al. 2015), which provides quantitative evidence about the emotional, cognitive, and structural components of verbal and written speeches.

STUDY 2—RESULTS

Human Coding

Given the type of task included in the experiment, as expected, the decision outcomes are generally similar across conditions. The two most frequently mentioned #1 priorities (personnel and breakfast) and the two most frequently mentioned #1 interventions (employee training and replacement) are the same across the two conditions. Yet the distributions of #1 priorities and #1 actions show a slightly higher level of dispersion in the treatment condition, suggesting that managers using text analytics aids exhibited, on average, a less convergent cognitive process with fewer alternatives. There is similar evidence when considering the decision-making time. In a weak-structured problem like that in our experiment, time can be considered a surrogate measure for cognitive effort (Jarvenpaa 1989) as subjects exhibiting a short decision time are more likely to terminate their cognitive processes as soon as they locate a plausible explanation. The median time to execute the task is higher in the treatment condition (15 minutes vs. 13 minutes; z=-2.11, p = .035), suggesting that subjects using text analytics tools may have expended more cognitive effort.

The key step in the protocol analysis is the analysis of the subjective scoring. We identify three main themes that emerged directly from subjects' words pertaining to their perceptions about: i) the task, ii) the provided material, and iii) the cognitive process to perform the task. Regarding the task, the investigation of the protocol supports that the subjects in both the control and treatment conditions mostly perceived the task as "easy" and "interesting," as the two illustrative quotes below show. This reflects our experimental choice to expose informants to a limited number of reviews in order to keep the task easier.

> Cinzia (Control Condition): *"The task is easy and, above all, useful."*
>
> Claudia (Treatment Condition): "*I think that the assigned task was easy and funny."*

In general, participants across the two conditions perceived the texts of the online customer reviews as "divergent and subjective" and, in turn, hard to interpret. Informants in the treatment condition instead positively assessed the text analytic aid they received. In particular, they perceived the text analytics tools as "objective" and "useful" in supporting them during the task. Furthermore, in contrast to the control condition, the treatment condition considered the material that they had received (i.e., the texts of the reviews and the text analytics tools) as being complete and sufficient to perform the task. Two informants in the control condition stated that they would have greatly benefited from having the results of a text analysis of the online reviews. Below are some illustrative quotes from the protocol analysis:

> Carlo (Control Condition): *"[Reviews' texts] are important sources of information that highlight more positive than negative opinions. However, they are not complete and sufficient to take a correct and comprehensive decision."*
>
> Claudia (Treatment Condition): *"The material is clear and sufficient [to perform the task]."*
>
> Francesco (Treatment Condition): "*The two analytical outcomes allowed me to get an accurate picture of the situation. I used the reviews' texts to deepen and argue the data of the graphs."*

Regarding the cognitive process to perform the task, all informants in the treatment condition (except for one participant) dedicated attention to both the texts of online reviews and the outcomes of text analysis. Therefore, subjects in the treatment condition did not appear to trade off information sources. This evidence precludes the possibility that participants who were exposed to the word cloud and the sentiment table over-weighted this information (i.e., local rationality and vividness bias), neglecting the textual data.

Although participants in both conditions mostly focused on the negative information (i.e., negative online reviews and negative sentiments in text analysis) to identify the most problematic aspects of the hotel's offering, the process to identify the critical issues appeared to be different between the two conditions. Indeed, participants in the treatment condition claimed that their decision was based only on the data in the material unlike the informants in the control conditions who

reported they based their decision mostly on subjective criteria. This evidence suggests that using text analytics tools reduces the emphasis on subjective and preconceived opinions corroborating the role of texts analytics tools as effective cognitive repairs. The quotes below illustrate this important differential situation across the two conditions:

> Alberto (Control Condition): "*I relied mainly on my personal experience. […] I have evaluated the most important aspects, which in my experience cannot be missing in a 3-star hotel.*"
>
> Silvia (Control Condition): "*I tried to think as a hotel manager, but I based [my decision] on my experience as a customer.*"
>
> Marco (Treatment Condition: "*Regarding my decision, I relied on the sentiment table. The numbers are colder (than the texts), but they are not misguided by the emotional content in of the reviews' texts.*"
>
> Pietro (Treatment Condition): "*Often customers' opinions change according to the period of the year, to the experience, to the expectations. I read the graphs (text analysis outcome) and I identified the main problem, which is the staff.*"

Automated Text Analysis

As discussed above, we also subjected our protocol data to an automated textual analysis using LIWC software (Pennebaker et al. 2015). In a nutshell, LIWC uses a validated dictionary of more than 6,000 words to capture the emotional, cognitive, and structural components of a text, providing quantitative scores for these dimensions based on word frequency. Since we collected the original data in Italian (the native language of the subjects), we used the LIWC software version based on the Italian dictionary.

The "cognitive process" section of the dictionary is the most salient category of interest to us, providing us the frequency of words related to the cognitive mechanisms employed by the respondent during the task (e.g., cause, know, ought). In this case, the protocols of managers treated with the text analytics tool reveal a greater use of cognitive mechanisms indicators with respect to: i) the sub-category "insight," with an average score of 5.2 compared to the 4.0 score of managers in the control condition ($F_{1,98}$=5.91; p = .017) and ii) the sub-category "certainty," with an average score of 1.90

compared to the 1.26 score of managers in the control condition ($F_{1,98}$=4.16; p = .044). These findings suggest managers exposed to text analytics tools (the treatment condition) showed some evidence of more introspection and less uncertainty in their cognitive processes compared to their counterparts in the control condition, corroborating the qualitative evidence that emerged in the subjective coding phase of the protocol analysis.

In sum, the findings of study 2 indicate that when two groups of marketing managers—one using text analytics tools and the other not—are confronted with the same task (i.e., analyzing online reviews for one hotel), their cognitive processes are different. Those using the text analytics tools appear to believe that they rely on a more complete set of information, used more objective criteria for assessment, and experienced more insightful and less uncertain cognitive processes. This lends further support to this research's central thesis that text analytics tools can be considered effective cognitive repairs for marketing managers involved with online reputation management goals.

IMPLICATIONS AND CONCLUSIONS

This research represents, to our knowledge, the first attempt to examine the impact of text analytics tools on online reputation management. The market for such text analytics tools is expected to double its size in the next few years, moving from about USD 4 Billion to USD 8.79 Billion by 2022 (Marketsandmarkets 2017), and business experts (McKinsey 2016) and policy makers (EU Commission 2016) expect that text analytic solutions will enable managers to exploit consumer insights embedded in textual data. Yet the effect of managerial use of online text analytics tools on firms' online reputation has been overlooked in the literature. Addressing this research gap, we use a multi-method interpretation approach that combines quantitative evidence from a RCT in the hotel industry with qualitative evidence from a protocol analysis in a lab experiment with marketing managers. The research's findings can be summarized as follows.

First, the use of text analytics tools significantly improves firms' online reputation, with an ATET ranging from +4.4% to 6.2% in terms of TripAdvisor score over six months. This is a sizable effect; in terms of the rounded scores used by TripAdvisor, it means treated hotels reached an average of 4.5 stars (crossing the rounding threshold of 4.25) while untreated ones maintained an average

rounded score of 4 stars. Second, the use of text analytics tools better enables managers to improve the weakest elements of their product offerings. Treated hotels were able to improve the sentiment score of the weakest spot of their product offerings by more than 16 percentage points while untreated ones did not experience any tangible increase in the sentiment score of their weakest spot. Third, the use of text analytics tools stimulates a more insightful, unbiased, and less uncertain cognitive process in managers involved in reputation management tasks. Yet untreated managers seem to use more subjective criteria to evaluate customer feedback, take less time to process the information, and develop a sense of greater uncertainty in the cognitive process. Fourth, the managerial use of text analytics tools is inefficient for online reputation if managers are already satisfied with their actual level of performance. The moderation effect is sizable since the (lack of) an attainment discrepancy effectively acts as a motivational constraint for managers in 45% of the hotels in our sample who indicated that they were satisfied with their hotel's current performance.

These findings generate implications for various literature streams. First, we extend the insights on online reputation management that so far have only focused on the response process by managers (Proserpio and Zervas 2017; Wang and Chaudhry 2018; Chevalier, Dover, and Mayzlin 2018). Our findings indicate that, over and beyond the act of responding to online feedback, marketing managers can improve their firms' online reputation by focusing on the integration process using text analytics tools to make sense of online reviews, fix weak spots in their product offering, and improve it. Our findings suggest that customer feedback should not be treated by managers as mere communication outcomes managed "in retrospective" by simply providing responses to *past* consumption episodes. Instead, customer feedback should also be considered as key managerial input, managed "in perspective" to change the existing offering and improve the online reputation for *future* consumption episodes. Our findings also suggest a possible explanation for the mixed evidence on the effect of managerial response so far accumulated in literature, which might have overlooked the key role played by the integration process of online reputation management.

Our work also contributes to the decision support systems literature that calls for empirical evidence on the mechanisms through which decisional aids, such as text analytics tools, affect managerial decision processes (van Bruggen and Wierenga 2010). We find that the use of text

analytics tools spur a different cognitive process for managers, who use more objective criteria in executing the task and perceive less uncertainty in their evaluations. Moreover, we find that the use of text analytics outcomes does not occur at the expense of attention to the raw text data and does not generate local rationality bias in managerial assessment (Lurie and Mason 2007). Despite their qualitative nature, our findings suggest managers "think differently" when helped with text analytics tools, gaining more confidence on the insights that can be captured from online customer feedback (i.e., text analytics tools can act as effective cognitive repairs).

The findings also extend the core thesis of behavioral management theory (Greve 1998) to the marketing context, showing that the cognitive repairs provided by text analytics tools do not work for online reputation management if managers lack the appropriate incentives to act. This suggests the importance of setting up adequate motivation for managers engaged in online reputation management (Powell, Lovallo, and Fox 2011). Otherwise, there may be the risk that inertia will trump the potential of cognitive repairs, generating opportunity costs and undesired outcomes (i.e., text analytics tools can be inefficient because of motivational factors).

Finally, the research, the first on the integration process of online reputation management offers interesting avenues for future research. First, the stimuli used in our RCT were based on a standard solution in the marketplace at the time of the RCT. It would be interesting to examine the extent to which the effects detected in this research replicate in the case of more sophisticated text analytics tools that are being used in practice (e.g., interactive platforms). The question is interesting as although sophisticated text analytical solutions are likely to have greater diagnostic power, they are also more cognitively demanding so that their "net" effect may not be straightforward. Second, our quantitative analysis focused on one specific online platform for online reviews (i.e., TripAdvisor). A useful research extension would be to examine if the effect of text analytics tools replicates on other online feedback channels including social media platforms as customer feedback may vary across online channels.

REFERENCES

Athey S, Imbens GW (2017) The econometrics of randomized experiments. Banerjee AV, Duflo E, eds. *Handbook of Economic Field Experiments,* Vol. 1 (Elsevier, North-Holland), 73-140.

Banerjee AV, Duflo E (2017) *Handbook of Economic Field Experiments,* Vol. 1 (Elsevier, North-Holland)

Bell PC, O'Keefe RM (1995) An experimental investigation into the efficacy of visual interactive simulation. *Management Sci.* 41(6):1018-1038.

Blei DM (2012) Probabilistic topic models. *Commun. ACM.* 55(4):77-84.

Büschken J, Allenby GM (2016) Sentence-based text analysis for customer reviews. *Marketing Sci.* 35(6): 953-975.

Carpenter JR, Kenward MG (2008) Missing data in clinical trials—a practical guide. *National health service coordinating centre for research Methodology. Birmingham*. UK: National Health Service Coordinating Centre for Research Methodology

Carpenter JR, Kenward MG, Vansteelandt S (2006) A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *J. R. Stat. Soc. A Stat.* 169(3):571-584.

Chevalier JA, Dover Y, Mayzlin D (2018) Channels of impact: User reviews when quality is dynamic and managers respond. *Marketing Sci.* 37(5):688-709.

Cohen J (1988), *Statistical Power Analysis for the Behavioral Sciences* (Lawrence Earlbaum Associates, Hillsdale, NJ).

Cyert RM, March JM (1963) *A behavioral theory of the firm* (Prentice-Hall, Englewood Cliffs, NJ)

Davis DF, Golicic SL, Boerstler CN (2011) Benefits and challenges of conducting multiple methods research in marketing. *J. Acad. Marketing Scie.* 39(3):467-479.

Dellarocas C (2003) The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Sci.* 49(10):1407-1424.

Ericsson KA, Simon HA (1993) *Protocol analysis: verbal reports as data* (MIT Press, Cambridge, MA).

EU Commission (2016), "Competence Centre Text Mining and Analysis", Workshop on Text Mining in Policy making, Brussels, Belgium, December 13 2016.

Fisher LD, Dixon DO, Herson J, Frankowski RK, Hear Ron MS, Peace KE (1990) Intention to treat in clinical trials. Peace KE, ed. *Statistical Issues in Drug Research and Development* (Marcel Dekker, New York), 331-350

Glennerster R (2017) The practicalities of running randomized evaluations: partnerships, measurement, ethics, and transparency. Banerjee AV, Duflo E, eds. *Handbook of Economic Field Experiments*, Vol. 1 (Elsevier, North-Holland), 175-243.

Greve HR (1998) Performance, aspirations and risky organizational change. *Adm. Sci. Q.* 43(1): 58-86.

Grilli L, Rampichini C (2010) Selection bias in linear mixed models. *Metron* 68(3):309-329.

Heath C, Larrick RP, Klayman J (1998) Cognitive repairs: How organizational practices can compensate for individual shortcomings. *Res. Organ. Behav.* 20:1–37.

Hsieh HF, Shannon SE (2005) Three approaches to qualitative content analysis. *Qual. Health Res.* 15(9):1277-1288.

Jarvenpaa SL (1989) The effect of task demands and graphical format on information processing strategies. *Management Sci.* 35(3):285-303.

Laird NM, Ware JH (1982) Random-effects models for longitudinal data. *Biometrics* 38(4):963-974.

Laureiro-Martínez D, Brusoni S (2018) Cognitive flexibility and adaptive decision-making: Evidence from a laboratory study of expert decision makers. *Strategic Manage. J.* 39(4):1031-1058.

Lee DS (2009) Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Rev. Econ. Stud.* 76(3):1071–1102.

Litvin SW, Goldsmith RE, Pan B (2008) Electronic word-of-mouth in hospitality and tourism management. *Tourism Manage.* 29(3):458-468.

Ludwig S, de Ruyter K, Friedman M, Brüggen EC, Wetzels M, Pfann G (2013) More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates. *J. Marketing* 77(1):87-103.

Lurie NH, Mason CH (2007) Visual representation: Implications for decision making. *J. Marketing* 71(1):160–177.

March JG, Shapira Z (1987) Managerial perspectives on risk and risk taking. *Management Sci.* 33(11):1404-1418.

Marketsandmarkets (2017) Text Analytics Market - Global Forecast to 2022, Business Report.

McKinsey (2016) The Age of Analytics: Competing in a Data-Driven World, Global Institute Report.

Miller D, Chen MJ (1994) Sources and consequences of competitive inertia: a study of the U. S. airline industry. *Adm. Sci. Q.* 39(1):1-23.

Moehring K, Schmidt-Catran AW (2013) MLT: Stata module to provide multilevel tools (Statistical Software Components S457577). Boston, MA: Boston College Department of Economics. Retrieved from http://ideas.repec.org/c/boc/bocode/s457577.html.

Niu RH, Fan Y (2018) An exploratory study of online review management in hospitality services. *J. Serv. Theor. Pract.* 28(1):79-98.

Pennebaker JW, Boyd RL, Jordan K, Blackburn K (2015) *The development and psychometric properties of LIWC2015* (LIWC.net, Austin, TX).

Powell TC, Lovallo D, Fox CR (2011) Behavioral Strategy. *Strategic Manage. J.* 32(13):1369-1386.

Proserpio D, Zervas G (2017) Online reputation management: Estimating the impact of management responses on consumer reviews. *Marketing Sci.* 36(5):645-665.

Puhani P (2000) The Heckman correction for sample selection and its critique. *J. Econ. Surv.* 14(1):53-68.

Rubin DB (1976) Inference and missing data. *Biometrika* 63(3):581-592.

Rust RT, Cooil B (1994) Reliability Measures for Qualitative Data: Theory and Implications. *J. Marketing Res.* 31(1):1–14.

Schafer JL, Graham JW (2002) Missing data: Our view of the state of the art. *Psychol. Methods* 7(2):147–177.

Seaman SR, White IR (2011) Review of inverse probability weighting for dealing with missing data. *Stat. Methods Med. Res.* 22(3):278–295.

Searle SR, Casella G, McCulloch CE (1992) *Variance components* (John Wiley & Sons, New York).

Seifert C, Sabol V, Kienreich W, Lex E, Granitzer M (2014) Visual analysis and knowledge discovery for text. Gkoulalas-Divanis A, Labbi A, eds. *Large-Scale Data Analytics* (Springer, New York),189-218.

Selya AS, Rose JS, Dierker LC, Hedeker D, Mermelstein RJ (2012) A practical guide to calculating Cohen's f^2, a measure of local effect size, from PROC MIXED. *Front. Psychol.* 3:1-6.

Simester D (2017) Field experiments in marketing. Banerjee AV, Duflo E, eds. *Handbook of Economic Field Experiments*, Vol. 1 (Elsevier, North-Holland), 465-497.

Skrondal A, Rabe-Hesketh S (2004) *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models* (Chapman and Hall/CRC, New York).

Tauchmann H (2013) Lee's treatment effect bounds for non-random sample selection—An implementation in Stata. Berlin, Germany: Sonderforschungsbereich (SFB) Discussion Paper 823.

Tirunillai S, Tellis GJ (2014) Mining meaning from online chatter: Strategic brand analysis of big data with latent dirichlet allocation. *J. Marketing Res.* 51(4):463-479.

Todd PA, Benbasat I (1987) Process tracing methods in decision support systems research: Exploring the black box. *Mis Quart.* 11(4):493-512.

Trope Y, Liberman A (1996) Social hypothesis testing: Cognitive and motivational mechanisms. Higgins ET, Kruglanski AW, eds. *Social psychology: Handbook of basic principles* (Guilford, New York), 239–270.

Van Bruggen GH, Wierenga B (2010) Marketing decision making and decision support: Challenges and perspectives for successful marketing management support systems. *Found. Trends. Marketing* 4(4):209-332.

Wang Y, Chaudhry A (2018) When and how managers' responses to online reviews affect subsequent reviews. *J. Marketing Res.* 55(2):163-177.

White IR, Horton NJ, Carpernter J, Pocock SJ (2011) Strategy for intention to treat analysis in randomised trials with missing outcome data. *BMJ*, 342, d40.

Winkels J, Withers S (2000) Panel attrition. Rose D, ed. R*esearching Social and Economic Change: The Uses of Household Panel Studies* (Routledge, London).

Wooldridge JM (2002) *Econometric analysis of cross-sectional and panel data* (MIT Press, Cambridge MA)

You Y, Vadakkepatt, GG, Joshi AM (2015) A meta-analysis of electronic word-of-mouth elasticity. *J. Marketing* 79(2):19-39.

TABLES AND FIGURES

Table 1 – Study 1: Linear Mixed Model Estimations: Intention-To-Treat and Non-Compliance Models

	Mod 1 Intention-To-Treat	Mod 2 Non-Compliance CAR (no correction)	Mod 3 Non-Compliance AR (Ipw correction)	Mod 4 Non-Compliance NAR (Heck Two-Step)
Fixed Effects	b se p	b se p	b se p	b se p
Treatment	-.061 (.087) .479	-.095 (.108) .382	-.124 (.121) .314	-.094 (.110) .391
Time	.037 (.054) .491	.055 (.069) .425	.045 (.061) .456	.051 (.069) .464
Treatment*Time (b_3)	.145 (.077) .059	.193 (.097) .046	.213 (.098) .030	.199 (.097) .041
[Boot 95% CI]	[.006 / .307]	[.015 / .381]	[.020 / .395]	[.008 / .404]
Constant	4.13 (.061) .000	4.14 (.077) .000	4.17 (.075) .000	4.23 (.182) .000
Inverse Mills Ratio				-.198 (.356) .579
Random Parameters	b se	b se	b se	b se
Var(constant) (hotel)	.227 (.031)	.254 (.041)	.256 (.057)	.253 (.042)
Var(residual) (hotel)	.144 (.015)	.163 (.020)	162 (.027)	.163 (.020)
N	201	145	145	(201) 145
Effect size (f^2)	.070	.091	.098	.091

Table 2 – Study 1: Probit Model to Estimate Non-Compliance

	b	se*	p
Treatment	-.058	.192	.763
Type (Budget)			
Mid-price/Suite	.015	.297	.961
Boutique	-.136	.371	.714
Location (Urban)			
Seaside	-.250	.338	.459
Mountainside	-.340	.240	.158
Lakeside	-.717	.374	.055
Size	.491	.238	.039
Pre-treatment performance	.049	.139	.725
Constant	-.806	.610	.187

** Robust Std. Error*

( ) = reference category

Figure 1 – Conceptual Framework of the Studies

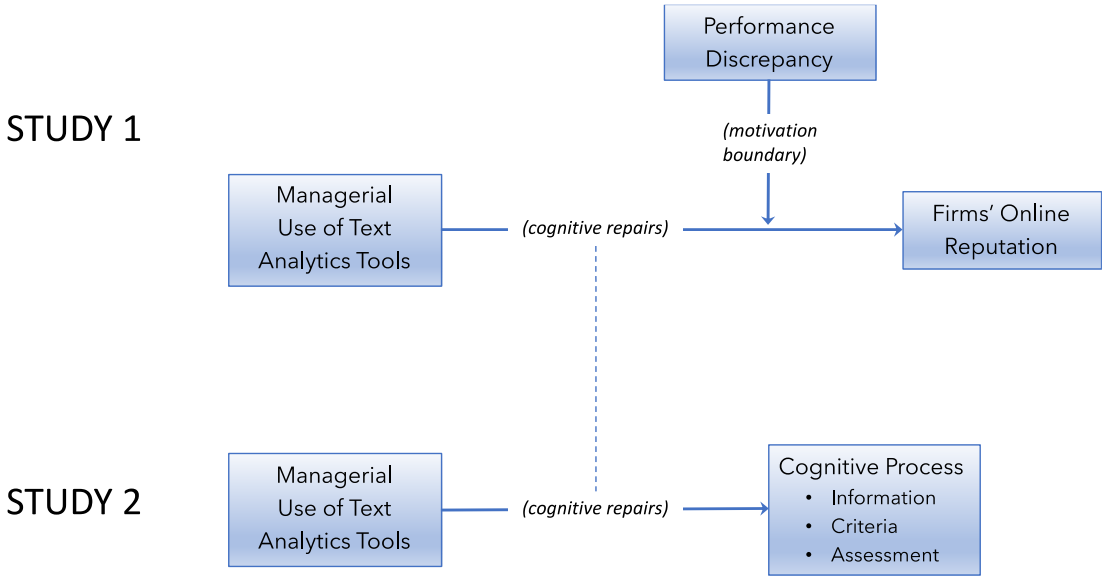


Figure 2 – Study 1: Plot of Single Hotels' Changes in Online Reputation after Treatment

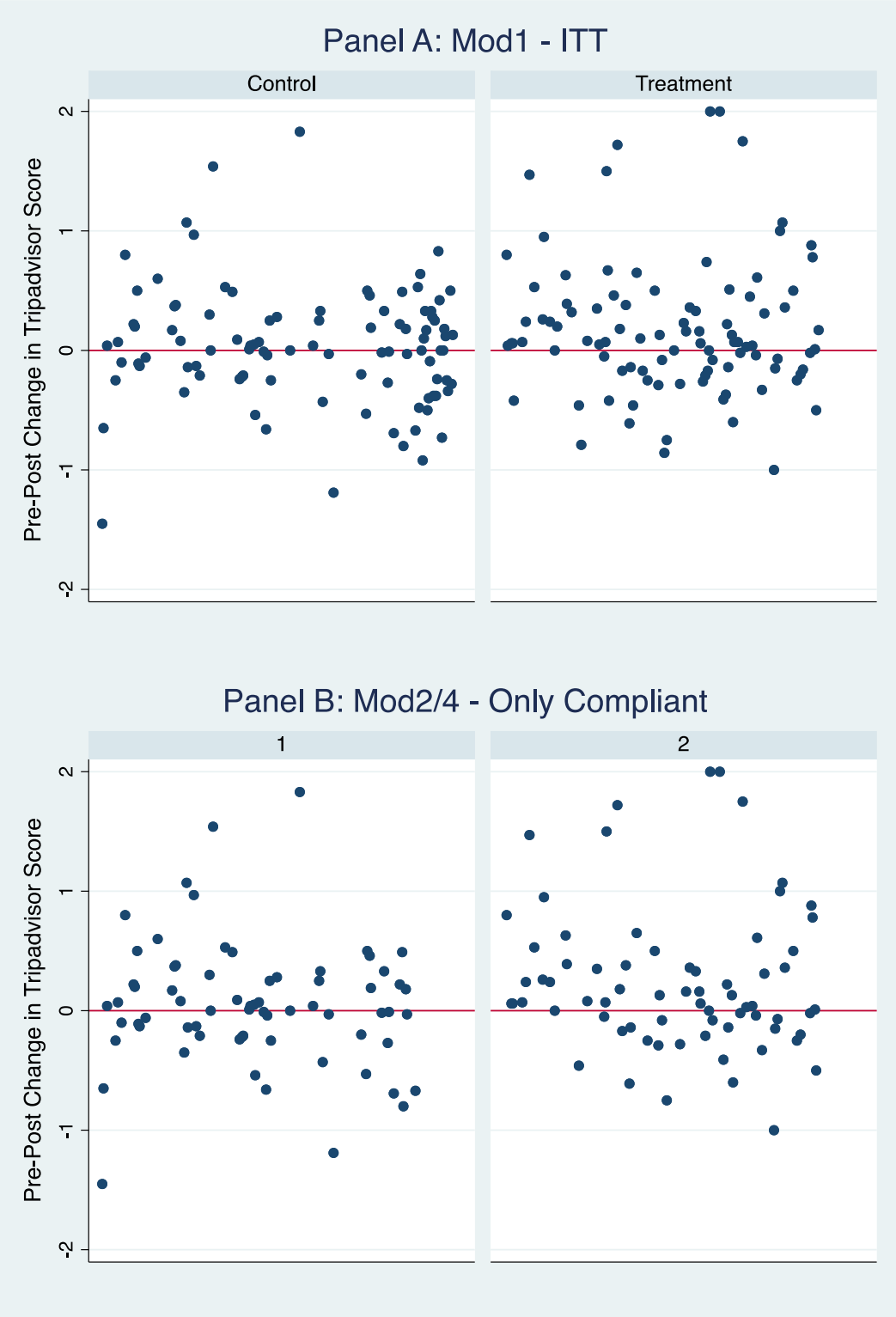


Figure 3 – Study 1: Average Treatment Effect of Text Analytics Tools on Online Reputation Across Models

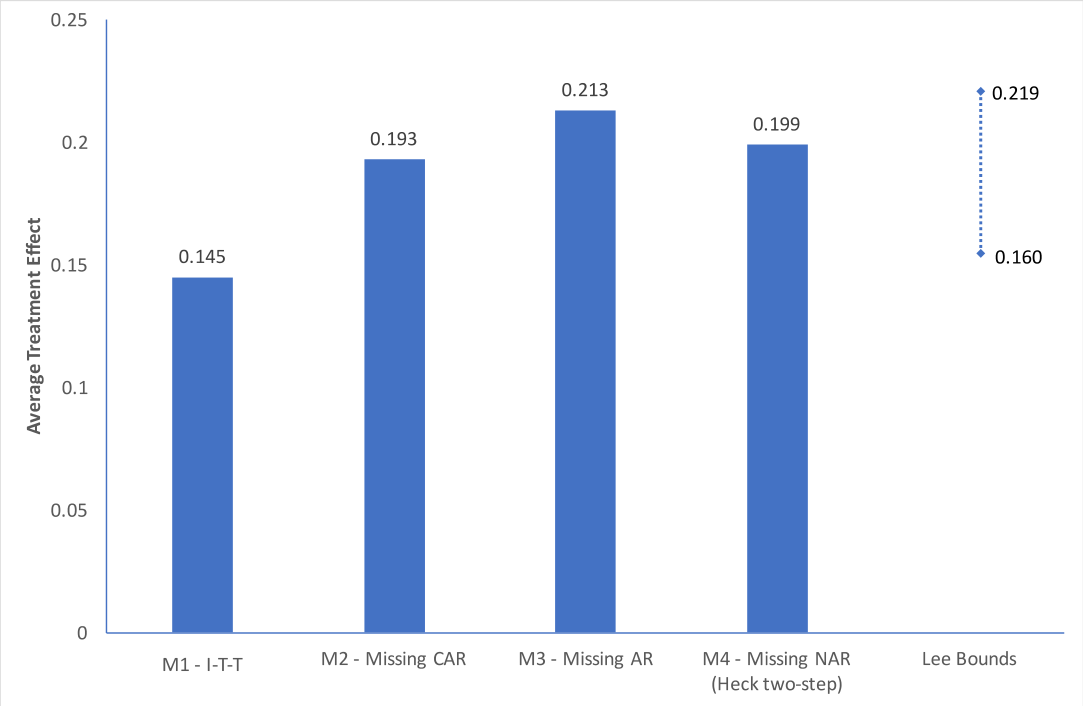
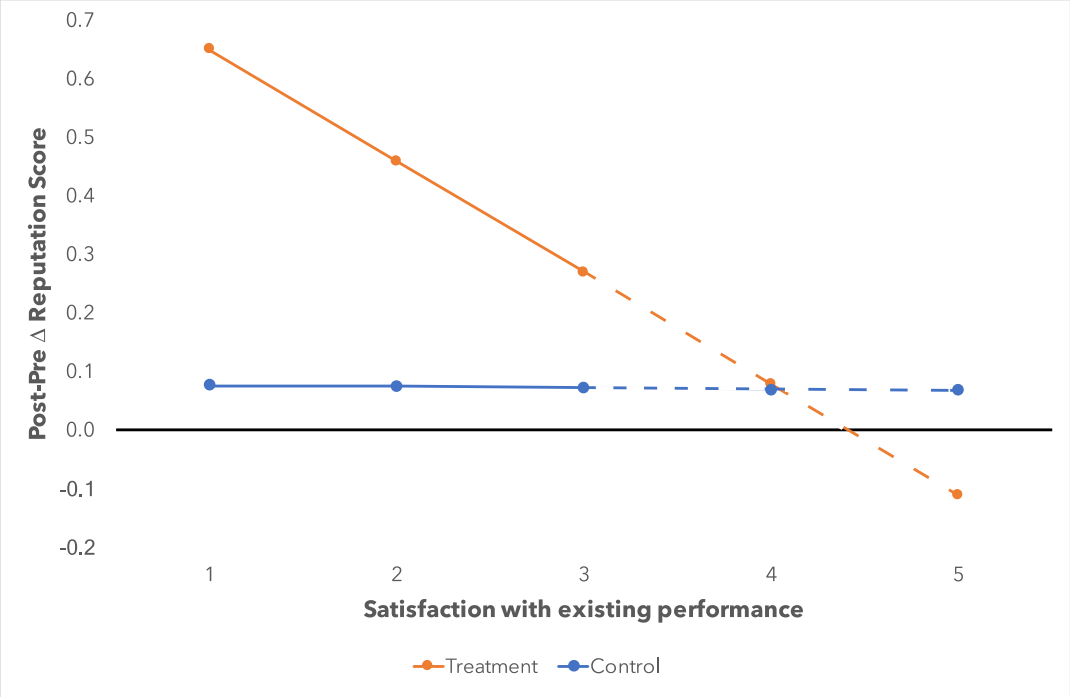


Figure 4 – Study 1: Pre-Post Changes in Online Reputation Scores across Conditions, at Different Levels of Attainment Discrepancy



The Effects of Artificial Intelligence on Productivity and Quality: A Field Experiment in Banking Services

Target Journal: Marketing Science

Anastasia Nanni

Doctoral Student in Marketing, PhD in Business Administration & Management Bocconi University, Via Rontgen 1, 20136 Milan, Italy
Phone +39 02 58363510
anastasia.nanni@unibocconi.it

In joint with:

Andrea Ordanini

BNP Paribas Professor of Marketing & Service Analytics Bocconi University, Via Rontgen 1, 20136 Milan, Italy, Phone +39 02 58363623
andrea.ordanini@unibocconi.it

P.K. Kannan

Dean's Chair in Marketing Science, Robert H. Smith School of Business, University of Maryland
pkannan@rhsmith.umd.edu

Abstract

We investigate the effects of the implementation of an AI solution using a unique database obtained by a global bank that also helped us to set up a field experiment concerning the rollout of such AI solution. The bank decided to implement an AI solution aiming to automate the responses to tickets opened by employees who, in its various branches, need support to solve issues related to mortgages raised by customers. While until last year all these tickets were handled by a task force of experts at the corporate level, in collaboration with the bank we set up temporary experimental conditions according to which, for two months, a group of branches started to operate assisted by the new AI solution, while another group of branches continued to work with the traditional method. Our findings from a series of diff-in-diff analyses reveal that the AI solution generates: i) a positive efficiency effects on the time to handle the ticket; ii) a negative effect on the satisfaction of the branch employees involved in the process; iii) an unexpected spillover on service offering (fewer mortgages granted), and iv) a positive effect on customer satisfaction for the assigned mortgages (mostly due to a perceived better process). Our analyses contribute to both literature and practice in shedding light on the important and often overlooked trade-off effects associated to the implementation of AI in service organizations.

Keywords: Artificial Intelligence; field experiment; service productivity; service quality; banking industry

Introduction

Increasing productivity has always been one of the main objectives of all companies. In recent years, finding a way to reduce time and costs of production has become an imperative for managers (Harward Business Review 2017; McKinsey 2020). This is a difficult task for all companies, but it is particularly sensitive for service companies since an increase in productivity can negatively affect the quality of the output offered to the customers. In fact, while for manufacturing firms productivity and quality are loosely related elements, as the former pertains to the use of inputs for output production while the latter regards how the output is evaluated by the market, in the case of service firms the two performance elements are not independent and typically involve trade-off relations (Calabrese 2012). In services, production and consumption cannot be separated, as the customers directly participate to the generation of the service output, and thus changing the inputs in a service operation inevitably also shapes the assessment of the output quality, and vice-versa (Grönroos and Osajalo 2004). Yet, according to the literature, spill-over effects between productivity and quality are often negative for service companies, as "better service typically requiring more labor intensity, lower productivity, and higher cost" (Rust and Huang 2012, p.47) and "what appears to be improved productivity in terms of better production efficiency may turn out to have a negative effect on perceived service quality, customer value and, in the final analysis, on the economic result of the firm" (Grönroos and Osajalo 2004, p.414).

One possible solution of productivity-quality tradeoff is to integrate technological solutions in the service provision process. In particular, companies are racing ahead in investing in Artificial Intelligence (AI) such as machine learning and Internet of Things to increase both front office and back office productivity (Forrester 2019). In a global survey conducted by McKinsey (2018), 75% of managers state that their companies have already begun to automate business processes trough AI or plan to do so within the next year. The

reason is that AI solutions are considered strategic assets for service companies, as they constitute both a resource capable to shape jobs tasks and organizational processes and, at the same time, a major source of market innovations that can increase customer satisfaction (Huang and Rust 2018). Considered as a sort of new factors of production, AI solutions carry big hopes as their application is expected to improve both the efficiency of internal processes and the effectiveness of customer offers, potentially resolving the tradeoff between service productivity and quality. Such possible double effect explains the current hype associated to the diffusion of AI, as its solutions might redefining the two core elements of a firm's successful business model: the supply model and the revenue model (Cachon 2018).

In terms of productivity, AI solutions promise to save 20%-40% of costs and streamline routine activities, thus generating tangible gains in terms of back-office and labor automation (McKinsey, 2020). Nonetheless, AI solutions are trained to handle standard and repetitive tasks, therefore the management of peculiar requests could increase both the times and costs of the service, mitigating the benefits of implementing AI (McKinsey 2019). Regarding quality, AI solutions are designed to reshape market offering and enhance firm-customers interactions with an estimated 15%-20% increase in customer satisfaction (Accenture 2018; McKinsey 2020). However, services marketing literature shows that the integration of technological solutions can have uninteded effects on the overall quality of the service. First, the implementation of AI solutions may reduce the scope of employees' actions, lowering their commitment and, consequently, their satisfaction (Cadwallader et al. 2010; DiMascio 2010). Second, reducing or eliminating the interaction between customers and employees, a fundamental aspect for the service to succeed, AI solutions may have negative effects also on customers satisfaction (Gremler and Gwinner 2008).

Based on the above arguments, understanding the overall effect of AI solutions on service performance represents a non-trivial, often over-simplified, critical research question.

In other words: will AI solutions be able to keep their promises and ensure both internal efficiency and external effectiveness when, as in the case of services, such goals are potentially in a trade-off?

In the paper we address this research question through a field experiment we undertook with the collaboration of a global bank. Banking is expected to face a massive transformation led by AI solutions in the upcoming years, and given its high levels of labor intensity, process complexity, and customer interaction (and especially its retail arm), banking seems a perfect candidate to investigate our research question on the trade-off between efficiency and effectiveness (Calabrese 2012). Indeed, industry reports estimate that AI is going to help financial institutions save $1 trillion in project cost savings and add $1.2 trillion in value to the financial industry by 2035 (Accenture 2019).

Specifically, our partner bank decided to introduce an AI solution to increase the productivity of one of its crucial activities (i.e., mortgage granting), and we set up a research design according to which only a portion of the bank branches were exogenously exposed to the AI solution for a certain period. This allowed us to generate treatment and control conditions during a specific time window. Being able to collect unique data streams regarding core performance metrics (i.e., process efficiency, employee satisfaction, offering, and customer satisfaction) before and after the treatment period, we can estimate the effect of AI on both internal efficiency and external effectiveness.

Drawing on the Service Profit Chain (SPC) model, we assess the overall effect of AI solution on a set of inter-related service outputs: process efficiency, employees, offering, and customers (Hogreve et al. 2017). More specifically, according to SPC model, a successful AI implementation should: (i) enable the internal process to deliver the service efficiently (i.e., process efficiency), (ii) improve employees' satisfaction, so they have positive attitudes, are more productive and engage in behaviors to support both the organization and the customers

(Harrison, Newman, and Roth 2006), (iii) improve the service offering, (iv) increase customer satisfaction that leads to customer loyalty and positive word of mouth (Fournier 1998; Rust and Zahorik 1993).

Institutional setting

The field experiment was conducted in collaboration with the Italian division of a global bank operating in 70 countries, which is among the largest financial institution of the country with 680 branches, 2.7 million customers and 13.000 employees. The national division (from here on simply the Bank) planned to renovate and reshape all their business processes (internal and external) through the integration of Artificial Intelligence (AI) solutions.

The first step of its renovation process regards their internal system of ticketing to support employees. While fundamental to ensure a meaningful customer interaction in the service encounter, the ticketing process represent an typical area that is ripe for efficiency gains, in terms of time and quality of the feedback Before the advent of the AI solution, the ticketing support process was entirely handled by human resources. When an employee needs any kind of support regarding a banking procedure (e.g., mortgage granting, information on stocks investments, deposits), she opens a ticket using an internal platform in which she explains what the problem is. Once the ticket is opened, the ticket is then sent to the manual execution task force, a group of 110 back office employees, for resolution. This unit is unique for the whole domestic market in response of all the inquires of all the branches and it manages more than 1M tickets per year.

In the last year the bank management decided to implement an AI solution in this ticketing support system, focusing first on the tickets that emanate from the *process of mortgages granting*, as this process is critical for bank's offering and it accounts for a significant amount of tickets. Given the complexity of the mortgage process, this

implementation is particularly interesting, as it puts to the test the cognitive capabilities of AI, that should solve also non-trivial requests. Our field experiment exactly focuses on this implementation that, according to managers, should increasing the automatization of the resolution of employees' tickets; reduce the resolution time of those tickets; and save costs on the ticketing support system.

This AI solution's way of working is based on its capability to understand, in a totally unsupervised way, the textual content of the tickets. The implemented AI solution is able to recognize content within images and documents (visual recognition), extract and classify the main entities present in texts (natural language classification), highlight elements of interest such as names and addresses (names and entities recognition) and analyze structured and unstructured textual content (natural language processing). Once the AI analyzes the content of the ticket, it searches for a match on a repository of solved tickets, bank's protocols and manuals, and provides the most appropriate response given the level of accumulated knowledge together with a level of confidence for the proposed solution. It is worth noting that, when the AI self-estimates that the level of confidence of its proposed solution is above 90%, the response is directly provided to the branch employee who raised the request. In cases in which the confidence level of the solution provided by the AI is below 90%, the response suggested by the AI is sent to the manual task execution group that is free to consider, integrate or discard the suggestion before sending back the feedback to the employee. Figure 1 shows the entire AI solution workflow.

Insert Fig 1 about here

The aim of the research is to understand the effects of this AI solution that handles the tickets opened by employees during the process of mortgages granting on bank's productivity metrics, such as resolution time (i.e., process efficiency) and employee satisfaction, but also on quality metrics, like mortgages granting (i.e. offering) and customers satisfaction.

Experimental Setting

The time window of our study ranges from January 2019 to March 2020 and involves the tickets concerning mortgages opened during this period. Before the real implementation occurred in December 2019, the AI solution went through different preparation stages (pilot, tuning and training)[3], instrumental for properly training the AI algorithm. These preparatory stages were fundamental to calibrate the solution and ensure the rollout started only when its cognitive capabilities reached a satisfactory level.

To ensure a clean test of the AI implementation effects, we excluded from the analysis these preparation stages occurred between August and November 2019. At the end, our dataset includes all 7,250 tickets issued:

- *before the treatment* period (Jan 2019-July 2019), when the AI solution was not present and all the tickets were traditionally handled by the manual execution task force (5,184 tickets);
- *during the treatment*-period, from (Dec 2019-Feb 2020), when the AI solution was in place and able to handle automatically all the tickets open by employees' branches (2,066 tickets).

Table 1 summarizes the number of tickets in the treatment and control groups during the pre-treatment and the treatment period.

Insert Table 1 about here

3 The pilot stage was a short period in which all the tickets were processed by the AI, to fix the operational activity of the solution. The tuning stage was another time window in which the tickets processed by the AI were then evaluated and handled anyway by the manual task execution group: this in order to explore the AI/human interaction for those tickets that may require it. Third, there was the classic training period in which the AI processed all the tickets in order for a period of time in order to improve its ability to provide solutions (see Appendix A).

Thanks to the collaboration with the Bank, we obtained to generate a temporary heterogeneity of the AI rollout in the network of branches. Specifically, for a two-month period (Jan 2020-Feb 2020), the AI solution was introduced only in some branches of the retail network while, in the same two-month period, the remaining branches continued to handle the tickets using traditional human interaction in place before the treatment period. Due to the complexities of the changes required in the bank's IT system to keep this double-system of ticket management in place, a random assignment of all the 638 branches across the two conditions was practically infeasible. Hence, the bank hierarchically allocated the branches to the treatment arms on the basis of their macro area geographical location. Specifically, 126 branches located in two macro areas were intentionally assigned to the control condition, while the remaining 512 branches located in the other eight macro areas were assigned to the treatment condition (see Tab 2 for details). While not random, this specific assignment regime was set up to minimize the threats to identification associated to the technique (DID) we use for empirically estimating the effects (see below for details). This choice aligns our analysis to a valuable type of natural experiment in which *verifiable* external factors determine treatment assignment (Titiniuk 2020).

Insert Table 2 about here

First, while each area could have been considered for inclusion in any of the treatment arms, the allocation ensured that, on average, the areas selected in the control group were comparable with those in the treatment group in terms of economic conditions, with GDP per capita equal to 32.6k € and 27.7k € respectively. The two areas included the control conditions, selected to act as a valid counterfactual, are also similar in terms of economic conditions and are relevant, as they absorb 25% of the national GDP. Considering that financial regulations and bank strategies operate at the national level and the treatment period is short, time-varying potential confounders at the area level are unlikely to emerge.

Second, while treatment is assigned at the area level, our effects of interest operate at the branch level. In this sense, as branches are hierarchically allocated to treatment conditions and they are unaware of the trial, this assignment regime avoids self-selection bias (branches cannot choose the treatment arm), helps ensuring SUTVA conditions (branches cannot switch treatment arm), and preclude strategic behaviors oriented to anticipate the expected effect (Lechner 2010).

Third, given the unique and homogeneous retail model implemented by the bank, the two treatment arms result comparable also in terms of pre-treatment outcome levels and branch features that can matter for the outcome (see Table 3). This situation facilitates the satisfaction of key assumptions for identification, as it increases the chance that potential effects are idiosyncratic to the treatment period (parallel trend) and that covariates that can affect the outcome trend are group invariant (common shocks) (Angrist and Pischke 2009).

Insert Table 3 about here

Finally, it should be pointed out that this asymmetric allocation scheme responded to a specific purpose. As the AI might not be able to address all the tickets assigned to it (see Fig.1), some of tickets in the treatment group will be handled by AI only, while others will be treated by AI + human intervention. The asymmetric allocation facilitates a follow-up balanced investigation of the treatment effect across three possible levels: “only human” (control condition) vs. “human after AI” (treatment condition) vs. “only AI” (treatment condition).

Dependent Variables and Modeling Approach

As anticipated, our analyses are rooted on the Service Profit Chain (SPC) model, which states service performance are multi-dimensional and should be assessed across a series of metrics regarding both productivity and quality: internal process quality, employees’

satisfaction, external offering, and customers' satisfaction (Hogreve et al. 2017). Based on the above, we tested the effect of our AI treatment on four different dependent variables related to the Bank's service performance, as shown in Fig. 2. In order to implement these analyses, we leverage on four different streams of data, pre and post treatment, provided by the bank: i) secondary data on ticket resolution processing; ii) survey on employees' feedback on ticket resolution; iii) secondary data on mortgage granting at branch level; iv) survey on customer satisfaction from clients who got the mortgage granted.

Specifically, Study 1 focuses on the *ticket resolution time*, our core measure of process efficiency: a quicker response to the tickets will streamline the internal process of mortgage granting, allowing employees to solve the problem and serve the customers more efficiently. In Study 2 we move our attention to the ticket *employee satisfaction*: employees' satisfaction is pivotal to the service provision process since employees that are satisfied with the ticket solution are more likely to have a positive attitude, to be more productive and to engage in behaviors to support both the organization and the customers (Harrison, Newman, and Roth 2006). In Study 3 we investigate the effect of our treatment on the breadth of the *service offering*, in particular our dependent variable is the ratio between the mortgages granted and the applications received by the branches. This ratio reflects the extent to which internal efficiency gains spill over to the external market offering. Lastly, understanding the effect of the AI solution on customers is important since *customers satisfaction* leads to customer loyalty and positive word of mouth (Fournier 1998; Rust and Zahorik 1993): hence, in Study 4 we focus on Net Promoter Score (NPS), customers evaluation of employees and perceived complexity of the mortgage granting process as DVs.

Insert Fig 2 about here

Given the non-random exogenous nature of our assignment mechanism and the availability of various data before and after the treatment period, we use a difference-in-

difference (DID) approach to estimate the AI effects. It should be noted that in three out of the four models portrayed in Fig 2, the criterion variable does not operate at the branch level but at a lower level: in Study 1 and 2 the effects emerge at the ticket level while in Study 4 at the customer interview level, in all cases within the branch. To account for the non-independence of observations in those cases, we cluster our analysis over space (Rabe Heskett and Skrondal 2008) and set up a DID model with a random intercept at the branch level (Study 1; Study 4) and employee level (Study 2). The formal notation for our DID model is the following:

$$[1] \quad Y_{ij} = aX + b_1 treat + b_2 post + b_3 treat*post + u_{0j} + e_{ij}$$

With: *i* = observation (ticket, interview); *j* = cluster (branch, employee); *treat* = treatment dummy; *post* = pre-post dummy; *Y* = outcome variable; *X* = vector of covariates (including constant); *u* = random effect (estimated); *e* = residual error

The specific functional form for estimating Y will depend on the nature of the dependent variables which, as we will explain below, in the first two studies are binary (hence Probit), in study 3 is a ratio (hence Fractional Probit), and in study 4 is metric (hence OLS). The parameter of our interest is the coefficient b_3, which still represents the average treatment effect on the treated for both linear and non-linear DID models (Karaca-Mandic, Norton, and Dowd 2012), and from which we derive marginal effects for interpretation.

Empirical Analysis and Results

Study 1: Resolution time (Process efficiency)

When a ticket is opened, the system assigns it an expected resolution time (SLA) on the base of the complexity of the request. The expected resolution time is 1 day for simple tickets (e.g. "a customer asks us about the possibility of mortgaging a property awaiting building permit. Can we proceed?") and 3 days for complex tickets (e.g., "The client requests commercial renegotiation of his mortgage. From the analysis of its credit position we verify that at the

moment the income parameters are not respected both for the installment ratio and for the subsistence threshold. Can the renegotiation still be possible?"). Taking into consideration the different complexity of the tickets, we first operationalized our measure of resolution time as actual resolution time minus expected resolution time (SLA). While the treatment (vs. pre-treatment) period saw a greater proportion of complex tasks (63% vs. 55%; $\chi^2_{(1)}$ =41.87; *p*=.001), a logit regression shows this difference does not change across treatment and control groups (b= 0.007, p>.1).

Given that the distribution of the resolution time measure is severely left skewed (Sk=31.8), highly kurtotic (K=1,755) and includes several important outliers, we decided to use a binary dependent variable equal to 0 if the actual resolution time is within the expected resolution time (i.e. within the SLA expected resolution time) and 1 if actual resolution time higher than expected resolution time (i.e., beyond the SLA expected resolution time). The dependent variable in our DID is thus the probability to handle the ticket in an inefficient way.

As anticipated, since the dependent variable is at ticket level within branch, we use a Probit regression with random intercept at branch level. A likelihood-ratio test, comparing pooled estimator (i.e., simple Probit) and panel estimator (i.e., Probit model with random effects at branch level), supports the use of random effects at branch level ($\chi^2_{(1)}$ =15.49; *p*=.001).

Results from Table 4 suggest that our treatment reduces the probability that the actual resolution time is greater the SLA threshold (b=-.169; *p*=.06).

Insert Table 4 about here

Marginal analyses (see Figure 3) reveal that the probability to miss the SLA increases in treatment group during the treatment period (Pr_{T0} = 16.7%; Pr_{T1} =22.0%) but significantly

less (-5.5 percentage points; p<.05) than what happens in the control group (Pr_{C0} = 17.5%; Pr_{C1} =28.3%).[4]

Insert Figure 3 about here

We then relax the assumptions of the random effect modeling, and replicate the effect using both a population-averaged Probit model (b=-.167; *p*=.06) and a conditional Logit model with branch fixed effect (b=-.383; *p*=.03). In both cases, results support the evidence that the AI treatment reduces the probability to miss SLA. We also replicate the analysis including a region fixed effect, and results still maintain that our treatment negatively affects the probability to miss the SLA (b=-.163; *p*=.07).

As anticipated, not all the tickets of the branches assigned to the treatment group were handled by AI only, but when the response from the machine is self-assessed to have a reliability below 90%, the human task force takes the AI response as a mere suggestion, but it has to take decisions and then deliver the final answer to the branch employee. Hence it is possible to run a follow-up analysis within the treatment group that reveals how the efficiency effect above described is entirely driven by the tickets managed by AI only (i.e., when the human intervention is substituted by technology). Specifically, the 100% of ticket processed only by AI are within the expected resolution time, while only 72% of the ticket processed by the manual task execution in the control group is within SLA ($\chi^2_{(2)}$ = 97.5; *p*=.001). Contrary, the percentage of tickets solved within SLA is not different between tickets processed by AI and integrated by human (AI+Human) in the treatment group and the tickets only processed by human in the control group (73% vs 72%; $\chi^2_{(2)}$ =.40; *p*<.1). It

4 We replicate the analysis using a three-level dependent variable: -1 resolution time<expected resolution time;0 resolution time=expected resolution time; 1 resolution time>expected resolution time. Results suggest that our treatment does not affect the probability to beat the SLA (b= -.052; p>.1), but negatively affects the probability to miss the SLA (b= -.30; p=.075).

suggests the effect we detected is almost entirely driven by the tickets handled by AI only, because when AI simply filters the request, the efficiency gain in terms of timing vanishes.

Pre-treatment parallel trend

The Difference in Difference (DID) identification first rests on the assumption that average change in the dependent variable for the treatment group in the pre-treatment period equals the average change in the dependent variable for the control group (Mora and Reggio, 2015). Hence, we check the trend of our dependent variable across treatment and control groups in the pre-treatment period for which we have 4 two-month periods. During the pre-treatment periods, the difference in our dependent variable is not significantly different from 0 ($\chi^2_{(3)} = 1.24$; $p>.1$). Figure 4 shows the Probabilities to miss the SLA of treatment and control group during the 4 periods of the pre-treatment period that are consistent with the assumption of parallel trend.

Insert Figure 4 about here

Common shocks

The DID model requires another assumption for identification, namely that anything else beyond the treatment affects subjects in both control and treatment groups in the same way (Zeldow and Hatfield 2019). While several covariates at the branch level (that are time invariant) are balanced between groups (see Table 3) and in any case do not show time-varying effect on the DV[5], we identify a possible time-varying confounder that could affect our dependent variable. Specifically, the number of tickets opened by each branch, which may affect the resolution time while varying over the treatment period. Hence, we replicate

5 In particular, we tested whether the number of employees and the structure of the branch could have some time-varying effect on our DV. Results reveal that neither the number of employees (b=.001; p>.1) nor the structure of the branch ($\chi^2_{(2)} = 2.46$; p>.1) have a time-varying effect on probability to miss the SLA. Moreover, we found that our treatment effect holds also controlling for the time-varying effect of number of employees (b=-.171; p=.06) and structure of the branch (b=-.167; p=.06).

our analysis including the interaction between the number of tickets and time (i.e. pre and post treatment period) (Zeldow and Hatfield 2019). We found that the interaction between number of tickets and time does not affect our dependent variable (b=-.007; *p*=>.1), while our treatment has a negative effect on the probability to miss the SLA (b=-.166; *p*=.07). Hence, our effect seems robust to this potential confounding effect.

Study 2: Employee Feedback (Internal Satisfaction)

When a ticket is solved, the employee who opened it chooses whether to give a feedback regarding its solution. Our dataset contains 1,134 employees' feedbacks. It is important to remind that employees did not know whether their tickets are processed by AI or by the human task force. In detail, the bank collects three feedback measures from the branch employee who opened the ticket: i) the accuracy of the solution, ii) fit between the problem and the solution and iii) timeliness of the solution.

All the measures are based on a scale from 0 to 10. A principal component analysis reveals that the three measures are unidimensional (eigenvalue= 2.75; 92% of variance explained) and show high reliability (α= 95.17%). However, the measure is left-skewed (skewness= -3.73), with a mean of 9.3, a standard deviation of 1.86, and a kurtosis index greater than 10. Substantially, being peer evaluations among employees, high scores for this type of feedback are expected: while low scores are unlikely, employees traditionally manifest their dissatisfaction by negating the maximum evaluation. For these empirical and substantial reasons, we thus dichotomize the original scale with a median split, such that we account for employees' feedback if it is 10 (73%). Despite the loss of information associated with such dichotomization, the practice is justified in the presence of extreme skewness in the distribution of the original metric, which can lead to significant bias (MacCallum et al. 2002).

The voluntarily nature of these feedbacks can generate self-selection bias, but we provide some evidences that can attenuate this bias. First, the proportion of employees who did not provide a feedback is equivalent across treatment and control group (91.36% vs. 92.10 %; $\chi^2_{(1)}$ =.833; p>.1). Second, a non-response bias test (Armstrong and Overton 1977) reveals that the feedbacks from the early (first quartile) and the late (fourth quartile) employees are not different ($\chi^2_{(1)}$ =.94; p>.1).

As for resolution time, also employee satisfaction is at ticket level, hence we again use a Probit regression with random effects in this case at the employee level to account for any idiosyncratic effect of the subject who provides the feedback. A likelihood-ratio test, comparing pooled estimator (i.e., simple Probit) and panel estimator (i.e., Probit model with random effects at employee level), supports the use of random effects at employee level ($\chi^2_{(1)}$ = 78.04; p=.001).

The findings suggest that our treatment does not provide a significant average effect on employee feedback (b=.012; p>.1). As did in Study 1, we conduct a follow-up analysis focusing only on post-treatment period: we estimate the effect on the dependent variable of our two levels of treatment: tickets processed by AI and tickets processed by AI+ Human. Our results suggest that resolution by AI lowers the probability the employee provides a 10 score (b=-2.43; p=.04), while AI + Human resolution does not affect the dependent variable (b=-1.04; p>.1) (Table 5). Specifically, while tickets processed by humans have 65% of probability get a score of 10, this probability decreases to 50% for tickets processed by AI and 54% for tickets processed by AI + Human (p<.05). These results replicate also using population-averaged Probit model (b=-.522; p=.05). It suggests that even for this service outcome, AI provides an effect that is entirely explained by the tickets handled by AI only. However, while for process efficiency the effect was positive, in this case, the effect of AI implementation on employee satisfaction is negative.

Insert Table 5 about here

Study 3: Mortgage Granting (service offering)

Our third dependent variable shifts the attention from the internal side to the external side of the service offering. The focus here is on the immediate consequence of the internal process, namely if the employees accept the customers' requests and close the mortgage deal: hence, the focal dependent variable here is the ratio between granted mortgages and applications. Contrary to the previous variables, the ratio between granted mortgages and applications operates at branch level, hence, it is not necessary to include any random effect in our analysis. We estimate the effect of our treatment using a Fractional Probit model since our dependent variable (i.e., a ratio) is bounded between 0 and 1.

Insert Table 6 about here

The results shown in Table 6 suggest that our treatment has a negative effect on the dependent variable (b= -.15; p=.008) (see Figure 5). We replicate the analysis using an OLS estimator and found the same somewhat unexpected negative effect we found in previous model (b= -.047; p=.002).

Insert Figure 5 about here

Pre-treatment parallel trend

As did before, we check whether the trend of our dependent variable across treatment and control groups in the pre-treatment period (for which we have 3 8-week periods) is consistent with the parallel trend assumption. During the pre-treatment periods, the difference of our dependent variable across treatment arms is not significantly different from 0 ($\chi^2_{(2)} = 1.12$; $p>.1$). Figure 6 shows the means of the ratio between granted mortgages and

applications for treatment and control group during the 3 8-week periods of the pre-treatment period[6].

Insert Figure 6 about here

Follow-up analysis: ticket solution text analysis

Our findings suggest that branches in which the tickets are processed by AI and AI + Human are more selective in granting mortgages. We discussed with two managers of the bank about the possible explanations for this unexpected effect. From the discussion, two possible explanations emerged. Since the AI partially bases its responses on bank's manuals it can provide i) more conservative/negative responses or ii) more strict, unambiguous responses (without proposing alternative procedures to follow). In order to investigate these possible explanations, we analyzed the textual content of the tickets' solutions. We use LIWC, an automated textual analysis software (Pennebaker et al. 2015). LIWC uses a validated dictionary of more than 6,000 words to capture the structural components of a text and provides quantitative scores for dimensions such as negative and positive sentiment based on word frequency. Since the solutions are in Italian, we used the LIWC software version based on the Italian dictionary.

First, we test whether the tickets processed by AI and AI + Human contain more negations. We found that the solutions processed by AI + Human contain the same amount of negations than the solutions processed by Humans (b= -.04; p>.1) but, contrary to

6 For mortgage granting we did not find a time-varying variable that could affect our DV. However, as for resolution time, we tested the possible time-varying effect of number of employees and branch structure. We found that neither the number of employees (b=-.001; p>.1) nor the structure of the branch ($\chi^2_{(2)} = 3.22$; p>.1) have a time-varying effect on the ratio between mortgages granted and applications. Moreover, we found that our treatment effect holds also including in the estimation the time-varying effect of number of employees (b=-.148; p=.01) and structure of the branch (b=-.150; p=.01).

expectations, solutions provided by AI only have significantly less negations than the ones provided by the manual task execution group (b= -.45; p=.001). Results replicate also using as dependent variable a dummy equal to 1 whether in the solution there are negations and 0 if there are not (b= -.74; p=.001). Marginal analysis reveals that solutions provided by AI have 25% less probability to contain negations than solutions provided by humans (p=.001). The findings suggest that the AI does not provide more negative solutions, which then is unlikely to explain the decrease of granted mortgages in treated branches.

Then, we explored whether the solutions provided by the AI contain less words related to possibility, so to generate stricter and less ambiguous indications. We found that, in line with the second explanation, the solutions processed by AI have significantly less words related to possibility than the ones provided by the manual task execution group (b= -.64; p=.001). As for negations, the results replicate using as dependent variable a dummy equal to 1 whether in the solution there are words related to possibility and 0 if there are not (b= -.40; p=.001). Marginal analysis reveals that solutions provided by AI have 15% less probability to contain words related to possibility than solutions provided by humans (p=.001) and AI+ Human (p=.002). In short, it seems that the type of feedback provided by AI is less prone to interpretations/adjustments that sometimes happens in the human-to-human interactions, and this seems associated to a more straightforward and restrictive decision on the offering.

Study 4: Customer Satisfaction

Our final analysis focuses on the terminal metric of the SPC model: customer satisfaction. The analysis is possible as bank regularly surveys customers who had their mortgages granted. The bank collected 1,797 interviews during pre-treatment and treatment period. Yet, contrary to what happened for the feedbacks from employees, we don't have information about the customers who did not fill the questionnaire as they have not been

surveyed. The bank collects from the customers three measure of satisfaction: Net Promoter Score (NPS), evaluation of the frontline employee and perceived complexity of the process. To increase the accuracy of our estimation, we include the number of visits to the branch as control variable to take into account the level of interaction between the customer and the bank.

We first test the effect of our treatment on NPS. NPS is based on the response to the question "How likely is it that you would recommend our bank to a friend or colleague?" with a score that can range from 0 to 10. Customers who assign a score from 0 to 6 are called *detractors* (i.e., they will advise against the bank), customers who respond 7 or 8 are *passives* (i.e., they will not recommend the bank), and customers who assign 9 or 10 are *promoters* (i.e., they will recommend the bank actively). We investigate the effect of our treatment on NPS using a mixed effect regression model including a random intercept at the branch level. Results shown in Table 7 (model 1) suggest that our treatment increases NPS (b=.72; p=.019) (Figure 7a). We further qualify our analysis using as dependent variable a categorical measure of NPS equal to -1 for *detractors*,0 for *passives* and 1 for *promoters*. We tested the effect of our treatment using a multinomial logit and we found that processing tickets through AI reduces the probability that customers are detractors (b= -1.33; p=.05) but it does not affect the probability to be promoters (b= -.14; p>.1). It means the positive effect is mostly explained by a reduction of dissatisfied customers.

Insert Table 7 about here

Our second dependent variable is the customer's evaluation of the frontline employee who closed the deal. The bank collects three measures regarding frontline employees: clarity of information provided, comprehension's capacity and problem-solving skill. These three measures are based on a 0 to 10 scale. A principal component analysis reveals that the three measures have only one component (eigenvalue= 2.81; 93.7% of variance extracted) with a

high level of reliability (α= 96.4%). Hence, we create a composite variable of evaluation of the frontline employee. The results (Table 7, model 2) from a mixed-effects regression model with random effects at branch level show that our treatment has a positive effect on evaluation of the frontline employee (b=.62; *p*=.05) (Figure 7b).

Insert Figure 7 about here

Our last dependent variable is the perceived complexity of the granting mortgage process. The bank collects a measure of perceived complexity of the process on a scale from 1 to 5. As for NPS and evaluation of frontline employee, we analyze the effect of our treatment on perceived complexity of the process using a mixed-effects regression model with random effects at branch level. The findings reported in Table 7 (model 3) show that our treatment reduces customers' perceived complexity of the process (b= -.23; *p*=.06) (Figure 7c).

Pre-treatment parallel trend

We check whether the trends of our dependent variables across treatment and control groups in the pre-treatment period (for which we have 3 8-week periods) are consistent with the parallel trend assumption. During the pre-treatment periods, the differences in NPS ($\chi^2_{(2)}$ = 1.55; *p*>.1), frontline employee evaluation ($\chi^2_{(2)}$ = 3.18; *p*>.1) and perceived complexity ($\chi^2_{(2)}$ = 1.44; *p*>.1) are not significantly different from 0. Figure 8 shows the predicted means of NPS (a), frontline employee evaluation (b) and perceived complexity (c) for treatment and control group during the 3 8-week periods of the pre-treatment period.

Insert Figure 8 about here

Common shocks

As for resolution time, we identify a possible time-varying confounder that could affect our dependent variable: the number of interviews collected in each branch. Hence, we replicate our analysis including the interaction between the number of interviews and time (i.e., pre

and post treatment period). We found that the interaction between number of interviews and time does not affect NPS (b=-.090; p>.1), frontline employee evaluation (b=-.091; p>.1) and perceived complexity (b=.024; p>.1). Moreover, results suggest that our effect is robust to this potential confounding effect. Specifically, our treatment has a positive effect on NPS (b=.705; p=.02) and frontline employee evaluation (b=.605; p=.06), while it reduces customers perceived complexity of the process (b=-.23; p=.07).

Discussion

Service literature has amply shown that, in services, an increase in productivity is often associated with a decrease in quality (e.g., Calabrese 2012; Grönroos and Osajalo 2004). In this paper we investigated whether the implementation of AI in service could solve the tradeoff between productivity and quality. We implemented a field experiment with the collaboration of a global bank that has recently implemented an AI solution to increase the productivity of solving tickets regarding mortgage granting. During the period of our experimentation, the tickets opened by the employees of the branches in the treatment group were processed by AI, while tickets of branches in the control group were processed by the manual task execution force. In treatment group, the AI estimates the level of confidence of the solution provided. If the level is above 90%, the solution is directly provided to the employee. In cases in which the confidence level of the solution provided by the AI is below the threshold, the response suggested by the AI is sent to the manual task execution group which integrate the response and send it to the employee. Drawing on the Service Profit Chain (SPC) model (Hogreve et al. 2017), we tested the effect of this AI solution on both internal (i.e., resolution time, employees' satisfaction) and external metrics of service performance (mortgage granting, customers' satisfaction).

Regarding process efficiency, we found that the implementation of AI decreases the resolution time of tickets. The effect size is about 5% percentage points reduction in the chance of missing expected resolution time. Yet the effect is entirely driven by tickets handled by AI only, while there is no difference in efficiency between handling tickets only by human or by human after an AI mere filtering. When the AI platform is not powerful enough to significantly substitute the human presence, efficiency gains remain limited: in our case, tickets managed by AI only represent a minority proportion (17%) and hence conditions of opportunity costs seem to arise. This result is in line with the AI solutions promise to streamline routine activities, thus generating tangible gains in terms of back-office and labor automation (McKinsey 2020), but mostly when it substitutes (vs. complement) humans.

We then tested the effect of AI implementation on employees' satisfaction. While our treatment does not affect the evaluation provided by employees on the average, a follow-up analysis reveals that resolution by AI-only has a significant negative effect on evaluations by employees. In other words, frontline employees find that solutions provided by AI are less accurate and appropriate than solutions provided by humans. This finding suggests that even if AI solutions can increase the productivity, it may have some drawbacks (e.g., handling special requests) that can be only overcome by humans' experience (McKinsey 2019). But, as said before, this would compromise process efficiency.

When we move from internal to external side of the activity, which cannot be separated in the case of services, our results show that the same AI solution makes treated branches more conservative in granting mortgages. To inspect a potential reason for this result, we analyzed the textual content of the solutions provided by AI (vs. humans) using LWIC. Our analysis reveals that solutions provided by AI have a lower probability to contain words related to "possibility". This result validates the argument that AI solutions are able to perform standard and repetitive tasks, but they are not effective in proposing alternative

solutions to complex problems (McKinsey 2019). In our case, the AI is trained to search a solution to the tickets in a repository of old tickets and bank's manuals, so it provides the most straightforward solution to the problem while humans are more inclined to propose multiple ways to solve the problem. What is likely to happen is that efficiency (satisfaction) gains from human substitution (complementation) translate into a more streamlined (more open) indications for market actions delivered through the retail network. Hence, depending on the side of the trade-off chosen, there is a different consequence brought by AI (expanding vs. constraining) on the service offering.

Moving to the market feedback in terms of quality of the offering, we found AI solution improves customer satisfaction but only in terms of reducing the proportion of brand detractors. It suggests the role of AI is to mainly to avoid problems generated by human inefficiencies in the relationship with frontline employees. In fact, customers of treated branches are happier with the personal relation they experienced at the encounter especially because they felt their service delivery process being less complicated. This result, being based only on customers who signed the mortgage, are consistent with the previous results: the AI solution tends to make customers who closed the deal more satisfied for the efficiency gains and the quality of the frontline interaction. However, we do not know what is the impact of the fewer deals closed on unserved customers and, in general, on the branch bottom line.

Going back to our research question, "will AI solutions be able to ensure both internal efficiency and external effectiveness when, as in the case of services, such goals are in a trade-off?", the set of our results suggest that, despite some positive effects, trade-offs are still there and may constrain the implementation of AI solutions. The interesting and overlooked thing is that trade-offs do not only involve the external/internal dimension of the service offering, but generate tensions also within the internal side (efficiency vs. employee

satisfaction) as well as the external side (breadth of offering vs. customer satisfaction). Our findings reveal that the AI solution generates a positive efficiency effect on the time to handle the ticket but a negative one on the satisfaction of the branch employees involved in the process. AI also produces positive spillovers on customer satisfaction but mostly on the efficiency of the process at the encounter, and, unexpectedly, makes treated branches more prudent/selective in granting mortgages. These findings aim to contribute to a better understanding of the multi-faceted effects of AI solutions, speaking to both theory and practice.

References

Accenture (2018) Redefine Banking with Artificial Intelligence. Report retrieved at: https://www.accenture.com/_acnmedia/pdf-68/accenture-redefine-banking.pdf

Angrist, Joshua D., and Jörn-Steffen Pischke. Mostly harmless econometrics: An empiricist's companion. Princeton university press, 2008.

Armstrong, J. Scott, and Terry S. Overton. "Estimating nonresponse bias in mail surveys." Journal of marketing research 14.3 (1977): 396-402.

Cadwallader, Susan, et al. "Frontline employee motivation to participate in service innovation implementation." Journal of the Academy of Marketing Science 38.2 (2010): 219-239.

Calabrese, Armando. "Service productivity and service quality: A necessary trade-off?." International Journal of Production Economics 135.2 (2012): 800-812.

Di Mascio, Rita. "The service models of frontline employees." Journal of Marketing 74.4 (2010): 63-80.

Forrester (2019) Transform The Contact Center For Customer Service Excellence. Report retrieved at: https://www.forrester.com/report/Transform+The+Contact+Center+For+Customer+Service+Excellence/-/E-RES75001

Fournier, Susan. "Consumers and their brands: Developing relationship theory in consumer research." Journal of consumer research 24.4 (1998): 343-373.

Gremler, Dwayne D., and Kevin P. Gwinner. "Rapport-building behaviors used by retail employees." Journal of Retailing 84.3 (2008): 308-324.

Grönroos, Christian, and Katri Ojasalo. "Service productivity: Towards a conceptualization of the transformation of inputs into economic results in services." Journal of Business research 57.4 (2004): 414-423.

Harrison, David A., Daniel A. Newman, and Philip L. Roth. "How important are job attitudes? Meta-analytic comparisons of integrative behavioral outcomes and time sequences." Academy of Management journal 49.2 (2006): 305-325.

Harvard Business Review (2017) When You've Got to Cut Costs—Now. Article retrieved at https://hbr.org/2010/05/when-youve-got-to-cut-costs-now

Hausman, Jerry A. "Specification tests in econometrics." Econometrica: Journal of the econometric society (1978): 1251-1271.

Hogreve, Jens, et al. "The service–profit chain: A meta-analytic test of a comprehensive theoretical framework." *Journal of Marketing* 81.3 (2017): 41-61.

Karaca-Mandic, Pinar, Edward C. Norton, and Bryan Dowd. "Interaction terms in nonlinear models." *Health services research* 47.1pt1 (2012): 255-274.

Lechner, Michael, and Ruth Miquel. "Identification of the effects of dynamic treatments by sequential conditional independence assumptions." *Empirical Economics* 39.1 (2010): 111-137.

MacCallum, Robert C., et al. "On the practice of dichotomization of quantitative variables." Psychological methods 7.1 (2002): 19.

McKinsey (2018) The automation imperative. Report retrieved at: https://www.mckinsey.com/business-functions/operations/our-insights/the-automation-imperative

McKinsey (2019) Winning in automation requires a focus on humans. Report retrieved at: https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/winning-in-automation-requires-a-focus-on-humans

McKinsey (2020) Service industries can fuel growth by making digital customer experiences a priority. Report retrieved at: https://www.mckinsey.com/business-functions/mckinsey-

digital/our-insights/service-industries-can-fuel-growth-by-making-digital-customer-experiences-a-priority

Mora, Ricardo, and Iliana Reggio. "didq: A command for treatment-effect estimation under alternative assumptions." The Stata Journal 15.3 (2015): 796-808.

Pennebaker, James W., et al. The development and psychometric properties of LIWC2015. 2015

Rabe-Hesketh, Sophia, and Anders Skrondal. Multilevel and longitudinal modeling using Stata. STATA press, (2008).

Rust, Roland T., and Anthony J. Zahorik. "Customer satisfaction, customer retention, and market share." Journal of retailing 69.2 (1993): 193-215.

Rust, Roland T., and Ming-Hui Huang. "Optimizing service productivity." Journal of Marketing 76.2 (2012): 47-66.

Titiunilik Rocio. “Natural experiments”, in *Advances in Experimental Political Science*, James Druckman and Donald Green (eds), Cambridge University Press.

Zeldow, Bret, and Laura A. Hatfield. "Confounding and Regression Adjustment in Difference-in-Differences." arXiv preprint arXiv:1911.12185 (2019).

Table 1

Tickets in the treatment and control groups during the pre-treatment and the treatment period.

	Treatment Group	**Control Group**	**Total**
Pre-treatment Period			
Human	4172	1012	**5184**
Treatment Period			
Human		375	
AI+Human	1401		
AI	290		**2066**
Total	**5863**	**1,393**	**7250**

Table 2

Socio-Economic Data on Macro Areas

	GDP (mln)	Inhabitants	GDP per capita (.000€)	% National GDP
Italy	1,736,601	60,244,639	**28.8**	
Control macro areas				
Piedmont, Liguria, Aosta Valley	188,780	6,010,003	31.4	11%
Triveneto	241,791	7,193,880	33.6	14%
Total	430,571	13,203,883	**32.6**	**25%**
Treatment macro areas				
Calabria and Sicily	121,112	6,893,111	17.6	7%
Campania and Basilicata	119,908	6,342,795	18.9	7%
Emilia-Romagna and Marche	199,144	5,985,518	33.3	11%
Lazio and Sardinia	76,020	3,176,018	23.9	4%
Rome	154,000	4,320,000	35.6	9%
Lombardy	385,347	10,103,969	38.1	22%
Puglia, Abruzzo, Molise	112,261	5,616,331	20.0	6%
Tuscany and Umbria	136,920	4,603,014	29.7	8%
Total	1,304,712	47,040,756	**27.7**	**75%**

Table 3

Balancing between treatment and control groups

Pre-treatment outcomes	Treatment	Control	Difference
Resolution time			
within SLA	84%	82%	$\chi^2_{(1)}$ =.96; p=.32
beyond SLA	16%	18%	
Employee feedback			$\chi^2_{(1)}$ =.76; p=.38
<10	13%	9%	
10	87%	91%	
Mortgages granted/applications	.85	.84	t=-.85; p=.40
NPS	8.1	8.4	t=1.4; p=.17
Customer evaluation of employees	8.5	8.8	t=1.3; p=.20
Customer perceived complexity	2.3	2.1	t=-1.4; p=.17
Branch characteristics	**Treatment**	**Control**	**Difference**
Number of tickets	8.5	8	t=-.82; p=.42
Number of employees	6.3	5.6	t= -1.8; p=.07
Format[7]			
1	65%	68%	$\chi^2_{(5)}$ =9.7; p=.09
2	7%	9%	
3	10%	6%	
4	1%	0%	
5	11%	15%	
6	6%	2%	
Structure			
Branches with subsidiaries	32%	33%	$\chi^2_{(2)}$ =.99; p=.61
Branches without subsidiaries	39%	42%	
Subsidiaries	29%	25%	
Employees characteristics	**Treatment**	**Control**	**Difference**
Gender			$\chi^2_{(2)}$ = 1.2; p=.28
Female	59%	62%	
Male	41%	38%	
Age	46.3	46.1	t=-.30; p=.76
Work experience (yrs)	17.5	18.5	t=1.5; p=.13

7 The bank classifies its branches (i.e., format) on the base of the services offered from 1 (basic services) to 6 (all the services).

Table 4
Probit model with random effects at branch level, Study 1.

VARIABLES	(1) Resolution time (dummy miss the SLA)	(2) /
1.treatment_group	-0.0344	
	(0.0570)	
1.period	0.367***	
	(0.0786)	
1.treatment_group#1.period	-0.169*	
	(0.0900)	
lnsig2u		-3.266***
		(0.310)
Constant	-0.951***	
	(0.0509)	
Observations	7,077[8]	7,077
Number of id_branch	635	635

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

[8] Some observations are omitted for complete separation

Table 5

Probit model with random effects at employee level, Study 2[9].

VARIABLES	(1) Evaluation of employee (dummy evaluation=10)	(2) /
1.treatment(AI)	-2.429**	
	(1.159)	
2.treatment (AI+Human)	-1.036	
	(0.900)	
lnsig2u		3.295***
		(0.433)
Constant	2.263***	
	(0.792)	
Observations	306	306
Number of id_employee	208	208

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

[9] This analysis is conducted in the treatment period

Table 6

Fractional Probit model, Study 3

VARIABLES	(1) Ratio Mortgages granted/applications
1.period	-0.463***
	(0.0509)
1.treatment_group	0.0533
	(0.0506)
1.period#1.treatment_group	-0.150***
	(0.0568)
Constant	0.994***
	(0.0454)
Observations	1,283

Robust standard errors in parentheses

***

Table 7

Mixed-effects Regressions, Study 4

	Model 1			Model 2 Mixed-effects Regression			Model 3 Mixed-effects Regression		
		Random Effect Parameters			Random Effect Parameters			Random Effect Parameters	
VARIABLES	NPS	var(cons)	var(residual)	Evaluation of Employee	var(cons)	var(residual)	Perceived Complexity	var(cons)	var(residual)
Treatment	0.638***			0.460**			0.224**		
	(0.188)			(0.184)			(0.0975)		
Post	0.624**			0.626**			0.152		
	(0.271)			(0.288)			(0.114)		
Treatment#Post	0.723**			0.618*			-0.235*		
	(0.307)			(0.320)			(0.126)		
Branch_visits	0.583***			0.612***			0.275***		
	(0.0509)			(0.0501)			(0.0168)		
var(M1[id_branch])									
var(M2[id_branch])									
cov(M1[id_branch],M2[id_branch])									
Constant	10.39***	0.865**	0.776***	10.53***	1.236*	0.738***	1.283***	1.585***	0.170***
	(0.190)	(0.371)	(0.0383)	(0.189)	(0.703)	(0.0367)	(0.0977)	(0.202)	(0.0260)
Observations	1,267	1,267	1,267	1,264	1,264	1,264	1,267	1,267	1,267
Number of groups	509	509	509	509	509	509	509	509	509

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure 1

Ticket resolution process

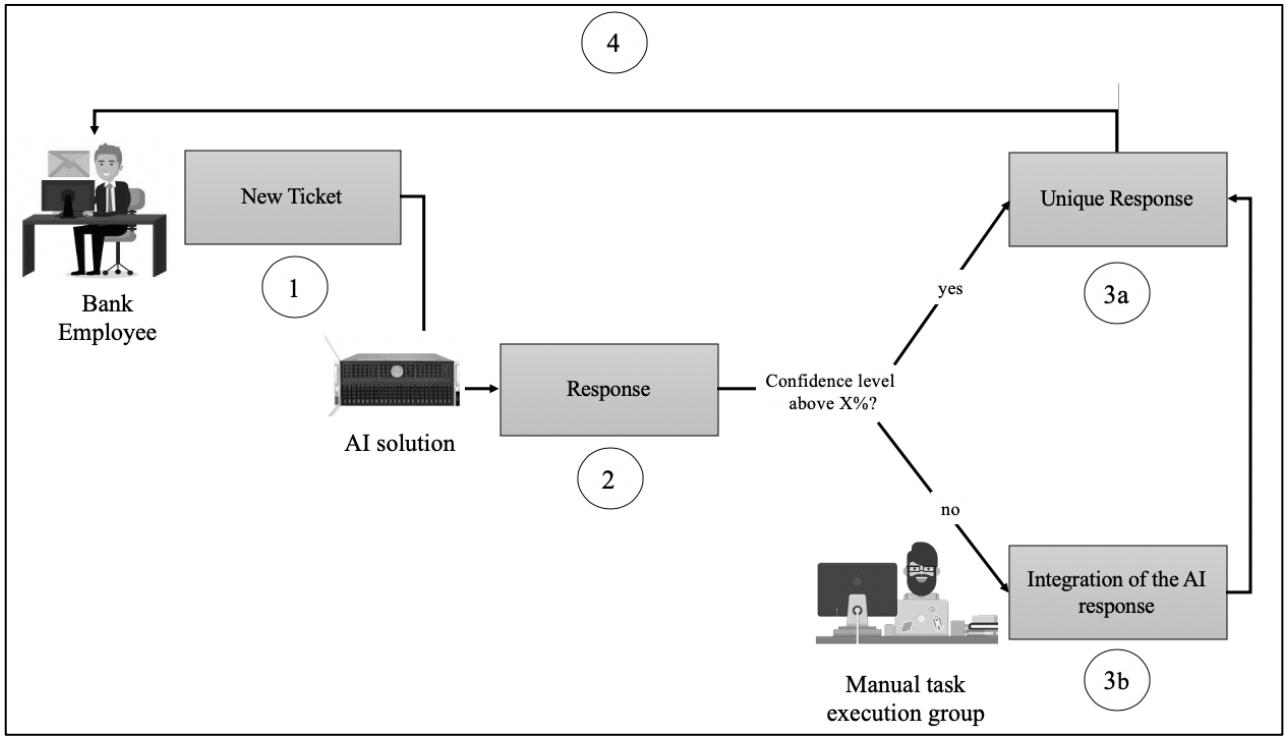


Figure 2

Overview of the empirical analysis

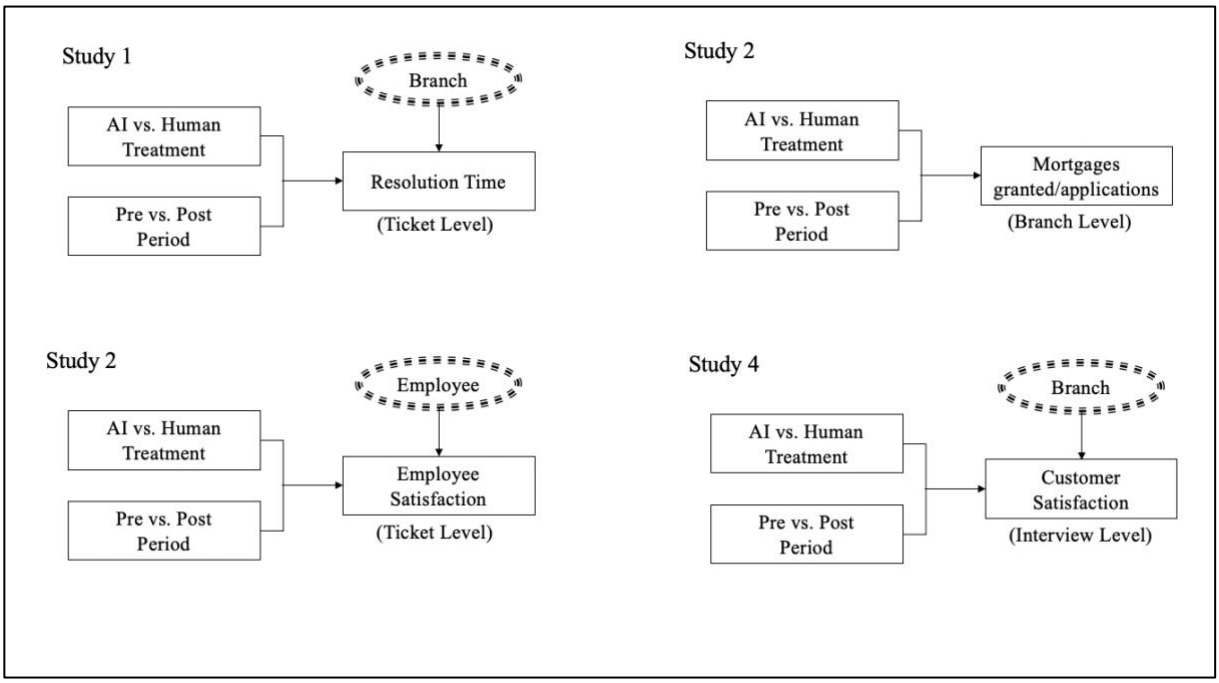


Figure 3

Probabilities to miss the SLA of treatment and control group during the pre-treatment and treatment period.

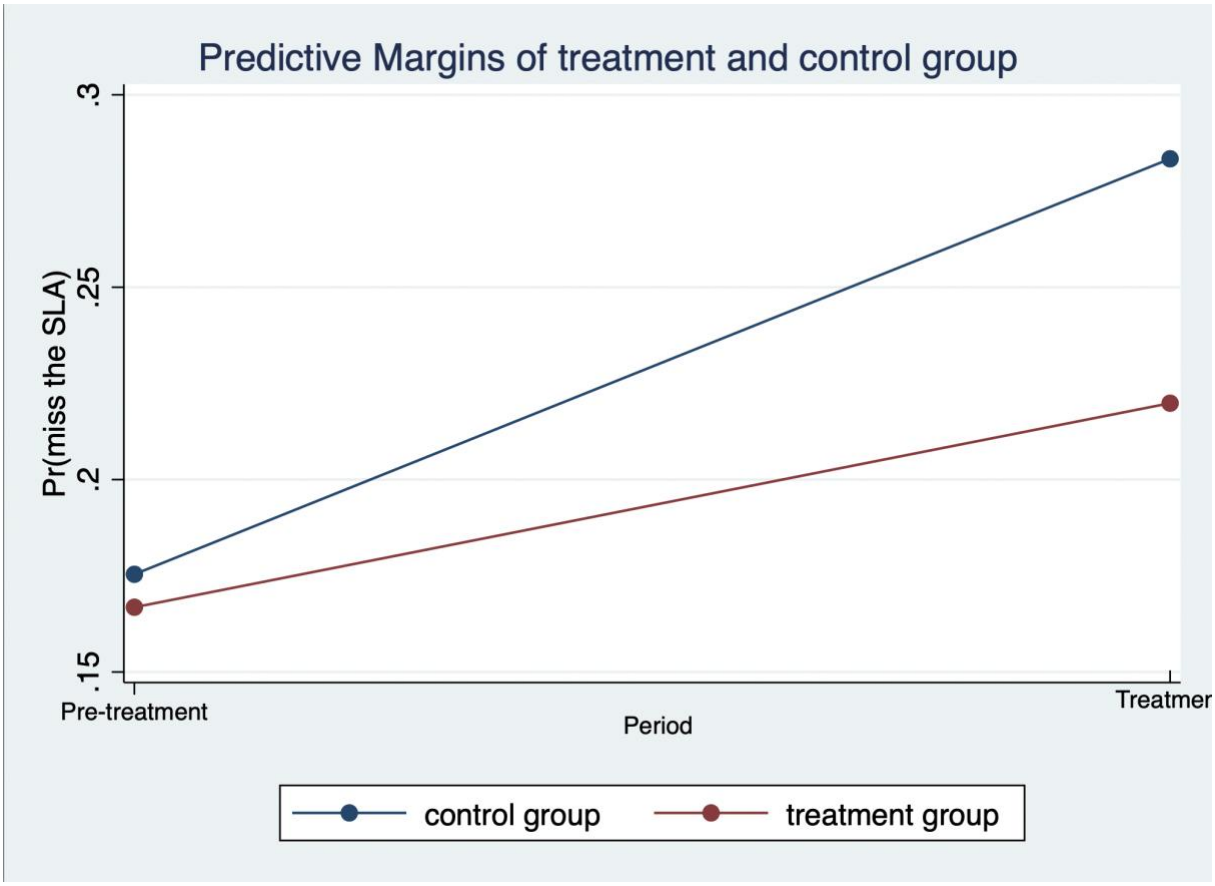


Figure 4

Marginal probabilities to miss the SLA of treatment and control group during the pre-treatment period (2-month periods)

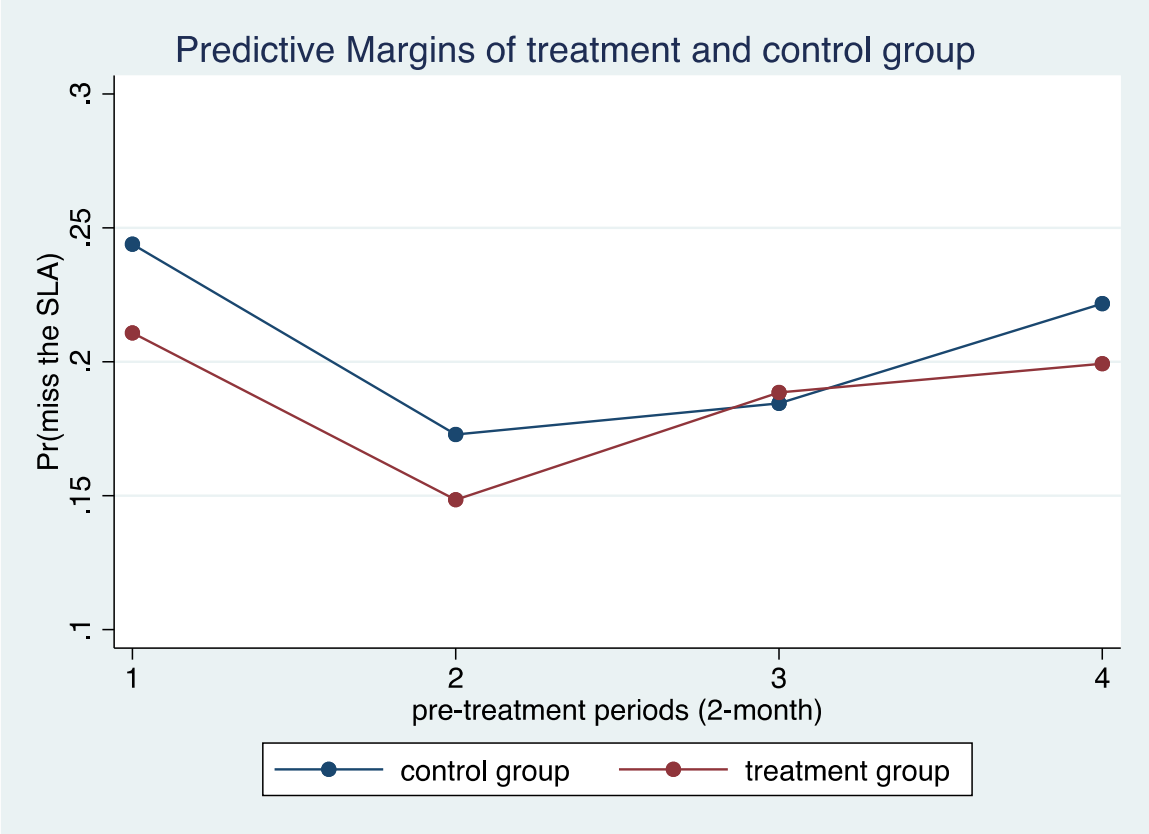


Figure 5

Predicted means of mortgages granting of treatment and control group during pre-treatment and treatment period

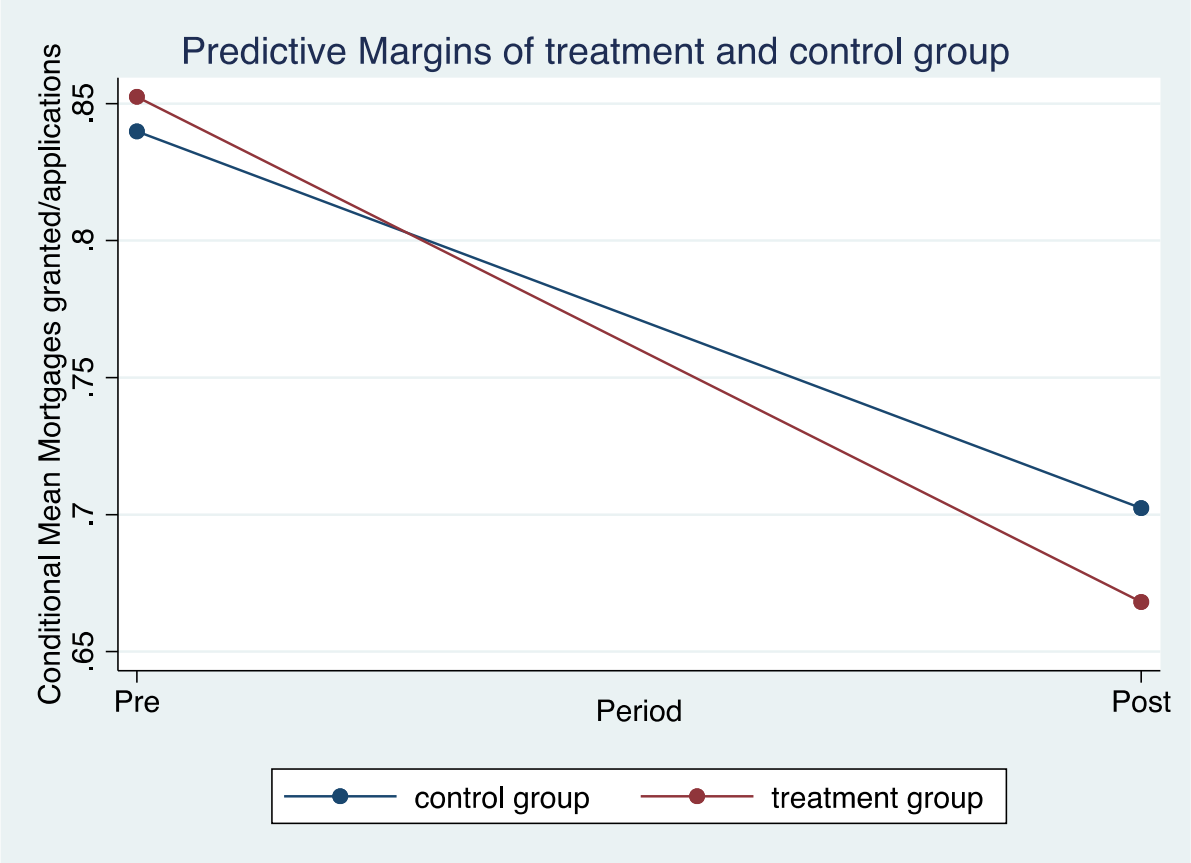


Figure 6

Predicted means of mortgages granting of treatment and control group during pre-treatment periods (8-week periods)

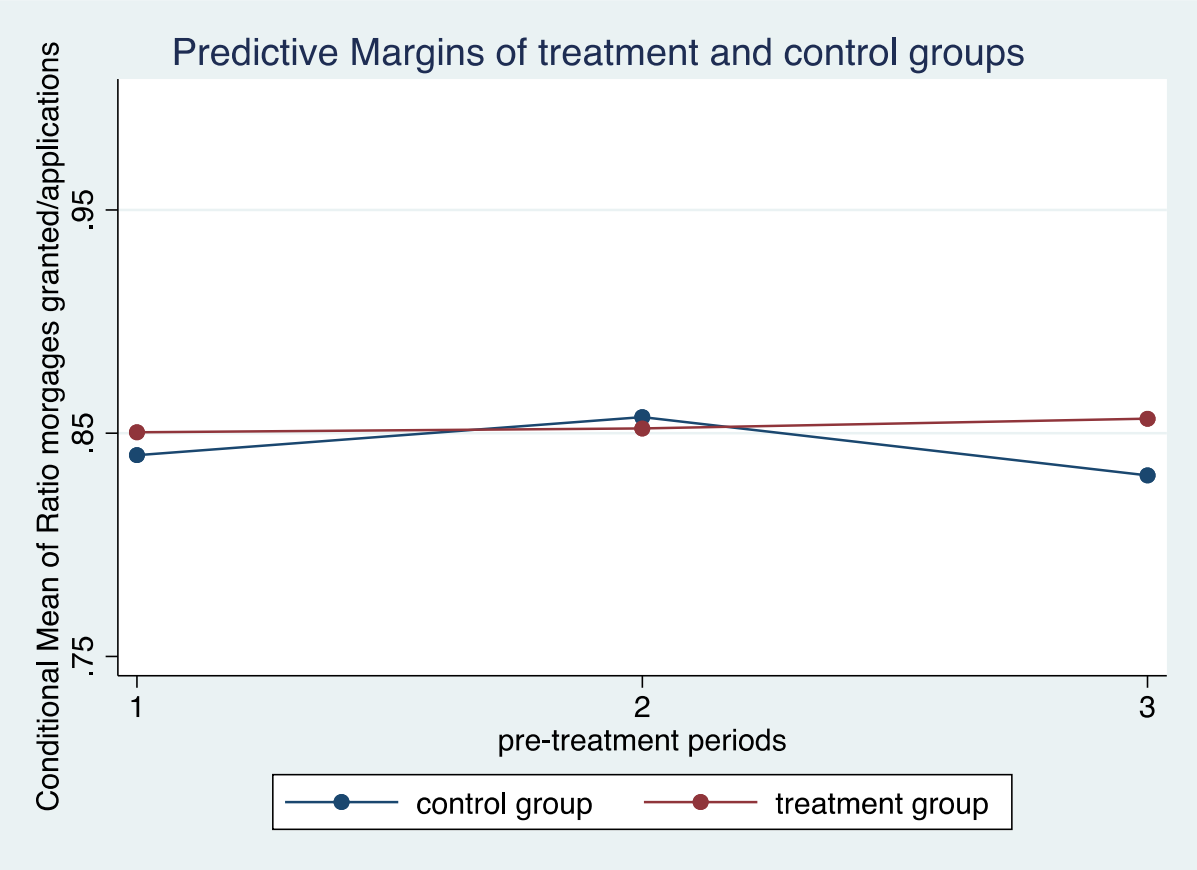
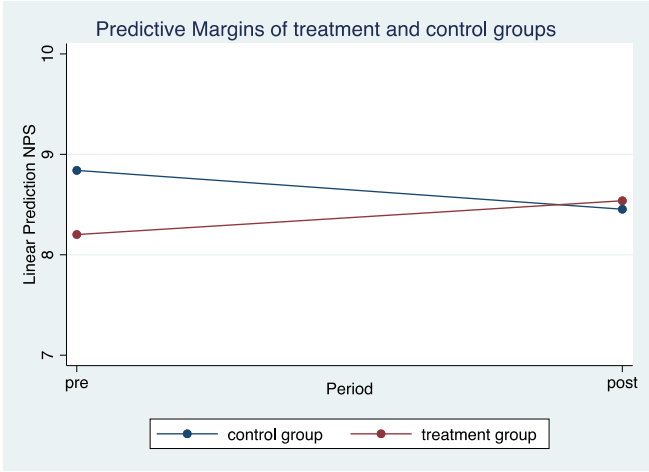


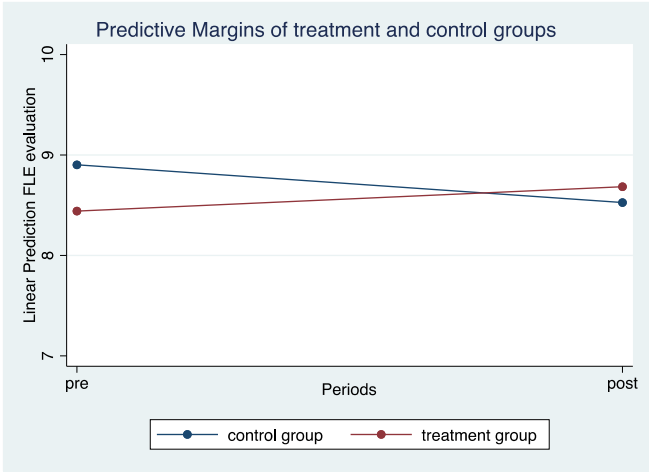
Figure 7

Predicted means of NPS, frontline employee evaluation and perceived complexity for treatment and control groups in pre-treatment and treatment periods

a)



b)



c)

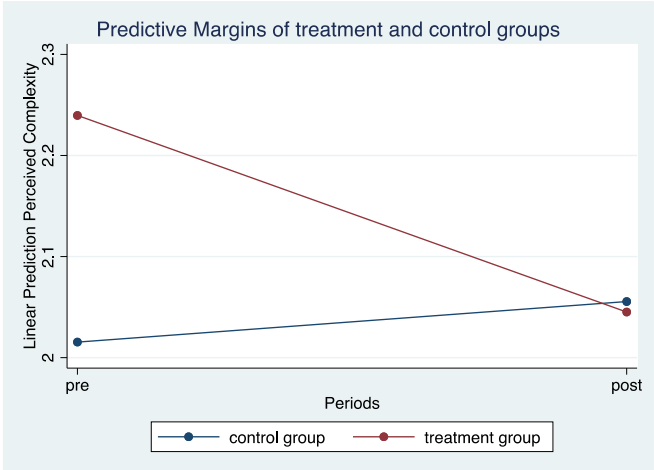
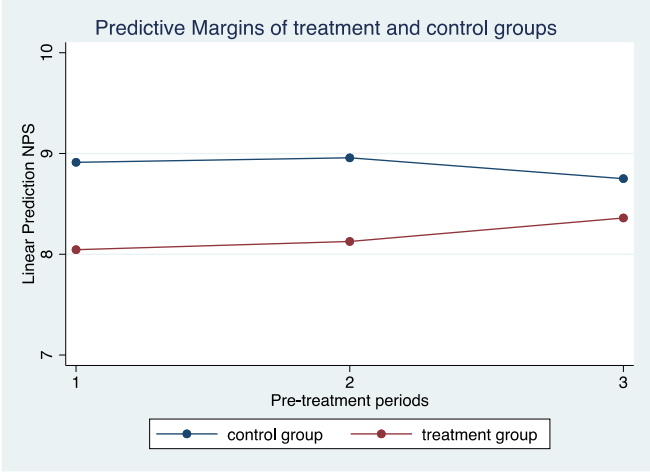


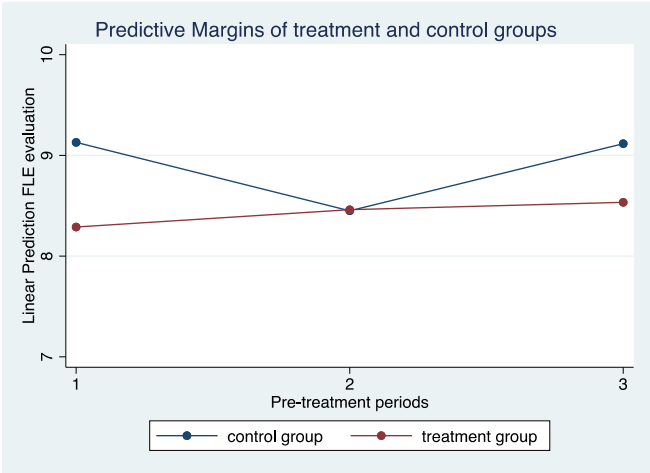
Figure 8

Predicted means of NPS, frontline employee evaluation and perceived complexity for treatment and control groups in pre-treatment periods (8-week periods)

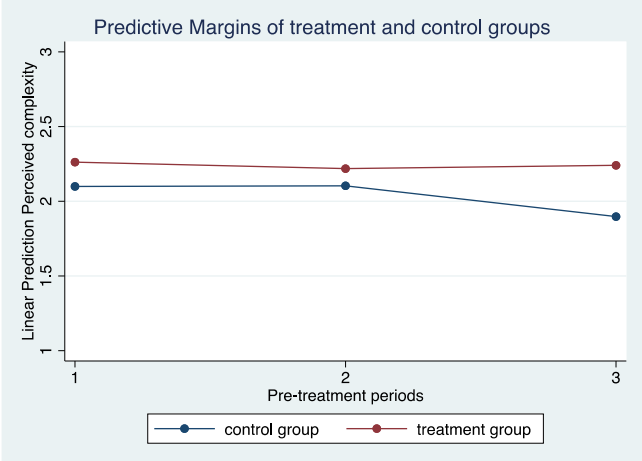
a)



b)



c)



Appendix A

Stages of the implementation of the AI solutions

Stage	Description	Start	End	Number of tickets
Pilot	First period in which the tickets went through the AI	01Jan19	14Jan19	180
Human- manual task execution (pre-treatment)	Tickets are processed only by human-back office	15Jan19	31Jul19	6,016
Tuning	AI proposes a solution to the human back office, who evaluates AI's suggestion and then processes the ticket	01Aug19	17Nov19	3,111
Training	During this phase, in order to train the AI, all the tickets went through the AI	18Nov19	04Dec19	845
Training (groups)	The same of the previous phase, but in this phase only the tickets of the treatment group were processed by AI	05Dec19	31Dec19	976
Study (treatment)	The period of our manipulation, AI processes the tickets of the treatment group and humans process the tickets of the control group	01Jan20	29Feb20	2,364
Post study	A month after the end of the study, AI processes all the tickets	01Mar20	31Mar20	1,306
			Tickets not closed	81
		Total		**14,879**