The Impact of Mobile App Failures on Purchases in Online and Offline Channels

Unnati Narang

Venkatesh Shankar

Sridhar Narayanan [1]

July 17, 2019

Most updated version available here

[1]Unnati Narang (email: unarang@mays.tamu.edu) is a Ph.D. student in Business Administration - Marketing and Venkatesh Shankar (email: vshankar@mays.tamu.edu) is Professor of Marketing and Coleman Chair in Marketing and Director of Research, Center for Retailing Studies at the Mays Business School, Texas AM University. Sridhar Narayanan (email: sridhar.narayanan@stanford.edu) is an Associate Professor of Marketing at the Graduate School of Business, Stanford University. We thank participants at the 2017 ISMS Marketing Science Conference, the UTDFORMS conference Ratchford Special Session, and University of California, Davis research seminar participants, the Editor, the AE, and the review team for valuable comments.

Abstract

The Impact of Mobile App Failures on Purchases in Online and Offline Channels

Over half of all shopping journeys start with the mobile channel. In particular, branded retailer mobile apps significantly influence shopping across channels. However, a majority of app users decrease app usage or even abandon an app, in part, due to app service failure(s), making app service failure prevention and recovery critical for retailers. Does an app failure influence purchases made within the online channel? Does it have any spillover effects across other channels? What factors moderate these effects? We leverage exogenous systemwide failure shocks in a large multichannel retailer's mobile app and related data to examine the impact of app failures on purchases in all channels using a difference-in-differences approach. We investigate shopper heterogeneity in the effects using a set of theoretically-driven moderators as well as data-driven machine learning methods. Our analysis reveals that although app failures have a significant overall negative effect on shoppers' frequency, quantity, and monetary value of purchases across channels, the effects are heterogeneous across channels and shoppers. Interestingly, the overall decreases in purchases across channels are driven by purchase reductions in brick-and-mortar stores and not in digital channels. Furthermore, we find that shoppers with a higher monetary value of past purchases, loyalty program members, and those with a greater number of stores in their neighborhood are less sensitive to app failures. The results suggest an app failure yields an annual revenue loss of $22 million for the retailer in our data, with 40% shoppers contributing to about 70% of the losses in revenues. We outline targeted failure prevention and service recovery strategies.

Keywords: service failure, mobile marketing, mobile app, retailing, omnichannel, difference-in-differences, natural experiment

1 Introduction

Mobile commerce has seen tremendous growth in the last few years, with the majority of shopping journeys starting with or involving interactions through mobile devices. Mobile applications (henceforth, apps) have emerged as an important channel for retailers because shoppers engage with apps and retailers can customize shopper experience through apps. However, failures experienced on retailers' mobile apps have the potential to negatively affect shoppers' satisfaction and their shopping outcomes within the mobile channel. In addition, app failures may potentially have spillover effects across the different channels due to both substitution of purchases across channels and impact on shoppers' overall satisfaction with the retailer.

In this study, we examine the impact of failures in a retailer's mobile app on shopping outcomes in its online and offline channels. We exploit a set of random systemwide failures in the app to estimate the causal effect of failures. Using unique customer-based data spanning online and offline retail channels, we study the spillover effects of such failures across channels. We investigate theoretically-driven moderators of these effects and find interesting differences in these effects across consumers based on shoppers' prior relationship strength with the retailer and digital channel use. We also analyze shopper heterogeneity in the effects using data-driven machine learning methods. Through our empirical analysis, we provide insights about the relationship between service failures in mobile apps and shoppers' subsequent interactions with focal firms in an omnichannel retail environment.

The mobile channel constitutes an important and growing part of a retailer's channel strategy. About 2.4 billion people use smartphones worldwide.[1] The mobile app environment provides a retailer the ability to offer an engaging shopping experience. Indeed, apps can uniquely influence shopping through both online and offline channels (Xu et al. 2016; Narang and Shankar 2019).

[1] Source: Mobile Marketing Magazine, 2017, `http://tinyurl.com/yb93lsf8`

On the flip side, app usage comes with the potential for service failures, some of which lie outside the control of the retailer. Preventing app failures is critical for managers to enhance shoppers' app experience because more than 60% shoppers abandon an app after experiencing failure(s) (Dimensional Research 2015). In 2016, app crashes were the leading cause of system failures, contributing 65% to all iOS failures (Blancco 2016). Given the potential damage that app failures could create for firms' relationships with customers, determining the impact of these failures is important in formulating service failure preventive and recovery strategies.

Despite the importance of app failures for firms and shoppers, not much is known about the impact of app failures on purchases. While app crashes in a shopper's mobile device have been shown to negatively influence subsequent engagement with the app (e.g., restart time, browsing duration, and activity level, Shi et al. 2017), the relationship between systemwide app failures and subsequent purchases has not been studied. Furthermore, a large proportion of shoppers use both online and offline retail channels. In such an environment, it is critical for retailers to understand the impact of failures on shopping behaviors not just within that channel, but also across channels (spillover effects). However, we do not yet know whether and how app failures influence purchases and spill over to other channels.

The effects of app failure may differ across shoppers. Shoppers may be more or less negatively impacted by failures depending on factors such as shoppers' relationship with the firm (Goodman et al. 1995; Hess et al. 2003; Chandrashekaran et al. 2007; Knox and van Oest 2014; Ma et al. 2015) and shoppers' prior use of the firm's digital channels (Cleeren et al. 2013; Liu and Shankar 2015; Shi et al. 2017). It is important for managers to better understand how the effects of failure vary across shoppers so that they can devise better preventive and recovery strategies and individually target shoppers with these actions. Yet, not much is known about heterogeneity in the effects of failure on purchase outcomes.

Our study fills these crucial gaps. We quantify and explain the impact of app failures on managerially important outcomes, such as the frequency, quantity and monetary value of purchases in the online and offline channels of a retailer. Specifically, we address four research questions:

- What is the effect of a service failure in a retailer's mobile app on the frequency, quantity, and monetary value of subsequent purchases by the shoppers?
- What is the effect of a service failure in an app on purchases in the online and offline channels?
- How do relationship strength and prior digital channel use moderate these effects?
- How heterogeneous is shoppers' sensitivity to failures?

Estimation of the effects of app failures on shopping outcomes is challenging. It is typically hard to estimate the impact of app failures on shopping behavior using observational data due to the potential endogeneity of app failures. This endogeneity may stem from an activity bias in that shoppers who use the app more frequently are more likely to experience failures than shoppers who use the app less frequently. Therefore, failure-experiencing shoppers may differ systematically from non-failure experiencers in their shopping behavior, leading to potentially spurious correlations between failures and shopping behavior. Panel data may not necessarily mitigate this issue because time-varying app usage/shopping activity is potentially correlated with time-varying app failures for the same reason. That is, shoppers are likely to engage more with the app when they are likely to purchase, potentially leading to more failures than in periods when shoppers engage less with the app. Additionally, the nature of activity on the app may be correlated with failures. For instance, a negative correlation between failures and purchases may result from a greater incidence of failures on the app's purchase page than on other pages. Thus, it is hard to make the case that correlations between app failures and shopping outcomes in observational data have a causal interpretation.

The gold standard among the methods available to uncover the causal impact of service failures is a randomized field experiment. However, such an experiment would be impractical in this context because a retailer will unlikely deliberately induce failures in an app even for a small subset of its shoppers for ethical reasons. Alternatively, we can use an instrumental variable approach to control for endogeneity. However, it is hard to come up with valid instrumental variables that exhibit sufficient variation to address the endogeneity concerns in this context.

We overcome the estimation challenges and mitigate the potential endogeneity of app failures using the novel features of a unique dataset from a large omnichannel retailer of video games, consumer electronics and wireless services. We observe a set of exogenous failure events that we exploit to estimate the causal effects of app failures. Specifically, we observe incidences of systemwide exogenous failure shocks in the retailer's mobile app due to server errors. App users who logged into the app on the day of the failure were randomly exposed to the failures depending on whether they logged in during the time window of the exogenous shock. We estimate the effects of app failures using a difference-in-differences procedure that compares the pre- and post- failure outcomes for the failure experiencers with those of failure non-experiencers. Through a series of robustness checks, we confirm that failure non-experiencers act as a valid control for failure experiencers, providing us the exogenous variation to find causal answers to our research questions. In the data, shoppers are tracked across both online and offline channels. This tracking feature allows us to investigate the effects not just within the online channel, but also to find spillovers to other channels and overall retailer-level impact.

We next investigate potential moderators of the effects of failures on shopping behavior by exploiting the panel nature of our dataset. Because we observe the shoppers in our sample for a relatively long period of time before the app failures, we test for the moderating effects of factors such as relationship with the firm and prior digital channel use on the effects of service failures. These factors have been explored for services in general (e.g., Ma et al. 2015; Hansen et al. 2018) but not in the digital or mobile context. In addition, we recover the heterogeneity of effects at the individual level using data-driven machine learning methods. This approach helps us fully characterize the heterogeneity in app failure effects. This is important for the retailer to devise targeted service failure recovery strategies.

Our results show that app failures have a significant overall negative effect on shoppers' frequency, quantity, and monetary value of purchases across channels, but the effects are heterogeneous across channels and shoppers. Interestingly, the overall decreases in purchases across channels are driven by purchase reductions in brick-and-mortar stores, rather than in digital channels. Furthermore, we find that shoppers with higher monetary value of past purchases, loyalty program members, and those with a greater number of stores in their neighborhood are less sensitive to app failures. Finally, most shoppers (69%) react negatively to failures, but about 40% of these shoppers contribute to about 70% of the losses in revenues that amount to $22 million for the retailer in the data.

In the remainder of the paper, we first discuss the literature related to service failures, cross-channel spillovers and consumer interaction with mobile apps, and develop the conceptual framework for our empirical analysis. Next, we discuss the data in detail, summarizing them and highlighting their unique features. Subsequently, we describe our empirical strategy, lay out and test the key identification strategy, and conduct our empirical analysis of the effects of app failures. We then conduct robustness checks to rule out alternative explanations. We conclude by discussing the implications of our results for managers.

2 Conceptual Background and Related Literature

2.1 Services Marketing and Service Failures

The nature of services has evolved considerably since academics first started to study services marketing. For long, the production and consumption of services remained inseparable primarily because services were performed by humans. However, in addition to people-enabled services, technology-enabled services have risen in importance, leading to two important shifts (Dotzel et al. 2013). First, services that can be delivered without human or interpersonal interaction have grown tremendously. In retailing, online and mobile commerce no longer require shoppers to interact with human associates to make purchases. Second, closely related to this idea is the fact that services are increasingly powered by technologies that allow anytime-anywhere access and convenience.

With growing reliance on technologies for service delivery, service failures are becoming more common. A service failure is defined as service performance that falls below customer expectations (Hoffman and Bateson 1997). Service failures are widespread and are expensive to mend. Service failures resulting from deviations between expected and actual performance damage customer satisfaction (Smith and Bolton 1998). Post-failure satisfaction tends to be lower even after a successful recovery and is further negatively impacted by the severity of the initial failure (Andreassen 1999; McCollough et al. 2000). In interpersonal service encounters, both failure and recovery are largely influenced by the human element involved (Meuter et al. 2000) and rely on employee behaviors (Bitner et al. 1990). In technology-based encounters, such as e-tailing or in self-service technologies such as automated teller machines (ATMs), the opportunity for human interaction is almost eliminated after experiencing failure (Forbes et al. 2005; Forbes 2008). However, there may be significant heterogeneity in how consumers react to failures (Halbheer et al. 2018).

In the mobile context, specifically for mobile apps, it is difficult to predict the direction and extent of impact of a service failure on shopping outcomes. First, mobile apps are accessible at any time and in any location through an individual's mobile device. On the one hand, because a shopper can tap, interact, engage, or transact multiple times at little additional cost on a mobile app, the shopper may treat any one service failure as acceptable without significantly altering shopping outcomes. Such an experience differs from that with a self-service technological device such as an ATM, for which may need the shopper to travel to a specific location. On the other hand, because a typical shopper uses multiple apps and can easily compare his/her experiences across them, a service failure in any one app may aggravate the shopper's frustration and annoyance with the app, leading to strong negative effects for the concerned app provider and therefore on subsequent outcomes such as purchases.

Second, a mobile app is one of the many touchpoints available to shoppers in today's omnichannel shopping environment. Thus, a shopper who experiences a failure in the app could move to the web-based channel or even the offline or store channel. In such cases, the impact of a failure on the app could be zero or even positive (if the switch to the other channel leads to greater engagement of the shopper with the retailer). By contrast, if the channels act as complements or if the failure impacts the overall evaluation of the retailer as a whole, a failure in one channel could impede the shopper's engagement in other channels. Thus, it is difficult to predict the effects of app failure, in particular, how they might spill over to other channels. We next discuss these channel dynamics.

2.2 Channel Choice and Channel Migration

A shopper's experience in one channel can influence his/her behavior in other channels. Prior research on cross-channel effects is mixed, showing both substitution and complementarity effects, leading to positive and negative synergies between channels (e.g., Avery et al. 2012; Pauwels and Neslin 2015). The relative benefits of channels determine whether shoppers continue using existing channels or switch to a new channel (Ansari et al. 2008; Chintagunta et al. 2012). When a bricks-and-clicks retailer[2] opens an offline store or an online-first retailer opens an offline showroom, the offline presence drives sales in online stores (Wang and Goldfarb 2017; Bell et al. 2018). This is particularly true for shoppers in areas with low extant brand presence prior to store opening and shoppers with an acute need for the product. However, the local shoppers may switch from online purchasing once an offline store opens, even becoming less sensitive to online discounts (Forman et al. 2009). In the long run, the store channel shares a complementary relationship with the Internet and catalog channels (Avery et al. 2012).

While the relative benefits of one channel may lead shoppers to buy more in other channels, the costs associated with one channel may also have implications for purchases beyond that channel. In a truly integrated omnichannel retailing environment, the distinctions between physical and online channels are blurry, with online representing a showroom without

[2]Bricks-and-clicks retailer refers to a retailer with both offline ("bricks") and online ("clicks") presence.

walls (Brynjolfsson et al. 2013). Mobile technologies are at the forefront of these shifts. More than 80% shoppers use a mobile device while shopping even inside a store (Google M/A/R/C Study 2013). As a result, if there are substantial costs associated with using a mobile channel (e.g., app failures), such costs may spill over to other channels. However, if shoppers treat the channels as substitutes, failures in one channel may drive the shoppers to compensate with higher purchases in another channel. If shoppers' satisfaction for the retailer as a whole is impacted, then there may be overall negative consequences of app failure across channels, without a compensatory effect. Overall, the direction of the effect of app failures on outcomes in other channels such as in brick-and-mortar stores and online channels depends on which of these competing and potentially co-existing mechanisms is dominant.

2.3 Mobile Apps

The nascent but evolving research in mobile apps shows positive effects of mobile app channel introduction and use on engagement and purchases in other channels (Kim et al. 2015; Xu et al. 2016; Narang and Shankar 2019) and for coupon redemptions (Andrews et al. 2015; Fong et al. 2015; Ghose et al. 2018) under different contingencies.

To our knowledge, only one study has examined crashes in a mobile app and shopper's app use. Shi et al. (2017) assess the impact of software crashes in a shopping app on app restart time, browsing duration, and activity level in the subsequent session. They find that while crashes have a negative impact on future engagement with the app, this effect is reduced for those with greater prior usage experience and less persistent crashes. However, while they look at subsequent engagement of the shoppers with the mobile app, they do not examine purchases. Thus, our research adds to Shi et al. (2017) in several ways: First, we exploit the random variation in systemwide failure incidences to determine the causal effects of failure. Second, we quantify the value of app failures' effects on subsequent purchases. Our outcome measures include the frequency, quantity, and value of purchases, while the key outcome in that study is app engagement. Third, we examine the cross-channel effects of mobile app failures, including in brick-and-mortar stores, while Shi et al. (2017) study subsequent engagement with the app provider only within the app context. Fourth, we explore moderating variables such as relationship with the retailer and prior digital use in shoppers' sensitivity to failures and explore heterogeneous effects using a machine learning approach. Finally, we focus on a broader conceptualization of app failures to include any negative app experience caused by technical disconnections rather than limiting them to crashes. This issue is relevant from a managerial standpoint because 61% users expect apps to load in four seconds or less, and abandon the app if it runs slowly (Dimensional Research 2015) and the probability of a mobile site visitor discontinuing usage of the app increases by 123% as page load time goes up from one to 10 seconds (Google 2018).

Unlike prior related studies, our study (1) focuses on the effect of mobile app failure on purchases, (2) quantifies the effects on multiple shopping outcomes such frequency, quantity, and monetary value of purchases, (3) addresses the outcomes in each channel and across all channels (substitution and complementary effects), and (4) uncovers the moderators of the effects of app failure on shopping outcomes and heterogeneity in effects across shoppers.

3 Research Setting and Data

3.1 Research Setting

We obtained the dataset for our empirical analysis from a large U.S.-based retailer of video games, consumer electronics and wireless services with 32 million customers. We first describe the retailer and provide relevant details about the mobile app. The retailer sells a variety of products, including software such as video games and hardware such as video game consoles and controllers, downloadable content, and gift cards. The gaming industry is large ($99.6 billion in annual revenues), and our data provider is an important retailer in this industry, offering us a rich setting to examine shopping behaviors.

The retailer is similar to Walmart, PetSmart, or any other brick-and-mortar chain with a relatively large offline presence. The retailer's primary channel is its store network comprising 4,175 brick-and-mortar stores across the U.S. Additionally, it has an ecommerce website. The retailer introduced its app in July 2014. The app allows shoppers to browse the retailer's product catalog, get deals and offers, order online through the mobile browsers, or locate nearby stores to buy offline. The app permits shoppers to learn about the retailer's stores, including locations, open hours, phone numbers, and driving directions. Initially, the app did not offer in-app purchase capability, that is, the ability to pay for and complete the purchases within the app. However, the app did allow shoppers to add products to an in-app cart linked to a checkout button. When shoppers clicked "checkout" to pay, the app redirected them to the retailer's mobile website in the mobile browser to complete the purchase as an online purchase. The app is typical of mobile apps of large retailers (e.g., PetSmart, Costco) in that these apps do not or did not at some point in the app's history allow purchase transactions but improve the shopping experience by providing information and shopping convenience in other channels.

In 2016, the focal app in our research added the functionality to allow in-app purchases. Our data cover the period both before and after in-app purchases were allowed. The growth in adoption of the app is also similar to that of many large retailers. App adoption rate starts small and grows over time. Between 2014 and 2018, the app logins increased by 67%, reflecting the increased popularity of the app. Figure 1 provides some screenshots of the app.

< Figure 1 about here >

The online and offline mix of the retailer in our data is typical of most large retailers. About 76% of the total sales for the top 100 largest retailers in the U.S. originated from similar large retailers with a large store network of 1,000 or more stores (National Retail Federation 2018). Most large retailers have a predominantly high brick-and-mortar presence. For these retailers, most of the transactions and revenues come from the offline channels and online sales exhibit rapid growth. For example, Walmart's online revenues constitute 3.8% of all revenues, Costco's online sales are 3.6% of all sales, 1.3% of all PetSmart's sales come from the online channel, Home Depot generates 6.8% of all revenues from ecommerce, and 5.4% of Target's sales are through the online channel.[3] For the retailer in our data, online sales comprised 10.2% of overall revenues during the data period. Thus, the proportion of online sales for this retailer is at least as much as if not greater than that for similar large retailers. Online sales for the retailers mentioned are growing in double digits, similar to our setting that also exhibited a 13% annual average growth in the last five years (Barrons 2018).

The online sales volume and value of the retailer in our data are substantial. Online sales contribute nearly $1.1 billion in annual revenues. Thus, the impact of failures in app on both online and offline channel is important to consider since the online channel contributes a significant absolute dollar revenue. Although the percentage of online sales is considerably lower than offline sales, we can see substantial cross-channel effects of a mobile app failure in the data.

3.2 Data and Sample

There were four systemwide app failures on October 10, November 3, and December 10 in 2014 and April 11 in 2018.[4] The firm provided us with mobile app use data and transactional data across all channels for the sample users. Our sample comprises all users who logged into the app on the failure days. The online channel represents purchases made via the retailer's website, including those using the mobile browser. Nested within the mobile app use data are data on events that shoppers experience with their timestamps. The app failure events are recorded as "server error" in the mobile data. These represent exogenous app breakdowns, slowdowns and unresponsiveness of the app. These data allow us to identify shoppers who logged in during and experienced the systemwide server failure.

Table 1 provides the descriptive statistics for the variables of interest. Over a period of 30 days pre- and post- failures, shoppers make an average of approximately one purchase, buy

[3] Source: eMarketer Retail, `https://retail-index.emarketer.com/`

[4] We verified that these failures were systemwide and exogenous through our conversations with company executives. The failure on April 11, 2018 is farther away from the holiday season and helps rule out concerns about any potential idiosyncratic effects around Black Friday or Holiday season promotions.

two items, and spend \$69. Furthermore, 48% of the shoppers experience the failure. In the 12 months preceding a failure, shoppers on average make 17 purchases worth \$958. Shoppers made an average of 0.47 purchases online in the 12 months before failure.

< Table 1 about here >

4 Empirical Strategy

4.1 Overall Empirical Strategy

As outlined earlier, any study of the causal effects of app failures on shopper behavior is challenging for several reasons. First, it is hard to convince any retailer to conduct a field experiment that induces failure for even a subset of shoppers for ethical reasons and business concerns inherent in making these shoppers experience a failure that they would have otherwise not experienced. Second, observational data are challenging to use in this context because of the potential for the correlation between usage and failures to be driven by activity bias as described earlier, as well as the potential for reverse causality.

To mitigate the challenges, we adopt a quasi-experimental approach, exploiting exogenous systemwide failures. The main idea behind our empirical approach is that conditional on the usage of the app on the day of the failure, the experience of a failure by a specific shopper is random. We test this assumption in the data using a set of pre-failure variables. We find that there is no systematic difference between shoppers who experienced failures and those who did not. We then conduct a difference-in-differences analysis, comparing the post-failure with the pre-failure behaviors of shoppers who logged in on the day of the systemwide failure and experienced it (akin to a treatment group) relative to those who logged in on that day but did not experience the failure (akin to a control group). The treatment vs. control analysis helps estimate the causal effects of app failures and the pre-post differences enables the estimates to be efficient.

To analyze the treatment effects within and across channels, we repeat this analysis with the same outcome variables separately for the offline and online channel. To analyze heterogeneity in treatment effects, we first perform a moderator analysis using a priori factors such as prior relationship strength and digital channel use, and then conduct a causal forest analysis to fully explore all sources of heterogeneity across shoppers. Finally, we carry out multiple robustness checks.

4.2 Exogeneity of Failure Shocks

To verify the exogeneity of the failure shocks, we examine two types of evidence. First, we present plots of the behavioral trends in shopping for both failure-experiencers and non-experiencers for each of the four failure shocks in the 15 days before the app failure. Figure 2 depicts the monetary value of daily purchases by those who experienced the failure and those who did not. The purchase trends in the pre-period are similar for the two groups ($p > 0.10$).

< Figure 2 about here >

Second, we compare the failure experiencers with non-experiencers across observed demographic variables, such as age, gender, membership in loyalty program, number of stores in their core-based statistical areas (CBSA), and mobile operating system (see Figure 3). We do not find any significant differences in these variables across the two groups ($p > 0.10$).

< Figure 3 about here >

4.3 Econometric Model and Identification

To estimate the effects of app failure on shopping outcomes, we rely on a quasi-experimental research design with a difference-in-differences approach (e.g., Angrist and Pischke 2009) that leverages systemwide failure shocks and compares app users who experience these shocks with those who do not, given that they accessed the app on the day of the failure. Note that we have already shown that the two groups of users do not differ systematically either on demographics or on behavioral variables in the period leading up to the app failure, providing face-validity to our quasi-experimental design.

Our two-period linear difference-in-differences regression takes the following form:

$$Y_{its} = \alpha_0 + \alpha_1 F_{is} + \alpha_2 F_{ts} + \alpha_3 F_{is} P_{ts} + \tau_s + \nu_{its} \tag{1}$$

where i is the individual, t is the time period, and s is the failure event, Y is the outcome variable, F is a dummy variable denoting treatment (1 if shopper i is experienced the systemwide app failure s and 0 otherwise), P is a dummy variable denoting the period (1 for the period after the systemwide app failure s and 0 otherwise), α is a coefficient vector, τ is the failure-specific fixed effect, and ν is an error term. We cluster standard errors at the shopper level (Bertrand et al. 2004). The coefficient of $F_{is}P_{ts}$, i.e., α_3 is the treatment effect of the app failure.[5]

The assumptions underlying the identification of this treatment effect are: (1) the failure is random conditional on a shopper logging into the app during the time window of the failure shock, and (2) the change in outcomes for the non-failure experiencing app users is a valid counterfactual for the change in outcomes that would have been observed for failure-experiencing app users in the absence of the failure.

[5]Because we analyze the short-term effect of a service failure (15 and 30 days), we do not have multiple observations per shopper post failure for us to include shopper fixed effects in our analysis.

5 Empirical Analysis Results

5.1 Relationship between App Failures and Purchases

We first examine the overall differences in post-failure behaviors between shoppers who experienced failures and those who did not using model-free evidence 15 days pre and post failure. We choose a 15-day window period for two main reasons. First, this window is close to the median interpurchase time of 19 days for the shoppers in our dataset. Second, it is not too long to include the occurrence of other events, including other systemwide failure shocks.[6]

Table 2 reports the model-free results for both failure experiencers (232,299 treated) and non-experiencers (252,886 control) given that they accessed the app during the day of the failure shocks. We find that for post failure, shoppers who experienced the systemwide failure had 0.06 ($p < 0.001$) lower purchase frequency, 0.13 ($p < 0.001$) lower purchase quantity, and \$4.57 ($p < 0.001$) lower monetary value than shoppers who did not experience failures. A simple comparison of shopping outcomes across the two groups shows that the average monetary value of purchases decreased by 1.4% (\$68.01 to \$67.09) for failure-experiencers, while it increased by 4.4% (\$68.66 to \$71.66) for non-failure experiencers post failure shock relative to the pre period ($p < 0.001$). Given our identification strategy, this significant difference in the monetary value of purchases comes from the exogenous failure shock. These effects appear to be stronger for in-store purchases than for online purchases.

< Table 2 about here >

5.2 Diff-in-Diff Model Results

The results from the difference-in-differences regressions in Table 3 show a negative and significant effect of app failure on the frequency ($\alpha_3 = -0.05$, $p < 0.001$), quantity ($\alpha_3 = -0.12$, $p < 0.001$), and monetary value of purchases ($\alpha_3 = -3.92$, $p < 0.001$) across channels. Relative

[6]We also estimated the models for a 30-day period and found similar results. We report these results in Web Appendix Tables A3 and B3.

to the pre-period levels for these variables for the control group, the treated group experiences a decline in frequency by 5.1% ($p < 0.001$), quantity by 5.4% ($p < 0.001$), and monetary value by 5.7% ($p < 0.001$) of purchases.

< Table 3 about here >

Next, we examine the channel spillover effects of app failures in greater depth. We split the total value of purchases into store-based purchases and website purchases. Table 4 reports the results for these alternative channel-based dependent variables. There is a negative and significant effect of app failure on the frequency ($\alpha_3 = -0.05$, $p < 0.001$), quantity ($\alpha_3 = -0.12$, $p < 0.001$), and monetary value of purchases ($\alpha_3 = -3.75$, $p < 0.001$) in the offline channel. Interestingly, we do not find a significant ($p > 0.10$) effect of app failure on any of the purchase outcomes in the online channel.

Because there is no corresponding increase in the online channel and because the overall purchases drop, we conclude that the decreases in overall purchases across channels are largely due to declines in in-store purchases. We next provide descriptive evidence to help explain this result through relative channel costs. Some of the failure-experiencers may not have been close to a transaction at the time of experiencing the failure or may not be using the app with the intent of making purchases. For such shoppers, the app failure may not have had an immediate effect on their purchases either online or offline. However, other failure-experiencers who were close to a purchase or had purchase intent, would have had to determine whether to complete the transaction, and if so, whether to do it online or offline. For shoppers who typically buy online, the cost of going to the retailer's website and completing a purchase transaction interrupted by the failure is smaller than the cost of going to the store to complete the purchase. Therefore, these shoppers will likely complete the transaction online by going to the retailer's website and not exhibit any significant decrease in shopping outcomes in the online channel post failure. In contrast, shoppers who typically buy in the brick-and-mortar stores and who experience the systemwide app failure when they near a purchase decision, will perceive fewer incentives and greater costs to buy more from the brick-and-mortar stores.

We provide additional descriptive evidence to examine these potential explanations. Indeed, a descriptive analysis shows that stronger negative effects in store purchases come from shoppers with fewer stores in their CBSA. Because they have less stores in their neighborhood and their cost of traveling to the store is higher, these shoppers experience greater hassles (Dukes and Zhu 2019). Similarly, we measure the time between failure and subsequent purchase in the online channel. Failure experiencers go to the website and make an online purchase much faster than non-experiencers ($p < 0.001$). This result suggests that after failures, shoppers may quickly complete purchases or seek additional information in the online channel when the app fails to assist them. Because the overall effects in the online channel are insignificant, perhaps the increases in purchases from these shoppers are not adequate to justify an overall increase. Together, these analyses suggest that cross-channel heterogeneities may be originating from the ease or hassle of using specific channels and the shoppers' willingness to use the channels after a failure. In addition to heterogeneity across channels, we also examine the heterogeneity across shoppers in their sensitivity to app failures.

< Table 4 about here >

5.3 Moderators: Relationship Strength and Prior Digital Use

The literature on relationship marketing and service recovery suggest two factors that could moderate the impact of app failures on shopping outcomes: relationship strength and prior digital channel use.

5.3.1 Relationship Strength

The service marketing literature offers mixed evidence regarding the moderating role of the strength of customer relationship with the firm in customer sensitivity to service failure. Some studies suggest that stronger relationship with firms may aggravate the effect of failures on product evaluation, satisfaction, and on purchases (Goodman et al. 1995; Chandrashekaran et al. 2007; Gijsenberg et al. 2015). Other studies show that stronger relationship attenuates the negative effect of service failures (Hess et al. 2003; Knox and van Oest 2014).

Consistent with the direct marketing literature (Schmittlein et al. 1987; Bolton 1998), we operationalize customer relationship with the firm based on past behavior. We use the RFM (recency, frequency, and monetary value) dimensions to characterize the strength of the relationship. We considered including each of the RFM variables as potential moderators of the treatment effect. However, because of high correlation between the interactions of frequency with (failure experiencers x post shock) and value of purchases with (failure experience x post shock) ($r = 0.90$, $p < 0.001$) and because value of purchases is more important from the retailer's perspective, we drop frequency of past purchases from further consideration in the moderator analysis. Therefore, we estimate a model with interactions of each of recency and value of past purchases with the interaction between failure experiencers and the post shock indicator ($F_{is}P_{ts}$) in the proposed model to capture the moderating effects of these relationship variables.

The moderating effect of each indicator of relationship strength, namely, value of past purchases and recency may be positive or negative. High value shoppers may place more importance in the quality of products and the overall experience with the retailer. Consequently, a service failure in the app may not elicit a significantly negative response for these shoppers. In contrast, low value shoppers may be more transactional in nature. For such shoppers, even a single failure in the mobile app may trigger a negative reaction for their subsequent purchases. Thus, the value of past purchases may have a positive moderating effect on the effect of app failure on shopping outcomes.

Recency of purchases may have either a positive or negative moderating effect. After experiencing a failure in the app, shoppers who bought recently from the retailer can afford to wait a little longer to make their next purchase. Thus, these shoppers could be more tolerant of the service failure, leading to a positive or insignificant moderating effect of recency on shopping outcomes. By the same token, it could also be argued that recent shoppers may find an app failure annoying, especially when they wish to continue patronizing the retailer by wanting to buy more. Their irritated reaction could lead to a negative moderating effect of recency on shopping outcomes. Therefore, the effect of recency of purchases on the effect of app failure on shopping outcomes is an empirical issue.

5.3.2 Prior Digital Channel/Online Use/Experience

A shopper's prior digital channel/online use or experience with the retailer may moderate the effects of service failure on shopping outcomes. On the one hand, more digitally experienced app users may be less susceptible to the negative impact of an app crash on subsequent engagement with the app than less digitally experienced app users (Shi et al. 2017) because they are conditioned to expect some level of technology failures. This reasoning is based on the product harm crises literature (Cleeren et al. 2013; Liu and Shankar 2015) and the broader expectation-confirmation theory (Oliver 1980). The more experience a customer has with a service, the less impact a single piece of new information (from failure) will have on service evaluation and usage (Tax et al. 1998; Cleeren et al. 2008). On the other hand, prior digital exposure and experience with the firm may heighten shopper expectations and make them less tolerant of service failures. What is the empirical effect of prior digital experience on how service failures affect shopper purchases? To answer this question, in the empirical analysis, we operationalize prior digital channel use of a shopper as the cumulative number of purchases that the shopper made from the retailer's website prior to experiencing a failure.

The results of the model with relationship strength and past digital channel use as moderators appear in Table 5. Consistent with our expectation, the monetary value of past purchases has positive interaction coefficients with the difference-in-differences variable across all outcome variables when significant ($p < 0.001$). Thus, more valuable customers attenuate the negative effect of app failures on shopper purchases, suggesting that app failures affect high value shoppers less. However, recency has negative coefficients ($p < 0.001$), suggesting that the annoyance phenomenon is likely at play. A failure shock primarily affects the purchase frequency ($p < 0.001$) of shoppers with greater digital channel or online purchase exposure or experience with the retailer. In our data, most shoppers make at most one past purchase online. Such shoppers typically do not expect an app to fail. If they experience a failure in the app, they reduce the number of times they shop. These results are robust and stronger for alternative measures of digital experience, such as prior app usage (Web Appendix Tables A6 and B6).

< Table 5 about here >

6 Heterogeneity in Shoppers' Sensitivity to App Failures

While the difference-in-differences regression in Equation (1) allows us to recover an average treatment effect on the treated and the effects of theoretically-driven moderator variables, we are also interested to further explore heterogeneity in treatment effects relating to managerially useful additional observed variables (e.g., age, gender, membership in loyalty program) not fully investigated by prior research. Insights about additional variables that may influence heterogeneity in treatment effects have both theoretical and practical value. Estimation of individual level treatment effects and knowledge of their drivers is particularly useful for managers for service failure prevention and recovery purposes. Unfortunately, including these variables as additional moderators will impose a huge burden on the DID analysis as the number of potential main and interaction effects will become unmanageably large.

Recent advances in causal inference using machine learning allow us to recover individual-level conditional average treatment effects (CATE) (Athey and Imbens 2016, Athey et al. 2017, Wager and Athey 2018). These methods use ideas from regression trees and random forests (Breiman 2001) to identify subpopulations of the data that differ in the magnitude of the treatment effect. These methods have been applied in marketing in the context of customer churn and information disclosure (Ascarza 2018; Guo et al. 2018). In our context, we estimate a causal forest model, an ensemble of causal trees that averages the predictions of treatment effects produced by each tree for thousands of trees. We next provide an overview of causal trees and describe the algorithm for estimating a single causal tree followed by bagging a large number of causal trees into a forest. We follow the causal tree estimation with a regression of the individual treatment effect on covariates, including the moderators and observed demographic variables.

6.1 Causal Trees: Overview

A causal tree is similar to a regression tree. The typical objective of a regression tree is to build accurate predictions of the outcome variable by recursively splitting the data into subgroups that differ the most on the outcome variable based on covariates. A regression tree has decision/internal/split nodes characterized by binary conditions on covariates and leaf or terminal nodes at the bottom of the tree. The regression tree algorithm continuously partitions the data, evaluating and re-evaluating at each node to determine (a) whether further splits would improve prediction, and (b) the covariate and the value of the covariate on which to split. The goodness-of-fit criterion used to evaluate the splitting decision at each node is the mean squared error (MSE) computed as the deviation of the observed outcome from the predicted outcome. The tree algorithm continues making further splits as long as the MSE decreases by more than a specified threshold.

The causal tree model adapts the regression tree algorithm in several ways to make it amenable for causal inference. First, it explicitly moves the goodness-of-fit-criterion to treatment effects rather than the MSE of the outcome measure. Second, it employs "honest" estimates, that is, the data on which the tree is built (splitting data) are separate from the data on which it is tested for prediction of heterogeneity (estimating data). Thus, the tree is honest if for a unit i in the training sample, it only uses the response Y_i to estimate the within-leaf treatment effect, or to decide where to place the splits, but not both (Athey and Imbens 2016; Athey et al. 2017). To avoid overfitting, we use cross-validation approaches in the tree-building stage. Importantly, the goodness-of-fit criterion for causal trees is the difference between the estimated and the actual treatment effect at each node. While this criterion ensures that all the degrees of freedom are used well, it is challenging because we never observe the true treatment effect.

6.2 Causal Tree: Goodness-of-fit Criterion

Following Wager and Athey (2018), if we have n independent and identically distributed training examples labeled $i = 1, ..., n$, each of which consists of a feature vector $X_i \in [0,1]^d$, a response Y_i, and a treatment indicator $W_i \in [0,1]$, the CATE at x is:

$$\tau(x) = \mathbb{E}[Y_i^1 - Y_i^0 | X_i = x] \tag{2}$$

We assume unconfoundedness, i.e., conditional on X_i, the treatment W_i is independent of outcome Y_i. Because the true treatment effect is not observed, we cannot directly compute the goodness-of-fit criterion for creating splits in a tree. This goodness-of-fit criterion is as follows.

$$Q_{infeasible} = \mathbb{E}[(\tau_i(X_i) - \hat{\tau}_i(X_i))^2] \tag{3}$$

Because $\tau_i(X_i)$ is not observed, we follow Athey and Imbens's (2016) approach to create a transformed outcome Y_i^* that represents the true treatment effect. Assume that the treatment indicator W_i is a random variable. Suppose there is a 50-50 probability for a unit i to be in the treated or the control group, an unbiased true treatment effect can be obtained for that unit by just using its outcomes Y in the following way. Let:

$$Y_i^* = \begin{cases} 2Y_i & \text{if } W_i = 0 \\ -2Y_i & \text{if } W_i = 1 \end{cases} \tag{4}$$

It follows that:

$$\mathbb{E}[Y_i^*] = 2.(\frac{1}{2}\mathbb{E}[Y_i(1)] - \frac{1}{2}\mathbb{E}[Y_i(0)]) = E[\tau_i] \tag{5}$$

Therefore, we can compute the goodness-of-fit criterion for determining node splits in a causal tree using the expectation of the transformed outcome (see Athey and Imbens 2016 for details). Once we generate causal trees, we can compute the treatment effect within each leaf because it has a finite number of observations and standard asymptotics apply within a leaf. The differences in outcomes for the treated and control units within each leaf produces the treatment effect in that leaf.

6.3 Causal Forest Ensemble

In the final step, we create an ensemble of trees using ideas from model averaging and bagging. Specifically, we take predictions from thousands of trees and average over them (Guo et al. 2018). This step retains the unbiased, honest nature of tree-based estimates but reduces the variance. The forest averages over the estimates from B trees in the following manner:

$$\hat{\tau}(x) = B^{-1} \sum_{b=1}^{B} \hat{\tau_b}(x) \tag{6}$$

6.4 Analysis and Results

Because monetary value of purchases is the key outcome variable of interest to the retailer, we estimate individual level treatment effect on value of purchases for each failure experiencer separately using the observed covariate data. These covariates include age, gender, loyalty program, mobile operating systems (OS) type, and the number of stores in the shopper's CBSA in addition to the three theoretically-driven moderators, namely, value of past purchases, recency of past purchases and online buying/digital experience. These individual attributes are important for identifying individual-level effects and for developing targeting approaches (e.g., Neumann et al. 2019). We use a random sample of two-thirds of our data as training data and the remaining one-third as test data for predicting CATE. We use half of the training data to maintain honest estimates and for cross-validation to avoid overfitting.

The estimates from causal forests using 1,000 trees appear in Table 6. The average CATE is

-1.91 for the test data, close to the overall average treatment effect on the treated (ATET) of -3.92 obtained from the regression model. Furthermore, 69% of the shoppers have a negative value of CATE; their average is -3.42. The distribution of CATE across shoppers appears in Figure 4. The shopper quintiles based on the levels of CATE reflects this distribution in Figure 5, which reveals that Segments 1 and 2 of the most sensitive shoppers also exhibit higher variance than the rest.

< Table 6 and Figures 4 and 5 about here >

Next, we regress the CATE estimate on the covariate space to identify the covariates that best explain treatment heterogeneity. The results from such an ordinary least squares (OLS) regression of CATE on covariates appear in Table 7. All the covariates except mobile operating system are significant ($p < 0.001$). We can explain heterogeneity through past shopping behavioral as well as demographic variables. Shoppers with higher past purchase value, loyalty program membership, with higher frequency of past online purchases, and access to a greater number of stores are less sensitive to app failures than others. However, older and female customers and those with more recent purchases are more sensitive to failures than others. Figure 6 provides illustrations of the relationships between select covariates and the treatment effect. The relationships of CATE with each of value of past purchases, loyalty program membership, and the number of stores in the CBSA are positive, while that with age is negative.

< Table 7 and Figure 6 about here >

The causal forest-derived CATE regression differs from the moderator DID regression in important ways. First, the goal of the moderator regression is to understand the theoretically-relevant moderators of the treatment effect, whereas the purpose of CATE regression is to identify managerially relevant variables that drive individual treatment effect. Second, the primary focus of the moderator regression is inference, while that of CATE regression is prediction. Third, the moderator regression uses the entire sample for estimation, while the causal forest, the basis for the CATE regression, uses a subset of the data (the training sample) for estimation. Fourth, the causal forest underlying the CATE regression splits the training data further to estimate an honest tree, estimating from an even smaller subset of the moderator regression sample. Fifth, relative to the linear moderator regression, the CATE regression can handle a much larger number of covariates and consider their different threshold values for optimization and their interactions without imposing a functional form.

Because of these differences, the results of the CATE regression model may not exactly mirror those of the moderator regression model. Nevertheless, we find that the results are largely consistent across the two regressions. Only the effect of past online frequency in the CATE regression is counter to that in the moderator regression. This difference in the result is likely due to the differences in the two models discussed earlier.

7 Robustness Checks and Alternative Explanations

We perform several robustness checks and tests to rule out alternative explanations for the effect of app failure on purchases.

7.1 Alternative Model Specifications

Although the failure in our data is exogenous, to be sure, in addition to our proposed difference-in-differences model, we also estimate models with propensity score matching and Poisson count data models for the frequency and quantity variables. The results from these models replicate the findings from Tables 3 and 4 and appear in the Web Appendix Tables A1-A2 and B1-B2, respectively. The coefficients of the treatment effect from Table A1 and B1 represent changes in outcomes due to app failures, conditioned on covariates through propensity scores. These results are substantively similar to those in Tables 3 and 4. The insensitivity of the results to control variables suggests that the effect of unobservables relative to these observed covariates would have to be very large to significantly change our results (Altonji et al. 2005).

Similarly, the results are robust to a Poisson specification, reported in Tables A2 and B2.

7.2 Alternative Time Periods

In addition to results from 15 days pre- and post- app service failure, we present results for 30 days pre- and post- models in Web Appendix Tables A3 and B3. These results are substantively similar. Our proposed model leverages a shorter 15-day period to avoid overlaps across pre- and post- periods between shocks that occur close to each other (e.g., a 30-day post period for the November failure shock would overlap with a 30-day pre period of the December failure shock).

7.3 Outliers

We re-estimate the models by removing outlier (extremely high) spenders (three standard deviations in monetary value of purchases in the pre-period) from our data. Web Appendix Tables A4 and B4 report these results. We find consistent and even stronger results.

7.4 Existing Shoppers

Another possible explanation for app failures' effect can be that only new or dormant shoppers are sensitive to failures, perhaps due to low switching costs. Therefore, we remove those with no purchases in the last 12 months to see if their behavior is similar to that of the existing shoppers. Indeed, Web Appendix Tables A5 and B5 report substantively similar results after excluding the new or dormant shoppers.

7.5 Alternative Measures of Digital Channel Use

In lieu of past online purchases dummy as a measure of prior digital channel use, we use a measure based on median splits in the number as well as the share of online purchases. Furthermore, we also use another measure based on prior app usage in the time between app launch and each server failure in the app. The results for alternative online purchase measures are almost the same as our proposed model results, except for online purchases. Similar results emerge from the app use measure in Web Appendix Tables A6 and B6.

7.6 Regression Discontinuity Analysis

To ensure that there are no unobservable differences between failure experiencers and non-experiencers based on time of login, we carry out a ‘regression discontinuity’ (RD) style analysis in the one hour before the start time of the service failure. For the RD analysis, we consider only app users in the neighborhood of this time, using as control group those users who logged in one hour before and after the failure period and as treated the users who logged in during the failure period. The results are substantively similar to our main model results and are reported in Web Appendix Tables A7 and B7.

7.7 Multiple Failures Analysis

To ensure that the effects are robust to shoppers who experience more than one of the four failure shocks, we estimate the main model after including multiple failure experiencers in the sample. The results are similar to our main model results and appear in Web Appendix Tables A8 and B8 with the exception of frequency and quantity of online purchases that also decrease for those with multiple failures. The results also show that app failure decreases purchases in the online channel, suggesting that the presence of multiple failures might make shoppers more sensitive to failures in their online purchases as well.

8 Discussion

8.1 Summary

In this paper, we addressed novel research questions: What is the effect of a service failure in a retailer's mobile app on the frequency, quantity, and monetary value of purchases in online and offline channels? How do shoppers' relationship strength and prior digital channel use moderate these effects? How heterogeneous is shoppers' sensitivity to failures? By answering these questions, our research fills a gap at the crossroads of three disparate streams of research in different stages of development; the mature stream of service failures, the growing stream of omnichannel marketing, and the nascent stream of mobile marketing. We leveraged a set of random systemwide failures in the app to measure the causal effect of failures. To our knowledge, this is the first study to causally estimate the effects of digital service failures using real world data. Using unique data spanning online and offline retail channels, we examined the spillover effects of such failures across channels and examined heterogeneity in these effects based on channels and shoppers.

Our results reveal that app failures have a significant negative effect on shoppers' frequency, quantity, and monetary value of purchases across channels. These effects are heterogeneous across channels and shoppers. Interestingly, the overall decreases in purchases across channels are driven by reductions in store purchases and not in digital channels. Furthermore, we find that shoppers with higher monetary value of past purchases, loyalty program members, and those with a greater number of stores in their neighborhood are less sensitive to app failures.

Overall, we find that not all shoppers are equally sensitive to app failures. In this way, our findings are consistent with the view that some customers may be tolerant and forgiving of technological failures (Meuter et al. 2000). Finally, our study offers novel insights into the cross-channel implications of app failures.

8.2 Managerial Implications

The effects of failures are sizeable for any retailer to alter its service failure preventive and recovery strategies. Based on our estimates, the economic impact of just one app failure in our data is an annual revenue loss of about $22 million for the retailer. Such an amount is strikingly large for any retailer. As sales through the mobile app and online sales are growing rapidly, this impact is only getting larger. Thus, the insights from our research better inform executives in managing their mobile app and channels. They offer several practitioner implications for service failure preventive and recovery strategies.

8.2.1 Preventive Strategy

The finding that app failures result in a 5.7% decrease in monetary value of purchases helps managers estimate the quantitative effect of an app failure. Managers can use this estimate to budget resources for their efforts to prevent or reduce app failures. The effects may be more pronounced for pure-play online retailers.

Managers should anticipate, monitor, regulate, and lower the likelihood of a systemwide failure that affects a majority of their customer base. Over time, these failures may be relatively easier to detect and timely remedial interventions through bug fixes and new app versions can minimize the effect of failure.

In particular, by identifying failure-sensitive shoppers based on relationship strength and prior digital use, managers can take proactive actions to prevent these shoppers from reducing their shopping intensity with the firm. They can issue customized assurances to these shoppers through other communication channels such as email or voicemail; warning them of likely disruptions in the app can preempt negative attributions and attitudes, and limit any impending damage to the brand and revenues due to app failure.

Finally, using the individual-level CATE estimates, managers can target shoppers for preventive strategy. Figure 7 represents the loss of revenues (spending) from each percentile of shoppers at different levels of failure sensitivity. About 70% of the losses in revenues due to failure arise from just 40% of the shoppers. Managers can identify these most failure-sensitive shoppers and manage these shoppers' expectations through email and app notification messaging channels. They can use the results of moderation analysis and causal forest to identify potential new failure insensitive shoppers for acquisition.

< Figure 7 about here >

8.2.2 Recovery Strategies

The finding that app failures result in reduced purchases across channels suggests that managers should develop interventions and recovery strategies to mitigate the negative effects of app failures not just in the mobile channel, but also in other channels, in particular, the offline channel. Thus, seamlessly integrating data from a mobile app with data from its stores and websites can help a multichannel retailer build continuity in shoppers' experiences.

Immediately after a shopper experiences an app failure, the manager of the app should provide gentle nudges and even incentives for the shopper to complete an abandoned transaction on the app. Typically, a manager may need to provide these nudges and incentives through other communication channels such as email, phone call, or face-to-face chat. These nudges are similar in spirit and execution to those from firms like Fitbit and Amazon, who remind customers through email to reconnect when they disconnect their watch and smart speaker, respectively. If the store is a dominant channel for the retailer, the retailer should use its store associates to reassure or incentivize shoppers. In some cases, managers can even offer incentives in other channels to complete a transaction disrupted by an app failure.

Managers should mitigate the negative effects of app failures for the most sensitive shoppers first. They should proactively identify failure-sensitive shoppers and design preemptive strategies to mitigate any adverse effects. We find that first-time digital channel users for a retailer and shoppers with weaker relationship with the provider are more sensitive to failures. Thus, firms should address such shoppers for recovery after a careful cost-benefit analysis. This is important because apps serve as a gateway for future purchases for these shoppers.

Finally, our analysis of heterogeneity in shoppers' sensitivity to app failures suggests that managers should selectively target shoppers for service recovery. Managers should satisfy first the shoppers with the highest values of CATE. Interventions targeted at the 40% of the shoppers who contribute to 70% of the revenue loss will likely lead to higher returns than other efforts.

8.3 Limitations

Our study has limitations that future research can address. First, we do not know the nature of each app failure, so we could not study the intensity of failure experience that could range from a temporary slowdown to a total shutdown. Second, our results are most informative for similar retailers that have a large brick-and-mortar presence with a small but growing online channel or app-induced checkouts and purchases. If data are available, future research could study app failures for primarily online retailers with an expanding offline presence (e.g., Bonobos, Warby Parker). Third, we do not have data on competing apps that shoppers may use. Additional research could study shoppers' switching behavior if data on competing apps are available. Fourth, our data contain relatively low number of purchases in the mobile channel. For better generalizability of the extent of spillover across channels, our analysis could be extended to contexts in which a substantial portion of purchase transactions are made within the mobile app. Fifth, we do not have data on purchases made through mobile apps vs. mobile web browsers. Examining the differences between these two mobile sub-channels is a fruitful avenue for future research. Sixth, in our data, the effect of app failure is strongest when sales are increasing (e.g., November failure). Future research can examine the effect of sales trends in exacerbating failure effects if data on a large number of failures are available. Finally, mobile apps may provide an effective means to resolve problems and recover from the adverse effects of service failures (Tucker and Yu 2018). Thus, the proposed preventive and service recovery strategies could be tested in ethically permissible situations if appropriate data can be gathered.

References

Andreassen, TW (1999) What drives customer loyalty with complaint resolution? *J. Service Res.* 1(4):324-332.

Andrews M, Luo X, Fang Z, Ghose A (2015) Mobile ad effectiveness: Hyper-contextual targeting with crowdedness. *Marketing Sci.* 35(2):218-233.

Angrist JD, Pischke JS (2009) *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton University Press, Princeton, NJ).

Altonji JG, Elder TE, Taber CR (2005) Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *J. Political Economy* 113(1):151-184.

Ansari A, Mela CF, Neslin SA (2008) Customer channel migration. *J. Marketing Res.* 45(1): 60-76.

Ascarza E (2018) Retention futility: Targeting high-risk customers might be ineffective. *J. Marketing Res.* 55(1):80-98.

Athey S, Imbens G (2016) Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sci.* 113(27):7353-7360.

Athey S, Imbens G, Pham T, Wager S (2017) Estimating average treatment effects: Supplementary analyses and remaining challenges. *American Econom. Rev.* 107(5):278-81.

Avery J, Steenburgh TJ, Deighton J, Caravella M (2012) Adding bricks to clicks: Predicting the patterns of cross-channel elasticities over time. *J. Marketing* 76(3):96-111.

Barrons (2018) Walmart: Can it meet its digital sales growth targets? Accessed May 20, 2019, https://tinyurl.com/y266gykm

Bell DR, Gallino S, Moreno A (2018) Offline showrooms in omnichannel retail: Demand and operational benefits. *Management Sci.* 64(4):1629-1651.

Bertrand M, Duflo E, Mullainathan S (2004) How much should we trust difference-in-difference estimates? *Quarterly Journal of Econ.* 119(1):249–75.

Bitner MJ, Booms BH, Tetreault MS (1990) The service encounter: Diagnosing favorable and unfavorable incidents. *J. Marketing* 71-84.

Blancco (2016) The state of mobile device performance and health: Q2. Accessed April 20, 2018, https://tinyurl.com/y9kd6bmz

Bolton RN (1998) A dynamic model of the duration of the customer's relationship with a continuous service provider: The role of satisfaction. *Marketing Sci.* 17(1):45-65.

Breiman L (2001) Random forests. *Machine learning* 45(1):5-32.

Brynjolfsson E, Hu YJ, Rahman MS (2013) Competing in the age of omnichannel retailing. *MIT Sloan Management Rev.* 54(4):23.

Chandrashekaran M, Rotte K, Tax SS, Grewal R (2007) Satisfaction strength and customer loyalty. *J. Marketing Res.* 44(1):153-163.

Chintagunta PK, Chu J, Cebollada J (2012) Quantifying transaction costs in online/off-line grocery channel choice. *Marketing Sci.* 31(1):96-114.

Cleeren K, Dekimpe MG, Helsen K (2008) Weathering product-harm crises. *J. Academy of Marketing Sci.* 36(2):262-270.

Cleeren K, Van Heerde HJ, Dekimpe MG (2013) Rising from the ashes: How brands and categories can overcome product-harm crises. *J. Marketing* 77(2):58-77.

Dimensional Research (2015) Failing to meet mobile app user expectations: A mobile app user survey. Accessed September 24, 2017, https://tinyurl.com/ybbnffcc

Dotzel T, Shankar V, Berry LL (2013) Service innovativeness and firm value. *J. Marketing Res.* 50(2):259-276.

Dukes, A, Zhu Y (2019) Why customer service frustrates consumers: Exploiting hassle costs by a tiered organizational structure. *Marketing Sci.* Forthcoming.

Fong NM, Fang Z, Luo X (2015) Geo-conquesting: Competitive locational targeting of mobile promotions. *J. Marketing Res.* 52(5):726-735.

Forbes LP (2008) When something goes wrong and no one is around: Non-internet self-service technology failure and recovery. *J. Services Marketing* 22(4):316-327.

Forbes LP, Kelley SW, Hoffman KD (2005) Typologies of e-commerce retail failures and recovery strategies. *J. Services Marketing* 19(5):280-292.

Forman C, Ghose A, Goldfarb A (2009) Competition between local and electronic markets: How the benefit of buying online depends on where you live. *Management Sci.* 55(1):47-57.

Ghose A, Kwon H, Lee D, Oh W (2018) Seizing the commuting moment: Contextual targeting based on mobile transportation apps. *Inform. Systems Res.* Forthcoming.

Gijsenberg MJ, Van Heerde HJ, Verhoef PC (2015) Losses loom longer than gains: Modeling the impact of service crises on perceived service quality over time. *J. Marketing Res.* 52(5):642-656.

Goodman PS, Fichman M, Lerch FJ, Snyder PR (1995) Customer-firm relationships, involvement, and customer satisfaction. *Academy of Management Journal* 38(5):1310-1324.

Google M/A/R/C Study (2013) Mobile in-store research: How in-store shoppers are using mobile devices. Google M/A/R/C. Accessed September 24, 2017, http://tinyurl.com/gr7ghpn

Google (2018) Find out how you stack up to new industry benchmarks for mobile page speed. *Google*. Accessed April 27, 2018, https://tinyurl.com/y7zu427b

Guo T, Sriram S, Manchanda P (2018) The effect of information disclosure on industry payments to physicians. *Working Paper.* Available at SSRN: https://ssrn.com/abstract=3064769.

Halbheer D, Gärtner DL, Gerstner E, Koenigsberg O (2018). Optimizing service failure and damage control. *Int. J. Res. in Marketing* 35(1):100-115.

Hansen N, Kupfer AK, Hennig-Thurau T (2018) Brand crises in the digital age: The short-and long-term effects of social media firestorms on consumers and brands. *Int. J. Res. in Marketing* 35(4):557-574.

Hess RL, Ganesan S, Klein NM (2003) Service failure and recovery: The impact of relationship factors on customer satisfaction. *J. Academy of Marketing Sci.* 31(2):127.

Hoffman KD, Bateson JE (1997) *Essentials of Services Marketing.* The Dryden Pres, USA.

Kim SJ, Wang RJH, Malthouse EC (2015) The effects of adopting and using a brand's mobile application on customers' subsequent purchase behavior. *J. Interactive Marketing* 31:28-41.

Knox G, Van Oest R (2014) Customer complaints and recovery effectiveness: A customer base approach. *J. Marketing* 78(5):42-57.

Liu Y, Shankar V (2015) The dynamic impact of product-harm crises on brand preference and advertising effectiveness: An empirical analysis of the automobile industry. *Management Sci.* 61(10):2514-2535.

Ma L, Sun B, Kekre S (2015) The squeaky wheel gets the grease: An empirical analysis of customer voice and firm intervention on Twitter. *Marketing Sci.* 34(5):627-645.

McCollough MA, Berry LL, Yadav MS (2000) An empirical investigation of customer satisfaction after service failure and recovery. *J. Service Res.* 3(2):121-137.

Meuter ML, Ostrom AL, Roundtree RI, Bitner MJ (2000) Self-service technologies: understanding customer satisfaction with technology-based service encounters. *J. Marketing* 64(3):50-64.

Narang U, Shankar V (2019) Mobile app introduction and online and offline purchases and product returns. *Marketing Sci.* Forthcoming.

National Retail Federation (2018) Stores top 100 retailer, Accessed January 5, 2019, https://stores.org/stores-top-retailers-2018/

Neumann N, Tucker CE, Whitfield T (2019) How effective is black-box digital consumer profiling and audience delivery?: Evidence from field studies. *Marketing Sci.* Forthcoming.

Oliver RL (1980) A cognitive model of the antecedents and consequences of satisfaction decisions. *J. Marketing Res.* 460-469.

Pauwels K, Neslin SA (2015) Building with bricks and mortar: The revenue impact of opening physical stores in a multichannel environment. *J. Retailing* 91(2):182-197.

Schmittlein DC, Morrison DG, Colombo R (1987) Counting your customers: Who are they and what will they do next? *Management Sci.* 33(1):1-24.

Shi S, Kalyanam K, Wedel M (2017) What does agile and lean mean for customers? An analysis of mobile app crashes. *Working Paper.*

Smith AK, Bolton RN (1998) An experimental investigation of customer reactions to service failure and recovery encounters: Paradox or peril? *J. Service Res.* 1(1):65-81.

Tax SS, Brown SW, Chandrashekaran M (1998) Customer evaluations of service complaint experiences: Implications for relationship marketing. *J. Marketing.* 62(2):60-76.

Tucker CE, Yu S (2018) Does it lead to more equal treatment? An empirical study of the effect of smartphone use on customer complaint resolution. *Working Paper.* Massachusetts Institute of Technology MA.

Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113(523):1228-1242.

Wang K, Goldfarb A (2017) Can offline stores drive online sales? *J. Marketing Res.* 54(5):706-719.

Xu K, Chan J, Ghose A, Han SP (2016) Battle of the channels: The impact of tablets on digital commerce. *Management Sci.* 63(5):1469-1492.

Table 1. Summary Statistics

Variable	Mean	St. Dev.	Min.	Max.
Frequency of purchases	0.99	1.48	0	79
Quantity of purchases	2.16	4.89	0	1,642
Value of purchases ($)	68.91	175.16	0	70,518
App failure (F)	0.48	0.5	0	1
Time period (P)	0.5	0.5	0	1
Recency of past purchase	-35.82	82.81	-2,725	0
Frequency of past purchases	17.06	18.91	0	964
Value of past purchases ($)	958.4	2,724.43	0	16,73,900
Past online purchase frequency	0.47	1.81	0	460

Notes: (1) These statistics are averaged over the 30-day period pre and post- 15 days of the failure with the exception of "past" variables that are calculated using 12 months before each failure (2) the results are robust even without an outlier spender with $1.6 million purchases (possibly a reseller) who bought several game consoles in bulk over 12 months. Recency is measured as the negative of the number of days since last purchase, such that higher recency implies more recent purchase. N = 970,370.

Table 2. Model-free Evidence: Means of Outcome Variables for Treated and Control Groups

Variable	Treated pre period	Treated post period	Control pre period	Control post period
Frequency of purchases	0.99	0.96	0.99	1.02
Quantity of purchases	2.23	2.03	2.23	2.16
Value of purchases ($)	68.01	67.09	68.66	71.66
Frequency of purchases – Online	0.04	0.03	0.04	0.04
Quantity of purchases – Online	0.06	0.06	0.06	0.05
Value of purchases – Online ($)	2.69	2.47	2.49	2.44
Frequency of purchases – Stores	0.95	0.95	0.95	0.98
Quantity of purchases – Stores	2.16	1.97	2.17	2.11
Value of purchases – Stores ($)	65.32	64.62	66.17	69.22

Notes: These statistics are based on pre- and post- 15 days of the failures. N = 970,370.

Table 3. DID Model Results of Failure Shocks for Purchases across Channels

Variable	Frequency of Purchases	Quantity of Purchases	Value of Purchases
Failure experiencers x Post shock (DID)	-0.054***	-0.123***	-3.918***
	(0.005)	(0.013)	(0.588)
Failure experiencers	0.004	0.009	0.626
	(0.004)	(0.014)	(0.528)
Post shock	0.025***	-0.075***	2.997***
	(0.003)	(0.009)	(0.454)
Intercept	0.984***	2.14***	58.056***
	(0.004)	(0.014)	(0.458)
R-squared	0.002	0.001	0.004

Notes: Robust standard errors clustered by shoppers are in parentheses; *** p < 0.001, ** p < 0.01, * p < 0.05. N = 970,370. DID = Difference-in-Differences.

Table 4. DID Model Results of Failure Shocks for Purchases by Channel

	Offline			Online		
Variable	Frequency of Purchases	Quantity of Purchases	Value of Purchases	Frequency of Returns	Quantity of Returns	Value of Returns
Failure experiencers x Post shock (DID)	-0.052***	-0.122***	-3.749***	-0.002	-0.001	-0.169
	(0.004)	(0.013)	(0.577)	(0.001)	(0.002)	(0.101)
Failure experiencers	0.004	0.008	0.432	0.000	0.001	0.194*
	(0.004)	(0.013)	(0.000)	(0.001)	(0.002)	(0.081)
Post shock	0.029***	-0.069***	3.049***	-0.004***	-0.007***	-0.051
	(0.003)	(0.009)	(0.449)	(0.001)	(0.002)	(0.064)
Intercept	0.937***	2.062***	55.313***	0.047***	0.078***	2.743***
	(0.004)	(0.013)	(0.451)	(0.001)	(0.002)	(0.062)
R-squared	0.0029	0.0012	0.0043	0.0001	0.0007	0.0007

Notes: Robust standard errors clustered by shoppers are in parentheses; *** p < 0.001, ** p < 0.01, * p < 0.05. N = 970,370. DID = Difference-in-Differences.

Table 5. DID Model Results of Failure Shocks for Purchases across Channels: Moderating Effects of Relationship with Retailer and Past Online Purchase

Variable	Frequency of Purchases	Quantity of Purchases	Value of Purchases
Failure experiencers x Post shock (DID)	-0.553***	-1.273***	-40.197***
	(0.048)	(0.282)	(9.284)
DID x Value of past purchases	0.000***	0.001***	0.034**
	(0.000)	(0.000)	(0.010)
DID x Recency of past purchases	-0.002***	-0.005***	-0.186***
	(0.000)	(0.001)	(0.024)
DID x Frequency of past online purchases	-0.053***	-0.075	-2.719
	(0.009)	(0.089)	(1.633)
Value of past purchases	0.000	0.000	0.011
	(0.000)	(0.000)	(0.010)
Recency of past purchases	0.004***	0.007***	0.225***
	(0.000)	(0.001)	(0.023)
Frequency of past online purchases	0.079***	0.277***	3.837*
	(0.01)	(0.055)	(1.479)
Failure experiencers	0.020**	0.074*	2.332*
	(0.006)	(0.031)	(1.000)
Post shock	0.022***	-0.082***	2.711***
	(0.003)	(0.010)	(0.456)
Intercept	1.059***	2.006***	57.349***
	(0.044)	(0.266)	(8.388)
R-squared	0.0954	0.0955	0.0652

Notes: DID = Difference-in-Differences. Robust standard errors are in parentheses; cohort fixed effects included; *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. N = 964,916. The number of observations includes observations of shoppers with at least one purchase in the past for computing recency.

Table 6. Causal Forest Results: Summary of Individual Shopper Treatment Effect for Value of Purchases

Variable	N_{test}	Mean	SD	Min	Max
$\hat{\tau}$	96,071	-1.91	3.35	-13.41	27.59
$\hat{\tau} \mid \hat{\tau} < 0$	66,611	-3.42	2.62	-13.41	-0.00007
$\hat{\tau} \mid \hat{\tau} > 0$	29,460	1.53	2.04	0.00002	27.59

Note: $\hat{\tau}$ represents the estimated individual Conditional Average Treatment Effect (CATE) in the test data.

Table 7. Causal Forest: Post-hoc CATE Regression for Value of Purchases

Variable	CATE (Standard Error) Purchases
Intercept	-1.313*** (0.041)
Recency of purchases	-0.004*** (0.000)
Past value of purchases	0.000*** (0.000)
Past online purchase frequency	0.333*** (0.006)
Age	-0.021*** (0.001)
Gender	-0.046*** (0.014)
Loyalty program	0.163*** (0.017)
Number of stores in CBSA	0.002*** (0.000)
Mobile operating system	-0.018 (0.015)
R-squared	0.61

Note: CATE = Conditional average treatment effect. Robust standard errors are in parentheses; cohort failure dummies are included in the model; gender = 0 (male), or 1 (female); CBSA = Core-Based Statistical Area; mobile operating system = 0 (iOS), or 1 (Android). *** $p < 0.001$. N = 96,071. The number of observations refers to the test sample used in causal forest.

Figure 1. App Screenshots

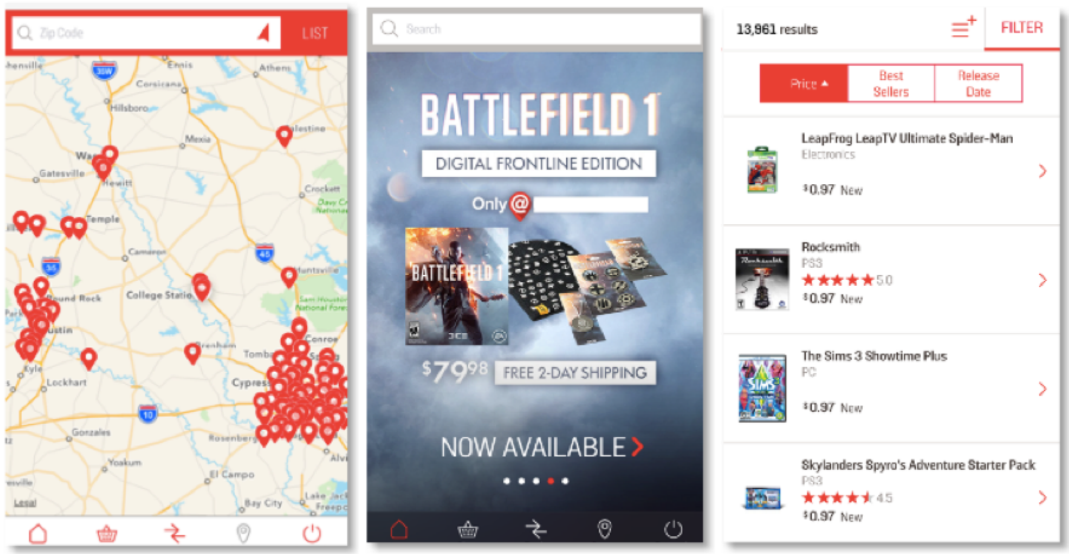
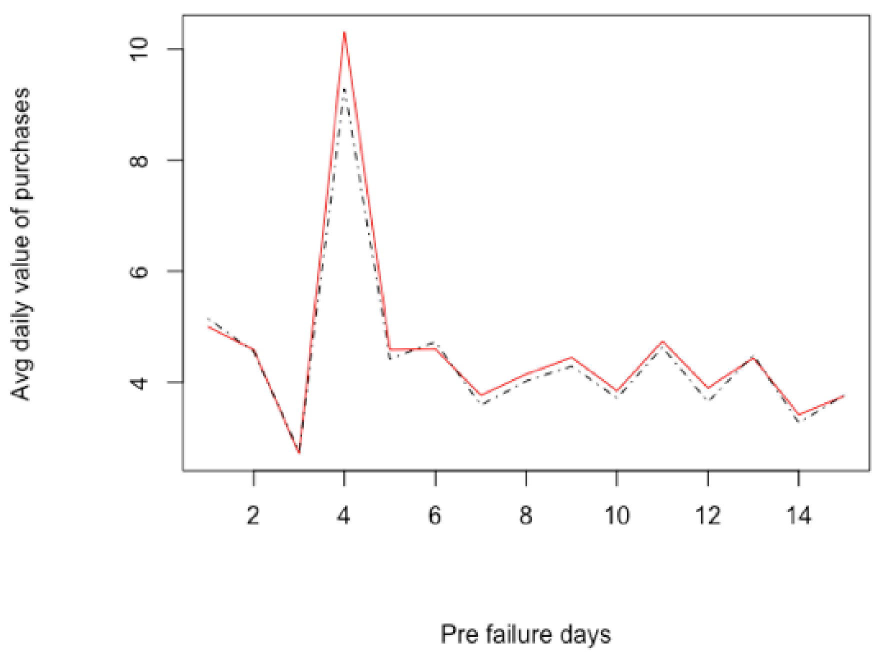
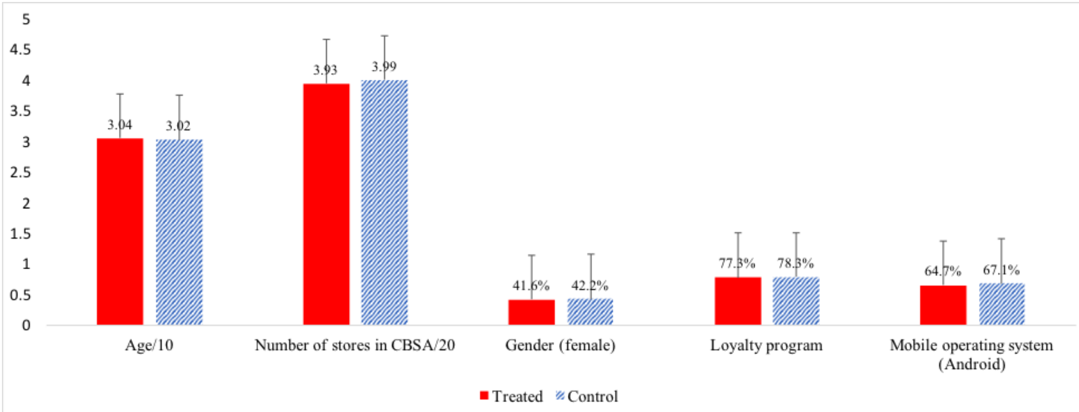


Figure 2. Comparison of Failure-experiencers' and Non-experiencers' Monetary Values of Purchases 15 days before a Failure Shock



Note: The solid red line represents the treated (failure experiencers) group, while the black dotted line represents the control (failure non-experiencers) group.

Figure 3. Comparison of Failure-experiencers and Non-experiencers by Demographics



Notes: CBSA = Core-Based Statistical Area; Gender = 0 (male), 1 (female); loyalty program level represents whether shoppers were enrolled (=1) or not (=0) in an advanced reward program with the retailer on the day of failure; mobile operating system = 0 (iOS), 1 (Android).

Figure 4. Causal Forest Results: Individual CATE

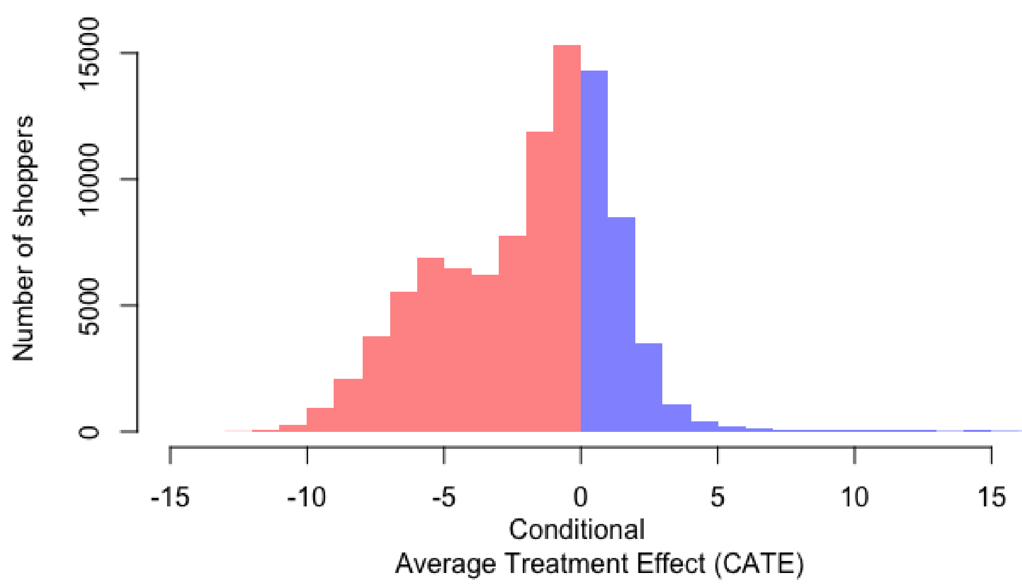
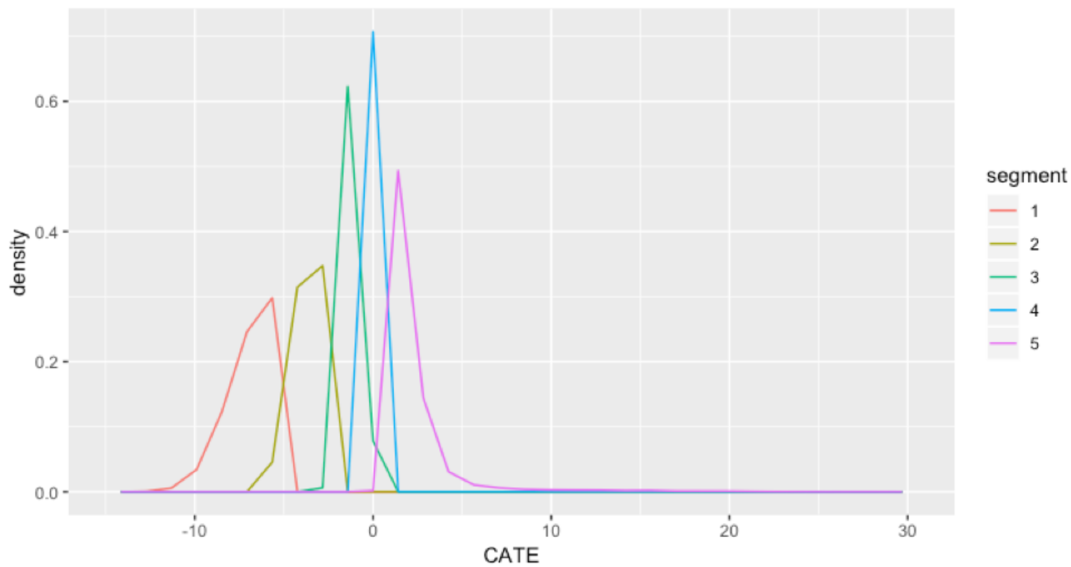


Figure 5. Causal Forest Results: Quintiles by CATE



Note: Segment 1 represents shoppers most adversely affected by failure while Segment 5 represents those who are least adversely affected.

Figure 6. Relationships Between Treatment Effect and Select Covariates

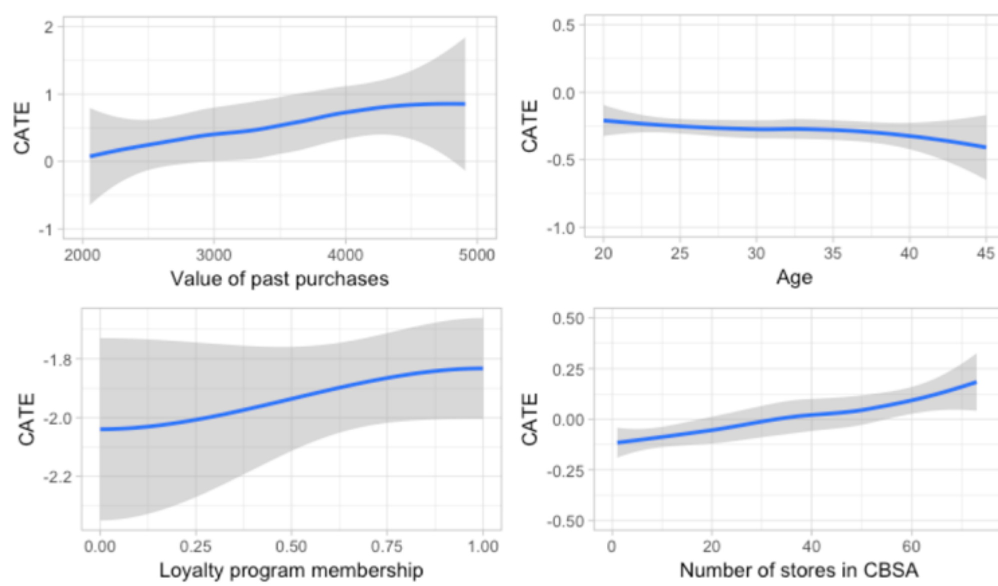
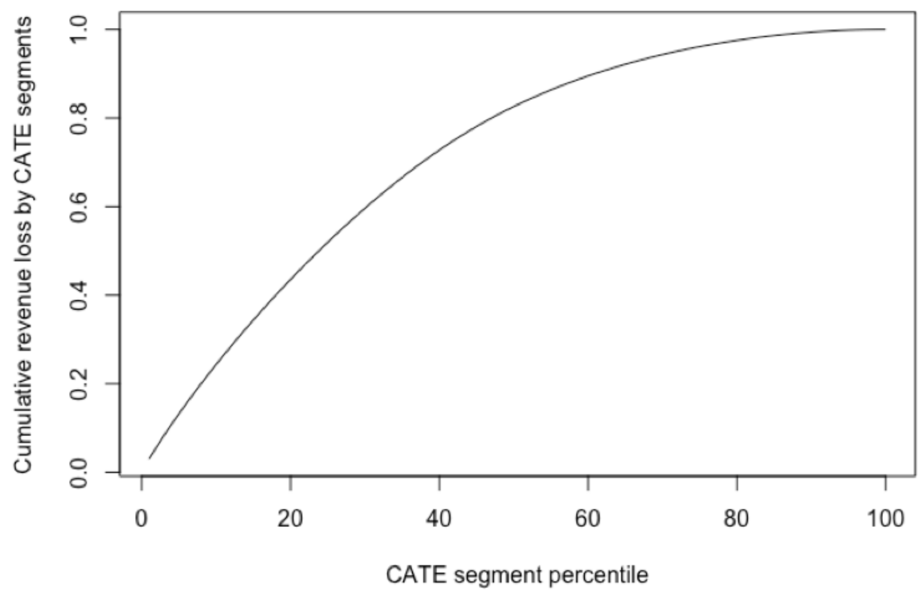


Figure 7. Retailer’s Revenue Loss by Percentile of Shoppers Experiencing Failure



A Web Appendix: Robustness Check for Table 3

In this section, we present the results for robustness checks for the main estimation in Table 3 relating to: (a) alternative models with Propensity Score Matching and using Poisson model (Tables A1-A2), (b) varying time periods (Table A3), (c) outliers (Table A4), (d) existing shoppers (Table A5), (e) alternative measures for prior use of digital channels (Table A6), (f) regression-discontinuity style analysis (Table A7), and (g) sample including multiple failure experiencers (Table A8).

Table A1. Robustness of Table 3 Results to Propensity Score Matching

Variable	Frequency of Purchases	Quantity of Purchases	Value of Purchases
Failure experiencers x Post shock (DID)	-0.041*** (0.005)	-0.100*** (0.016)	-3.117*** (0.683)
Failure experiencers	0.044*** (0.005)	0.106*** (0.016)	4.279*** (0.613)
Post shock	-0.028*** (0.004)	-0.117*** (0.011)	-2.839*** (0.467)
Intercept	0.554*** (0.006)	1.003*** (0.015)	32.789*** (0.556)
R-squared	0.010	0.006	0.008

Notes: Number of observations is 537,772 for 134,443 treated shoppers with complete demographic data matched 1:1 with replacement out of a pool of 158,328 control shoppers with complete demographic data. Robust standard errors clustered by shoppers are in parentheses; *** p < 0.001, ** p < 0.01, * p < 0.05. DID = Difference-in-Differences.

Table A2. Robustness of Table 3 Results to Poisson Specification

Variable	Frequency of Purchases	Quantity of Purchases
Failure experiencers x Post shock (DID)	-0.054*** (0.006)	-0.059*** (0.009)
Failure experiencers	0.005 (0.004)	0.004 (0.006)
Post shock	0.025*** (0.004)	-0.034*** (0.007)
Intercept	-0.016*** (0.004)	0.760*** (0.006)
Log pseudo-likelihood	-1,463,816	-2,985,536

Notes: Number of observations is 970,370. Robust standard errors are in parentheses; *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. DID = Difference-in-Differences.

Table A3. Robustness of Table 3 Results to 30-Day Time Period

Variable	Frequency of Purchases	Quantity of Purchases	Value of Purchases
Failure experiencers x Post shock (DID)	-0.092*** (0.01)	-0.175*** (0.032)	-8.051*** (1.025)
Failure experiencers	-0.044*** (0.007)	-0.126*** (0.023)	-1.805* (0.726)
Post shock	-0.099*** (0.007)	-0.429*** (0.022)	-7.473*** (0.709)
Intercept	2.039*** (0.006)	4.493*** (0.02)	113.289*** (0.627)
R-squared	0.110	0.049	0.052

Notes: Number of observations is 970,370. Robust standard errors clustered by shoppers are in parentheses; *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. DID = Difference-in-Differences.

Table A4. Robustness of Table 3 Results to Outlier Spenders

Variable	Frequency of Purchases	Quantity of Purchases	Value of Purchases
Failure experiencers x Post shock (DID)	-0.051***	-0.108***	-3.033***
	(0.006)	(0.014)	(0.458)
Failure experiencers	0.008*	0.022*	0.305
	(0.004)	(0.010)	(0.325)
Post shock	0.077***	0.092***	18.09***
	(0.004)	(0.010)	(0.317)
Intercept	0.897***	1.848***	42.074***
	(0.003)	(0.009)	(0.280)
R-squared	0.003	0.002	0.012

Notes: Number of observations is 934,638. Robust standard errors clustered by shoppers are in parentheses; *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. DID = Difference-in-Differences.

Table A5. Robustness of Table 3 Results to Existing Shoppers

Variable	Frequency of Purchases	Quantity of Purchases	Value of Purchases
Failure experiencers x Post shock (DID)	-0.053***	-0.116***	-3.81***
	(0.006)	(0.020)	(0.724)
Failure experiencers	0.003	0.001	0.490
	(0.004)	(0.014)	(0.513)
Post shock	0.028***	-0.063***	3.675***
	(0.004)	(0.014)	(0.501)
Intercept	0.992***	2.143***	57.566***
	(0.004)	(0.012)	(0.445)
R-squared	0.002	0.001	0.005

Notes: Number of observations is 942,822. Robust standard errors clustered by shoppers are in parentheses; *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. DID = Difference-in-Differences.

Table A6. Robustness of Table 3 Results to Alternative Measures of Digital Channel Use based on App Usage Before Failure

Variable	Frequency of Purchases	Quantity of Purchases	Value of Purchases
Failure experiencers	-0.474***	-1.107***	-35.032***
x Post shock (DID)	(0.007)	(0.022)	(0.817)
DID x Value of	0.000***	0.001***	0.032**
past purchases	(0.000)	(0.000)	(0.000)
DID x Recency of	-0.002***	-0.005***	-0.175***
past purchases	(0.000)	(0.000)	(0.005)
DID x Frequency of	-0.019***	-0.052***	-1.169***
past online purchases	(0.001)	(0.002)	(0.064)
Value of	0.000***	0.000***	0.011***
past purchases	(0.000)	(0.000)	(0.000)
Recency of	0.003***	0.006***	0.200***
past purchases	(0.000)	(0.000)	(0.002)
Frequency of	0.064***	0.146***	3.851***
past online purchases	(0.000)	(0.001)	(0.032)
Failure experiencers	0.064***	0.171***	5.106***
	(0.004)	(0.013)	(0.487)
Post shock	0.022***	-0.082***	2.711***
	(0.004)	(0.013)	(0.475)
Intercept	0.791***	1.466***	40.48***
	(0.004)	(0.013)	(0.466)
R-squared	0.1480	0.1149	0.0797

Notes: Number of observations is 964,916. Robust standard errors clustered by shoppers are in parentheses; Each moderator interacts with the difference-in-differences (DID) term failure experiencers x post shock; *** p < 0.001, ** p < 0.01, * p < 0.05. The observations include those of shoppers with at least one purchase in the past for computing recency.

Table A7. Robustness of Table 3 Results to Regression Discontinuity Style Analysis

Variable	Frequency of Purchases	Quantity of Purchases	Value of Purchases
Failure experiencers x Post shock (DID)	-0.069***	-0.183***	-2.606
	(0.008)	(0.027)	(1.479)
Failure experiencers	0.061***	0.137***	2.740
	(0.008)	(0.031)	(1.432)
Post shock	0.065***	0.019	5.115***
	(0.006)	(0.024)	(1.382)
Intercept	0.911***	1.974***	53.976***
	(0.007)	(0.027)	(1.182)
R-squared	0.001	0.001	0.003

Notes: Number of observations is 345,708. Robust standard errors clustered by shoppers are in parentheses; *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. DID = Difference-in-Differences.

Table A8. Robustness of Table 3 Results for Shoppers with Multiple Failures

Variable	Frequency of Purchases	Quantity of Purchases	Value of Purchases
Failure experiencers x Post shock (DID)	-0.049***	-0.086***	-3.603***
	(0.004)	(0.012)	(0.552)
Failure experiencers	0.047***	0.112***	2.960***
	(0.004)	(0.013)	(0.513)
Post shock	0.025***	-0.075***	2.997***
	(0.003)	(0.009)	(0.454)
Intercept	0.950***	2.048***	56.110***
	(0.004)	(0.014)	(0.456)
R-squared	0.002	0.002	0.005

Number of observations is 1,190,056. Robust standard errors clustered by shoppers are in parentheses; *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. DID = Difference-in-Differences.

B Robustness Check for Table 4

In this section, we present the results for robustness checks for the cross-channel estimation in Table 4 relating to (a) alternative models with Propensity Score Matching and using Poisson model (Tables B1-B2), (b) varying time periods (Table B3), (c) outliers (Table B4), (d) existing shoppers (Table B5), (e) alternative measures for prior use of digital channels (Table B6), (f) regression-discontinuity style analysis (Table B7), and (g) sample including multiple failure experiencers (Table B8).

Table B1. Robustness of Table 4 Results to Propensity Score Matching

	Offline			Online		
Variable	Frequency of Purchases	Quantity of Purchases	Value of Purchases	Frequency of Returns	Quantity of Returns	Value of Returns
Failure experiencers x Post shock (DID)	-0.039***	-0.098***	-2.941***	-0.001	-0.003	-0.176
	(0.006)	(0.016)	(0.668)	(0.001)	(0.002)	(0.136)
Failure experiencers	0.041***	0.100***	3.838***	0.003**	0.006**	0.441***
	(0.005)	(0.016)	(0.599)	(0.001)	(0.002)	(0.116)
Post shock	-0.019***	-0.101***	-2.121***	-0.009***	-0.016***	-0.718***
	(0.004)	(0.011)	(0.459)	(0.001)	(0.002)	(0.08)
Intercept	0.522***	0.959***	31.028***	0.032***	0.043***	1.762***
	(0.006)	(0.015)	(0.546)	(0.001)	(0.002)	(0.101)
R-squared	0.009	0.005	0.008	0.002	0.002	0.001

Notes: Number of observations is 537,772 for 134,443 treated shoppers with complete demographic data matched 1:1 with replacement out of a pool of 158,328 control shoppers with complete demographic data. Robust standard errors clustered by shoppers are in parentheses; *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. DID = Difference-in-Differences.

Table B2. Robustness of Table 4 Results to Poisson Specification for Count Outcomes

	Offline		Online	
Variable	Frequency of Purchases	Quantity of Purchases	Frequency of Returns	Quantity of Returns
Failure experiencers x Post shock (DID)	-0.055***	-0.060***	-0.044	-0.014
	(0.006)	(0.009)	(0.029)	(0.051)
Failure experiencers	0.0045	0.004	0.003	0.016
	(0.0043)	(0.006)	(0.019)	(0.029)
Post shock	0.0302***	-0.032***	-0.098***	-0.116**
	(0.0042)	(0.007)	(0.02)	(0.035)
Intercept	-0.0646***	0.722***	-3.050***	-2.5420***
	(0.0039)	(0.006)	(0.016)	(0.0273)
Log pseudo-likelihood	-1,432,743	-2,941,566	-162,877	-254,328

Notes: Number of observations is 970,370. Robust standard errors in parentheses; *** p < 0.001, ** p < 0.01, * p < 0.05. DID = Difference-in-Differences.

Table B3. Robustness of Table 4 Results to 30-Day Time Period

	Offline			Online		
Variable	Frequency of Purchases	Quantity of Purchases	Value of Purchases	Frequency of Returns	Quantity of Returns	Value of Returns
Failure experiencers x Post shock (DID)	-0.087***	-0.167***	-7.652***	-0.004*	-0.009	-0.399**
	(0.009)	(0.031)	(1.008)	(0.002)	(0.006)	(0.142)
Failure experiencers	-0.042***	-0.126***	-1.942**	-0.002	0.000	0.136
	(0.007)	(0.022)	(0.714)	(0.001)	(0.004)	(0.101)
Post shock	-0.098***	-0.429***	-7.494***	-0.001	0.001	0.021
	(0.006)	(0.022)	(0.697)	(0.001)	(0.004)	(0.098)
Intercept	1.943***	4.313***	108.151***	0.096***	0.18***	5.138***
	(0.006)	(0.019)	(0.617)	(0.001)	(0.004)	(0.087)
R-squared	0.108	0.049	0.05	0.007	0.002	0.004

Notes: Number of observations is 970,370. Robust standard errors clustered by shoppers are in parentheses; *** p < 0.001, ** p < 0.01, * p < 0.05. DID = Difference-in-Differences.

Table B4. Robustness of Table 4 Results to Outlier Spenders

	Offline			Online		
Variable	Frequency of Purchases	Quantity of Purchases	Value of Purchases	Frequency of Returns	Quantity of Returns	Value of Returns
Failure experiencers x Post shock (DID)	-0.049*** (0.005)	-0.107*** (0.014)	-2.960*** (0.448)	-0.002 (0.001)	-0.001 (0.002)	-0.073 (0.085)
Failure experiencers	0.008* (0.004)	0.021* (0.01)	0.232 (0.317)	0.000 (0.001)	0.001 (0.002)	0.073 (0.06)
Post shock	0.078*** (0.004)	0.094*** (0.01)	17.691*** (0.31)	-0.002* (0.001)	-0.003 (0.001)	0.399*** (0.059)
Intercept	0.854*** (0.003)	1.781*** (0.009)	39.821*** (0.273)	0.043*** (0.001)	0.067*** (0.001)	2.254*** (0.052)
R-squared	0.0024	0.0017	0.0118	0.0011	0.001	0.0008

Notes: Number of observations is 934,638. Robust standard errors clustered by shoppers are in parentheses; *** p < 0.001, ** p < 0.01, * p < 0.05. DID = Difference-in-Differences.

Table B5. Robustness of Table 4 Results to Existing Shoppers

	Offline			Online		
Variable	Frequency of Purchases	Quantity of Purchases	Value of Purchases	Frequency of Returns	Quantity of Returns	Value of Returns
Failure experiencers x Post shock (DID)	-0.051*** (0.006)	-0.115*** (0.020)	-3.628*** (0.713)	-0.001 (0.001)	-0.001 (0.003)	-0.182 (0.108)
Failure experiencers	0.003 (0.004)	0.000 (0.014)	0.306 (0.505)	0.000 (0.001)	0.001 (0.002)	0.184* (0.076)
Post shock	0.032*** (0.004)	-0.055*** (0.014)	3.743*** (0.493)	-0.004*** (0.001)	-0.007*** (0.002)	-0.067 (0.075)
Intercept	0.945*** (0.004)	2.064*** (0.012)	54.784*** (0.438)	0.048*** (0.001)	0.079*** (0.002)	2.783*** (0.066)
R-squared	0.0019	0.0012	0.0045	0.0012	0.0007	0.0007

Notes: Number of observations is 942,822. Robust standard errors clustered by shoppers are in parentheses; *** p < 0.001, ** p < 0.01, * p < 0.05. DID = Difference-in-Differences.

Table B6. Robustness of Table 4 Results to Alternative Measure of Digital Channel Use based on App Usage Before Failure

	Offline			Online		
Variable	Frequency of Purchases	Quantity of Purchases	Value of Purchases	Frequency of Returns	Quantity of Returns	Value of Returns
Failure experiencers	-0.455***	-1.027***	-33.232***	-0.019***	-0.08***	-1.799***
x Post shock (DID)	(0.007)	(0.022)	(0.805)	(0.001)	(0.003)	(0.127)
DID x Value of	0.000***	0.001***	0.031***	0.000***	0.000***	0.002***
past purchases	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
DID x Recency of	-0.002***	-0.005***	-0.164***	0.000***	0.000***	-0.01***
past purchases	(0.000)	(0.000)	(0.005)	(0.000)	(0.000)	(0.001)
DID x Frequency of	-0.018***	-0.051***	-1.128***	-0.001***	0.000	-0.041***
past online purchases	(0.001)	(0.002)	(0.064)	(0.000)	(0.000)	(0.01)
Value of	0.000***	0.000***	0.011***	0.000***	0.000***	0.000***
past purchases	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Recency of	0.003***	0.006***	0.193***	0.000***	0.000***	0.007***
past purchases	(0.000)	(0.000)	(0.002)	(0.000)	(0.000)	(0.000)
Frequency of	0.060***	0.139***	3.602***	0.004***	0.007***	0.248***
past online purchases	(0.000)	(0.001)	(0.032)	(0.000)	(0.000)	(0.005)
Failure experiencers	0.061***	0.164***	4.695***	0.003***	0.007***	0.411***
	(0.004)	(0.013)	(0.48)	(0.001)	(0.002)	(0.075)
Post shock	0.026***	-0.075***	2.769***	-0.004***	-0.007**	-0.058
	(0.004)	(0.013)	(0.468)	(0.001)	(0.002)	(0.074)
Intercept	0.758***	1.417***	38.674***	0.033***	0.049***	1.806***
	(0.004)	(0.013)	(0.46)	(0.001)	(0.002)	(0.072)
R-squared	0.1423	0.1115	0.0771	0.0117	0.009	0.0058

Notes: Number of observations is 964,916. Robust standard errors clustered by shoppers are in parentheses; *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. DID = Difference-in-Differences. The observations include those of shoppers with at least one purchase in the past for computing recency.

Table B7. Robustness of Table 4 Results to Regression Discontinuity Style Analysis

	Offline			Online		
Variable	Frequency of Purchases	Quantity of Purchases	Value of Purchases	Frequency of Returns	Quantity of Returns	Value of Returns
Failure experiencers x Post shock (DID)	-0.071*** (0.008)	-0.189*** (0.027)	-2.587* (1.465)	0.002 (0.002)	0.006 (0.004)	-0.019 (0.181)
Failure experiencers	0.06*** (0.007)	0.135*** (0.031)	2.478 (1.423)	0.001 (0.001)	0.002 (0.003)	0.262 (0.148)
Post shock	0.071*** (0.006)	0.031 (0.023)	5.110*** (1.374)	-0.006*** (0.001)	-0.012*** (0.003)	0.006 (0.134)
Intercept	0.866*** (0.007)	1.903*** (0.027)	51.516*** (1.175)	0.044*** (0.001)	0.071*** (0.003)	2.461*** (0.109)
R-squared	0.001	0.0007	0.0026	0.0006	0.0006	0.0002

Notes: Number of observations is 345,708. Robust standard errors clustered by shoppers are in parentheses; *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. DID = Difference-in-Differences.

Table B8. Robustness of Table 4 Results to Multiple Failures

	Offline			Online		
Variable	Frequency of Purchases	Quantity of Purchases	Value of Purchases	Frequency of Returns	Quantity of Returns	Value of Returns
Failure experiencers x Post shock (DID)	-0.046*** (0.004)	-0.083*** (0.012)	-3.310*** (0.544)	-0.003** (0.001)	-0.003 (0.002)	-0.292** (0.089)
Failure experiencers	0.045*** (0.004)	0.109*** (0.013)	2.669*** (0.507)	0.002** (0.001)	0.004* (0.002)	0.291*** (0.069)
Post shock	0.029*** (0.003)	-0.069*** (0.009)	3.049*** (0.449)	-0.004*** (0.001)	-0.007*** (0.002)	-0.051 (0.064)
Intercept	0.904*** (0.004)	1.972*** (0.013)	53.41*** (0.045)	0.046*** (0.001)	0.076*** (0.002)	2.700*** (0.06)
R-squared	0.0017	0.0015	0.0046	0.001	0.0007	0.0007

Notes: Number of observations is 1,190,056. Robust standard errors clustered by shoppers are in parentheses; *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. DID = Difference-in-Differences.