PERSONALIZED GAME DESIGN FOR IMPROVED USER RETENTION AND MONETIZATION IN FREEMIUM GAMES

Eva Ascarza
Harvard University

Oded Netzer
Columbia University

Julian Runge
Northwestern University

December 2024

Accepted at the *International Journal of Marketing Research*

Eva Ascarza is the Jakurski Family Associate Professor of Business Administration, Harvard Business School (email: eascarza@hbs.edu). Oded Netzer is the Arthur J. Samberg Professor of Business at Columbia Business School (email: onetzer@gsb.columbia.edu), and Julian Runge is Assistant Professor of Marketing at Northwestern University's Medill School of Journalism, Media and Integrated Marketing Communications (email: julian.runge@northwestern.edu). The authors are grateful to Hengyu Kuang for excellent research assistantship.

Personalized Game Design for Improved User Retention and Monetization in Freemium Mobile Games

Abstract

One of the most significant levers available to gaming companies in designing digital games is setting the level of difficulty, which essentially regulates the user's ability to progress within the game. This aspect is particularly significant in free-to-play (F2P) games, where the paid version often aims to enhance the player's experience and to facilitate faster progression. In this paper, we leverage a large randomized control trial to assess the effect of dynamically adjusting game difficulty on players' behavior and game monetization in the context of a popular F2P mobile game. The results highlight the intertwined dynamics of customer retention and monetization in such settings. As expected, offering players an easier game significantly decreases purchases in the specific round played — faced with an easier game, users do not need to resort to in-game purchases to make progress. However, because lowering the game difficulty increases both immediate engagement and long-term retention, lower difficulty levels result in a significant increase in customer spending both in the short and long run. We find substantial heterogeneity in the strength of these effects. Customers who are more prone to making progress in the game exhibit stronger effects in both the short and long run, whereas customers who previously spent money on the game exhibit stronger effects in long-term monetization. We leverage these insights to demonstrate how the focal firm can use game difficulty adjustment to further increase revenues from both advertising and premium services and to recommend personalized product design strategies for freemium apps more broadly.

Keywords: gaming, game difficulty, dynamic difficulty adjustment (DDA), freemium, retention, monetization, field experiments

1. INTRODUCTION

Mobile gaming is one of the fastest-growing sectors in the digital realm, accounting for over 70% of consumer spending on mobile apps.[1] By 2027, mobile gaming revenue is projected to surpass $222 billion worldwide, as per Statista (2023). Free-to-play (F2P) games stand as a dominant force within this burgeoning market. F2P is the application of freemium pricing to games. In such games, users can play without incurring any cost but also have the option of accessing premium services through in-app purchases (IAP) or "add-ons". These can be in the form of virtual tools, puzzle cues, extra time, and the like, which enhance the game, facilitate progress, or provide an edge against other players.

Central to the appeal of the F2P model is its dual-revenue stream. One part of a company's income is derived from advertising, capitalizing on the large user base engaged with the free version of the game. Simultaneously, a significant portion of firms' revenues comes from players making premium purchases to augment their experience. This setup presents a multifaceted strategic challenge for game companies. They not only view user retention and monetization as pivotal revenue drivers but also acknowledge that these two aspects can sometimes clash in terms of game design strategy. On the one hand, the allure of the game's free version must remain intact, as its quality and attractiveness are fundamental to attracting and maintaining a large user-base, and in turn advertising revenue and pipeline for IAP-based monetization. On the other hand, there is a pressing need to steer players toward premium purchases, which often involves restricting the range or simplicity of the freely available game content.

One important lever that game designers can use to strike the balance between the "free" and "paid" aspects of a game is dynamic difficulty adjustment (DDA), which involves tailoring

[1] https://www.appannie.com/en/insights/market-data/games-accounted-for-70-of-consumer-spend-in-apps-in-q3-2019/

the game's difficulty over time based on the user's performance or current level of play. This mechanism is widely regarded as influential by academics (Hunicke 2005) and consumers (Reddit 2023) alike and has received ample scholarly attention in human-computer interaction as a method of modifying the game so that players do not feel bored when the game is very simple, or frustrated when it is very difficult (Zohaib 2018). DDA is an important and broadly applicable tool for gaming companies to increase player engagement and retention. It can be viewed as achieving for single-player games and game modes (when single and multi-player modes coexist, which is often the case) what matchmaking achieves in multi-player gaming (Lopez-Vargas, Runge, and Zhang 2022): personalizing the level of challenge to individual players' motivation and skill level.

From a marketing perspective, DDA is a personalization of the product. By hitting the right mix of reward and challenge, DDA aims to provide a more motivating game experience, enhancing customer's engagement and retention (Huang, Jasin, and Manchanda 2019). Unlike financial levers, such as promotional offers or discounts for IAP (Runge, Levav, and Nair 2022), or communication levers, like in-app notifications (Runge et al. 2014; Bashirzadeh, Mai, and Faure 2022), which require the customer to respond to company-initiated interactions, DDA is more subtle as it does not interrupt gameplay. Instead, users encounter a seamlessly adjusted difficulty level based on their current performance. Furthermore, the ease of customization in digital products (Ansari and Mela 2003) enables game designers to vary the difficulty both across users and dynamically over time.

This research leverages experimental variation from the rollout of a DDA system in a popular F2P mobile game to analyze its impact on customer behavior and game monetization. We demonstrate that, although reducing game difficulty often substitutes for IAPs and

immediate game monetization, its positive effect on customer engagement and retention leads to a long-term increase in monetization. Specifically, we find that enhancements to engagement and retention among low-activity users, through adjustments to the game's difficulty level, yield a boost in in-game spending. From a scholarly perspective, to the best of our knowledge, we are the first to document how a DDA system can alleviate the commonly assumed tension between the 'free' and monetized aspects of freemium games (Halbheer et al., 2014; Lambrecht and Misra, 2017; Lee, Kumar, and Gupta, 2017). From a practical standpoint, our research explores an important personalization strategy for game designers—namely, dynamically and individually adjusting game difficulty. Personalization is particularly vital in the F2P domain, as these games often experience high churn rates (Runge et al. 2014; Ceci 2023). Our empirical findings indicate that modifying the game's difficulty can transform previously inactive or 'at-risk' players into profitable customers.

We collaborated with a large game company that ran a field experiment to evaluate the rollout of a DDA system in a popular F2P mobile game. Specifically, we track the behavior of over 300,000 players over the course of 12 weeks through which game difficulty was exogenously manipulated. This dynamic adjustment is compared to a fully randomized holdout group without any adjustments to difficulty. In the treatment condition, users who were identified to be at higher risk of churning received increasingly easier game puzzles. In the control condition, game difficulty remained the same irrespective of risk of churn. From this exogenous and randomized variation in the level of difficulty, we measure the short- and long-term effect of the DDA intervention on customer engagement, retention, and purchases.

We identify two types of positive synergies between retention and monetization under DDA. Firstly, while reducing the game's difficulty decreases the immediate propensity to

purchase extras—often utilized to ease gameplay and hasten progress—the beneficial impact of reduced difficulty on same-day engagement and future retention leads to an overall positive effect on game monetization, resulting in additional spending on IAPs. This positive effect on monetization is apparent even on the first day of the intervention. Secondly, the increased retention resulting from the lowered difficulty under DDA is particularly pronounced in IAPs for items that provide access to more game content, such as gates (repeated paywalls that gate access to further levels of a game), both in the short and long term.

We explore the heterogeneity of these effects, examining how individual player characteristics, such as prior in-game spending and progress, intensify or mitigate the observed impacts. We observe significant heterogeneity in the effect of reducing game difficulty. Customers more inclined towards making progress in games, especially during moments of frustration or when they can achieve more progress, exhibit a stronger positive response to reduced game difficulty. Additionally, those who previously spent money on the game, display the strongest effect on IAP spending. These findings align with the notion that, though users are less inclined to spend on making the game easier when difficulty is lowered, they're more willing to invest in removing other obstacles (such as paywalls) that hinder their continued consumption and progress (Lambrecht and Misra 2017; Aral and Dhillon 2021).

The paper proceeds with a brief review of relevant literature before presenting empirical results. The paper concludes with a discussion and overview of contributions.

2. LITERATURE REVIEW

Substantively, this paper contributes to the literature on games in marketing and the marketing of and in games (Nair 2007; Hofacker et al. 2016; Appel et al. 2020; Haenlein, Libai, and Muller 2023; Runge and van Dreunen 2024). While games were historically distributed on

physical media such as discs, CDs, and DVDs and sold in retail stores (Nair 2007), digitization, the freemium business model, and the proliferation of mobile phones have deeply transformed the games industry (Runge, Levav, and Nair 2022). Gaming on handheld devices and phones now accounts for three times the revenue of gaming on all other platforms such as consoles, PCs, and laptops combined. More than 3 billion people play digital games worldwide (Statista 2023). Games are distributed to phones as mobile apps that are downloaded in real-time from app stores such as Apple's Appstore and Google's Playstore. Most gaming apps make use of freemium pricing with the ability to obtain in-game goods in IAP or through advertising exposure.

Despite the economic significance of mobile games, the role of marketing strategy in this industry has not received substantial attention in the academic literature. In a study of mobile apps and games, Appel et al. (2020) observed that "the question of choosing the correct business model and marketing mix becomes more pressing." Runge, Levav, and Nair (2022) urge marketing scholars to revise their perspectives on the price and promotion dimensions of the marketing mix in this context. Utilizing a large-scale field experiment in a F2P mobile game, they find that price reductions are remarkably profitable, both in the short and in the long run.

One of the key strategies available to publishers of mobile games to address the product dimension of the marketing mix is DDA, i.e., the data-driven adaptation and personalization of the game's difficulty to a specific user's needs and preferences (Xue et al. 2017; Zohaib 2018; Huang, Jasin, and Manchanda 2019). The focus of our paper is to explore the role of DDA in affecting both customer behavior and game monetization. We do so by leveraging experimental manipulation of the DDA system of a popular F2P mobile game.

The present work is also related to literature investigating customer engagement and progress in video games (e.g., Huang, Jasin, and Manchanda 2019; Rutz, Aravindakshan, and Rubel

2019). Huang et al. (2019) find that gamers in different engagement states exhibit varied responses to motivations, such as the need for challenge. Rutz et al. (2019) examined user behavior across 193 mobile games, modeling user engagement following the initial download. They identified significant usage heterogeneity across these games. While these studies offer invaluable insights into player behavior within games, their reliance on observational data curtails their ability to explore the effects of game design alterations on user behaviors and game outcomes. Hofacker et al. (2016) suggested that achieving an optimal balance between game difficulty and the correlation between difficulty and reward can yield positive game outcomes. Amabile and Kramer (2011) posit that, for optimal player engagement, game designers should ensure players 1) experience daily progress, 2) attain small victories even amidst setbacks, and 3) progress in diverse manners. The DDA system we study manipulates precisely that: progress and incremental wins achieved and realized by players through dynamically personalized game difficulty.

Several computer science and information systems papers (e.g., Hadiji et al. 2014; Lee et al. 2016; Runge et al. 2014; Viljanen et al. 2020) employed machine learning classifiers to predict churn in F2P games. In a context similar to ours, Runge et al. (2014) studied the impact of firm communications on at-risk customers. In this paper, we explore a strategy to prevent churn not through communication but through "behind the scenes" product adjustment by changing the level of difficulty. The notion of DDA is well-established in the Human-Computer Interaction (HCI) literature in Computer Science. Scholars in this field have crafted advanced artificial intelligence tools to enhance the incorporation of DDA in game design (Hunicke 2005; Xue et al. 2017; Zohaib 2018). However, to the best of our knowledge, the current literature has

not yet examined the implications of DDA systems on customer behavior and game monetization – a gap that we address.

This work is also related to the literature on freemium monetization (Halbheer et al. 2014; Lambrecht and Misra 2017; Gu, Kannan, and Ma 2018; Aral and Dhillon 2021). Despite its potential significance, the impact of the design of a freemium offering on customer retention is not well understood. A notable exception is Appel et al. (2020) who develop an analytical model in which satiation (a parameter that captures the customer's likelihood of becoming satiated with the content and thus churning) can influence the design of the freemium product. More recently, Haenlein, Libai, and Muller (2023) also explored the role of satiation in a specific game, examining it in the context of cross-promotions intended to transition players to other games. Both studies consider satiation as an external aspect of the product, while we study one of the most impactful product strategies available to game publishers (DDA systems) that has the potential to fundamentally change a product's satiation and retention dynamics (Xue et al. 2017; Zohaib 2018). Our findings bear relevance for the effects of personalized freemium product design on customer retention more generally. While existing accounts (e.g., Lee, Kumar, and Gupta, 2017) commonly posit a substitutive relationship between monetization and retention — suggesting that more of the game for free (i.e., lower monetization) will increase retention at the cost of lower monetization — our results indicate that product personalization strategies can instead turn this relationship synergistic, both in the short and in the longer run.

In the literature on freemium design, several studies have explored the economic viability of the freemium model in comparison to its two extremes: exclusively free or solely paid (e.g., Li, Jain, and Kannan 2019; Liu et al. 2014; Shi, Zhang, and Srinivasan 2019; Deng, Lambrecht, and Liu 2020). Other research has centered on determining the quantity of free products to offer

(Lambrecht and Misra 2017) or the effects of broadening the product range (Gu, Kannan, and Ma 2018). The majority of the aforementioned studies examine feature-limited versions of freemium, where customers access a restricted, free version of the product, and a subset transition to a paid variant. These studies generally emphasize the factors influencing customer conversion from free to premium (e.g., Lee, Kumar, and Gupta 2017). In contrast, in this work, we delve into the IAP version of freemium, sometimes termed noncontractual freemium. Here, customers can use the app (or play the game) for free but have the option to enrich their experience by purchasing add-ons at any moment. In this model, because premium usage is on a "per transaction" basis (as opposed to switching to a paid subscription), other elements of customer lifetime value, such as usage, monetary purchases, and retention, become vital in comprehending the profitability of the freemium design. Due to the noncontractual nature of purchases and due to purchases at many price levels being available, this setting further promises to allow for a more nuanced study of associations between user monetization and retention.

3. EMPIRICAL APPLICATION

In this section, we introduce the empirical context and provide model-free evidence of the impact of game difficulty on game engagement, retention, and monetization. We also detail the field experiment and the data employed in our empirical study.

3.1 Business context: A free-to-play (F2P) game

We collaborated with a company specializing in mobile gaming applications. The game utilized for our empirical study is a F2P puzzle mobile game, reminiscent of the popular Candy Crush Saga. In this game, players advance through levels by matching three or more pieces of the same color arranged on a game board. When these pieces are matched successfully, they vanish, allowing new pieces to occupy the vacant spaces on the board. The more pieces a player matches

in a single move, the higher the points they earn. The game features numerous levels, which players must tackle sequentially. For instance, one cannot access level 33 without successfully completing level 32. Each level poses distinct objectives/challenges, such as matching a particular number of specific colored pieces or reaching a set score. These objectives must be achieved within a set number of moves or a designated time. Failing to achieve the goal means the player loses a “life” and must retry the level to progress. Upon successfully meeting a level’s objectives, players are awarded one to three stars, determined by their performance, and can then advance to the subsequent level.

The game has inbuilt mechanisms that periodically hinder players' progress. Firstly, players have a limited life count at any given time, starting with a stock of five lives. If all lives are depleted, players must wait for new lives to be generated to continue playing. Specifically, the game replenishes one life every 30 minutes, with the stock never exceeding five lives at any time. Secondly, there is a “gate” every 20 levels, starting from level 40. This implies that after completing levels 40, 60, 80, and so on, players face a five-day waiting period during which they cannot attempt higher levels and thus cannot progress. Regardless of their position in the game, be it halted by a gate or otherwise, players can replay previous levels as long as they have lives.

3.1.2 Monetization via in-app-purchases (IAP)

The game monetizes its users through the sale of in-game currency, commonly referred to as “coins” which can be purchased using real money. Upon starting the game, users receive an initial amount of 70 coins. This amount can be increased or replenished by spending real money. Coins can be redeemed for one of three benefits. Firstly, if a user depletes all their lives, they can buy a new life with coins to continue playing. Secondly, when faced with a gate, players can

expend coins to bypass it and advance to higher levels.[2] Lastly, coins can be used to purchase "extras" that assist in achieving level objectives. For instance, a player might purchase a booster that reduces the number of pieces' colors in a puzzle, facilitating easier matches and thereby aiding progress towards level goals and achieving higher scores. Essentially, these "extras" allow users to pay to reduce the difficulty of the level they are playing. All purchasable items (extras, gates, lives) essentially enhance a user's ability to progress in the game.

Like most freemium products and mobile games in the market, a significant portion of players never spend money. However, those who do often make multiple purchases, contributing a substantial portion of the firm's revenue. In our game, the majority of IAP revenue stems from users purchasing extras and bypassing gates rather than buying additional lives. To provide some context regarding the amounts spent by players, the average purchase amount is approximately $4 per player, and the cost to bypass a gate is equivalent to $1.

3.1.3 Monetization via in-app advertising

The mobile game also earns revenue from advertising, which is directly tied to the number of active players at any given moment ("eyeballs"). The in-app advertising market has been growing rapidly, with projections indicating it will amount to $315B in 2023 (Statista 2023a). The most prevalent form of advertising in this game consists of brief (10-30 second) videos that users must watch before starting another round. This method is widely adopted across mobile games, with 94% of game developers stating they used in-app advertising in 2019. While we lack specific data on advertising exposure and revenue for our focal company (as this information was not shared), all other factors being equal, the more rounds a user plays, the

[2] Users can add one life to their "stock" or "break" a gate by sending requests to friends via Facebook. However, at the time of the data, this feature was rarely used—less than 10% of our uses even linked their game to Facebook—and therefore we ignore the activity on Facebook.

greater their exposure to advertisements, leading to higher advertising revenue. Consequently, player engagement is a critical metric for the firm and is consistently monitored and optimized (Seufert 2013).

3.2 User behavior

3.2.1 Retention and monetization in F2P games

Like most other freemium services, including F2P games and mobile apps, retention rates tend to be low. **Figure 1**(left) displays the proportion of users who play the game X days after installation (up to 28 days), based on a random sample of users (N=10,000). From the figure, it is evident that only 47.54% of users continue playing the game the day after installation, and this number drops to just 9.02% by the 28th day. Additionally, among those who continue playing, the level of engagement with the game diminishes over time. **Figure 1**(right) illustrates the average number of rounds played, provided that the user plays at all on that specific day. This significant decline in both retention and engagement underscores the importance for firms to focus on retaining and engaging their customers.

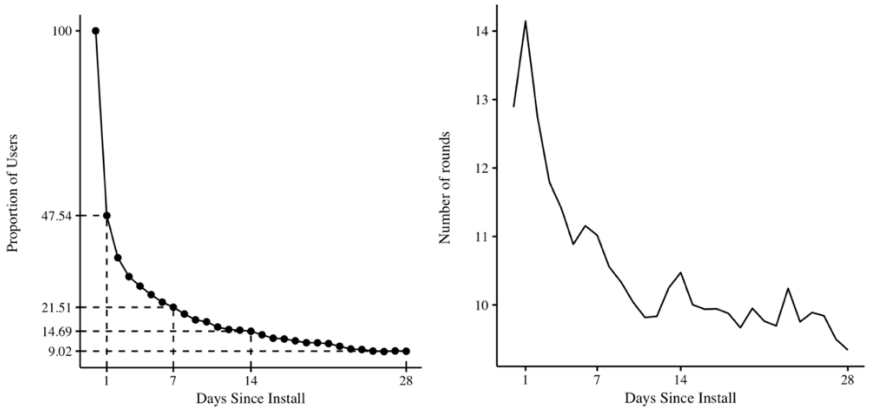

Figure 1: Retention and Engagement by day after installation*. The figure on the left shows the proportion of users who play up to 28 days after having installed the game. The figure on the right shows the average number of rounds played up to 28 days after installation, conditional on the user playing on that particular day.*

Consistent with patterns observed in most F2P games, many users never spend money on the game. For instance, of the customers who churned during their first 28 days of play after installation, over 60% did not use any coins (keep in mind every user begins the game with an endowment of 70 coins), and over 99% did not spend any money. Of the users who remained active past their first 28 days (meaning we observed them playing subsequently), 64% used at least one (free) coin, and 4.1% spent money within those initial 28 days. While this discrepancy in monetization rates between early churners and users who continue playing can be partially attributed to the game's increasing complexity in later stages, it also underscores the significance of user retention in achieving high in-game expenditures.

We will now delve into how specific aspects of the game design impact user engagement, retention, and monetization. It is important to note that these model-free analyses are not intended to accurately gauge the effect of game characteristics and progress on customer behavior. For such precision, we would require a clear and exogenous measure of progress. Instead, our goal is to showcase evidence in our data linking game design and progress to user engagement, retention, and monetization. We will further bolster this with data from a large-scale field experiment.

3.2.2 Progress and game design

In most puzzle-type games, the primary motivation for playing is to progress through new levels. Consequently, the capacity to make progress influences user retention, engagement, and monetization. **Figure 2** shows the relationship between the progress achieved — measured by the number of new levels passed on one day of play — and the likelihood that the user will play again on the subsequent day. A distinct positive correlation is evident: users are more inclined to return and play if they have made significant progress in the game the prior day. Moreover, given

that they have lives remaining to continue the game, users are 14% more likely to pause playing for at least 10 minutes after failing to clear a level in the current round compared to when they have successfully completed a level for the first time.

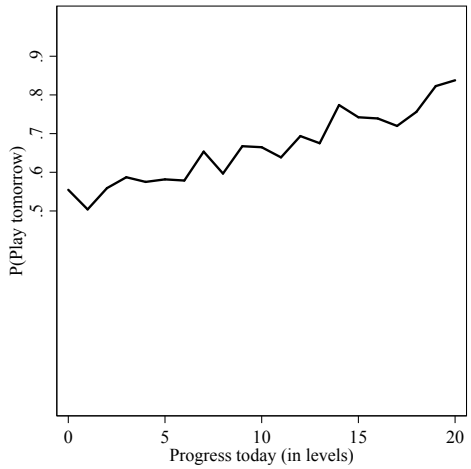

Figure 2: Probability of playing again tomorrow as a function of progress today*. Progress today is measured as the number of new levels that the user passed in a particular day (levels that the user did not pass in the past).*

Despite the positive relationship between game progress and user engagement and retention, game designers often introduce mechanisms that deliberately impede progress to create a sense of challenge and monetize the game. As expected, these mechanisms influence user retention and engagement. One such mechanism that directly hinders users from progressing is the introduction of gates. Here, users must wait for 5 days to play higher levels unless they use coins to bypass the gate. **Figure 3**-left shows that users are more likely to churn (defined as not playing in the next 30 days) after surpassing a gate level — namely, levels 40 and 60 — compared to non-gate levels. Additionally, we note that users tend to stop playing immediately after clearing a "gate" level (even though they can still engage with lower levels) more than after surpassing a non-gate level (as seen in **Figure 3**-right). Here, "stop playing" means not initiating another round for at least 10 minutes, assuming they have lives left to continue. These illustrations offer clear, model-free evidence that users' interest wanes when they realize they

cannot progress further in the game, even after completing the most recent level and having lives remaining.

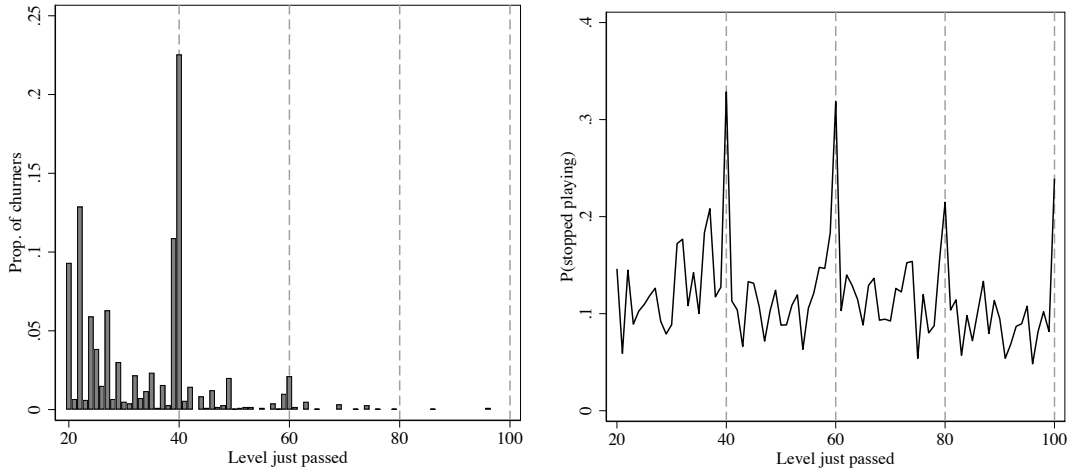

Figure 3: Retention and engagement when stopped at gates. *The left figure displays a histogram of the highest level achieved by churners (defined as users who have not played for the next 30 days). The figure on the right illustrates the proportion of users who stop playing right after clearing a level for the first time (here, "stop playing" is defined as not initiating another round for at least 10 minutes, despite having lives available). Vertical dashed lines indicate "gate" levels. Only users who surpassed level 20 are considered in this analysis.*

One may wonder why the firm would introduce gates given their higher churn rate. Gates represent a juncture where the free portion of the game intersects with the company's monetization strategy, enticing players to spend money. In fact, one of the primary uses for coins is to enable users to bypass a gate, allowing them to access higher levels without enduring the five-day wait. This behavior is corroborated in **Figure 4**, which displays the average likelihood of spending money based on the highest level reached in the game. The figure clearly demonstrates that users are indeed prone to using coins immediately after clearing a gate level, precisely when they have the option to bypass a gate using coins.

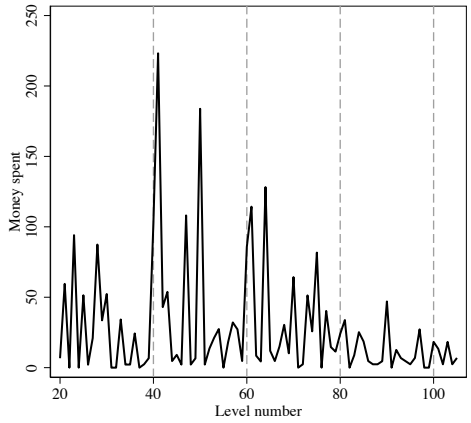

Figure 4: Total money spent by level. *Total amount of money (multiplied by an unannounced factor to disguise firm's revenues) collected in each level of the game. Vertical dashed lines mark "gate" levels. The pattern shows clear peaks right after a user has passed a "gate" level (e.g., levels 40, 60), indicating that the game obtains meaningful revenue from monetizing the gates. Note that this analysis does not control for the number of users playing each level.*

To examine the correlation between game progress and monetization, we zero in on instances when users are halted at a gate. **Figure 5** displays the average number of newly completed levels (referred to as progress) on the day a gate is reached, segmented by whether users spend money or coins at that gate. Across both scenarios, users who achieve more progress on a given day are more inclined to spend money or coins to bypass a gate without the five-day wait. Consequently, we deduce that a user's recent progress impacts their willingness to spend money or coins at a gate. Specifically, users with substantial progress during the day have a heightened desire to advance to higher levels and are, thus, more willing to pay for it. This implies that facilitating user progress, perhaps by easing game difficulty, can not only reduce their chances of stopping at a gate but also increase their propensity to pay to move past it.

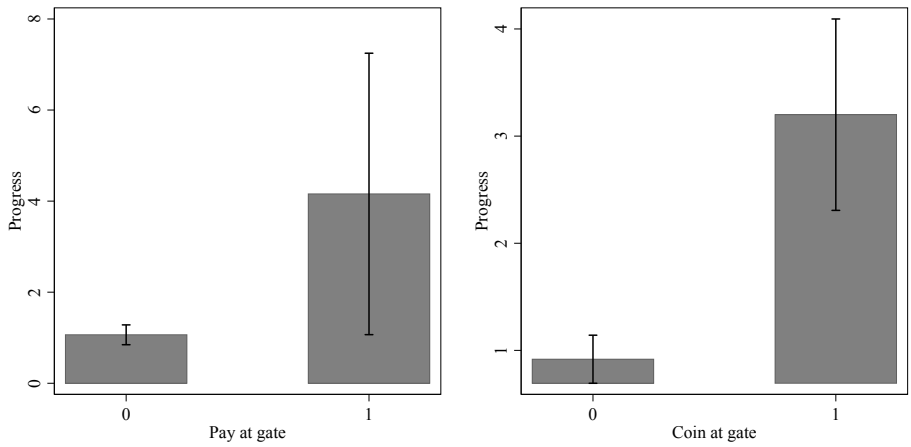

Figure 5: Monetization in gates and progress made*. Progress is measured as the number of new levels that a user has passed on that playday. The progress (number of new levels played) is significantly higher for users who paid at the gate relative to users who did not (left-figure; t=–3.82, p-val<0.001) and for users who used coins at the gate relative to users who did not (right-figure; t=–5.63, p-val<0.001). Only users stopped at a gate are considered in these figures.*

3.2.3 The role of game difficulty

Another factor that directly impacts how much progress users can make in the game is the difficulty of each round. The total difficulty of a round arises from two components: (1) the *goal* of the level played, which captures the specific requirement to pass each level (for example, some levels necessitate players to reach a particular score, while others demand a certain number of piece combinations within a fixed timeframe), and (2) the *allocation of pieces* on the game board. Whereas the goal of each level is pre-specified by the game and is the exact same for all users, the allocation of pieces has a stochastic component whereby the co-occurrence of same-colored pieces as well as the number of "special" pieces, is allowed to vary across users and across time. As mentioned previously, a "move" in this game entails matching three or more pieces of the same color. When the game design clusters more pieces of identical colors together, accomplishing the level's goal becomes more straightforward. Similarly, there are "special chips" that empower users to clear multiple pieces simultaneously, thereby elevating the score and increasing the likelihood of completing the level. Even though the specific placement of

each piece is randomized — with pieces getting shuffled every round — the probability of encountering two or more pieces of the same hue remains steady across all attempts at the same level. The same consistency applies to the appearance of "special chips."

The game design encompasses a broad spectrum of difficulty levels. Even though there's a general trend of increasing difficulty as players advance, the challenge can fluctuate significantly between consecutive levels. Some levels are effortlessly cleared, typically within two to three tries, while others present a steeper challenge, often demanding 20 or more attempts for success. Notably, numerous blogs and forums exist, centering their discussion on infamously challenging levels in popular games. Aligning with the observed relationship between game progress and user behavior, we discern a strong correlation between a level's difficulty and the retention and monetization of users at that specific level. We define a level's difficulty based on the average number of attempts users require to surpass it. Subsequently, we determine the correlations between this difficulty metric and both monetization and churn metrics. To ensure a uniform difficulty metric across levels — one that accounts for survival bias (meaning weaker players might abandon the game sooner, resulting in a skewed representation in the initial levels) — we assess game difficult considering only users who have reached level 100.

	P(coin)	P(money)	% users who churn	P(churn)
Correlation with difficulty	**0.302**	**0.553**	**0.880**	**0.394**
P-value	(0.000)	(0.002)	(0.000)	(0.012)
# obs (game levels)	100	100	100	100

Table 1: Correlation between level difficulty and user behavior*. % users who churn is the proportion of users who ever reached a particular level but did not make it to the next (during our observation window of at least 30 days). P(churn) is the propensity to churn in each round played. We include levels 1-100 as they have a large number of observations. Results are robust to exclude gate levels, which would avoid the possible confound with the use of coins/money for passing the gate.)*

Table 1 illustrates that the likelihood of a user using coins or spending money during a round at a specific level is positively correlated with the difficulty of that level, with correlation

values of 0.302 and 0.553, respectively. Difficulty also has a pronounced correlation with customer churn, which we evaluate in two ways. Firstly, we calculate the percentage of customers who reach a level but then churn (refraining from playing in the subsequent 30 days). The correlation between this metric and level difficulty is notably high at 0.880. However, this metric can be intertwined with game difficulty since users engage in numerous rounds at more challenging levels. As a result, we also determine the proportion of rounds at a given level that ended up being a user's "final round" (with "final round" being characterized by the user not playing again for a minimum of 30 days). We observe a positive correlation of 0.394 between the likelihood of churning during a specific round and the level's difficulty, even after accounting for the number of rounds a player engaged in during that concluding level.

Our model-free evidence demonstrates that game design elements like gates and elevated game difficulty can boost immediate monetary expenditures, but at the same time, they may reduce game engagement and escalate churn rates. We also show that the connection between game difficulty and churn is, to some extent, tied to the player's capacity to progress. When players encounter roadblocks in their advancement (due to either increased difficulty or gates), they are more prone to disengage. Our preliminary analyses further indicate that even in the short run, reducing game difficulty might bolster game monetization at gates. This is because players are more inclined to make payments at gates if they have experienced substantial progress before encountering them.

In the subsequent sections of this paper, we primarily focus on the role of game difficulty—more so than the role of gates. This emphasis is due to difficulty's direct connection to game progression. Additionally, game difficulty is a unique feature that can be tailored to individual players without disrupting the collective experience for the broader player community.

The level of difficulty can also be adjusted for an individual player across different play sessions. This adaptability empowers firms to intervene dynamically, either maximizing impact when anticipated to be most effective or averting looming player drop-offs. Consequently, difficulty becomes a powerful tool, facilitating the external control of a player's progression within the game.

3.3 Field experiment

Building upon the model-free analyses, we now aim to empirically and causally investigate the interplay of the impact of game difficulty and game progression on trade-offs between retention and monetization in freemium environments. We are particularly interested in understanding the short- and long-term impact of product design on customer retention and monetization. Such insights could offer firms direction in balancing customer retention and monetization through strategic freemium product or service design. To this end, we utilize a field experiment that manipulates a user's capacity for game progression. Specifically, we collaborated with a company that independently ran a field experiment that adjusted the difficulty each user encountered, thereby influencing their game progression potential.

Notably, this difficulty adjustment used in the experiment is user-centric, keeping other game aspects consistent. Specifically, players interact with the same game levels, in their intended sequence, but certain users enjoy a more favorable piece allocation on the board. As detailed in Section 3.2.2, the game's initial design ensured pieces were randomly positioned with consistent color co-occurrence probabilities for users at an equivalent game level. However, our partner company later introduced technology allowing user-specific, day-to-day difficulty modulation (or dynamic adjustment). Essentially, each day's onset saw a subset of users allocated a modified difficulty level, making it likelier to find adjacent pieces of the same color

or obtain "special chip" than usual. This technological adoption aimed at curbing churn among seasoned players.[3] The company's policy entailed reducing difficulty (increasing the chances of color co-occurrence by 20% and special chips by 10%) for users who played under 20 rounds in the previous week. Fewer rounds in the past week led to further difficulty decreases. If a user played any between zero to five rounds in the preceding week, difficulty was reduced but increasing color co-occurrence chances up by 100% and special chip acquisition by 45%. More details on manipulation intensity and game visuals are in Appendix A1.

Significantly, this policy was trialed in a controlled randomized experiment (A/B test), with half the eligible players unaffected during the 50-day period. From the start of the experiment, qualifying players (those past level 20 and with less than 20 rounds in the past week) were randomly assigned to the modified difficulty group (41.8% of users become treatment), with others retaining default difficulty (control). This arrangement presents an advantageous scenario, where similar players face varying difficulty levels while attempting the same game level.

It is worth noting that randomization was induced at the user level. Once assigned to a group, users remained there for the experiment's duration (from June 11th to August 3rd). However, placement in the modified difficulty group did not guarantee consistently lowered difficulty. Instead, every day at midnight (00:00 UTC), a user's past week play was assessed, and those in the modified group received adjusted difficulty based on their activity in the last seven days. This difficulty persisted for 24 hours. This experiment structure significantly influences our analysis approach.

Like other longitudinal studies employing user-level randomization, observations from days subsequent to the initial intervention must be handled carefully since game difficulty in

[3] For the focal company, a "seasoned" player was a player who had already passed level 20, thus showing a certain level of commitment for the game.

these periods may not mirror the control group. While we can measure the treatment's long-term causal effect by contrasting the overall *cumulative* behaviors across experimental groups, we should be wary when ascertaining game progress's causal effect on retention and monetization at any specific point in time beyond the initial round. However, comparing cumulative behavior up to a certain point in time in the future between the treatment and control group is causally valid.

3.4 Data

Our dataset encompasses the complete gameplay history for the 330,000 players who were eligible for the experiment. To be eligible, players needed to have surpassed level 20 in the game, and at some point between June 11th and August 3rd (period in which the experiment was running), they had to show low levels of engagement, defined as having played fewer than 20 rounds within the preceding week. Only players who downloaded the game between May 1st and July 3rd were considered, ensuring that each user is observed from their initial round and for at least a month after they started playing the game. During this period, these 330,000 users embarked on a cumulative 79 million rounds, from which we observe:

- Level played: Denoting the Level the customer is playing on that round
- Outcome of the round:
    - `Win`: did the player pass the round?
    - `Stars`: # stars (from 0 to 3) obtained in the round
    - `Points`: numerical score obtained in the round
    - `Combination size`: maximum number of pieces that were connected to form a snake.
    - `Moves`: number of cell combinations formed in total
    - `Time`: # seconds that the round lasted
- Other behaviors of interest:
    - `Coins`: did the player use any coins before finishing the round?

- `Extras`: did the player use any extra (to make the board more favorable) during this round?
- `Purchase`: did the player spend real money to buy game currency before finishing the round?

Table 2 shows the descriptive statistics for these per-round attributes.

Table 2: Descriptive statistics of round-level data. *N = 79,030,000 rounds.*

	Mean	SD	p5	p25	p50	p75	p95
Level played this round	34.0	22.8	7.0	17.0	29.0	43.0	77.0
Win	0.39	0.49	0	0	0	1	1
Stars	1.10	0.79	0	1	1	1	3
Points	35,006	19,280	11,650	22,450	31,400	43,050	69,900
Combination size	5.14	1.18	4	4	5	6	7
Moves	15.63	6.94	6	10	16	19	28
Time	121.1	71.5	47	76	106	146	248
Coins	0.007	0.081	0	0	0	0	0
Extras	0.028	0.204	0	0	0	0	0
Purchase	0.001	0.033	0	0	0	0	0

We also create a set of metrics capturing user heterogeneity in terms of their skills and level of engagement prior to treatment. Those metrics are based on the users' activity prior to passing level 20—as all qualifying users were required to have passed that level. We define `Level20` variables as the number of days/rounds played before level 20, total number of stars and coins collected before level 20, and whether the user had used coins/extras before level 20. We also use the round-level panel data to capture user-level variables that change over time and that will be relevant for the analysis. These include `Age` variables (e.g., maximum level achieved, tenure with game), `RFM` variables (e.g., amount of play, days since last play), `Stuck` metrics (e.g., # rounds in the current level, proportion of wins in the last day), and `Skill` variables (e.g., average # rounds per level, average # stars per level). **Table 3** shows the summary statistics for the most relevant variables (see Appendix A2 for the full list of variables).

		Mean	SD	p5	p25	p50	p75	p95
Level20	# rounds	45.83	36.91	23	29	36	49	99
	Did use coins	0.513	0.5	0	0	1	1	1
Age	# rounds	189.6	217.7	35	62	113	225	615
	Max level achieved	37.74	16.97	20	24	39	40	72
RFM	# days since last play	13.88	17.9	1	3	7	16	54
	# rounds last week	5.728	6.68	0	0	2	12	18
Stuck	# rounds in this level	26.72	62.45	0	2	7	25	115
	# playdays in this level	3.054	3.804	0	1	2	4	10
	Prop. wins yesterday	0.308	0.341	0	0	0	1	1
Skill	Avg. # rounds/level	4.410	3.653	2	2	3	5	11
	Avg. # stars/level	1.988	0.257	2	2	2	2	2

Table 3: Descriptive statistics for the user-level variables*. Level20 variables were measured at the moment the customer passed level 20. The rest of the variables are dynamic (i.e., values change over time) and are summarized using the values on the day each user qualified for the experiment. N = 329,999 users.*

Users exhibit substantial heterogeneity in their gameplay patterns. For instance, on average, users necessitate 45.8 rounds to advance beyond level 20, though this exhibits a standard deviation of 36.9. Additionally, 51.3% of users deploy some coins prior to reaching level 20. The gameplay level at the point of intervention also displays marked variation; users, on average, have completed 189.6 rounds by the time they are subjected to treatment or assigned to the control condition. The highest level attained spans from level 20 (at the 5th percentile) to level 72 (at the 95th percentile). There is a discernible variation in the degree to which players are entrenched within a level — the number of rounds engaged in their current level can range from a mere 0 to an extensive 115. Moreover, there are differences in their gameplay proficiency — the mean rounds per level is 4.4, but this can range from a minimum of 2 to a maximum of 11. Such variation offers a valuable framework to investigate the potential effects of moderating game difficulty across diverse user segments.

4. ANALYSES AND RESULTS

4.1 Randomization and manipulation tests

We first confirm that the randomization was well executed by comparing the distributions of the user-level variables at the moment of the intervention. We verify that there are no systematic differences across conditions (see Appendix A2 for the full set of results). Second, we corroborate that the intervention caused the intended effect of making the game easier for treated users. To do so, we look at four outcomes that are directly affected by the difficulty of the round namely `Win, Stars, Points,` and `CombinationSize` at the observations (rounds) of treatment and control players in the first day of the intervention. Specifically, we run a linear model for each outcome against a treatment dummy and cluster the standard errors at the user level (unit of randomization). See **Table 4** for the results. All outcomes show a substantial and (statistically significant) positive change for users in the treatment condition; users have a higher chance to win, collect more stars, get more points, and are able to create longer snakes. For example, users in the treatment condition on average earn 7,382 points and 0.297 star more than those in the control condition. All these outcomes were expected as treated customers faced more favorable board allocations.

	Win	Stars	Points	Combination
Treatment	**0.141**	**0.297**	**7,382**	**0.172**
	(0.002)	(0.003)	(64.7)	(0.004)
Constant	0.368	1.048	35,412	5.157
	(0.001)	(0.002)	(36.0)	(0.002)
# obs	2,009,966	2,009,966	2,009,966	2,009,966

Table 4: Manipulation checks. *OLS of the round outcome against a treatment dummy using all rounds on the first day of the experiment. Standard errors (in parenthesis) are clustered at the user level. Bold numbers indicate that p-value<0.01.*

We also run separate regressions for each level of difficulty reduction in the first day of the intervention—recall that the intensity of difficulty adjustment changed by the amount of play in the previous week. As expected, the impact of the difficulty adjustment on these outcomes is stronger for users with lower playing levels in the seven days prior to treatment (more pronounced difficulty decrease). For example, looking at the effect of treatment on the

probability of winning and the number of stars collected in each round (Table 5), we observe that the magnitude of the effect increases monotonically as the difficulty adjustment increases (where level 1 is the lowest difficulty level).[4] We therefore conclude that the randomization was well executed and was successful at manipulating game difficulty as intended.

	Win (difficulty 5 = default)				Stars (difficulty 5 = default)			
	4	3	2	1	4	3	2	1
Treatment	**0.030**	**0.076**	**0.133**	**0.190**	**0.052**	**0.135**	**0.249**	**0.419**
	(0.004)	(0.004)	(0.004)	(0.002)	(0.005)	(0.006)	(0.006)	(0.003)
Constant	0.327	0.347	0.358	0.393	0.983	1.012	1.030	1.090
	(0.002)	(0.002)	(0.003)	(0.001)	(0.003)	(0.004)	(0.004)	(0.002)
# obs	370,137	302,836	281,827	1,055,166	370,137	302,836	281,827	1,055,166

Table 5: Manipulation checks by degree of difficulty. *OLS of Win and Stars against a treatment dummy using all rounds on the first day of the experiment. Standard errors are clustered at the user level. Treatment variable in bold indicates p-value<0.01.*

4.2 Evaluating the impact of the intervention on player behavior

We explore the impact of the intervention on customer behavior, focusing on the behavioral outcomes relevant to understanding retention and monetization in this context. For every user, we tabulate the cumulative metrics of playdays, rounds, money expenditure, and coin utilization during the 30 days after the intervention. As **Figure 6** shows, the intervention significantly changed customer behavior. As anticipated, gameplay increased among the treated cohort, with the magnitude of this increase becoming more pronounced over time. Specifically, within a month following the intervention (or assignment to the control condition), treated participants, on average, have engaged in an additional playday and ten more rounds compared to their counterparts in the control group. Additionally, they demonstrate accelerated progression in the game, attaining level 44 as opposed to level 42. For the focal gaming company, these effects are

[4] This pattern is consistent across outcomes (results for `Points` and `CombinationSize` are shown in Appendix A3). In the remainder of the analyses, we report the average treatment effect, i.e., across the four treatment intensities and present the results conducted by groups in Appendix A4.

of considerable significance, translating to an approximate 20% uptick in overall customer engagement.

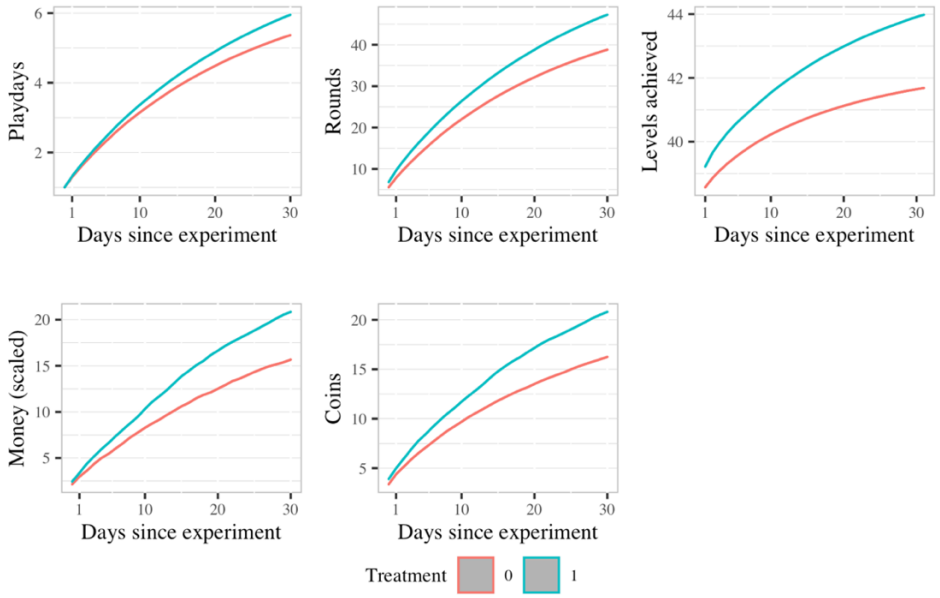

Figure 6: Cumulative behavior across experimental conditions. *These figures show the cumulative number of playdays, rounds, money spent (scaled by an unknown amount) and coins used in the game. Standard errors around the lines are reflected in the graphs but are not detectable due to their small magnitude.*

The intervention not only increases gameplay and engagement, but it also has a positive impact on monetization, with treated users spending more money and more coins than those in control (see last two charts in Figure 6). Not only does this finding go against the company's prior expectations (i.e., that customers would spend more money and coins when the game is more difficult), but it seems at odds with the preliminary evidence (Section 2) where we show that users tend to spend money and coins when progress is prevented (gates and difficult levels). This difference is already visible on the first day of the intervention and increases monotonically over time. Thus, in the tradeoff between offering more of the free product (easier game) and encouraging consumers to purchase in-app goods, we find that by providing an easier game, the firm increases, rather than decreases, long-term IAPs.

While the above cumulative analysis allows us to make causal claims about the overall increase in behavior due to treatment, it does not allow us to separate whether this increased behavior stems directly from an increase in the user's tendency to spend coins and money or because lower difficulty decreases churn or makes it less likely to run out of lives, and therefore allows the user to play more rounds and make more progress, having the chance to use more extras, coins However, from this aggregate and cumulative analysis, we cannot definitively conclude that a reduction in game difficulty—or the increased capacity to progress within the game— *directly* affects the user tendency to spend coins and money. This is because lower difficulty makes it less likely to run out of lives, and therefore allows the user to play more rounds and make more progress, etc. Furthermore, if the ability to make progress in the game increases retention and engagement — as we expect given the preliminary evidence — treated users will be more likely to play again after being treated than control users. As such, the increased use of money or coins might not be a direct consequence of ability to make progress but merely an artifact of treated users' extended gameplay.

Moreover, the experimental design was such that whether a treated user receives difficulty level 1, 2, 3, 4, or 5 depends on the number of rounds they have played in the previous seven days (recall that control users always get difficulty 5). Therefore, it is possible that treated users get different intensities of difficulty adjustment when they play the game after the first day in which they were treated.[5] This means that while the allocation to the treatment group is fully randomized, the degree to which the ability to make progress is manipulated, is not constant over time, and hence can be endogenous *after* the first day of the experiment.

[5] See Appendix A1 for an illustration of how difficulty adjustment intensity vary over time depending on the user amount of play.

To make causal statements about the direct impact of *progress* on user retention and monetization, we focus on the first day of play after treatment. Recall that treatment adjustments are only done at the end of each day, so the first day provides a clear causal measure.

4.2.1 Impact of treatment on user engagement and retention

Focusing *on the first day of the experiment*, we conduct intent-to-treat (ITT) analyses running linear regressions with the dependent variable being a behavior of interest and using treatment as a dummy variable:

$$y_i = \alpha + \beta\, Treatment_i + \epsilon_i \quad (1)$$

Utilizing this ITT approach ensures that the coefficient for the 'treatment' variable across all regressions offers an unbiased comparison between treated and control users. Note that the 'treatment' is binary and refers to 'whether the user was allocated to the dynamic difficulty condition;' it does not represent the exact level of difficulty faced in a particular round. We examine the following dependent variables (y_i):

- `#Rounds`, denoting the number of rounds played on the day of the experiment.
- `Progress made`, defined as the number of (new) levels passed on the day of the experiment.
- `Engagement`, calculated at the conclusion of each round and signifies whether a user opts to play another round within the subsequent 10 minutes, provided they have lives remaining (thus, we exclude observations when users deplete their lives).
- `Retention 1`, measuring if the user plays the following day.
- `Retention 7` measures if the user plays at least once in the subsequent 7 days.
- `Retention 14` measures if the user plays at least once in the subsequent 14 days.

Table 6 shows the parameters from the ITT analyses, with "Treatment" presenting the estimated difference between treated and control observations. We observe that treated users engage in more rounds and achieve greater progress than users for whom the difficulty level remained constant, with an average of 1.25 additional rounds played and 0.746 more levels attained. Admittedly, these metrics may be influenced to some extent by whether a player retains lives to continue their gameplay, which is more likely to have lives when users face easier levels during the first day of treatment. As a result, we investigate differences in "engagement", which only considers rounds in which a user has remaining lives. We find that users engage in more rounds (0.0213 on average) when the level of difficulty is low, offering compelling evidence that reducing the game's difficulty, thus facilitating user progress, markedly enhances engagement.

	# Rounds played	Progress made	Engagement	Retention 1	Retention 7	Retention 14
Treatment	**1.2470**	**0.7460**	**0.0213**	**0.0270**	**0.0246**	**0.0199**
	(0.0251)	(0.0088)	(0.0007)	(0.0016)	(0.0017)	(0.0016)
p-val	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*
Constant	5.5650	0.7840	0.8180	0.2950	0.6430	0.7430
	(0.0132)	(0.0039)	(0.0005)	(0.0011)	(0.0011)	(0.001)
# obs	329,999	329,999	1,867,849	326,472	316,257	308,003

Table 6: Impact of the intervention on engagement and retention on the day of the experiment. *The table presents the parameters from the intent-to-treat analyses in (1), with "Treatment" presenting the estimated difference between treated and control observations The number of observations for "Engagement" corresponds to the rounds in which a user has lives to continue playing. Standard errors are clustered at the user level (unit of randomization). All other regressions are at the individual level. The number of observations decreases for retention metrics because some users (both treatment and control) started their treatment shortly before our observation period ended, so we didn't have enough time to observe the retention outcome. Robust standard errors (in parenthesis) and p-values are reported. Treatment variable in bold indicates p-value<0.01.*

Similarly, the treatment significantly increases the proportion of users who play again the day after (Retention 1), 7 days later (Retention 7), and 14 days later (Retention 14).[6] As discussed

[6] We replicate the analysis using retention in the next 3 days, 28 days, and 45 days, obtaining similar results for all metrics, with diminishing effects as the retention window is longer. We also replicate the analysis including different set of control variables: (1) 'time-varying' factors related to the day in which the intervention took place, including day of the week, month, and holiday, (2) 'max level of play' at the moment of the intervention, and (3)

earlier, this significant increase in retention rates might have caused the positive impact on monetization rates (Figure 6). That is, treated users may have generated more expenditures in the long run not because greater progress directly increases monetization, but rather because progress delays churn, giving treated users more opportunities to spend money in the game.

4.2.2 Impact of treatment on user monetization

To disentangle the direct impact of the intervention on customer spending from its effect on increased retention, we assess the consumption of extras and IAPs at two distinct levels: *per-day* monetization and *per-engagement* monetization. The former evaluates the total number of rounds in which a user consumes extras, coins, and money on the first day—corresponding to the initial data points in Figure 6. The latter represents the propensity to utilize premium content within a specific round during the first day. Since this metric is conditional on a user's active participation in that round, any observed effect on monetization inherently accounts for variations in engagement or retention across users.[7] Essentially, this measure captures the *direct* influence of the treatment on the likelihood of spending within the game, excluding changes in expenditure attributed to heightened engagement. The outcomes of our regression analyses are presented in **Table 7**.

'degree of difficulty adjustment', which is determined by the activity during the 7 days prior to the intervention. Results are presented in Web Appendix A4.

[7] One might still worry about potential bias induced by the type of users who continue playing; we discuss that possibility in Appendix A4 and present a set of analyses demonstrating robustness of our results to that potential concern.

	Per-day monetization			*Per-engagement monetization*		
	# Rounds with extras	# Rounds with coins	# Rounds with money	Prob(Use extras)	Prob(Use coins)	Prob(Use money)
Treatment	**0.0075**	**0.0051**	**0.0012**	**–0.0025**	**–0.0004**	0.0000
	(0.0010)	(0.0007)	(0.0003)	(0.0003)	(0.0002)	(0.0001)
p-val	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0051*	*0.6340*
Constant	0.0922	0.0390	0.0057	0.0229	0.0078	0.0013
	(0.0007)	(0.0004)	(0.0002)	(0.0002)	(0.0001)	(0.0000)
# obs	329,999	329,999	329,999	2,009,966	2,009,966	2,009,966

Table 7: Impact of the intervention on monetization on the first day of the experiment*. For per-day metrics, the number of observations correspond to the number of users participating in the experiment. Robust standard errors are reported in parentheses. For per-engagement metrics, the number of observations corresponds to the rounds used for each regression. Standard errors (in parentheses) are clustered at the user level. Treatment variable in bold indicates that p-value<0.01.*

Notably, treated players utilize more extras, spend more coins, and invest more money on the first day of their intervention compared to control group participants. Yet, when we dive deeper and adjust for engagement/retention effects, a contrasting pattern emerges. Given that a user is actively playing a round, their probability of deploying extras or coins in that round is diminished under the treatment condition. This reduction in game difficulty's influence on the usage of extras and coins aligns with expectations; making the game easier *substitutes* for the need to use extras to ease the gameplay (enhancing the user's ability to combine board pieces).

Thus, our results demonstrate that even though easing game difficulty substitutes for the need for monetary investment in any particular round played, contrary to expectations, even in the short run (on the first day of intervention), the overall effect of reducing game difficulty on monetization becomes positive due to increased engagement and retention. This challenges the firm's strategy of increasing difficulty to spur short-term spending, though the likelihood of using real money is not significantly influenced by game difficulty. While these variations might seem minor, they bear significant implications for the firm, especially since such behaviors are infrequent in the actual gaming context.

To gain deeper insights into monetization behavior, we further dissect per-engagement expenditures on the first day post-treatment—both coins and money—distinguishing between expenditures on extras (to surpass levels) and on gates (to bypass them). Results are presented in Table 8.[8] Another interesting finding emerges, even though *reducing* the *difficulty* decreases using coins and money for extras to pass the level, it significantly *increases* the *propensity to expend* coins and actual currency *on gates*. It is important to emphasize that the spending coin and money to pass gates are contingent upon users being halted at a gate—meaning this effect is not merely because users more frequently encounter gates. Hence, this treatment effect solely corresponds to an elevated likelihood of spending coins and real money once faced with a gate. The progress that customers made due to the easier game experience in the treatment condition enticed customers to spend coins and money once they got to non-difficulty-related obstacle (gate). This finding is particularly intriguing as it underscores the paradox that offering more of the free game component—anticipated, at face value, to diminish purchases of premium features—unexpectedly propels users to spend more.

[8] To ensure the robustness of our results, we replicated the results from Tables 7 and 8 by adding controls (including day-level controls and past levels of activity on the intervention) as to the regressions as well as using subsets of data. See results in Appendix A4.

	P(Coin extra)	P(Coin gate)	P(Money extra)	P(Money gate)
Treatment	–**0.00059**	**0.00082**	–0.00007	**0.00019**
	(0.00012)	(0.0002)	(0.00005)	(0.00008)
p-val	*0.00000*	*0.00003*	*0.13000*	*0.02540*
Constant	0.00449	0.00548	0.00069	0.00105
	(0.00449)	(0.00548)	(0.00069)	(0.00105)
# obs	2,009,966	652,574	2,009,966	652,574

Table 8: Impact of the intervention on the type of expenditure. *OLS of the behavior of interest against a treatment dummy. Standard errors (in parentheses) are clustered at the user level. The number of observations corresponds to the rounds used for each regression on the first day post treatment. The gate-related outcomes are conditioned on the user being at a gate Treatment variable in bold indicates that p-value<0.03.*

To summarize, while much of the literature on freemium design has centered on the notion that "enhancing" the free product boosts customer engagement at the cost of diminishing monetization, our findings challenge this in the context of F2P games. Specifically, for users at the risk of churning—which accounts for roughly 50% of users of our focal firm at any given moment—the tradeoff does not materialize. Instead, we observe a beneficial synergy between enhancing the free product and monetization. Modifying the product design to offer an enhanced (easier) game not only bolsters user engagement and retention but also amplifies the likelihood that users will use premium features, both immediately (on the initial day of intervention) and in the longer run (Figure 6).

4.3 Heterogeneity in treatment effects

Our results provide actionable insights to firms as to how personalization in the product design can be leveraged to manage customer retention efforts as well as to increase the revenue of each customer. Specifically, the firm can use these insights to identify players who will be most likely to increase revenue — from advertising, premium purchases, or both — via a temporal difficulty reduction. Arguably, there is no strong need to leverage the heterogeneity in treatment effect in this case because, as it was shown in Section 4.2, the overall impact of the intervention was

overwhelmingly positive. Nevertheless, we believe that this exercise is still of interest for two main reasons. First, investigating heterogeneity in treatment effects can further shed light on the account that progress is the underlying mechanism for the observed effect of game difficulty on monetization. If individuals with a higher need or sensitivity for progress exhibit stronger treatment effects. Second, given the ease of product design personalization in online games, as demonstrated by the experimental design of our study, an analysis that identifies individuals with greater benefit from game personalization, illustrates the more general potential value of personalization in the product design, which has been understudied in the marketing literature.

Building on the analysis from Section 4.2, we anticipate that users who tend to seek more progress or situations wherein users are inclined to seek progress will display the most pronounced treatment effects. By examining user behaviors *prior* to the intervention, we identify specific players (termed "target" users) whose long-term value is most likely to increase following a game difficulty reduction. We subsequently assess the intervention's impact on the future behaviors of these identified players, comparing it against similarly profiled individuals in the control group.

4.3.1 Identifying targets: Heterogeneity in treatment effect

In this section, we assess how the treatment effect of making the game easier varied between users. Such an analysis can help the firm identify potential targets for the intervention. We identify potential targets in two different ways. First, we identify "types" of users who would be more responsive to the game difficulty reduction treatment. We do so by capturing intrinsic differences across users that are of substantive importance across games. We select progress and spending as two key behavioral dimensions in games from a marketing perspective. Progress is a unifying characteristic of games. Progress and skill have been shown to be crucial inputs for game engagement and monetization, sometimes facilitated by social elements such as competition and

collaboration (Lopez-Vargas, Runge, and Zhang 2022). Making progress in and becoming skillful at the game are further economically meaningful as they are likely to increase switching cost (Klemperer 1987; Caminal and Matutes 1990), making it less likely for players to turn to other activities.

Second, we identify circumstances when it might be a good moment to intervene. Even for users who might not generally benefit from reduction in difficulty, there might be circumstances (e.g., when they are increasingly frustrated because of lack of progress) that make them more responsive to the intervention. Conceptually, the moment of frustration can be framed as a moment of satiation with the app (Appel et al. 2020) and relatively lower switching cost, making the player more at-risk of disengaging (Caminal and Matutes 1990; Hartmann and Viard 2008). Similarly, users who strive for a more distant goal (e.g., are far from a gate), may have lower attainment expectations and experienced self-efficacy (Manderlink and Harackiewicz 1984; Latham and Seijts 1999), which may lead to lower switching cost as they are further from obtaining a perceived reward (Hartmann and Viard 2008). It is worthwhile to note that, given our experimental design, we cannot pin down the exact moment when the company should intervene, but rather investigate if general moments, such as when user is more frustrated, are characterized by a stronger treatment effect.

For the first type of metrics, the "*who*", we rely on the behavior up to level 20. This is in the spirit of Padilla and Ascarza (2021) who use the first transaction of users to identify customers with higher responsiveness to marketing interventions. We use the behavior up to level 20, in part, because every user in our sample reached that level and therefore everyone is observed up to that point. But most importantly, because it allows for apples-to-apples comparison of the degree of complexity and difficulty of the game across users. If, on the contrary, we used the full history or only the most recent history of each player, users would be playing at very different levels of the game, and therefore the behavior observed would not necessarily reflect intrinsic differences (i.e., customer heterogeneity) but rather differences in the characteristics of the levels being played.

For the second type of metrics, the "*when*", we use the most recent behavior prior to the moment of first intervention. That way, we can identify characteristics of the current play (e.g., related to the progress made or the need to spend money at a certain moment) that are associated with a stronger sensitivity to the intervention *at that moment*. This type of information is of great value to developers in the context of gaming and other online/mobile services as they can not only modify the product characteristics individually, but also alter the product features dynamically to better match users' variable needs as they evolve in their use of the product/service.

We consider the following types of customers (who) and situations (when):

- Players who seem to enjoy (or seek) progress more than the average player. We expect the intervention will be more impactful to customers who tend to seek progress in the game. We create two variables that serve as a proxy for progress:

  Who:Early_Progress is determined by the number of days a user played before reaching level 20. The idea is that users keen on progression will advance quickly, especially in initial levels with lesser difficulty. While this may also reflect a player's skill (since skilled players might reach level 20 faster), we introduce another metric to address this.

  Who:Progress_Prone captures a user's inclination to continue playing based on their progress. Using rounds played prior to level 20, we calculate the frequency a user plays after passing a level versus stopping after achieving it.[9] Users who value progress are more likely to play again upon level completion.
- *Who:Spender*: Players more inclined to spend money in the game. At first, it may seem counterproductive to ease the game for spenders as it might reduce sales. However, because

[9] By definition right after passing a level the user has at least one life and hence is eligible to continue playing. Continue playing equals 1 when the user plays another round within 10 minutes of ending the current round.

of the dual impact on retention, these users are likely to increase spending provided that the intervention extends their lifetime. While we do not anticipate these players to engage more than the typical user, the sheer increase in their engagement and retention will elevate their spending compared to others. We measure this using a binary variable indicating if a user made any purchase before reaching level 20.

- *When:Frustrated*: Players who, at the time of first intervention, might be more frustrated than usual, due to lack of progress in recent rounds. We posit that these players need progress more than usual, and therefore their reaction to the intervention might be stronger. We operationalize this variable using the number of rounds that a user has attempted to pass (unsuccessfully) the current level (which is the level they were at when first receiving the lower difficulty treatment).
- *When:Distance-to-gate*: Players currently at a point distant from the next gate are likely to enjoy more progress before encountering a gate. Our findings indicate that users are more inclined to purchase complementary products when they have recently progressed. We hypothesize that the intervention's impact on complementary product purchases is amplified when users are further from a gate. This allows them to experience more progress before hitting the gate. Since these players can enjoy more uninterrupted progress, they are also likely to engage more in terms of rounds played and days played. We measure this using the count of levels remaining until the upcoming gate.

4.3.2 Impact of the intervention on customer targets

We assess the long-term implications of the intervention for these customer groups by examining the variation in treatment effects across users. This is achieved through a set of linear regressions (extension of the model used in previous analyses) where the dependent variable represents the

outcome of interest 30 days post-experiment. As independent variables we include the treatment variable, *all* five moderator variables along with their interaction effects with treatment, and all possible combinations of 3-way interactions among the moderators.[10] The model also controls for the amount of gameplay during the 7 days preceding the experiment, which captures the degree of difficulty reduction players would have encountered on the experiment's initial day. This control ensures that observed differences between groups are not due to potential correlations between heterogeneity variables and the degree of difficulty adjustment (we standardize these variables to ensure coefficient magnitude comparability). Lastly, for each regression, we calculate the ratio of the interaction term to the treatment effect, providing a metric of "the extent to which the treatment effect amplifies for each customer type" (as depicted in Figure 7). Comprehensive regression outcomes can be found in Table A12 and Table A13 of Appendix A5.

Consistent with the discussion earlier, we find that *early_progress* users (one standard deviation above the mean on that metric) exhibit treatment effects that are approximately 15-17% stronger than those of the average customer in terms of retention and engagement, and between 40% and 60% stronger in terms of monetization outcomes. The interaction is marginally weaker, albeit still meaningful and significant, when progress is characterized as *progress_prone*. The group of *spenders* reveals a distinct pattern: Predictably, these users do not exhibit enhanced effects in retention metrics (for instance, comparable levels of playing days, rounds, and progress). However, due to their inclination to spend, they display a significant and strong impact on monetization outcomes.

[10] Results are robust to running excluding the three-way interactions and to running separate regressions for each moderator. We favor the joint model with all interactions to capture possible correlations as well as interactions effects among moderators.

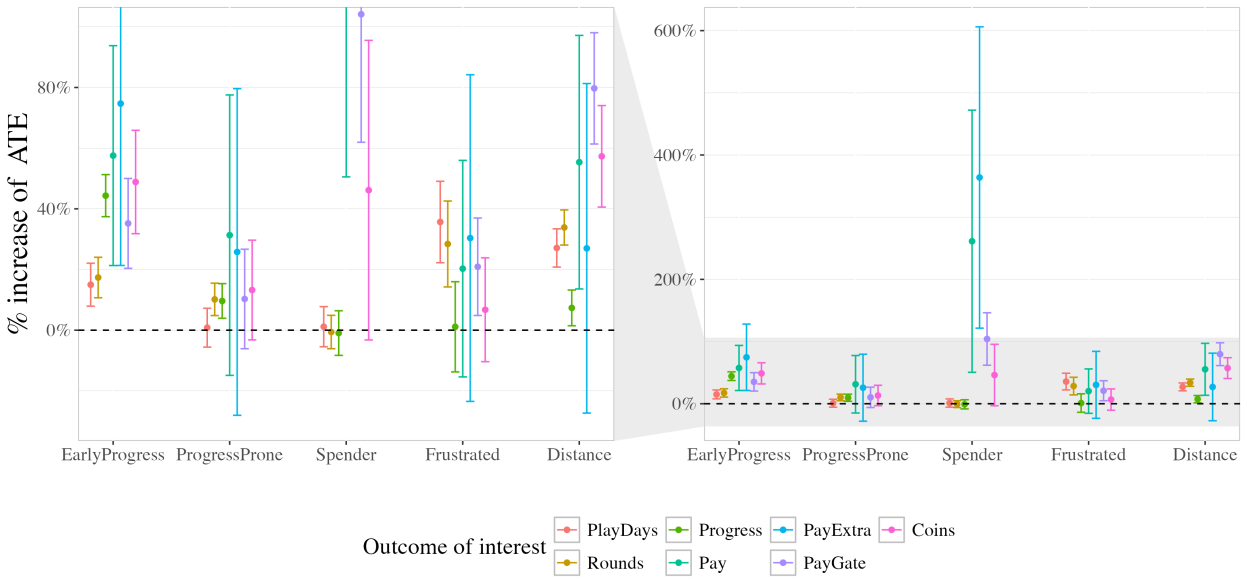

Figure 7: Increase in treatment effect for selected group of customers. *This figure shows the percentage increase in the average treatment effect (ATE) of the intervention (on each outcome) for a standard deviation change in the heterogeneity variable. For example, the first (red) whisker point means the intervention's impact increases by 15% (in playdays over the next 30 days) when "early_progress" is one standard deviation above average. The left panel is a zoomed-in view of the right panel, making differences easier to see.*

In contrast, *frustrated* users exhibit the reverse pattern. They react more favorably to the intervention in terms of engagement (evident in the increased number of rounds, playdays, and progress), but this heightened engagement does not yield more potent effects on monetization outcomes (with the exception of 'paying at gates'). Additionally, the variable *distance_to_gate* not only moderates the intervention's effect on engagement outcomes (with a 27% increase in playdays and a 34% rise in rounds) but also exerts an even stronger influence on monetization, leading to a 55% surge in purchases and a 57% growth in coin usage. These findings underscore a secondary impact of retention: When users are further from a gate, they derive greater benefits from progress, resulting in heightened engagement. Simultaneously, the additional progress they achieve, enhanced by the intervention, spurs them to expend more in their pursuit of continued

progress. Thus, game designers should pay particular attention to players who are still further away from gates, because they can benefit more from game alterations that enhance game progress.

The three-way interactions (presented in **Table A13** of Web Appendix 5) provide a more nuanced understanding of these moderation effects. For example, when we delve deeper into the moderation effect of "early progress," we observe that exhibiting progress along two dimensions further intensifies the treatment effect as users who are identified as both “early progress” and "progress prone" exhibit stronger effects across all behaviors. Likewise, the moderation effect of "early progress" becomes more significant among frustrated customers, especially concerning monetary outcomes like spending and coin usage, suggesting that relieving frustration among early progress users is particularly valuable. Moreover, the effect of users who are prone to spending on the effect of reduced game difficulty on spending behavior is particularly pronounced for users who are far from gates, possibly suggesting a useful timing consideration for the firm in targeting the intervention. On the other hand, frustration negatively moderates the effect of user spending on spending behavior.

Overall, our targeting analysis aligns with the notion that the primary mechanism underlying the effects we observe is game progress. We demonstrate that groups of players, who are theoretically most likely to be affected by the treatment due to a need for progress or a tendency to spend, indeed demonstrate the strongest treatment effects. From a managerial perspective, we highlight that certain players, at specific times, show a much stronger response (dozens, and sometimes hundreds, of percent stronger) to reductions in game difficulty. This underscores the opportunity for personalizing product design in freemium settings, where IAPs are a significant source of revenue.

5. DISCUSSION

In this study, we investigate the effects of dynamic difficulty adjustment on user retention and monetization leveraging a field experiment conducted by a popular F2P mobile game during the rollout of such a system. Results highlight the intertwined dynamics of customer engagement, retention and monetization in such settings. As expected, giving customers an easier game significantly decreases purchases in the specific round played. However, lowering the game difficulty not only increases short-term play and subsequent retention, but also increases customer spending on premium services both in the short and in the longer run. We explore this synergistic relationship between customer retention and monetization within the framework of F2P games in greater depth. Enhancing progress in the free segment of the game for users at risk of churn does not merely enhance engagement and retention. It also heightens game monetization via IAP, given that retained users encounter more opportunities to expand. Financially, this translated to an additional revenue of $0.07 per user, with approximately 82% of this surplus stemming from the monetization of premium services.

Whereas we find overall highly positive effects from the DDA system, we observe considerable heterogeneity in the outcomes. Those customers with a predisposition towards game progression manifested more pronounced effects. Notably, users who had previously made purchases within the game showed the most significant impact on IAP. Surprisingly, providing these “spender” users with added free benefits (e.g., simplifying gameplay) led to a surge in their spending, exceeding twice the uptick observed in the average user. Furthermore, during moments of frustration or when players were distanced from gates (in-game barriers), the impact of decreased difficulty was magnified. The nuances discerned from our heterogeneity in treatment

effect analyses corroborate that the observed patterns are predominantly driven by consumers' motivation for game progression.

In the discussion, we want to shed light on potential impacts of the DDA system on ad monetization and discuss limitations and avenues to generalize findings to non-gaming settings.

5.1. Impact on IAP and ad monetization

We have illustrated the impact of reduced difficulty on engagement, retention, and monetization. A comprehensive approach to understanding these effects is to convert them into monetary values, representing revenue for the firm. While the proportion of players who spend money on IAP tends to be relatively small (around 4%) in our data, their contribution to overall revenue is often high. IAP have been estimated to account for nearly 50% of all mobile app earnings and more than three times the revenue from advertising.[11] In this section, we quantify the net monetary impact of the intervention and decompose it into additional revenue derived from both the free (i.e., advertising) and premium (i.e., additional purchases) components of the service. We further contrast the effect decomposition of the average player with that of the "target" users.

Given that we do not have access to the advertising revenue data from the focal company, and the precise monetization rates (from premium services) were obscured for confidentiality reasons, we rely on average industry figures and reports from the focal company for this analysis. Specifically, we assume that the revenue for an ad exposure is 1.40¢ per round, and that the average expenditure, once a purchase is made, stands at $4.21 per transaction.[12]

[11] https://www.businessofapps.com/guide/in-app-purchases/#:~:text=In%2Dapp%20purchases%20account%20for,worldwide%20on%20in%2Dapp%20purchases.

[12] We corroborate with members of the focal firm that these figures are representative of their business. The average expenditure per paying round is obtained from our data and the advertising revenue can be thought as an average of $7 per thousand exposures (CPM) and a single ad being shown every 5 rounds.

We combine the monetization estimates with the gameplay data and our treatment effect projections to calculate both the net and disaggregated impacts of the treatment, expressed in USD. Specifically, given our observations of the rounds played and purchases made by each player over the 30 days post-experiment, we multiply the ad revenue per round by the increase in the number of rounds played. Similarly, we multiply the average purchase amount by the uplift in the number of purchases attributable to the intervention. The findings are depicted in Figure 8.

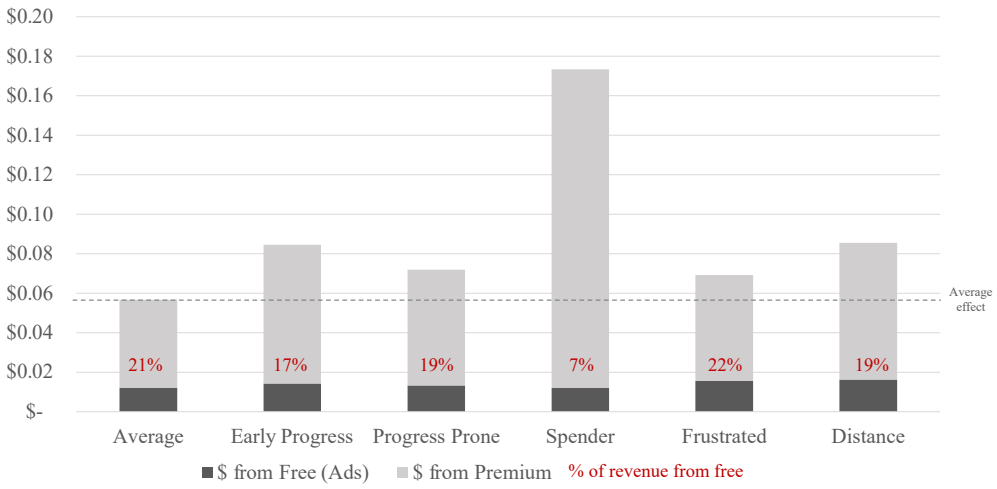

Figure 8: Decomposition in treatment effect for the average customer and for selected groups of customers. *This figure shows the additional revenue generated by the intervention. The numbers (in red) correspond to the proportion of additional revenue that comes from the advertising side. To be read: for an average customer, the intervention increased total revenue by $0.07. 18% of this revenue increase comes from advertising. The heterogeneity user groups reflect one standard deviation above the mean for that heterogeneity metric.*

From these back-of-the-envelope calculations, the intervention added a net revenue of $0.08 per customer within 30 days. Considering the millions of customers eligible for this intervention within the firm, this increase is of substantial financial magnitude. Of this revenue, 21% comes from additional revenue from ads served to engaged customers. Conversely, 79% comes from their added purchases. The predominant treatment effect is from in-app spending. This might seem

counterintuitive since the intervention made the game easier, which could have reduced IAP—especially as players often buy in-app extras to overcome challenges. Yet, consistent with our findings, the main revenue uplift from retention is due to increased purchases.

Comparing these figures across "target" groups, we can further "monetize" the heterogeneity in treatment effect analysis. The group that offers the least benefit with respect to the average user is the frustrated type, whereas "spenders" generate more than twice the additional revenue. For those users, 94% of the extra revenue is coming from the purchase of premium items. While these "back-of-the-envelope" calculations are only approximative in that they are based on general industry revenue figures, we believe that this analysis highlights the importance of incorporating the impact of retention on short- and long-term monetization. Furthermore, it also allows us to combine together the differential effects of the game design to better understand differences across customers.

5.2. Avenues to generalize beyond gaming

Substantively, we introduce DDA as a key lever in the marketing of mobile games (Xue et al. 2017; Zohaib 2018; Huang, Jasin, and Manchanda 2019), and as a lever that relates to the product aspect of the marketing mix (Appel et al. 2020). As the system was evaluated in a randomized control trial, we were able to document the causal, incremental impact of such a system on player behavior.

DDA can be thought of as reducing satiation with a gaming app (Appel et al. 2020) by infusing a burst of progress in players' experience through easier gameplay. This easier gameplay is timed to set on when players exhibit behavior consistent with satiation, i.e., when they display relatively lower use of the app. In this way, DDA counters hedonic decline, i.e., the fact that

repeated exposure to a stimulus can reduce the hedonic response such as enjoyment (Galak and Redden 2018). It automatically and intrinsically changes the stimulus for users with lower enjoyment as induced from the observation that they play relatively less.

Appel et al. (2020, p. 106) provide recommendations on how such effects could be achieved beyond games; e.g., by encouraging users to change their consumption rates, imagining future changes and variety in experience, or managing the similarity of experiences. Firms, thereby, need to carefully manage experiences such that they are neither too similar nor too dissimilar (Lasaleta and Redden 2018). Based on the reported results, it appears that the DDA system studied in this paper achieves a good and appropriately targeted balance of variety and similarity.

Levers other than difficulty available to app publishers to modulate similarity and variety of experiences can be the amount and type of free content; various informational popups (Galak, Kruger and Loewenstein 2013; Appel et al. 2020); or frequency, type and length of advertising. While the F2P gaming industry is substantial, these levers can serve to fuel beneficial outcomes in freemium settings outside the gaming industry. For example, learning apps like Duolingo can adjust exercise types and add hints dynamically to improve engagement and learning outcomes, potentially boosting revenue. Similarly, streaming platforms like Spotify or YouTube can reduce the number or length of ads at strategic moments, such as during drops in activity, to encourage user engagement and future upgrades. Nascent work by Agrawal et al. (2023) reports similarly beneficial effects on engagement and retention for "ed tech" (educational technology) software, as do Huang et al. (2021) for massively online open courses (MOOCs). One distinction of the system we investigate, compared to Huang et al. (2021), is that the DDA system intrinsically rewards users, whereas the incentives in Huang et al. (2021) likely work both intrinsically and extrinsically, or mostly extrinsically. The fact that DDA systems implicitly and intrinsically change incentives

may help explain why we observe relatively strong treatment effects (Gneezy, Meier, and Rey-Biel 2011; Levitt et al. 2016).

More broadly, our research demonstrates that systems for data-driven personalization of freemium product experiences may be able to induce a dual positive effect for customer retention in navigating the trade-off between the free and premium facets of freemium products. By failing to consider this dual effect of retention, firms may make myopic decisions when designing the free aspect of the service, potentially leading to lost long-term monetization opportunities. While our analyses are rooted in online gaming and game difficulty, we believe they provide strong motivation for further study of personalization systems in freemium apps and virtual experiences more generally to drive such synergistic and dual positive effects.

Firms may overlook significant revenue opportunities by not adequately valuing retention's possible dual impact in freemium contexts more generally. This study offers compelling empirical evidence that tailoring the freemium product individually and dynamically can bolster financial profitability. While our empirical analysis did not identify individuals adversely impacted by reducing game difficulty—potentially due to focusing on high-risk customers—we anticipate that excessively simplifying gameplay for challenge-seeking users might have counterproductive effects (Ascarza 2018). In such scenarios, assessing and paying close attention to the heterogeneity in treatment effects, as discussed in Section 4.3, is of paramount importance.

5.3. Limitations and future research

Our research has several limitations that suggest promising directions for future exploration. First, while our collaboration with a game publisher during the rollout of a DDA system provided us with a unique dataset and high ecological validity, it also restricted our control over treatment

design and protocols. As a result, difficulty assignment was partly endogenous—a feature welcomed by our industry partner—but this, coupled with the DDA system's effectiveness across user segments, limits our ability to explore boundary conditions. Specifically, the DDA system's method of reducing game difficulty for users who exhibited lower engagement in the previous week had positive effects overall, but we cannot assess its impact on users with higher levels of engagement. Another limitation is that users could receive multiple difficulty adjustments of varying intensity over time. This has constrained our ability to measure the impact of different levels of difficulty adjustment and instead led us to measure the overall cumulative impact of the DDA system. Future research could address these limitations by implementing a field experimental approach that initially randomizes different global difficulty levels to develop an optimal policy. This approach would also clarify boundary conditions where difficulty reduction might negatively impact player engagement.

Second, we do not directly observe ad revenue but instead impute it using usage data. To generate treatment effect estimates with this approach, we assume that the rate of ad consumption—ads consumed per time spent in-app—is unaffected by the treatment. However, this assumption could be violated. For instance, in some mobile games, players may watch a short video ad to gain a power-up. If the game's difficulty is reduced, players' willingness to engage with such rewarded ads might decrease. Since the partnering game primarily used non-rewarded interstitial ads, we believe this assumption is reasonable in our case. Future research focused specifically on the financial implications of DDAs could measure directly both IAP and Ad revenues.

A further promising research trajectory, building upon this work and recent theoretical work on the sequencing of game experiences (Li, Ryan, and Sheng 2023), might explore the long-term

effectiveness of difficulty adjustments as a strategic tool. Key questions include how users' sensitivity to adjusted difficulty evolves over time and how treated users behave if difficulty levels are restored. Would their behavior resemble that of never-treated users, or does the intervention hinder skill development, affecting future performance? Beyond online gaming, this research avenue could be relevant to fields like education, where adjusting difficulty could influence engagement and learning outcomes.

Moreover, future research could investigate potential interactions between DDA and systems for social engagement in freemium settings. Social interactions, such as messaging, collaborating, competing, sharing, and trading, are critical to engagement in many F2P games and freemium apps. These social systems could complement DDA systems to enhance engagement, retention, and monetization (Lopez-Vargas, Runge, and Zhang 2022). Future work could explore how individual and social personalization systems might be orchestrated.

In conclusion, our research emphasizes the importance of considering heterogeneity and dynamics when balancing free products and monetization within the freemium pricing model. As digital gaming and freemium models continue to evolve, our findings encourage firms to adopt a broader perspective that considers both the multifaceted dynamics of player engagement and the overarching ramifications for revenue generation.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT 4 in order to correct grammar. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

REFERENCES

Agrawal, K., Athey, S., Kanodia, A., & Palikot, E. (2022). Personalized recommendations in EdTech: Evidence from a randomized controlled trial. *arXiv preprint arXiv:2208.13940*.

Amabile, T., & Kramer, S. (2011). *The progress principle: Using small wins to ignite joy, engagement, and creativity at work*. Harvard Business Press.

Ansari, A., & Mela, C. F. (2003). E-customization. *Journal of Marketing Research*, *40*(2), 131-145.

Appel, G., Libai, B., Muller, E., & Shachar, R. (2020). On the monetization of mobile apps. *International Journal of Research in Marketing*, 20, 79.

Aral S., Dhillon, P. S. (2021). Digital paywall design: Implications for content demand and subscriptions. *Management Science*, 67(4), 2381-2402.

Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research*, *55*(1), 80-98.

Bashirzadeh, Y., Mai, R., & Faure, C. (2022). How rich is too rich? Visual design elements in digital marketing communications. *International Journal of Research in Marketing*, *39*(1), 58-76.

Caminal, R., & Matutes, C. (1990). Endogenous switching costs in a duopoly model. *International Journal of Industrial Organization*, *8*(3), 353-373.

Ceci L. (2023). Mobile app user retention rate worldwide Q3 2022, by category. Statista report, last accessed on June 12, 2023 at https://www.statista.com/statistics/259329/ios-and-android-app-user-retention-rate/.

Deng, Y., Lambrecht, A., & Liu, Y. (2022). Spillover effects and freemium strategy in the mobile app market. *Management Science*, 69(9), 5018-5041

Galak, J., Kruger, J., & Loewenstein, G. (2013). Slow down! Insensitivity to rate of consumption leads to avoidable satiation. *Journal of Consumer Research*, *39*(5), 993-1009.

Galak, J., & Redden, J. P. (2018). The properties and antecedents of hedonic decline. *Annual Review of Psychology*, *69*(1), 1-25.

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives*, *25*(4), 191-210.

Gu, X., Kannan, P. K., & Ma, L. (2018). Selling the premium in freemium. *Journal of Marketing*, 82(6), 10-27.

Hadiji, F., Sifa, R., Drachen, A., Thurau, C., Kersting, K., & Bauckhage, C. (2014). Predicting player churn in the wild. *In 2014 IEEE Conference on Computational Intelligence and Games.* 1-8.

Haenlein, M., Libai, B., & Muller, E. (2023). Satiation and cross promotion: Selling and swapping users in mobile games. *International Journal of Research in Marketing*, 40(2), 342-361.

Halbheer, D., Stahl, F., Koenigsberg, O., & Lehmann, D. R. (2014). Choosing a digital content strategy: How much should be free?. *International Journal of Research in Marketing*, 31(2), 192-206.

Hartmann, W. R., & Viard, V. B. (2008). Do frequency reward programs create switching costs? A dynamic structural analysis of demand in a reward program. *Quantitative Marketing and Economics*, *6*, 109-137.

Hofacker, C. F., De Ruyter, K., Lurie, N. H., Manchanda, P., & Donaldson, J. (2016). Gamification and mobile marketing effectiveness. *Journal of Interactive Marketing*, 34(1), 25-36.

Huang, Y., Jasin, S., & Manchanda, P. (2019). "Level Up": Leveraging skill and engagement to maximize player game-Play in online video games. *Information Systems Research*, 30(3), 927-947.

Huang, N., Zhang, J., Burtch, G., Li, X., & Chen, P. (2021). Combating procrastination on massive online open courses via optimal calls to action. *Information Systems Research*, 32(2), 301-317.

Hunicke, R. (2005). The case for dynamic difficulty adjustment in games. In Proceedings of the 2005 *ACM SIGCHI International Conference on Advances in Computer Entertainment Technology*. 429-433.

Klemperer, P. (1987). Markets with consumer switching costs. *The Quarterly Journal of Economics*, *102*(2), 375-394.

Lambrecht, A., & Misra, K. (2017). Fee or free: when should firms charge for online content?. *Management Science*, 63(4), 1150-1165.

Lasaleta, J. D., & Redden, J. P. (2018). When promoting similarity slows satiation: The relationship of variety, categorization, similarity, and satiation. *Journal of Marketing Research*, *55*(3), 446-457.

Latham, G. P., & Seijts, G. H. (1999). The effects of proximal and distal goals on performance on a moderately complex task. *Journal of Organizational Behavior*, *20*(4), 421-429.

Lee, C., Kumar, V., & Gupta, S. (2017). Designing freemium: Strategic balancing of growth and monetization. Working paper.

Lee, S. K., Hong, S. J., Yang, S. I., & Lee, H. (2016, October). Predicting churn in mobile free-to-play games. *In 2016 International Conference on Information and Communication Technology Convergence* (ICTC), 1046-1048.

Levitt, S. D., List, J. A., Neckermann, S., & Sadoff, S. (2016). The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy*, *8*(4), 183-219.

Li, H., Jain, S., & Kannan, P. K. (2019). Optimal design of free samples for digital products and services. *Journal of Marketing Research*, *56*(3), 419-438.

Li, Y., Ryan, C. T., & Sheng, L. (2023). Optimal sequencing in single-player games. *Management Science*, 69(10), 6057-6075.

Liu, C. Z., Au, Y. A., & Choi, H. S. (2014). Effects of freemium strategy in the mobile app market: An empirical study of Google Play. *Journal of Management Information Systems*, *31*(3), 326-354.

López Vargas, K., Runge, J., & Zhang, R. (2022). Algorithmic assortative matching on a digital social medium. *Information Systems Research*, *33*(4), 1138-1156.

Manderlink, G., & Harackiewicz, J. M. (1984). Proximal versus distal goal setting and intrinsic motivation. *Journal of Personality and Social Psychology*, *47*(4), 918.

Nair, H. (2007).Intertemporal price discrimination with forward-looking consumers: ApplicationTo the US market for console video-games. *Quantitative Marketing and Economics,* 5, 239-292.

Padilla, N., & Ascarza, E. (2021) Overcoming the cold start problem of CRM using a probabilistic machine learning approach. *Journal of Marketing Research* 58(5), 981-1006.

Reddit (2023). Adaptive difficulty in 2022. *Reddit Community*. Last accessed on October 23, 2023 online at https://www.reddit.com/r/gamedesign/comments/10jaquv/adaptive_difficulty_in_2022/.

Runge, J., Gao, P., Garcin, F., & Faltings, B. (2014). Churn prediction for high-value players in casual social games. In *2014 IEEE Conference on Computational Intelligence and Games*, 1-8.

Runge, J., Levav, J., & Nair, H. S. (2022). Price promotions and "freemium" app monetization. *Quantitative Marketing and Economics*, *20*(2), 101-139.

Runge, J., & van Dreunen, J. (2024). How to use games to build relationships with your customers. *Harvard Business Review*, Digital Article. Last accessed on December 24, 2024 online at https://hbr.org/2024/11/how-to-use-games-to-build-relationships-with-your-customers.

Rutz, O., Aravindakshan, A., & Rubel, O. (2019). Measuring and forecasting mobile game app engagement. *International Journal of Research in Marketing*, 36(2), 185-199.

Seufert, E. B. (2013). *Freemium economics: Leveraging analytics and user segmentation to drive revenue*. Elsevier.

Shi, Z., Zhang, K., & Srinivasan, K. (2019). Freemium as an optimal strategy for market dominant firms. *Marketing Science*, *38*(1), 150-169.

Statista (2023). Video game market revenue worldwide from 2017 to 2027. Last accessed on September 18, 2023 at https://www.statista.com/statistics/1344668/revenue-video-game-worldwide/.

Statista (2023a). In-App Advertising - Worldwide. Last accessed on November 23, 2023 at https://www.statista.com/outlook/amo/advertising/in-app-advertising/worldwide#ad-spending

Xue, S., Wu, M., Kolen, J., Aghdaie, N., & Zaman, K. A. (2017). Dynamic difficulty adjustment for maximized engagement in digital games. *In Proceedings of the 26th International Conference on World Wide Web Companion*, 465-471.

Zohaib, M. (2018). Dynamic difficulty adjustment (DDA) in computer games: A review. *Advances in Human-Computer Interaction*.

ONLINE APPENDIX

Personalized Game Design for Improved User Retention and Monetization in Freemium Mobile Games

A1 Experimental design – Further details about the degree of difficulty

Table A1 shows the difficulty adjustment intensity for each group of players, depending on the number of rounds that they have played in the previous 7 days. If a user played 20 or more rounds, the difficulty remained intact, at the default level denoted by 5. If a user played between 15 and 19 rounds, the difficultly was set to 4, which implies that the probability of connectivity in the board increased by 10% (i.e., the game was made easier for that player). As the user played fewer rounds in the last 7 days, the chances of connecting pieces were made easier.

# rounds in last 7 days	Difficulty level	Increase of "connectivity"
>= 20	5 (more difficult)	0%
< 20	4	10%
< 15	3	20%
< 10	2	30%
< 5	1 (least difficult)	45%

Table A1: Difficulty adjustment intensity by level of play during the past 7 days.
Difficulty 5 is the default option in the game design.

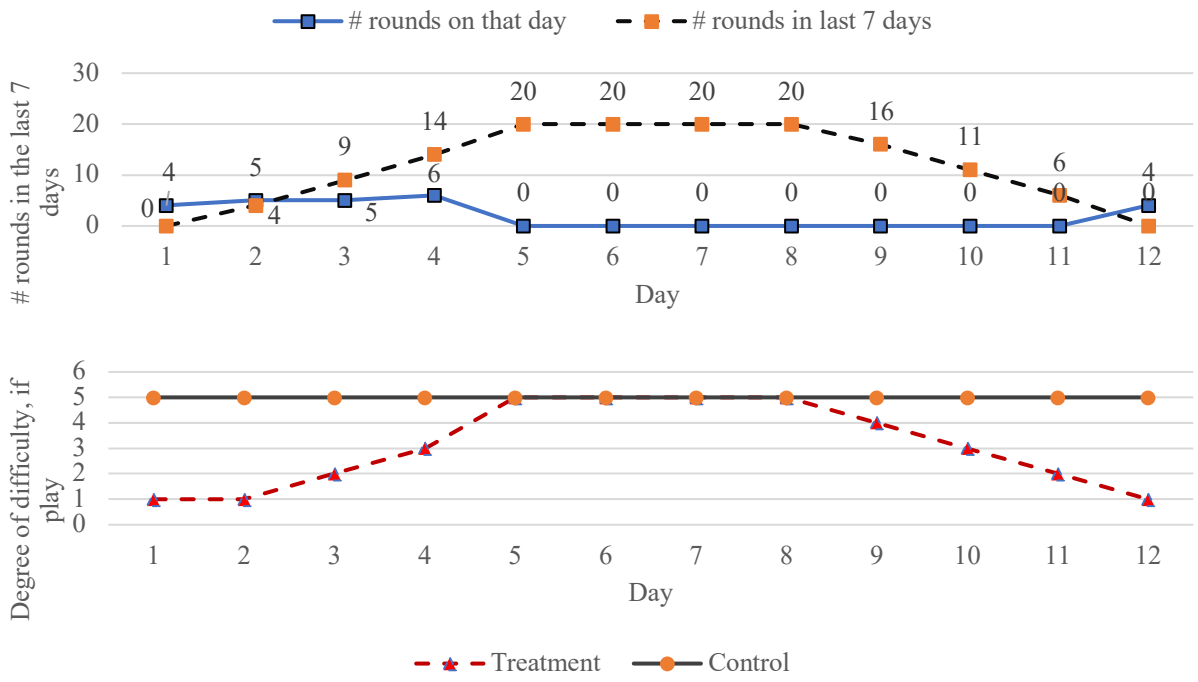

Figure A1: Hypothetical example to illustrate the level of difficulty (bottom figure) that correspond to different levels of past play. *The red line (dashed) corresponds to treated users, whose difficulty was altered based on the amount of play in the previous 7 days, and the back line (solid) corresponds to control users whose level of difficulty never changed.*

Figure A2 shows two difficulty scenarios. On the left the player is facing the highest level of difficulty ("default" in the game design) whereas the player on the right is facing the easiest scenario with low difficulty levels.

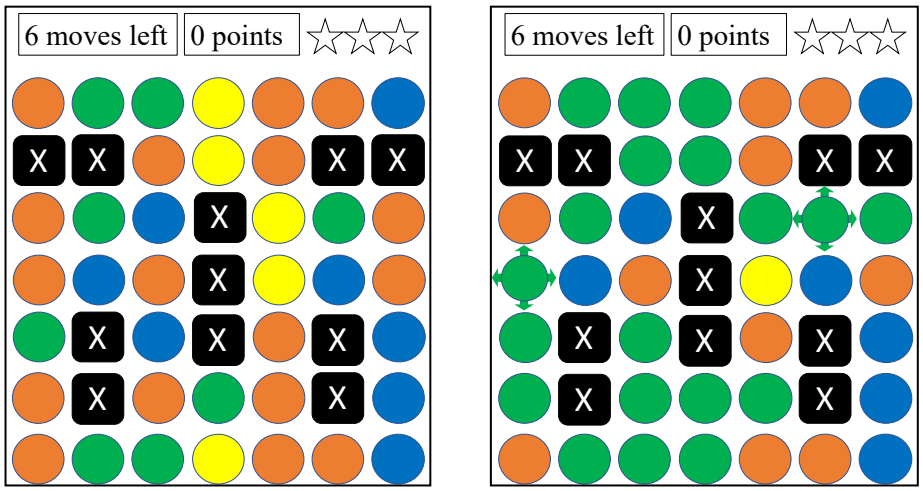

Figure A2: Example of difficulty manipulation. *Image of a level of the game, with two different levels of difficulty. On the left, the most difficult scenario which corresponds with the default difficulty of the game. On the right, the player is facing an easier scenario in which there are much higher chances of creating long combinations (of green color) and has two "special chips," identified by the arrows around the two pieces.*

A2 Data – Descriptive statistics and randomization checks

Table A2 shows the descriptive statistics and randomization checks for the full set of user-level variables. Level20 variables are measured when users pass level 20 (and remains constant over time) and all other variables are measured at the moment of the intervention.

	N=	329,999					190,863	139,136		
		Mean	SD	p25	p50	p75	Control	Treatment	Difference	p-value
level20_rounds		45.83	36.91	29	36	49	45.85	45.81	-0.04	0.810
level20_days		8.691	15.74	1	3	9	8.67	8.719	0.049	0.373
level20_stars		39.80	4.497	37	39	42	39.81	39.78	-0.03	0.110
level20_coins_collected		40.75	75.93	0	0	70	40.82	40.64	-0.18	0.499
level20_did_use_extra		0.926	0.261	1	1	1	0.927	0.926	-0.001	0.668
level20_did_use_coin		0.513	0.5	0	1	1	0.513	0.514	0.001	0.343
level20_did_purchase		0.011	0.103	0	0	0	0.011	0.011	0.000	0.080
age_rounds		189.6	217.7	62	113	225	190.0	189.2	-0.8	0.315
age_days		45.85	31.91	19	37	66	45.78	45.94	0.16	0.168
age_level		37.74	16.97	24	39	40	37.78	37.69	-0.09	0.101
age_stars		75.19	36.68	49	67	87	75.3	75.04	-0.26	0.045
age_coins		24.0	74.0	0	0	30	24.04	23.96	-0.08	0.752
age_distance_to_gate		6.353	6.561	0	4	11	6.484	6.174	-0.31	0.000
rfm_rec		13.88	17.9	3	7	16	13.85	13.92	0.07	0.276
rfm_week		5.728	6.68	0	2	12	5.735	5.718	-0.017	0.476
rfm_ratio		3.719	2.952	1.667	3.077	4.836	3.721	3.716	-0.005	0.605
stuck_rounds		26.72	62.45	2	7	25	26.79	26.61	-0.18	0.403
stuck_days		17.25	21.59	4	9	21	17.22	17.3	0.08	0.347
stuck_playdays		3.054	3.804	1	2	4	3.054	3.055	0.001	0.930
stuck_days_in_gate		23.47	23.38	8	15	32	23.32	23.67	0.35	0.033
stuck_broke_gate		0.0218	0.146	0	0	0	0.0217	0.022	0.0003	0.645
yesterday_progress		1.225	2.564	0	0	2	1.223	1.227	0.004	0.703
yesterday_win_prop		0.308	0.341	0	0.2	0.545	0.309	0.307	-0.002	0.289
skill_rounds_per_level		4.41	3.653	2.227	3.286	5.318	4.413	4.406	-0.007	0.568
skill_stars_per_level		1.988	0.257	1.821	1.957	2.149	1.988	1.987	-0.001	0.036

Table A2: Descriptive statistics (left-most columns) and randomization checks (last 4 columns). *We acknowledge that four variables (out of 25) show significant differences across conditions. Given the experimental set up, we attribute these differences to chance, and not to any systematic intrinsic difference between treatment and control users.*

A3 Manipulation checks by difficulty adjustment intensity

Table A3 shows the remaining metrics examined in the manipulation checks.

	Points (by level of difficulty)				Snake Length (by level of difficulty)			
	4	3	2	1	4	3	2	1
Treatment	**1275**	**3225**	**6126**	**10741**	**0.018**	**0.027**	**0.112**	**0.269**
	(128.1)	(137.9)	(157.4)	(91.2)	(0.009)	(0.009)	(0.009)	(0.005)
Constant	35818	35862	35648	35047	5.119	5.148	5.144	5.179
	(82.7)	(87.6)	(95.8)	(50.8)	(0.006)	(0.006)	(0.007)	(0.003)
N	370137	302836	281827	1055166	370137	302836	281827	1055166
R-squared	0.1%	0.6%	2.0%	4.9%	0.0%	0.0%	0.2%	1.1%

Table A3: Manipulation checks by degree of difficulty. *OLS of the round outcome against a treatment dummy using all rounds on the first day of the experiment. Standard errors are clustered at the user level. Treatment variable in bold indicates p-value<0.05.*

A4 Robustness checks for the main analyses

Day of intervention

One potential concern with the main results presented in **Table 6** is that the intervention does not occur simultaneously for each treated player. **Figure A3** shows the number of users assigned to each condition over time.

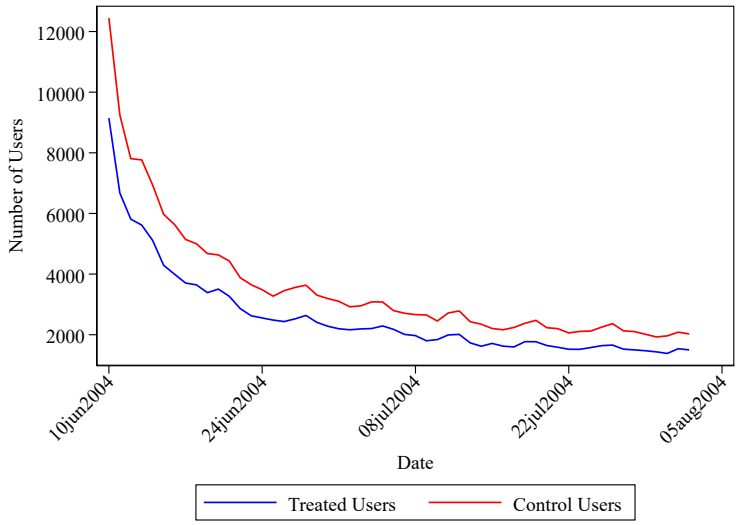

Figure A3: Number of users, by condition, qualifying for the experiment in each day.

Although time-varying factors (e.g., July 4th, weekday vs. weekend) are not confounds due to our experimental design—for every treated user, we have control users who would have been treated on the exact same day but were allocated to control instead—these factors might introduce some variation that could potentially reduce the efficiency of our ITT estimates. To address this, we re-run the main regressions, including time-related variables as controls to capture the potential differences arising from the non-uniform treatment timing. Specifically, we added 'day of the week', 'month', and 'July 4th' as controls. In **Table A4**, we replicate the results presented in Table 6 of the main manuscript, now controlling for variables that capture the day on which users received the treatment. This includes 'day of the week', 'month of the year', and 'July 4th', all incorporated as dummy variables. All results are consistent with those obtained in the main analysis.

	# Rounds played	Progress made	Engagement	Retention 1	Retention 7	Retention 14
Treatment	**1.247**	**0.746**	**0.0212**	**0.027**	**0.0244**	**0.0196**
	(0.0251)	(0.00882)	(0.000729)	(0.00164)	(0.0017)	(0.00156)
p-val	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Constant	5.256	0.65	0.812	0.279	0.632	0.721
	(0.0326)	(0.0108)	(0.00109)	(0.0023)	(0.00244)	(0.00225)
# obs	329,999	329,999	1,867,849	326,472	316,257	308,003

Table A4: Robustness check including day-level controls. *OLS of the behavior of interest against the treatment variable and 'day of intervention' dummies. Standard errors are clustered at the user level. Treatment variable in bold indicates p-value<0.05.*

Maximum level achieved prior to the intervention

It is possible that users who have previously achieved higher levels in the game respond differently to the DDA intervention compared to newer players at lower levels (recall that in order to be eligible for treatment (or control) users had to pass at least level 20). If this is the case, the results in the main manuscript might reflect a subset of users rather than the entire user base. In **Table A5**, we replicate the results from **Table 6**, incorporating the interaction between

treatment and the maximum level achieved by the user prior to the experiment (the variable 'Max Level' has been standardized). The treatment effect results remain consistent with those in the main manuscript. The main effect of 'Max Level' aligns with expectations: more experienced users play more rounds, achieve less progress, and exhibit higher engagement and retention levels. All interaction effects are positive, indicating that more experienced users (i.e., those who had reached higher levels before the intervention) react more strongly to the difficulty adjustment. Importantly, the magnitude of the interaction effect is much smaller than the treatment effect, confirming that lowering difficulty positively impacts user outcomes regardless of their previous maximum level.[13]

	# Rounds played	Progress made	Engagement	Retention 1	Retention 7	Retention 14
Treatment	**1.263**	**0.748**	**0.0211**	**0.0277**	**0.0251**	**0.0202**
	(0.0253)	(0.00877)	(0.000729)	(0.00164)	(0.0017)	(0.00156)
p-val	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Max Level	**0.136**	**-0.288**	**0.0059**	**0.04**	**0.0465**	**0.0363**
	(0.0146)	(0.00384)	(0.000487)	(0.00114)	(0.00106)	(0.000944)
p-val	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Treatment * Max Level	**0.332**	**0.093**	**0.00329**	**0.0102**	**0.00506**	**0.0035**
	(0.0272)	(0.00945)	(0.000699)	(0.00178)	(0.00162)	(0.00142)
p-val	0.0000	0.0000	0.0000	0.0000	0.0017	0.0139
Constant	5.57	0.773	0.818	0.297	0.644	0.744
	(0.0132)	(0.00375)	(0.000501)	(0.00105)	(0.00111)	(0.00103)
# obs	329,999	329,999	1,867,849	326,472	316,257	308,003

Table A5: Robustness check including max-level achieved prior to intervention. *OLS of the behavior of interest against the treatment variable and (standardized) 'maximum level achieved' prior to the intervention. Standard errors are clustered at the user level Standard errors are clustered at the user level. Treatment variable in bold indicates p-value<0.05.*

[13] Recall that all users in our sample had achieved level 20 prior to the intervention.

Degree of difficulty adjustment

The DDA condition sets a difficulty level based on the number of rounds a user played in the last 7 days. Consequently, the level of difficulty adjustment varies among users, both on the first day of the intervention and in subsequent periods. It is expected that the treatment effect is stronger in groups where the difficulty adjustment was more significant and weaker where the adjustment was minimal. Although the degree of difficulty adjustment is an endogenous metric determined by the number of rounds in previous days (which is likely influenced by earlier adjustments), we can utilize this metric on the first day of the experiment since prior consumption could not have been affected by the intervention. We conduct this analysis in two ways: First, we create a standardized continuous variable named 'Intensity,' which increases as the level of difficulty adjustment grows (i.e., when users are allocated to difficulty=1) on the first day of the intervention. We replicate the results in **Table 6**, adding both 'Intensity' and its interaction with 'Treatment' (**Table A6**). All results are as expected: Average treatment effects are consistent with those presented in **Table 6**, the main effect of intensity of difficulty adjustment is negative for all outcomes of interest, consistent with the notion that lower usage in the last 7 days predicts lower usage in the future, and the interaction terms are all positive, corroborating that stronger difficulty adjustment leads to stronger treatment effects. See results in **Table A6.**

	# Rounds played	Progress made	Engagement	Retention 1	Retention 7	Retention 14
Treatment	**1.219**	**0.725**	**0.0217**	**0.0269**	**0.0246**	**0.0202**
	(0.0249)	(0.00868)	(0.000732)	(0.00162)	(0.00161)	(0.00149)
p-val	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Intensity	**-0.578**	**-0.0925**	**-0.0104**	**-0.106**	**-0.15**	**-0.123**
	(0.0137)	(0.00409)	(0.000487)	(0.00109)	(0.000985)	(0.000892)
p-val	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Treatment * Intensity	**0.515**	**0.361**	**0.0101**	**0.00924**	**0.0136**	**0.0112**
	(0.0245)	(0.00823)	(0.00072)	(0.0017)	(0.00151)	(0.00135)
p-val	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Constant	5.597	0.79	0.818	0.301	0.65	0.747
	(0.0132)	(0.0039)	(0.000505)	(0.00103)	(0.00105)	(0.000982)
# obs	329,999	329,999	1,867,849	326,472	316,257	308,003

Table A6: Robustness check including difficulty adjustment intensity on the day of the intervention. *OLS of the behavior of interest against the treatment variable and difficulty adjustment 'intensity' (standardized) on the first day of the intervention. Standard errors are clustered at the user level Standard errors are clustered at the user level. Treatment variable in bold indicates p-value<0.05.*

Second, we treat difficulty adjustment as a categorical variable and replicate the results of the main manuscript separately for each level of difficulty adjustment cohort (see **Table A7** for levels of adjustment). As expected, and consistent with the manipulation analyses (**Table 5**), the effect of the intervention is stronger when the difficulty adjustment is larger.

	Intensity of Difficulty Adjustment			
	Largest (Difficulty = 1)	Large (Difficulty = 2)	Small (Difficulty = 3)	Smallest (Difficulty = 4)
# Rounds played	**1.639**	**1.279**	**0.781**	**0.334**
	(0.0352)	(0.0674)	(0.0619)	(0.0554)
p-val	0.0000	0.0000	0.0000	0.0000
Progress made	**1.041**	**0.659**	**0.381**	**0.154**
	(0.0132)	(0.0215)	(0.0196)	(0.0169)
p-val	0.0000	0.0000	0.0000	0.0000
Engagement	**0.0307**	**0.0204**	**0.0113**	**0.00582**
	(0.00104)	(0.0019)	(0.00181)	(0.00164)
p-val	0.0000	0.0000	0.0000	0.0004
Retention 1	**0.0333**	**0.0353**	**0.0181**	**0.0095**
	(0.00199)	(0.00459)	(0.00463)	(0.00431)
p-val	0.0000	0.0000	0.0001	0.0276
Retention 7	**0.0367**	**0.0258**	**0.00513**	0.00579
	(0.00242)	(0.00417)	(0.00368)	(0.00302)
p-val	0.0000	0.0000	0.1640	0.0558
Retention 14	**0.031**	**0.016**	0.00531	**0.00552**
	(0.00236)	(0.00354)	(0.00303)	(0.00243)
	0.0000	0.0000	0.0802	0.0232

Table A7: Impact of difficulty adjustment by cohorts, determined by levels of difficulty adjustment on the first day of the intervention. *OLS of the behavior of interest against the treatment variable by group of intensity of difficulty adjustment on the first day of the intervention. This table presents the parameter estimates of the variable of interest ('Treatment'). Standard errors are clustered at the user level. Treatment variable in bold indicates p-value<0.05.*

User persistence in the game

A potential limitation of the round-level analysis (as seen in **Table 7** and **Table 8**) is that the users who persist in playing for extended durations—and hence feature more prominently in this analysis—might exhibit systematic differences between the treatment and control groups. This phenomenon represents a survival bias, where users who continued gameplay after being treated with a reduced difficulty level during the initial round (or subsequent few rounds) may differ inherently from those in the control group who persevered despite facing more challenging game settings. To address this concern, we conducted two robustness checks. Firstly, we replicated the

round-level analyses considering only the initial five rounds post-treatment for each player.[14] The rationale for choosing the first five rounds is that it offers a conservative approximation of the number of rounds most users would likely play, irrespective of their treatment assignment. Given that all users start each day with a set of five lives, even those who encounter the most challenging difficulty levels can engage in up to five rounds without having to pause or purchase additional extras.

Secondly, we replicated all ITT analyses while controlling for all user attributes outlined in Section 3.4. It is important to note that these attributes capture historical user competencies in gameplay, activity levels, and prior tendencies to utilize extras, coins, and monetary resources. A detailed breakdown of the outcomes from these robustness checks can be found in Appendix A4. In brief, barring the exceptions of coins and money expended on extras within the inaugural five rounds (where notable differences were not identified), all other results align with those delineated in **Table 7** and **Table 8**.

	Continue playing	Use extras	Use coins	Use Money	Coin extra	Coin gate	Money extra	Money gate
Treatment	**0.0145**	**0.0012**	**0.0004**	**0.0002**	-0.00026	**0.00235**	0.00002	**0.00066**
	(0.0008)	(0.0004)	(0.0002)	(0.0001)	(0.00014)	(0.00036)	(0.00006)	(0.00016)
p-val	0.0000	0.0016	0.0269	0.0155	0.06340	0.00000	0.74800	0.00004
Constant	0.7920	0.0261	0.0095	0.0015	0.00503	0.00875	0.00078	0.00163
	(0.0005)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0)	(0.0001)
# obs	1136686	1177102	1177102	1177102	1177102	315600	1177102	315600

Table A8: Robustness check using first 5 rounds per user. *OLS of the behavior of interest against a treatment dummy using the first five observations per user. Standard errors are clustered at the user level. The gate-related outcomes are conditioned on the user being at a gate Treatment variable in bold indicates p-value<0.05.*

Similarly, **Table A9** shows the results when we add multiple controls per user. Specifically, we control for the activity-related variables that are observed at the moment in which the user

[14] One way to overcome this concern entirely would be to replicate the analysis using only the very first round of data. However, doing so discards a lot of useful information, dramatically reducing the sample size and thus the power to capture the observed effects.

receives the experiment for the first time. These include recency, frequency, past purchases (if any), number of round in current level, whether the user is stuck at a gate, number of days since they broke the previous gate, and various summaries of the level of activity on the previous day they played the game. All results are consistent with those presented in **Tables 6**, **Table 7**, and **Table 8** of the main analyses.

	Continue playing	Use extras	Use coins	Use Money	Coin extra	Coin gate	Money extra	Money gate
Treatment	**0.0202**	**-0.0016**	**-0.0006**	-0.0001	**-0.00062**	-0.00010	**-0.00012**	0.00001
	(0.0007)	(0.0003)	(0.0001)	(0.0001)	(0.00011)	(0.00019)	(0.00005)	(0.00008)
p-val	*0.0000*	*0.0000*	*0.0001*	*0.0660*	*0.00000*	*0.58000*	*0.01040*	*0.93200*
Constant	0.7850	0.0033	0.0021	-0.0109	-0.00821	0.03160	-0.00889	-0.00976
	(0.0083)	(0.004)	(0.003)	(0.0013)	(0.0021)	(0.00478)	(0.00103)	(0.00347)
# obs	1867811	2009921	2009921	2009921	2009921	652565	2009921	652565

Table A9: Robustness check adding controls. *OLS of the behavior of interest against a treatment dummy using all rounds on the first days of the intervention. Standard errors are clustered at the user level. The gate-related outcomes are conditioned on the user being at a gate Treatment variable in bold indicates p-value<0.01.*

A5 Implications for user targeting: Full set of results

In this appendix, we provide further details on the analysis conducted to explore the groups of customers who would be more responsive to the intervention (Section 4.3). First, we describe how each of the variables has been computed and report their summary statistics and correlations among them. Then, we present the full set of results.

Table A10 provides the summary statistics and the operationalization of each variable, while **Table A11** presents the pairwise correlations. The variables 'early progress,' 'progress prone,' 'spender,' and 'progression decline' are computed based on behavior prior to surpassing Level 20, whereas the 'frustrated' and 'distance' variables are derived from data at the time of the intervention.

Variable	Measure	Mean	St.Dev	P25	P50	P75
Early progress	# days to achieve Level 20	-8.69	15.74	-9.00	-3.00	-1.00
Progress prone	Proportion of rounds the user kept playing after winning	0.92	0.08	0.90	0.95	1.00
Spender	Whether user spent anything before Level 20	0.01	0.10	0.00	0.00	0.00
Frustrated	# unsuccessful rounds in current level	27.27	62.43	3.00	8.00	25.00
Distance	# levels until next gate	6.50	6.69	0.00	4.00	12.00

Table A10: Summary statistics for moderating variables.

Variables	(1)	(2)	(3)	(5)	(6)
(1) Early progress	1.000				
(2) Progress prone	0.206*	1.000			
	(0.000)				
(3) Spender	0.002	0.016*	1.000		
	(0.279)	(0.000)			
(4) Frustrated	0.095*	0.095*	0.003	1.000	
	(0.000)	(0.000)	(0.058)		
(5) Distance	0.146*	0.082*	0.020*	-0.110*	1.000
	(0.000)	(0.000)	(0.000)	(0.000)	
** shows significance at p<.05*					

Table A11: Pairwise correlations among moderating variables.

Next, we report the results from the full model. Given the large number of variables in the regressions, we split the results by treatment and interaction effects (**Table A12**), and all sets of three-way interactions (**Table A13**).

	Playdays	Rounds	Progress	Purchases	Pay extra	Pay gate	Coins	Coin extra	Coin gate
Treatment	**0.631**	**8.664**	**2.097**	**0.0106**	**0.00415**	**0.00483**	**0.026**	**0.00919**	**0.631**
	(0.0203)	(0.249)	(0.0632)	(0.0023)	(0.0011)	(0.0004)	(0.0021)	(0.0018)	(0.0203)
p-val	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Early progress	**0.0946**	**1.504**	**0.93**	**0.0061**	**0.0031**	**0.0017**	**0.0127**	**0.0075**	**0.0946**
	(0.0228)	(0.295)	(0.0742)	(0.002)	(0.0011)	(0.0004)	(0.0023)	(0.0019)	(0.0228)
p-val	0.000	0.000	0.000	0.002	0.006	0.000	0.000	0.000	0.000
Progress prone	0.0052	**0.88**	**0.202**	0.00332	0.00107	0.000497	0.00344	0.0025	0.0052
	(0.0206)	(0.236)	(0.0611)	(0.0025)	(0.0011)	(0.0004)	(0.0022)	(0.0018)	(0.0206)
p-val	0.800	0.000	0.001	0.184	0.346	0.219	0.114	0.165	0.800
Spender	0.00722	-0.0527	-0.0199	**0.0277**	**0.0151**	**0.00503**	0.012	**0.0125**	0.00722
	(0.0213)	(0.243)	(0.0787)	(0.0114)	(0.0051)	(0.001)	(0.0066)	(0.006)	(0.0213)
p-val	0.735	0.828	0.800	0.015	0.003	0.000	0.067	0.036	0.735
Frustrated	**0.225**	**2.462**	0.0234	0.00215	0.00126	**0.00101**	0.00175	**0.00413**	**0.225**
	(0.0432)	(0.626)	(0.159)	(0.0019)	(0.0011)	(0.0004)	(0.0023)	(0.0019)	(0.0432)
p-val	0.000	0.000	0.883	0.264	0.269	0.011	0.441	0.027	0.000
Distance	**0.171**	**2.933**	**0.154**	**0.00587**	0.00112	**0.00385**	**0.0149**	0.00216	**0.171**
	(0.0203)	(0.256)	(0.063)	(0.0023)	(0.0012)	(0.0005)	(0.0022)	(0.0018)	(0.0203)
p-val	0.000	0.000	0.014	0.009	0.328	0.000	0.000	0.231	0.000
# obs	329999	329999	329999	329999	329999	329999	329999	329999	329999

Table A12: Results for the (simple) interaction effects between each moderator variable and the treatment effect. *OLS with treatment, all moderators, their interaction effects, and all possible three-way interactions. All moderator variables have been standardized before computing the interaction terms. Robust standard errors reported in parentheses. Interaction variable in bold indicates that p-value<0.05.*

	Playdays	Rounds	Progress	Purchases	Pay extra	Pay gate	Coins	Coin extra	Coin gate
Early progress x	**0.0742**	**1.11**	**1.213**	**0.00323**	**0.00149**	**0.000611**	**0.00513**	**0.00376**	**0.0742**
Progress prone	(0.0113)	(0.125)	(0.0342)	(0.0008)	(0.0005)	(0.0002)	(0.0011)	(0.0009)	(0.0113)
p-val	0.000	0.000	0.000	0.000	0.001	0.001	0.000	0.000	0.000
Early progress x	-0.0238	-0.302	0.132	0.0148	0.00688	-0.000552	0.00251	0.0033	-0.0238
Spender	(0.0164)	(0.181)	(0.0798)	(0.0093)	(0.0056)	(0.0011)	(0.0081)	(0.0074)	(0.0164)
p-val	0.147	0.095	0.098	0.113	0.219	0.628	0.756	0.657	0.147
Early progress x	0.00357	0.782	**1.494**	**0.00414**	**0.00234**	0.000641	**0.00808**	**0.00479**	0.00357
Frustrated	(0.0434)	(0.621)	(0.144)	(0.0014)	(0.0008)	(0.0004)	(0.0023)	(0.002)	(0.0434)
p-val	0.934	0.208	0.000	0.004	0.005	0.081	0.001	0.016	0.934
Early progress x	**-0.108**	**-0.552**	**-0.191**	**0.00446**	**0.0021**	**0.00147**	**0.00612**	**0.00471**	**-0.108**
Distance	(0.0192)	(0.247)	(0.0595)	(0.0018)	(0.001)	(0.0004)	(0.0022)	(0.0018)	(0.0192)
p-val	0.000	0.025	0.001	0.013	0.043	0.000	0.005	0.007	0.000
Progr. prone x	0.0144	**0.405**	**0.13**	0.00125	-0.00431	0.00144	-0.00214	-0.0021	0.0144
Spender	(0.0161)	(0.165)	(0.0559)	(0.0091)	(0.0045)	(0.0009)	(0.0054)	(0.0051)	(0.0161)
p-val	0.372	0.014	0.020	0.891	0.342	0.112	0.690	0.679	0.372
Progr. prone x	**-0.202**	**-1.699**	**-1.06**	-0.00102	-0.00099	-0.000225	-0.00187	-0.00149	**-0.202**
Frustrated	(0.0312)	(0.44)	(0.104)	(0.0014)	(0.0009)	(0.0004)	(0.0024)	(0.002)	(0.0312)
p-val	0.000	0.000	0.000	0.478	0.278	0.540	0.437	0.455	0.000
Progr. prone x	**-0.0879**	**-0.719**	**-0.671**	0.00105	0.000451	0.000801	-0.00037	-0.0012	**-0.0879**
Distance	(0.0172)	(0.202)	(0.0493)	(0.0018)	(0.0009)	(0.0004)	(0.0018)	(0.0014)	(0.0172)
p-val	0.000	0.000	0.000	0.552	0.619	0.049	0.836	0.398	0.000
Spender x	**-0.0454**	**-0.601**	**-0.163**	**-0.0157**	**-0.00832**	**-0.00103**	**-0.011**	**-0.0111**	**-0.0454**
Frustrated	(0.0197)	(0.275)	(0.0359)	(0.0061)	(0.0033)	(0.0004)	(0.0043)	(0.0043)	(0.0197)
p-val	0.021	0.029	0.000	0.010	0.011	0.019	0.011	0.010	0.021
Spender x	0.0186	0.385	**0.229**	0.0174	**0.00941**	**0.00288**	**0.0179**	**0.0153**	0.0186
Distance	(0.0178)	(0.242)	(0.0666)	(0.0109)	(0.0048)	(0.001)	(0.0063)	(0.0059)	(0.0178)
p-val	0.296	0.112	0.001	0.108	0.047	0.005	0.004	0.010	0.296
Frustrated x	**0.122**	**1.365**	0.0603	0.00329	0.00199	**0.000873**	0.00124	**0.00375**	**0.122**
Distance	(0.0294)	(0.493)	(0.0981)	(0.0019)	(0.0011)	(0.0004)	(0.0019)	(0.0015)	(0.0294)
p-val	0.000	0.006	0.539	0.084	0.079	0.020	0.506	0.012	0.000
# obs	329999	329999	329999	329999	329999	329999	329999	329999	329999

Table A13: Results for the three-way interaction. *OLS with treatment, all moderators, their interaction effects, and all possible three-way interactions. All moderator variables have been standardized before computing the interaction terms. Robust standard errors reported in parentheses. Interaction variable in bold indicates p-value<0.05.*