OULUN AMMATTIKORKEAKOULU

Oona-Maria Karppinen

Mobile Gaming: Advertising and User Experience

Impact of Ad Source Quality on User Experience

OULUN AMMATTIKORKEAKOULU

Mobile Gaming: Advertising and User Experience

Impact of Ad Source Quality on User Experience

Oona-Maria Karppinen
Bachelor’s Thesis
Autumn Term of 2021
Bachelor’s Degree Programme in International Business and Administration
Oulu University of Applied Sciences

ABSTRACT

Oulu University of Applied Sciences
Degree programme in International Business and Administration

---

Author: Oona-Maria Karppinen
Title of the thesis: Mobile Gaming: Advertising and User Experience
Thesis examiner: Jaakko Sinisalo
Term and year of thesis completion: Autumn 2021 Pages: 56

The commissioner for this thesis is Fingersoft, an Oulu-based company that was founded in 2012. The company operates in mobile gaming industry and has developed the very successful Hill Climb Racing. With all its titles combined, Fingersoft has bypassed a milestone of 1.5 billion combined installs across platforms. The area of study is digital marketing, and the research question goes as follows; *How does the quality of ads impact the user experience?* The idea came up during an internal meeting in a discussion on how ad-related operations can be optimized to improve the user experience.

The aim of the project was to find out how much effect does the quality of ads have on user experience by comparing data gathered with A/B testing, which was a primary tool for conducting research. As the games developed by Fingersoft don't necessarily require the user to spend any actual money to access the apps, ads play a significant role in gaining revenue. Therefore, optimizing ads operations is vital. As having balance between user experience and revenue is vital for a business and its profitability, the thesis will also touch on a sub-question of how the set of networks associated with an application impact the overall performance.

Theoretical base for the research was put together by exploring online resources such as articles and academic writings, as well as interviewing professionals working with ads, user experience and A/B testing. The test itself was run with a dedicated A/B testing tool.

Results received from the experiment were not unambiguous and the benefit of changes applied in the test was not obvious. The analysis of outcomes indicated that Variant B should not necessarily be applied for the targeted audiences, but the results gained could be utilized for further testing and investigation regarding related issues. In coming to this conclusion, aspects considered were user experience related metrics and events and impacts on revenue – however, the development was not prominent on any scale. Which option seemed more profitable varied and depended on which metrics were under observation.

Even though the hypothesis of high-tier audiences being more sensitive to low-quality content could not be proved either true or false and an answer to the research question was not clearly stated, the research was not useless by any means. As the purpose of ad monetization team at Fingersoft is to continuously seek new solutions and best practices to optimize their operations, finding answers to these types of questions is necessary. Now that the question has been widely explored, the process can be continued by asking more detailed questions about metrics and what impacts them.

Keywords: Ads, user experience, mobile gaming, monetization, optimization

CONTENTS

1. INTRODUCTION

In today's digitalized world, the userbase of mobile devices is continuously growing. Consumers across the globe increasingly benefit of mobile applications for various purposes. According to recent studies, gaming remains in the lead of mobile application categories and is accountable for the majority of downloads with a share of 33%. (Perez 2019.) Last year, the global mobile gaming industry was valued at 165 billion USD, and the number is projected to keep on growing in the years to come (Knezovic 2021.)

1.1. Commissioner

Fingersoft is a Finnish, Oulu-based company operating in the mobile gaming industry. The company was founded in 2012 as one man creating different sorts of camera filters for smart phones but had its big break later in that same year with Hill Climb Racing – a game that quickly became an international hit. Thanks to Hill Climb Racing, Fingersoft has since become one of the biggest mobile game publishers in Finland.

The company has continued to grow and develop throughout the years and in addition to the Hill Climb Racing sequel, it has released a handful of other games as well. As of today, the company has totalled over 1.5 billion installs and more than 150 million euros in net revenue across different titles.

The research conducted for this thesis focuses on the Android version of Hill Climb Racing 2, which is a game that followed in late 2016 after Fingersoft's breakthrough. Hill Climb Racing 2 was released for Android, iOS, and later for Amazon devices. Both games in the series are physics-based and fall under racing-category. (Fingersoft 2021.)

1.2. Ad Monetization

Monetization is a core concept of this thesis. In short, monetization refers to gaining financial benefit from the content published. In mobile gaming, a good monetization strategy is an essential part of building and maintaining sustainability, since sufficient revenue is a bare necessity for a for-profit organization. In mobile games and applications overall, there are two main branches in monetization - IAPs, short for in-app purchases, are a way to have the user spend real-life currency on in-game currency, extra lives, special skins or an ad-free version of the mobile application in question, to mention a few. Ads on the other hand, are a stream of revenue that makes it possible for a game to be profitable without having users make purchases. (ElHady 2020.)

1.3. Research

The idea for the thesis topic came up during an ad-related meeting at Fingersoft. In the meeting, it was agreed that it would be highly beneficial for the teams' future operations to investigate the impact of ads' quality on user experience. As one of the main purposes of the ad monetization team at Fingersoft is to continuously find solutions to optimize their operations and revenue, the topic was found necessary to study further.

The objective of the research is to find an answer to the following question: *What is the impact of ads' quality on user experience?* The main tool used to reach a conclusion is A/B testing ads among two groups of users from the same target segment. In Fingersoft's games, ads are provided by *third-party ad networks*. Ad networks are external partners mediating ads from their pools of campaigns to *publishers*, like Fingersoft is. (Adjust GmbH 2021.) By observing each partner's performance, the higher- and lower-quality networks will be separated from each other and displayed for different groups in the A/B test. The testing is used to measure the correlation between ads and user behaviour, primary components compared being quality of ads and retention rate, which represents the percentage of users returning to the application after a certain number of days from first open (Draganov 2014, 163). To support the research, the project will also explore the overall impact of the set of networks used has on the overall performance.

In this research, user experience refers to the gamer's evaluation and perception of the target product (ProductPlan 2021). When implementing ads in a mobile game, user is the number one stakeholder to consider as a positive experience can result in a longer relationship with the player and therefore contribute to the sustainability of the product. On the flip side, a bad experience with ads can drive a user away rather quickly, especially if issues arise constantly. (Manjrekar

2018.) Problems, such as crashes, can expose a poor network in terms of ad quality (admonsters, 8).

The testing and researching will not be implemented globally. The decision to target so-called tier 1-markets is based on a hypothesis that users from these areas can be more sensitive to malfunctioning and low-quality content. In case of most mobile applications, tier 1-markets include countries with higher disposable incomes and higher standards of living, such as the United States, Canada, Germany and Norway. However, the tier 1-markets for different products can vary greatly. The idea and logic behind geographical user segmentation are further explained in section 2.7 of this thesis.

As the whole mobile gaming industry is constantly developing with new trends occurring after another, printed content on the subject is most likely lacking behind. Therefore, it is vital to support the theoretical base with up-to-date articles and papers, as well as with interviews. As was brought up earlier in this introduction, the method used for gaining results and answering the research question is A/B testing.

At the end of this project, the main point of interest is to find what correlation there might be between the ads' quality and the users' experience on the application, and what possible side effects are exposed through testing. The connection between the two will be explored by observing data on metrics related to user experience and overall performance. The A/B test provides a wide range of useful data in form of results, such as retention and estimated earnings per user. These metrics are significant for the commissioner in terms of profitability.

To enhance the readers' understanding on concepts dealt with in in this research and the structures around them, it might be necessary to browse through the concept map below before reading further. Figure 1.3 shows various concepts related to ad monetization in mobile gaming and how they intertwine with one another.

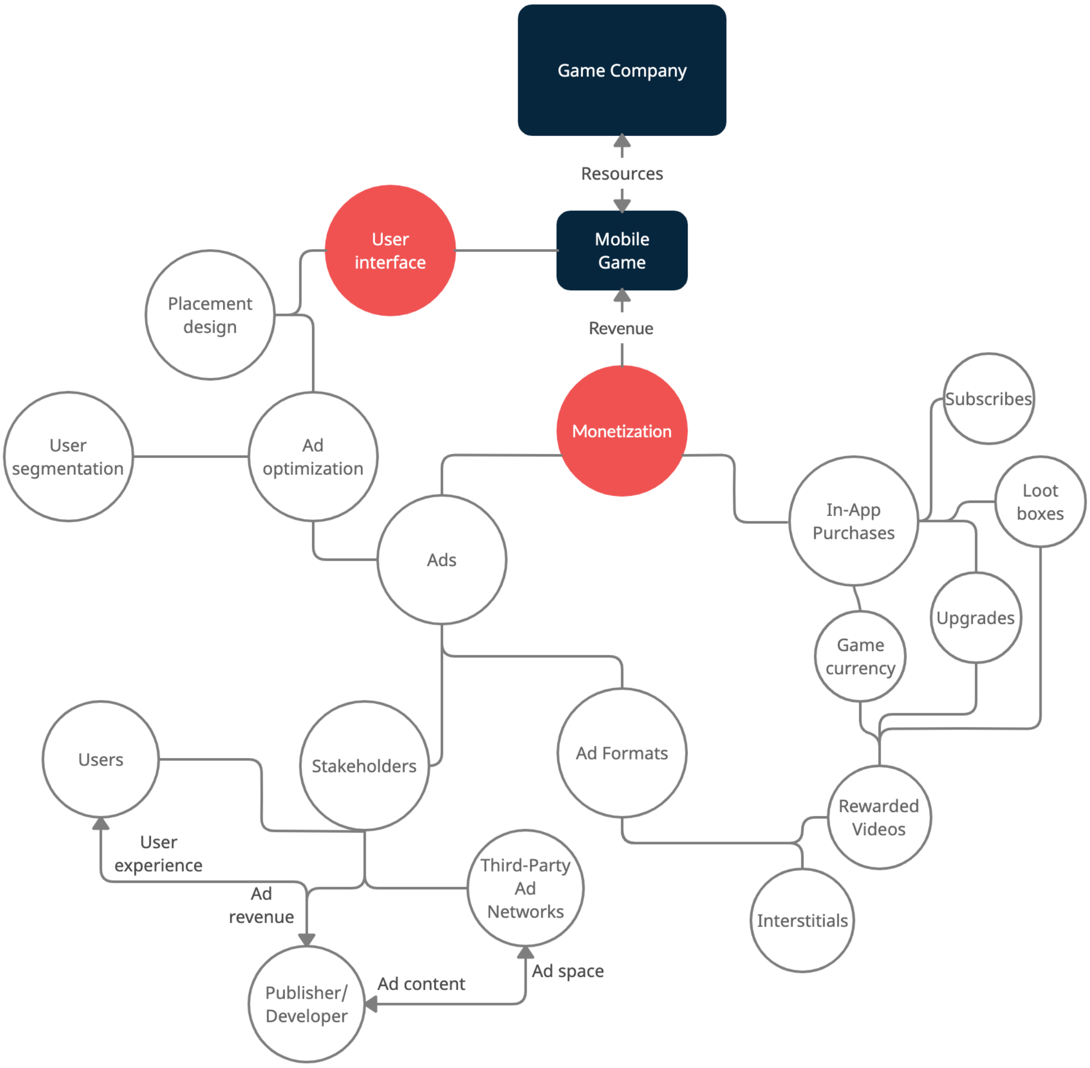

Figure 1.3: A concept map presenting themes and concepts related to thesis' topic.

2. ADVERTISING IN MOBILE GAMES

2.1. Market Size and Structure

The mobile gaming industry is changing and reforming at a fast phase as new trends and technologies are entering the market continuously. Despite the unpredictability, one statement about the industry can be confirmed: it’s growing. An article exploring mobile gamer statistics for 2021 shows that downloads experienced 18% of year over year growth in 2020. According to the writings, an average user spends up to 3 and a half hours of their daily time on mobile, and 11% of that time is spent on games.

Making purchases in mobile apps is on an upward trend as well. In the post, the author states that consumer spend on mobile games reached 100 billion USD in 2020 and is predicted to hit 138 billion USD a year by 2025. Currently, the majority of the revenue in 2021 is predicted to come from China.

The market growth is not originating only from increasing revenue per user, but also from the expanding user base. In 2020, the number of gamers globally was 2.69 billion, while the figure is estimated to reach 3.07 billion by 2023. The growth can be explained by the popularity of hyper casual genre – mobile games of this type are very simple, and the low threshold to playing them attracts users not necessarily identifying as gamers. (Knezovic 2021.)

One of the reasons why optimizing ads is so vital is the increasing engagement level of users. A great indicator of engagement is number of impressions per daily active user, which has in some cases tripled in the past 4 years. (Knezovic 2021.) Daily active user, often abbreviated to DAU, is a term used for referring to a consumer that engages with the product on the given day (Baremetrics). Therefore, the number of impressions per DAU is calculated by dividing the total number of ad impressions by the number of users opening and using the application during a day.

2.2. Freemium Model

Hill Climb Racing 2 is *freemium* in nature, which means that the content is monetized in two different ways: in-app purchases and ads. In-app purchases include subscriptions, ad-free application versions, game currency, special skins on vehicles and other premium content. However, the fact that the application itself is free and can be played without spending any money makes it freemium. The model's name is very self-descriptive as it combines the words "free" and "premium". In freemium, the premium content is only available in exchange for spending money or watching ads, meanwhile the free standard content is available for all users, regardless of their spending habits. (Lifewire 2020.) Freemium is a very common strategy for publishers to use, and games utilizing this model were already back in 2014 accountable for more than 90% of revenue generated from mobile apps (Draganov 2014, 12).

2.3. Hill Climb Racing 2

In reports such as one mentioned above, games are typically divided in different genres. The genre of focus in this thesis is racing, which mimics real-life racing, normally on various motor vehicles. Racing is not a very large genre in terms of consumer spend, since according to an article by Game Refinery, sports and racing together only generated slightly more than 5% of all in-app purchase revenues in 2019. In the top grossing-charts of three of the biggest mobile gaming markets - the US, Japan and China - only a few racing games made it to the top 500 list. However, out of all racing games, in the US, Fingersoft's Hill Climb Racing 2 was positioned at 7. (Kiiski 2019.)

The product this project is constructed around is Hill Climb Racing 2. The Hill Climb Racing sequel is a physics-based mobile game that falls under racing-category. The game was released in the end of 2016 and has since been downloaded more than 100 million times. Hill Climb Racing 2 has two different modes, which are adventure and cups. In addition, there are features where the player can join a team if they wish or take part in multiplayer events. (Fingersoft 2021.)

2.4. Ad Formats

Implementing ads is a very useful way of financially benefitting from non-spending users in an application, and they are also a very significant part in the monetization strategy of Hill Climb Racing 2. In this particular game, there are two different ad types, so called formats, in use. First of the two, rewarded videos, are always user-initialized and after completing, offer some sort of an advantage for the user. Rewarded videos work well for games in which game currencies, tickets or lives are a core component as they are some of the most common rewards offered. (The AdMob Team 2017.)

Users can be lured into watching rewarded videos by creating attractive *placements*. Placement refers to how and where the rewarded video prompt is shown to the user. A good ad placement is strategically designed and offers the user enough value to have them spend their time into watching one. Ad placements are either tied to physical locations in the user interface or events taking place when a user takes action within the game. A good example of a placement tide in a physical location is rewarded video offers in Hill Climb Racing 2's garage, where the user is occasionally allowed to upgrade vehicles by watching ads instead of spending game currency. Ad placements attached with events include for instance pop-ups appearing before a race and encouraging the user to watch the video to receive double coins from the upcoming round. (Cooper 2017.)

Second of the two is interstitials, sometimes called full-screen or display ads; they cover the whole screen and are usually videos or displays. Interstitials do not require initialization from the player; unlike rewarded videos they appear on the screen involuntarily. They do not offer a reward or any other sort of benefit to the user, but according to the guidelines, the user should be able to close the interstitial after a few seconds of watching. (Borgdon 2016.)

Interstitial ads don't require a similar placement design process as rewarded videos do since their nature as forced ads is to appear without prompt. However, their timing is crucial for user experience, and they shouldn't be displayed whenever. The general rule of thumb is to place them in transitions between levels or screen switches. This makes them feel less intrusive and a more integrated part of the flow. (Karnes 2019.)

2.5. Ad Networks

Third-party networks, also known simply as ad networks, are external ad providers often used in ad monetization. They normally operate by gathering a pool of advertisers seeking publicity for their products and services and buy ad space for their inventory from publishers. This process happening between an ad network and a publisher is called mediation. (Applovin 2021.)

Role of a publisher is to distribute the product, which in this context is a mobile application. The publisher isn't necessarily a synonym for developer, although it is common for a publisher to be the creator for the content they distribute. Publisher is also the party distributing ad content between users and third-party networks. (Adjust GmbH 2021.)

In the case of many publishers, as well as Fingersoft, the providing of ads is outsourced to third-party ad networks. Their role in monetization is to deliver ad campaigns from businesses advertising their products or services to publishers and therefore, to mobile users. The ad sources used by publishers are placed on a mediation platform where they compete against each other for ad impressions, which are events where a user has seen an ad. The implementation of an ad network typically requires a code integration to the target application in the publisher's end. (Applovin 2021.)

In most cases, ads are integrated in the application with SDKs, which is short for software development kit. They are external particles added to the code of the game and connect the game to external tech. (admonsters, 8.) Once an SDK has been added, the ad network needs to be added to a mediation platform.

2.6. Mediation

How mediation works in practice and is set up can vary depending on the platform used. The service provider used by Fingersoft is AdMob by Google, which will therefore be used as an example to explain the process. In AdMob, publishers can create mediation groups for each application and ad format. Mediation groups are linked with the publisher's application and ad networks with unique ad unit ID's. When a player is about to watch an ad, the information goes to AdMob, which sends an ad request to the mediation group corresponding the ad unit ID. See figure 2.6.1 for an illustration of how ad requests move between the application and the mediation platform.

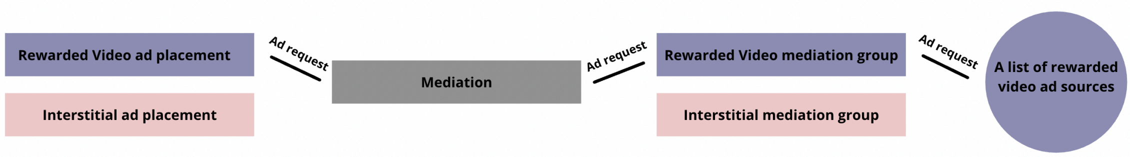

Figure 2.6.1: An illustration of the mediation process between an application, mediation platform and third-party ad networks.

Mediation groups consist of waterfalls, which are lists of ad source instances. Figure 2.6.2 displays a simplified version of what a waterfall might look like. An ad network can appear in a waterfall more than once as the idea in mediation waterfall is to set different price points to compete against one another – in waterfalls, the instances have fixed values, which means that their position is determined by their price. Highest bids are prioritized at the top of the list and the instances descend according to their fixed prices. After receiving a request, each ad network responds whether they're able to display an ad for the user in question or not. The bidding competition begins from the top of the waterfall. The first instance to be able to serve an ad will win the auction.

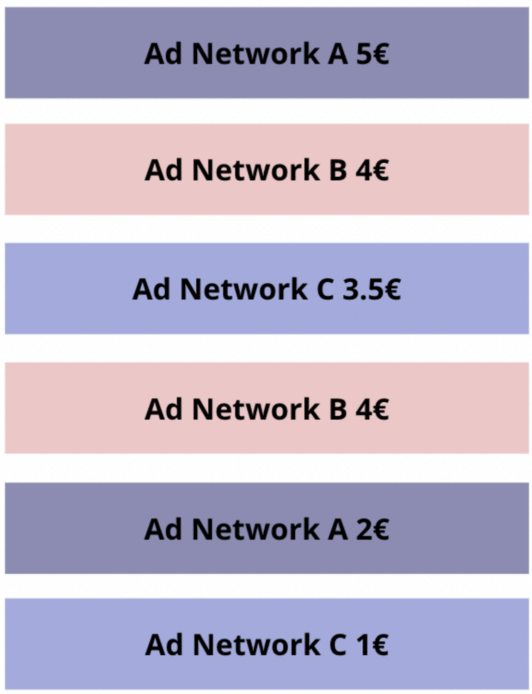

Figure 2.6.2: A demonstration of how networks and their bids are arranged in a waterfall.

Even when given the opportunity, ad networks can't always fill ad requests they receive. Typically, the match rate, which refers to the portion of ad requests an ad source instance is able to fill, is influenced by how well an ad network matches with the application in question. If an application has a big audience in a certain country, it's recommended to make sure potential partners target enough campaigns at that very location. If the networks' supply doesn't meet the publishers demand, there can be times when users don't see ads at all – due of this, it can be a good idea to have a variety of different ad partners. (Cohen 2018.)

In addition to the traditional waterfall mediation, some ad networks take part in bidding, which is a real-time auction method. As bidding happens in real-time, the networks included respond with their best possible offer at any given moment. In an auction, bidding networks are prioritized above waterfall so that once they have responded, the winner bid of this competition is placed in the waterfall with the priority determined by their offer. If the highest bidding offer is better than the highest available waterfall ad source instance, the bidding network wins the auction and fills the ad request. (Google AdMob Help Center.)

Figure 2.6.3 below shows the auction process. The bidding networks, which are placed on the far left in the image, are first given an opportunity to give their bid on the ad request received. The winner among bidding networks is then placed in the waterfall in a position defined by its value, and if it is higher than any of the waterfall ad sources, the bidding network wins and gets the ad impression.

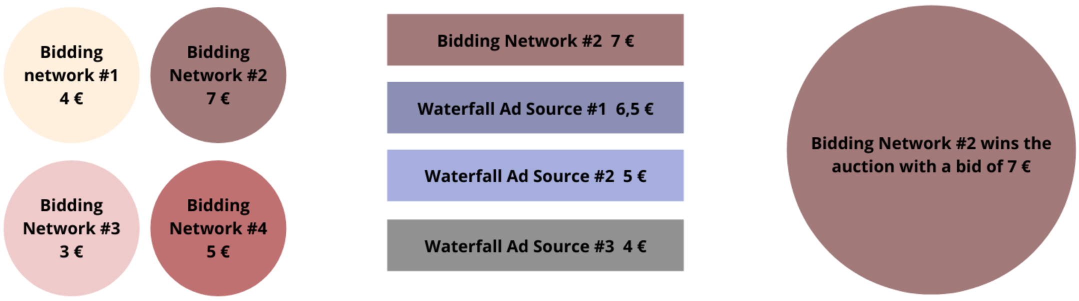

Figure 2.6.3: Image shows how bidding networks participate an auction. Network with the highest bid is placed in the waterfall to complete with waterfall ad sources.

If, however, there are items in the waterfall with higher offers than the winner of bidding networks, thy are first given an opportunity to fill the requests. If one of the waterfall ad sources placed above the bidding network is able to serve an ad, they win the auction. The figure 2.6.4 placed below describes the process in this case.

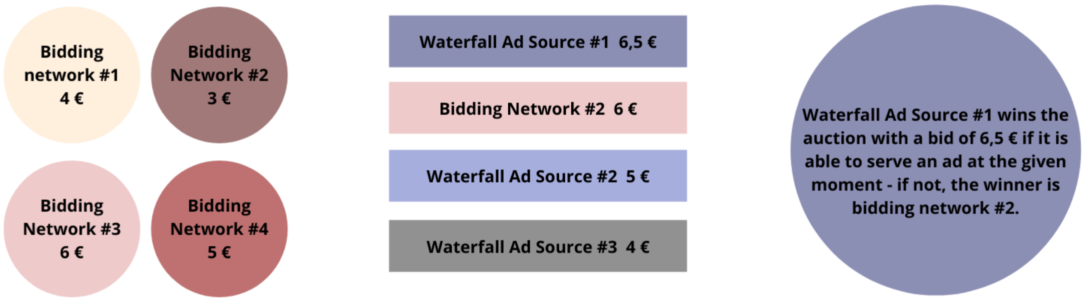

Figure 2.6.4: Image shows an auction process where a bidding source does not position at the top of the waterfall.

2.7. Ads' Optimization

There are various tools to optimize ads' serving with, and the two main principles are maximizing revenue from an impression and targeting the right audience with the right assets. The first one can be done by setting a suitable CPM. The offers in bidding methods are expressed in CPMs, which stands for cost per mille and shows the cost per 1000 impressions for a third-party ad network. For instance, if the CPM of an instance is 10 USD, thousand ad impressions cost 10 USD for the network, and one impression costs 0.01 USD.

There are various factors in digital advertising influencing the CPM level of a product. Geographical location and the economic state of the location are typically ones with a heavy impact. Countries with higher GDP and purchasing power are typically ones with more valuable audience for advertisers since the probability to spend is greater, which also results in higher CPMs. Geographical user segmentation is very significant for this study since the experimenting will be carried out for certain countries only.

Ad's targeting does also play a big role in the average CPM of the product. Publishers typically collect user data that can be used to serve ads matching the users' interests and therefore improve a chance of conversion. (Morrisroe 2020.)

Ad size impacts the CPM since physically larger ads are easier to notice and engage with, and the impressions are normally more expensive for advertisers compared to ads taking less space on the user interface - there are various other formats available to use for mobile advertising and not all of them are full-screen. (Morrisroe 2020; Borgdon 2016.) However, as the focus of this thesis is on types of ads used in Hill Climb Racing 2, it is not necessary to introduce all ad formats.

AdMob mediation platform offers a possibility to create multiple mediation groups for the same application and format. The benefit of having several groups for the same application is the capability to optimize, and an efficient way to optimize is to divide audience by their geographic location. When a publisher creates separate mediation groups with different geographical targets, it allows setting different CPM levels in different areas and thus, optimize every ad impression.

It is crucial for the reader to understand the logic behind geographical division on mediation groups, since the A/B testing used to acquire results will be implemented in particular markets. In the world of mobile ad monetization, dividing countries in different groups is quite a common practice. Third-party ad networks might deliver ads at higher CPMs in some countries, but when all countries are squeezed into one single mediation group, they might not reach their full potential in competition. (Wolf 2015.)

The geographical segment targeted in this process is so called tier 1-markets, which refers to countries with high CPM levels. The CPMs of a country can and most probably will always vary according to the product, format and mobile platform. The logic for building the list of tier 1-countries for Hill Climb Racing 2 will be introduced in the chapters describing the A/B testing process.

3. USER EXPERIENCE

User experience refers to a branch of study that aims at understanding how a user evaluates a product and its features. There are various tools available to measure user experience with and help in optimizing it - as can be noted from a report written about measuring user experience, some of the UX design tools are quite formal. The report identifies at least three widely known questionnaire tools that collect user feedback. (Djamasabi et al. 2014.)

3.1. Ads and User Experience

As the purpose of an ad is to be noticed and get engaged, their impact on user experience is significant. Among mobile users, the common attitude towards ads can be quite reserved. This is a repetitive theme appearing in the results from research conducted by Gui et al. exploring the aspects of ads that users care about. One of the research questions regards the distribution of star ratings among ad-related reviews compared to non-ad related reviews among apps included in the experiment. The report reveals that more than 33% of reviews related to ads were 1-star ratings, meanwhile the figure for non-ad related ratings was only 12.2%.

In the study, the authors investigate the type of negative, ad-related reviews target applications received. A relatively large portion of them were non-descriptive, meaning the user did not communicate any specific issue, but rather expressed general frustration towards ads. Among the descriptive reviews, typical subjects of complaints included crashes caused by ads, timing of ads, frequency, sizing, intrusiveness, or the nature of content. (Gui et al. 2017.)

However, as the industry has evolved, the position of ads as a part of mobile gaming experience has established. Nowadays, various organizations working with ads are continuously working to find best practices for the parties involved and keep on sharing their learning processes. The efforts have not been in vain since gamers increasingly engage with mobile ads. (Knezovic 2021.)

In mobile platforms and applications, the user interface is ideally designed player-first, and the same goes for mobile ads - having ads placed appropriately is of primary importance when it comes to user experience. It is smart for example to include ads in transitions, for instance when moving from main menu to another section of the app, or after a round is finished. If an ad pops up on the screen without prompt in the middle of browsing, it can be perceived as very irritating; interruptive ads are one of the most significant reasons for mobile users to use ad-blocking extensions. For rewarded videos, good placements are the ones seen as most valuable – when a user decides to watch a rewarded video, they exchange their time for the reward promised once the ad is finished. (Olker 2017.)

For the purpose of gaining better understanding and perspective on user experience and ads' quality, a quality assurance (QA) team representative was interviewed on the matter. The role of quality assurance in game development is testing the games and their features by playing them, and then evaluating and forwarding feedback inside the company. With ads, the testing process for each update is to go through each ad network with a couple of different operating software versions for each platform.

When asked about an ideal implementation of ads, they replied that in their opinion, ads are at their best when they're a well-integrated part of the game. The interviewee thinks that ad placements in game should be designed logically, and if the placement doesn't offer enough value, the users might end up not watching it or even getting annoyed with it.

"Nowadays, players are for sure used to seeing ads and accept them as a part of the game that has its own purpose," the interviewee says. "Players acknowledge that ads do also serve them – a user can save time by watching ads and progress faster. However, they're also used to the fact that ads don't always work perfectly. Some players say that if they've made a new record in the game and then get prompt to watch ad to get double coins, they probably won't end up watching them even though the benefit is good. They're scared to lose the progress if the ad happens to crash the game."

Considering the research question, the quality and proper functionality of the ad is emphasized here as opposed to designing ad placements. If there is a bug in the ad and it causes a crash or doesn't give the promised reward for some reason, the user can be very disappointed. In the worst case, issues of this type can cause users to stop using the application and decrease overall retention.

3.2. Metrics

Retention is a metric measuring the portion of users returning to the application after a certain number of days. Retention is presented with the number of days passed since installation or first app open event. For instance, D7 retention indicates percentage of users coming back to use the application after 7 days have passed from installation. The formula to calculating retention goes as displayed below in figure 3.2.1.

$$\frac{\text{\# OF USERS WHO OPEN THE APP THE N}^{\text{TH}}\text{ DAY AFTER DAY 0}}{\text{\# OF USERS WHO FIRST USED THE APP ON DAY 0}} = \text{CLASSIC RETENTION RATE}$$

Figure 3.2.1: Formula to calculate a game retention rate (Braze 2016.)

Retention is an excellent metric to use in testing user experience of different features in the product, not only ads. A great, non-ad related example in gaming context can be adding or changing a tutorial to the beginning of the first gaming session after installing the application. For instance, if the duration of the tutorial is shortened and D1 retention rate increases, it might indicate that the initial tutorial was too long for most of the players. Their first impression of the game might not have been that good and it turns out that the briefer tutorial offers just enough guidance for the user to start playing sooner (Braze 2016.) Figure 3.2.2 shows how a retention curve might look like for an application.

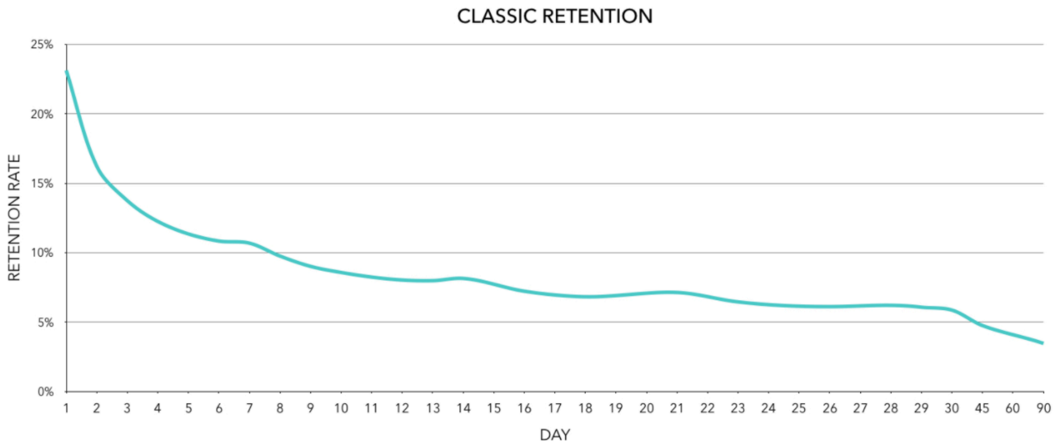

Figure 3.2.2: A graph displaying how game retention evolves as days pass (Braze 2016.)

Retention is not the only metric to measure user experience with, although it is an important one. There are a few others used in this research as well, which include for instance rewarded video completion rate and match rate. Rewarded video completion rate is calculated by dividing the number of total impressions with the total number of completed rewarded videos and it shows the percentage of how many started rewarded videos were finished. Meanwhile a low completion rate can expose an issue with an ad network, match rate rather shows how good of match the ad networks' campaigns are with the audience available. In both cases, a higher rate is a good sign. (AdMob Help; Guanziroli 2020.)

4. QUALITY OF AD NETWORKS

4.1. Definition of Ad Network Quality

Among mobile ads, quality is a very subjective term, and how it is measured is simply a matter of from whose perspective the quality if observed. As the main purpose of this research is to explore ads' impact on user experience, the ads are viewed from the user's point of view. Therefore, a high-quality ad is one that runs without errors, without causing crashes or other disruptions in the gameplay.

An ad with a nice design and visuals may also be considered high-quality from the user's point of view. However, since ad networks typically provide their publisher partners with multiple ad campaigns at once, there is no certainty of what individual ads are shown to the users. Of course, there is often a possibility to opt out from unwanted categories (for instance alcoholic beverages or dating apps), but it is not the typical case to browse through ad campaigns and manually pick ones to include from each partner. The publisher is not always in control of which campaigns are shared with their users before they're displayed. Of course, if certain ads are receiving a ton of complaints and continuously cause problems in the game, ad networks can be contacted and asked to either fix or filter out the campaign.

The QA representative from Fingersoft interviewed also on user experience, was asked questions about ad quality and ad-related issues. When asked about ad- and ad network quality overall, the interviewee states the following: "It's important that the quality of the ads would meet the quality of the whole game, since your average user does not know that the ad is externally sourced – from their point of view, it's still a part of our game. The SDKs, adapters, and all of those are external parts, but the player doesn't know that. Players see that the ads are ours and if they don't work, our game is low in quality."

4.2. SDKs

SDKs play a prominent role in the quality delivered by third-party ad networks. SDK is an acronym for software development kit, and it is a tool for connecting an app to external tech. In mobile applications, SDKs are used to connect apps with ad networks and enable them to deliver ads. A mobile app ad quality report constructed by admonsters does a great job describing the issues SDKs can cause and how integrating them gives away control over managing them. According to their playbook, problems possibly caused by SDKs impacting user experience include increased crash potential and latencies in loading ads. (admonsters, 8.)

4.3. Ad Issues and Guidelines

Unfortunately, issues with ads are quite common in mobile gaming. Sometimes, for instance, rewarded video ads might fail to offer a reward once completed, or an interstitial might be missing an exit-button. When these types of issues arise, it's likely that they're due to a problem in the ad network's end instead of publishers.

"The worst type of problems are the ones that crash or freeze the whole game, or the ads that freeze themselves," Fingersoft's QA representative says. "In these cases, the player has to turn off the whole application in order to proceed and those are the issues receiving the most complaints. Another typical issue is that the player watches a rewarded video but doesn't receive the reward they're entitled to. In my opinion, these are the most severe issues from the user's perspective."

The interviewee says that there can also be some random crashes, or even situations where the application doesn't start at all because of an ad network issue. This can happen because even if the user hasn't seen an ad yet, the ad is already being loaded beforehand in the background. In these incidents, it's virtually impossible for the user to associate the problem with ads, so they're likely to assume that problem is in the application itself. Events like these can have a very negative impact on the overall user's perception of the game.

According to the QA representative, ad networks do not follow the same standards with the ads they show, so it can be overwhelming for the player to see an ad that lasts much longer than what they've used to. Audio settings can also be an issue – for example on iOS, even when the device is on mute, some ad networks don't obey the device setting but the ad needs to be muted from the ads' settings. It can be an unpleasant surprise when a video ad starts playing on a very high volume even if the game itself is on silent.

The interviewee explains, that Fingersoft hasn't set official standards for their partners, but the evaluations are rather done case by case. The main objective is to only include ads that don't crash or freeze the game and that give out rewards as promised. The QA representative express that they still do receive complaints about issues such as too long durations, especially with interstitials as they are forced ads. "5 seconds has always been pretty much a standard duration before skipping and something users are used to, so a longer ad can easily upset a player."

4.4. Ad Network Performance

When evaluating ad networks, metrics such as CPMs and match rate are usually taken into consideration. It's important for the publisher to have ad networks that can deliver competitive CPMs, but they only indicate good performance from the publishers' perspective. The metric is not directly linked with user experience, but its contribution to the overall ad performance of a company is noteworthy. Even though the CPM levels and earnings are not the main metrics measured, results on their fluctuation are a very welcome by-product of the A/B test.

The networks' ability to deliver ads is important for user experience. Match rate shows the potential of a network to fill ad requests, meanwhile fill rate tells the actual percentage of ad requests the network in question has filled and therefore indirectly communicates the competitiveness of the network. There can be differences in how fill rates are calculated and how they're called – AdMob, for instance, uses the term *show rate.* In this thesis, fill rate is calculated by dividing the number of requests received by an ad source with the total number of impressions. (Digital Limbo 2019)

As it was described in part 2.6 explaining the mediation and auction processes, a poor ability to match ad requests does not always indicate that the network is not good, but rather that it might not be a good match to the application due to different audiences. Even though low fill- and match rates do not directly mean bad quality, they can still negatively influence the user experience, especially in case of rewarded videos. If there are no ads to serve at the given moment, it might feel disappointing to a user who wants to speed up their progress in the game by receiving rewards from watching ads.

In addition to metrics, performance in terms of functionality of the ads served should be noted. According to the QA expert, a good ad network serves ads that work intuitively. In case of ads, intuitively means that the ads work however the user most likely expects the to. For instance, if the user is indicated to swipe or tap the screen, they're not redirected to an app installation page as a result. Another aspect would be the duration – users are familiar with certain lengths in ads, and anything longer can cause frustration.

4.5. High-quality Ad Networks

To reach a conclusion and find an answer to the research question, one of the tasks is to evaluate Fingersoft's partners. The intention is to find ad networks that would classify as high-quality and experiment how using them only would impact metrics observed. There are some universal ad network comparisons available in online sources, but instead of using a general ranking, it can be more efficient to use a list of networks that have been recognized as well-functioning by Fingersoft.

As the ad networks included in the research are actual partners with Fingersoft and their performance is confidential, they will not be announced, but referred to as network 1, network 2, network 3 and so on. The decision of which partners to include was based on the experience from the publisher's side on working with ads. Altogether, there are 5 networks included in this premium-category built for the purpose of this research.

5. A/B TESTING

5.1. Definition of A/B Testing

A/B testing is a tool which compares two variants of a product, service or a feature and finds out which one works better. Depending on the feature tested, better can mean greater revenue, retention rate, number of ads watched per user or number of in-app purchases made. To get the best and most confident results, there should only be one aspect altered: if there are multiple differences between the variants used, it can be hard or even impossible to identify which one has yielded results.

In an interview with a data engineer from Fingersoft, they stated that especially when the metric tested is predicted to remain quite close to the baseline, it is necessary to have an audience big enough to reach sufficient exposure. Very small changes can be difficult to detect if the audience is too narrow – the audience size required for a confident test result is often tens or hundreds of thousands.

“It should be kept in mind that the result is never 100% certain”, they say. “In practice, there should be an infinite number of users to be entirely confident with the outcome, which is of course not possible. Normally, it is accepted at the end of these tests that there is a 5% possibility that the outcome is opposite of the actual test result. This is based on statistical probability. If the test does not gain enough audience or if the test is not run for long enough, the confidence rate calculated by Firebase might be 80, for instance. In this case, the assumption is that with a probability of 80%, the outcome of implementing the change is same as during the test, but there is a 20% chance that our conclusion is incorrect. Patience is important when A/B testing, as it takes time to get enough data from players.”

Depending on the product, A/B testing can vary a lot, and in mobile applications, it can be done with an A/B Testing tool. The tools used for this test are provided by Google.

5.2. A/B Testing Tools

5.2.1. Firebase

Firebase is a platform used by app publishers and developers. It is a Google-owned product, made for developing applications in mobile and web platforms. It offers tools for building the application, monitoring releases, and observing and improving engagement. A/B-testing with Firebase falls under the engagement-category and is done taking advantage of remote config. (Google Firebase Documentation.)

5.2.2. Remote Config

Remote config is a tool used to apply fixes, changes, and other updates in almost real time. The difference to major updates and releasing them is that there is no need for republishing the application, neither a need for the user to update the application in their app store. For instance, if there is a mistake in translation or other small error in the app, it can be corrected in a matter of minutes. With remote config, apps can also be customized for different user segments. (Google Firebase Documentation.) When running A/B tests with remote config, some of the results can be seen directly from Firebase, but a broader selection of data can be fetched afterwards from BigQuery.

5.2.3. BigQuery

In an interview, Fingersoft's data engineer explains that BigQuery is a Google-owned service and a part of Google Cloud, a bigger assembly of various products. "BigQuery is an SQL (Structured Query Language) based database, which consists of the database itself, as well as a tool for conducting searches and generating results by utilizing SQL. We get the data in BigQuery from FireBase, which comes directly from our games. The data in BigQuery is so called raw data, which goes through a process at SuperScale to reform the data into a readable format," they describe.

According to the data engineer, SuperScale is an external service provider in possession of their own orchestration too for data processing. After the process, the data can be filtered in BigQuery to find users and events corresponding the criteria. Some examples of these events and their details include how long did a player drive, what vehicle did they use, why did the race end and so on. SuperScale is used for the games in Hill Climb Racing-series.

5.3. Preparing for an A/B Test

In practice, the A/B testing requires a new ad unit ID for the B-variant's mediation group and setting up the technical side of remote config so, that the testing is possible. Once this technical part is done, the specifics of the test need to be defined in the Firebase console. The information required at this stage were the target countries, ad unit IDs, exposure and metrics tested. In this context, the exposure means the portion of users in the target segment that will be included in the test group.

Before the actual test, there was a preliminary test to ensure a smooth testing process. The mini test was targeted at smaller audience with a 20 % exposure and ran for a week, from Wednesday the 1st of September until Tuesday the 9th of September. The metrics tested were estimated total revenue and different retention rates: D1, D4-7 and D14 retention.

The preliminary experiment was run tun ensure that the technical aspects of the test were properly set up and did not cause any issues and that the test will deliver sufficient results to reach a confident outcome in the end. The test ran for a bit more than a week and everything went as hoped. The exact date range for the mini test was from 1st of September to 9th of September. The exposure of the test was quite low on purpose. In addition, the duration of the test was rather short. Typically, A/B test are recommended to run for several weeks, if not months, to reach satisfactory outcomes.

Before the actual test, a notification was sent to a contact person at SuperScale in order to have the data available later in BigQuery. Once they have been informed on the schedule of the test, it's good to go. Some of the test results from the A/B test will be updated to Firebase in real time, and the rest of the data can be accessed later using SQL.

5.4. Starting and A/B Test

The actual A/B test was started on the 14th of September, and the exposure for this test was 50% of all new users in the countries included. The list of countries is confidential information and will not be revealed, but by reading about geo-segmentation in part 2.7 and the description below, the reader can build a good understanding of what types of countries may be included.

The process of building the list of countries included observing the CPM levels and impression amounts in each country. All countries in the list above reached a certain limit of impressions in the past 7 days. The minimum impression amount was set because some countries, even if well off in terms of CPMs, might have such a small user base for Hill Climb Racing 2 that their overall input is not as significant. From the countries left, the top 15 of highest CPMs were included in the tier 1-list.

The idea of A/B test is to compare the impact of using different sets of ad networks. Baseline, which is the waterfall setup the mediation group currently has, will be put against variant B - a similar waterfall for the same audience, but with only so-called premium networks included.

Variant A (Baseline)		Variant B	
Premium networks	Network 1 (Bidding)	Premium networks	Network 1 (Bidding)
	Network 2 (Bidding)		Network 2 (Bidding)
	Network 3 (Bidding)		Network 3 (Bidding)
	Network 2		Network 2
	Network 4		Network 4
	Network 5		Network 5
Other networks	Network 6 (Bidding)		
	Network 7		
	Network 8		
	Network 9		
	Network 10		

Figure 5.4: An illustrative list of networks in both variants.

As can be detected from the figure 5.4, some of the networks are bidding only, meanwhile there are networks that appear in both waterfall and bidding. Among the networks taking part in the A/B test, Network 2 is the only one included in both categories. Network 1 and Network 3 are solely competing on the bidding side. It is crucial for the objective of the study to have Variant A set up similarly as the waterfall normally would be – if the setup was changed from the current and for instance only included the non-premium networks, the changes detected would not be of any use. This approach would probably more efficiently display the single ad networks' impact on metrics measuring performance and user experience but would not be of help with resolving whether the set of premium networks is better than the standard selection applied.

The A/B test will be run for the Android version of Hill Climb Racing 2. The reason for choosing this application is a greater audience -compared to the first Hill Climb Racing game, HCR2 has a bigger portion of its audience in tier 1-regions. Out of Android, Amazon, and iOS platforms, Android has the highest potential to deliver confident results, thanks to the largest userbase. In the target application the ratio of revenue from in-app purchases and ads is around 50/50, which means that role of ads is quite significant.

Testing will run for one format only, which is rewarded video. Rewarded videos were chosen as the target because quality issues in rewarded videos can have more impact on user experience. Since there is an incentive to watch an ad and the user is looking forward to receiving a reward, a crash or other error might lead to greater frustration compared to interstitials.

Rewarded videos can also potentially reach more users: all in-app purchases offer an interstitial-free experience, meanwhile rewarded videos can be watched by all users. Considering that only new users are included in the experiment, they might be more inclined to rely on rewarded videos to boost their progress as opposed to experienced long-term players who can be more likely to make in-app purchases.

6. RESULTS

The A/B test ran for a total of 30 days. The experiment was under observation on a daily basis to keep an eye on the development of the results throughout the testing period. After the 30 days, it was deduced that it no longer serves the purpose of this research to keep the test running.

During the 30 days, the test gained a total audience of around 462 thousand users with the exposure of 50% in the selected countries. For both variants, the audience was gathered evenly from each area – for instance, the number of users targeted at the United States is quite similar for both variants. As the purpose of this section is to solely present the results received from the A/B test, they will be analysed and explained more thoroughly in part 7.

6.1. Division of Traffic Among Networks

Among the networks, traffic allocation is highly dependent on the competitiveness of each ad network and how well their supply meets the demand. The share of voice-comparison below was conducted to compare networks with each other's is based on the total number of impressions during the testing period. The data used to make the calculations was acquired from AdMob reports. To see the share of voice for the networks included in Variant A, see figure 6.1.

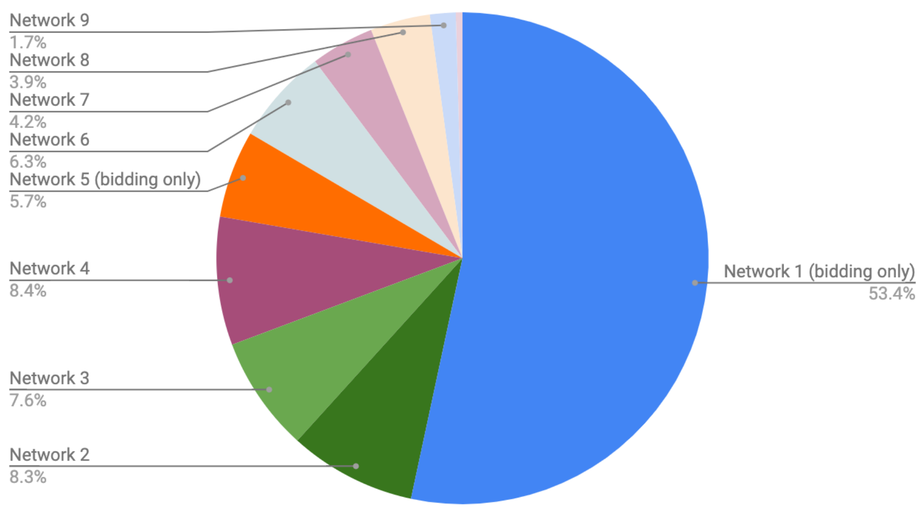

Figure 6.1.2: Traffic split among set of networks in Variant A, sorted by impressions received by each ad source.

Similar data for Variant B is presented in figure 6.1.2. The reader should note that in these graphs, the Network 2, which appears in both waterfall and bidding, only has a single slice in which the impressions from both categories are summed as a total.

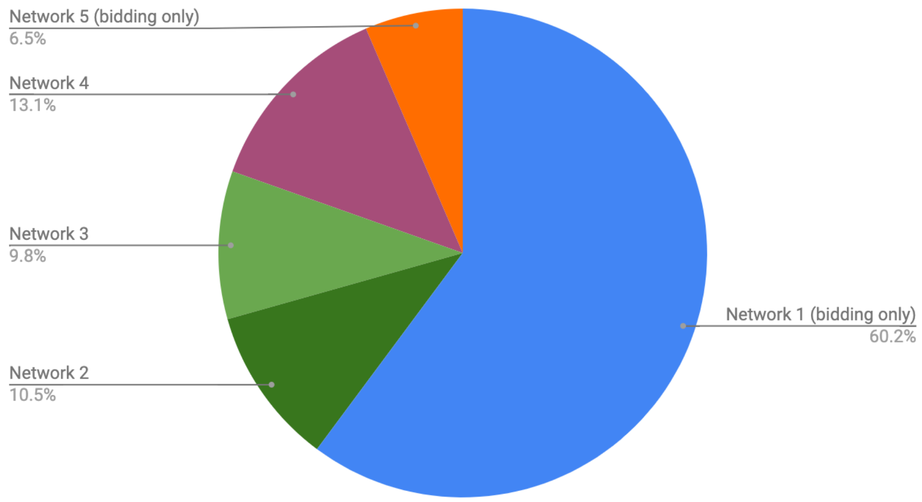

Figure 6.1.2: Traffic split among networks in Variant B.

6.2. Bidding/Waterfall Split

As there are both bidding and waterfall ad sources included in the process, it can be useful to see how the split between the two goes for both variants. For Variant A, the split is around 40/60. The allocation between traffic coming through waterfall ad networks and bidding networks is displayed below in figure 6.2.1.

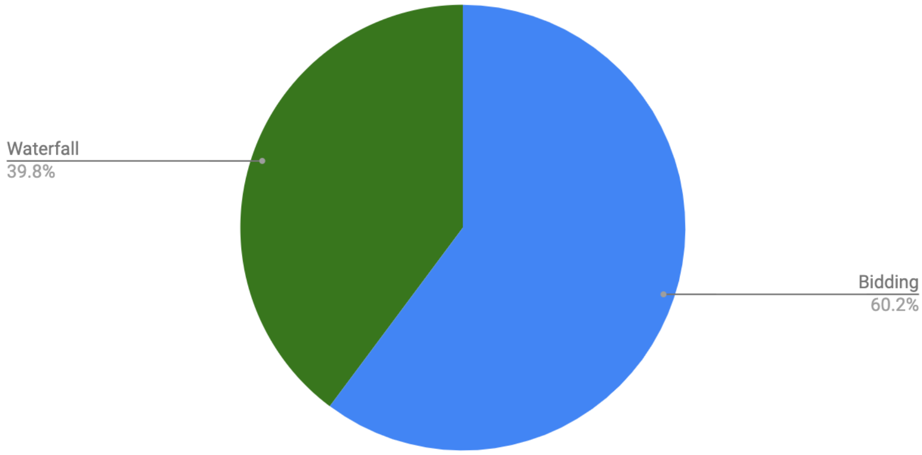

Figure 6.2.1: Share of voice in impressions comparison between waterfall- and bidding sources for Variant A.

Compared to Variant A, ads' traffic in Variant B seems to have more emphasis on the bidding side. See figure 6.2.2 below to detect the development in how traffic is split between these two categories. As there is a clear difference in how real-time bidding networks and waterfall ad sources are filling ad requests in variants, it can be worthwhile to observe what differences there are between the two bidding methods in other variants as well.

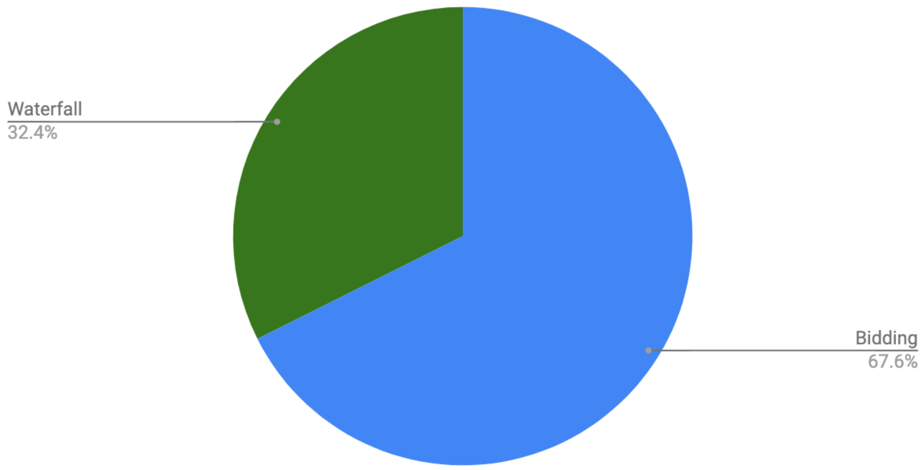

Figure 6.2.2: Share of voice in impressions comparison between waterfall- and bidding sources for Variant B.

6.3. Metrics Measuring User Experience

User experience is the main point of interest in the research, and therefore the results showing development related to it are of primary importance. This section will reveal what type of changes there were between the variants during A/B testing on aspects potentially improving or degrading user experience on Hill Climb Racing 2.

6.3.1. Rewarded Video Completion Rate

Rewarded video completion rate represents the number of times users competed watching a rewarded video they opted in for. If a user decides to exit a rewarded video mid-watching, it could be because they initially clicked themselves to the ad by accident, or perhaps they changed their mind during. There are various reasons to why a user would change their mind – perhaps the ad froze, maybe the duration is too long or another ad-related issue. Therefore, it's reasonable to believe that this metric is also a useful indicator of user experience with the ad displayed.

To calculate the rewarded video completion rate, daily data on impressions and completed rewarded ads were fetched from AdMob mediation report. Once the data is gathered and arranged on the sheet, the number of completed rewarded ads were divided by the number of impressions on the given day and the converted into percentages.

Table presented in figure 6.3.1, Variant B-column represents the rewarded video completion rate among users that were directed to the segment with only high-quality networks included. The data on the right-hand side column shows the rewarded video completion rate for mediation groups with all ad networks included. For clarification, data on both columns is from Hill Climb Racing 2 Android's rewarded video and from the same audience geographically.

Date	Variant A	Variant B
2021-09-14	80.42%	97.41%
2021-09-15	79.45%	98.81%
2021-09-16	80.14%	97.38%
2021-09-17	79.04%	94.85%
2021-09-18	78.95%	95.01%
2021-09-19	80.60%	96.88%
2021-09-20	79.41%	90.50%
2021-09-21	79.11%	82.34%
2021-09-22	77.48%	80.82%
2021-09-23	77.88%	80.86%
2021-09-24	80.13%	81.28%
2021-09-25	79.79%	81.44%
2021-09-26	77.85%	80.04%
2021-09-27	78.13%	82.91%
2021-09-28	77.87%	82.06%
2021-09-29	77.88%	81.56%
2021-09-30	76.44%	79.88%
2021-10-01	76.30%	82.14%
2021-10-02	77.03%	83.96%
2021-10-03	77.83%	84.56%
2021-10-04	79.34%	85.17%
2021-10-05	78.71%	83.19%
2021-10-06	77.29%	83.45%
2021-10-07	77.17%	83.69%
2021-10-08	76.09%	84.00%
2021-10-09	75.24%	82.34%
2021-10-10	75.34%	82.65%
2021-10-11	77.18%	85.28%
2021-10-12	78.90%	86.16%
2021-10-13	82.60%	84.36%
2021-10-14	82.29%	84.24%

Figure 6.3.1: Rewarded video completion rate for Variant A and B throughout the A/B test.

The table is further analysed and processed in section 7.3 and visualized into a graph for a clearer picture of how the metric developed through the testing period. For clarity on how significant the difference is, the average rewarded video completion rate for Variant A was 78%, meanwhile the same figure for Variant B was 85%.

6.3.2. Retention

Retention has fluctuated throughout the experiment – once the test started receiving enough data to put any results together, the Variant B with high-quality networks seemed to be outperforming baseline, but the roles have switched by the end of the test. However, the differences have remained very small throughout the period. The absolute retention figures will not be presented as it is not necessary for the purpose of the thesis and the aspects playing the key role are the differences between test groups.

Day 1 retention rate represents the number of users who returned to the application one day after they first started using it. During the test, the day 1 retention of variant B remained very close to the baseline, and the difference between the two was less than 1% the whole time. In the end of the testing period, the percentual difference of variant B from baseline (variant A) was -0.49%. During the experiment, the difference varied between -1.26% and +0.31%. On most days, variant A performed slightly better compared to variant B.

The graph in figure 6.3.2.1 shows how the day 1-retention developed in the process. The curve for variant A is grey and included in the same image, but as can be seen, the variants are so close to one another that it's hard to detect separate lines in the picture.

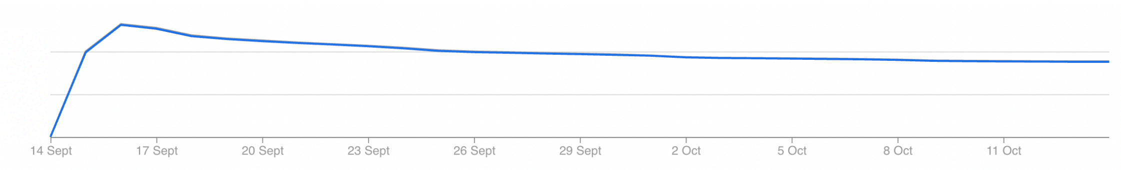

Figure 6.3.2.1: Day 1 retention.

The second metric represented in Firebase is retention for 2-3 days since first launch of the app, which is shown in figure 6.3.2.2. Like in day 1-retention, the differences between variants were very small. On most days, variant B outperformed the baseline with a tiny gap – the variation between groups was from –0.49% to +1.15%. Again, the gap between variants is so small that it is virtually impossible to detect in the graph below.

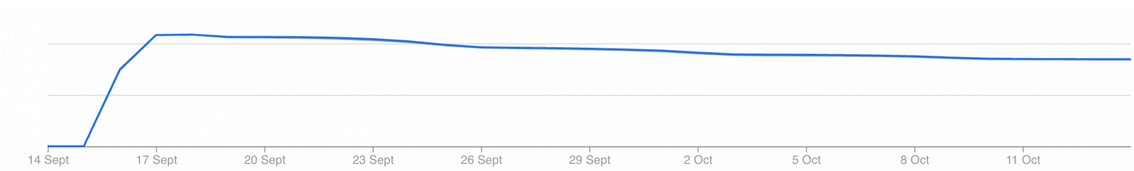

Figure 6.3.2.2: Day 2-3 retention.

As the curve below (figure 6.3.2.3) shows, retention for 4-7 days had a slightly different development in comparison with the two previous ones. To clarify, as all graphs display retention data from the same time period, the flat spot in the beginning lengthens as we proceed to a higher number of days. This is because in the first three days, there are no users who have retained on the 4th day after first launch. In the graph, logically the first retained users are visible on the 4th day of testing.

4-7-days retention rate has experienced the same phenomena as 2-3-days retention – throughout almost the period, variant B has been performing slightly better, the difference from baseline ranging between –1.19% and +0.53%. However, in the last 4 days of the test, retention on variant A squeezed past variant B.

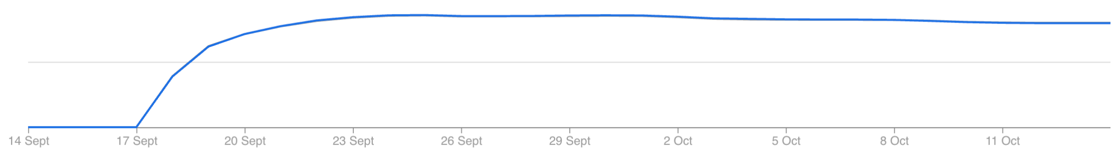

Figure 6.3.2.3: Day 4-7 retention.

Results for 8-14-days (in figure 6.3.2.4) retention were very similar compared to other figures – variant B performed slightly better than variant A, and on the last day of testing the difference from baseline is +0.11%. The range in difference has fluctuated from –0.82% to +1.05%.

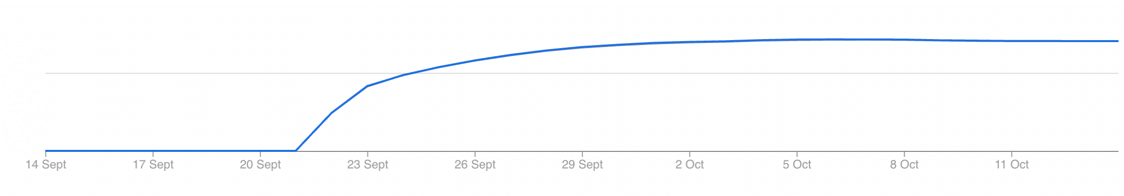

Figure 6.3.2.4: Day 8-14 retention.

Last in the retention category is the users retained after 15 days. Like with other retention figures, the gap has been very small throughout the testing period. The difference from baseline has varied between –0.88% and +1.37%, and on the last day of testing, the difference was 0.24%.

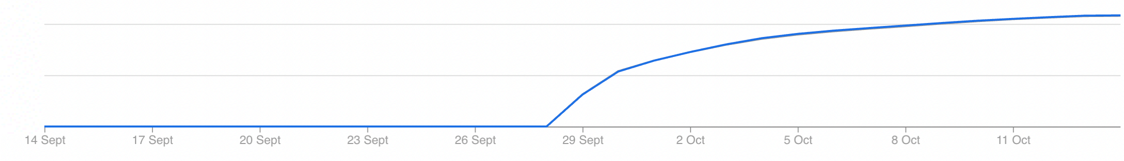

Figure 6.3.2.5: Day 15+ retention.

6.3.3. Fill- and Match Rate

When looking at changes in impressions, it is worthwhile to peek at *fill rate* and *match rate.* Match rate has already been introduced in this research, but to revise; it shows what percentage of ad requests it would be able to serve if they won an auction. Figures in fill rate column are calculated by dividing the number of impressions with the total number of requests received by the network. Purpose of the tables presented below is to demonstrate the change in networks' performance once numerous networks have been removed from the waterfall. The networks 6-10 are not included in the comparison even in Variant A as they do not exist in Variant B. Therefore, their performance cannot be compared to their past performance and enclosing their match rates would be irrelevant.

Figure 6.3.3.1 below shows figures for Variant A, and figure 6.3.3.2 displays corresponding values for Variant B. As one can see, both fill- and match rates are slightly higher for bidding networks in comparison to waterfalls, and the impact can also be detected in the distribution of impressions among the two bidding types. However, there are no individual networks standing out with their development to a direction or another.

Variant A	Fill rate (%)	Match rate (%)
Network 1 (bidding only)	48.89%	85.61%
Network 2	0.89%	1.95%
Network 3	0.94%	1.97%
Network 4	1.14%	2.16%
Network 2 (bidding)	0.85%	1.94%
Network 5 (bidding only)	7.37%	11.88%

Figure 6.3.3.1: Fill- and match rates for premium-classified networks in Variant A, where also networks 6-10 are included in the waterfall.

Variant B	Fill rate (%)	Match rate (%)
Network 1 (bidding only)	49.40%	86.12%
Network 2	0.78%	1.74%
Network 3	0.99%	1.95%
Network 4	0.89%	1.59%
Network 2 (bidding)	1.12%	2.45%
Network 5 (bidding only)	7.89%	12.50%

Figure 6.3.3.2: Fill- and match rates for premium-classified networks in Variant B.

When calculating the changes in average fill- and match rates between the two variants, there can be a pattern detected in their trends when networks are separated in waterfall and bidding. It looks like on both aspects, waterfall ad sources' rates are decreasing, meanwhile the bidding sources are growing their performance. Once numerous ad networks were removed from the setup, the change in fill rate is 2.28% for bidding and –10.6% for waterfall ad sources. The figures for match rate are +1.65% for bidding and –13.16% for waterfall.

6.4. Metrics Measuring Overall Performance

As the testing results gained do not only regard user experience related issues but extend to a great deal of additional metrics as well, it is worthwhile to take a look at the overall impact on target product. If the A/B testing indicated that Variant B clearly outperformed baseline in terms of retention, changes should probably not be implemented if results show that it harms the profitability of the product, for example. Even if the results on user experience were neutral, it might be more effortless of an approach to drop some networks – in this case as well the changes could end up in decreased revenues. It's vital to look at the overall impacts the experiment has revealed as a factor left undiscovered could harm other aspect of the game.

6.4.1. CPM Fluctuation

Changes in CPM are necessary to look at, as they play a key role in the economy of the whole company. The table in figure 6.4.1 shows how much the average CPM of each network has changed in percentages, and by comparing networks with each other, it's seen that the pattern has been quite different for bidding- and waterfall ad sources. In average, the change in bidding CPMs has been –11.22%, meanwhile the CPMs for waterfall items increased by 1.6%. The changes have been calculated from absolute values acquired from AdMob reports, which unfortunately cannot be displayed in the thesis and therefore there is no similar table for Variant A as figure 6.4.1.

Variant B	Change in CPM (%)
Network 1 (bidding only)	-12.83%
Network 2	18.37%
Network 3	-8.39%
Network 4	-5.18%
Network 2 (bidding)	-22.06%
Network 5 (bidding only)	1.23%

Figure 6.4.1: Changes in CPMs in Variant B in comparison to Variant A.

6.4.2. Revenue per User and Lifetime Value

The graph below in figure 6.4.2 shows the development of total revenue gained per new user. Total revenue per user takes all revenue sources into account, not only the revenue gained from ads. Due to confidentiality, the exact numbers will not be presented, but the focus will rather be on the difference. Compared to retention data, in revenue per user the gap between variants is a bit wider in the beginning of the test, with Variant B taking the lead. However, the curves seem to approach each other more and more the longer the test runs. On the last day, variant A has surpassed variant B, which now has a difference of –0.47% from baseline. The range of difference from baseline has fluctuated from –2.2% to +1.3%.

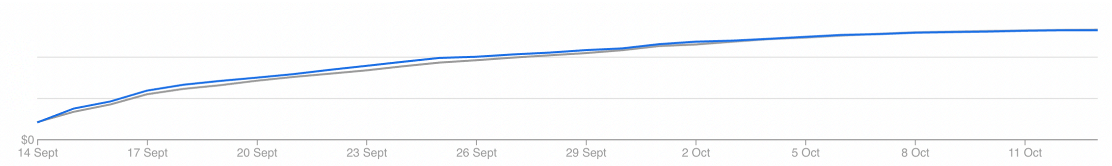

Figure 6.4.2: Average revenue per user curve.

Moving on from test results received directly from Firebase and AdMob, let's look at the figures extracted from BigQuery. As explained earlier BigQuery utilizes SQL (Structured Query Language) for generating tables and composing queries. SQL is designed to handle data by querying tables, and a query should return a result set containing data that was requested. The table of A/B test result was constructed by a data engineer at Fingersoft, and the table consists of a big number of rows including information on the actions users have taken inside the game. The data accessible is limited in the parameters displayed in the table used.

The objective of the first query was to find out the average of LTV revenue. LTV stands for *lifetime value,* and it is a key performance indicator in mobile gaming referring to an estimate of the revenue a user is predicted to bring in. The formula to calculating LTV is ARPDAU * Lifetime – in the equation, ARPDAU stands for average revenue per daily active user (LTV, Applovin.)

As with the previous results, neither are any actual LTV values going to be displayed. Instead, the difference is expressed in percentage difference between the two variables. The LTV revenue for Variant B was 8.06 % greater when compared with Variant A. One should bear in mind that this is not ads LTV, but an overall figure taking also IAP's into account, which means that several factors can play a role in why there is a difference.

6.4.3. Events

In addition to the differences in LTV, *events* can deliver important results from the experiment. Events provide the publisher with information on what is happening in the application. Examples of events include user actions or errors. Events can be defined and created by the developer according to what they perceive as relevant.

In terms of Hill Climb Racing 2, the events explored to potentially benefit from results were for such as network error, reward received, IAP purchased, app removed, and ad failed to load. For example, if Variant A (baseline) had 3000 network error events during the experiment but Variant B only had 1500, it would indicate that with the set of networks in Variant B generate less errors and might be able to deliver a better user experience. However, none of these relevant metrics showed any significant differences between the two variants; the split in terms of a single event between Variant A and Variant B was very even. Therefore, the event data gathered from the experiment so far will not be considered a relevant source for results.

7. ANALYSIS

The purpose of this section is to discuss the results gained from the Firebase experiment. The following paragraphs focus on interpreting the findings presented in part 6, one by one, and proposing more or less potential reasons for the outcomes. Staring with the traffic allocation among networks and referring to the graphs illustrating how the split of traffic changes when number of networks are reduced, it looks like the share of voice per network is growing in the same proportion. However, the charts displaying the split between waterfalls and bidding, it seems that bidding is taking the larger portion of traffic.

7.1. Traffic Split Among Ad Networks and Bidding Types

Considering shares on all aspects, some of the ad networks included are very dominant in mediation, and filling and matching ad requests are very significant contributors to a good performance. A high fill rate can communicate a good win rate in auctions, meanwhile match rate indicates ability to show suitable ad campaigns to the target audience.

7.2. CPM Fluctuation and Fill- and Match Rates

In Variant B, total share of bidding grew by 12.2%. Considering that fill- and match rates also increased for bidding in this variant, this could potentially positively impact the performance in terms of user experience. As explained earlier, having no ads available when a user wishes to watch a rewarded video can result in a feeling of disappointment, which is the reason why a high rate equals a good performance for both metrics.

Even though Variant B is looking very promising in terms of matching and filling requests, the winner of the two variants is not entirely clear. Looking into the impact of this experiment on the CPMs, it seems like there has been some changes in Variant B compared to the baseline. Main aspect to point out is, that CPMs in bidding have dropped by 11.22% meanwhile the average CPM in the waterfall increased by 1.6%.

Should be noted, that the portion of bidding in comparison to total traffic was larger before and even grew for Variant B in the experiment. Therefore, the drop in CPM would have a much bigger impact the total revenue than the increase in waterfall ad sources, even if the drops were equal. For demonstration, figure 7.2.1 shows the revenue split between bidding and waterfall for Variant A. Figure 7.2.2 displays the corresponding data for Variant B.

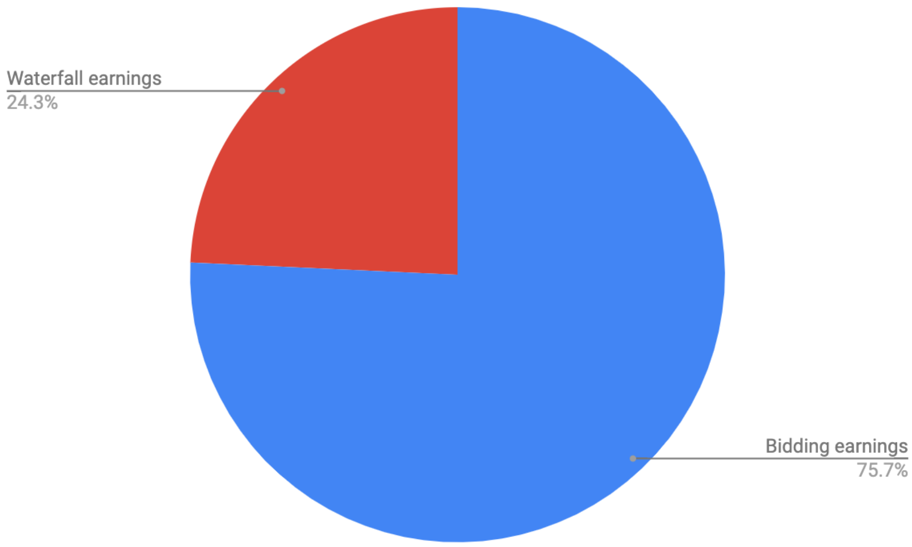

Figure 7.2.1: Revenue split among waterfall- and bidding ad sources in Variant A.

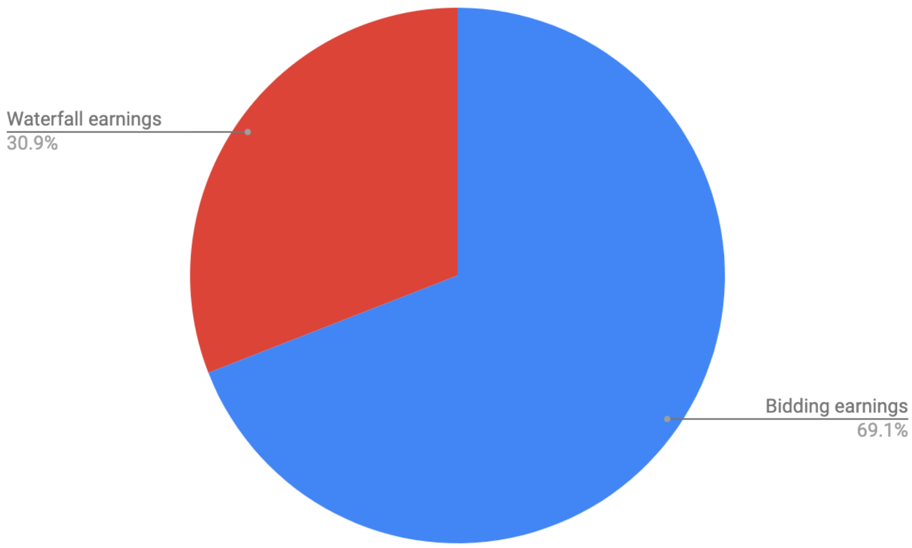

Figure 7.2.2: Revenue split among waterfall- and bidding ad sources in Variant B.

Compared to the pie charts constructed of impression split among the two categories in section 6, logically thinking this would mean that the total revenue in Variant B has dropped. The decrease in Variant B's bidding CPM has been so significant, that even though its portion of traffic grew by over 12%, the total earnings still dropped by 8.7%. An increase of 1.6% in waterfall CPMs is not sufficient to cover the loss.

Once numerous ad sources were removed, specially from the waterfall side, it means that there is less competition – even though a single network has several instances applied, it does not automatically mean that their chances of winning an auction are better. The fact that not all campaigns match all users and not all impressions are perceived as equally valuable still applies. This could be a factor to why waterfall traffic has decreased, and the average CPM has increased for Variant B: perhaps the users linked with ad requests don't match with ad networks in waterfalls that were and therefore they're less eager to serve ads.

Assuming that the networks included in waterfalls are not filling the requests for users valued lower, these impressions are won by bidding sources. As a result, the traffic for bidding increases, meanwhile the average CPM drops. This theory would also explain the fluctuations of fill rate as less competition would lower the threshold for willing auctions with lower bids. However, it remains unclear what impacted the match rates in Variant B as the targeted segments are similar as in Variant A.

7.3. Rewarded Completion Rate

Rewarded completion rate described in the results-section can be an indicator for either good or bad user experience. A low completion rate means that many of rewarded videos started were exited before the ending, meanwhile a higher rate means that a bigger portion of the rewarded videos started were finished. There can be numerous reasons to why a user would quit watching the video before receiving the reward.

Some of them have to do with user interface – perhaps the rewarded video prompt is poorly placed, and a user ends up clicking by accident. However, as the purpose of the research is to evaluate ads' quality rather than the user interface in Hill Climb Racing 2, it's not necessary to consider this. The user interface remained the same for both variants, so this can't play a role.

Instead, the ad network can have a huge role. The video could've caused a crash, been very long in duration or froze. These issues can be incidental, but as there is a clear pattern over the entire 30-day testing period, it could be assumed that removing certain ad networks has contributed to the improvement.

A clear difference between the two columns can be seen, and the completion rates with high-quality networks are higher every day during the experiment. The calculation from including all data in the table above reveals that the average completion rate for the entire period is 85.53% for high-quality networks, meanwhile the figure for all networks combined is 78.30%. If a mediation group totalled 300 000 impressions a day, a completion rate of 85.53% would mean 43,410 unfinished rewarded videos a day. If the completion rate is 78.30% the number would be increased by 21,690 incomplete rewarded video views. For a better overview of how the completion rates developed throughout the period, see the graph below.

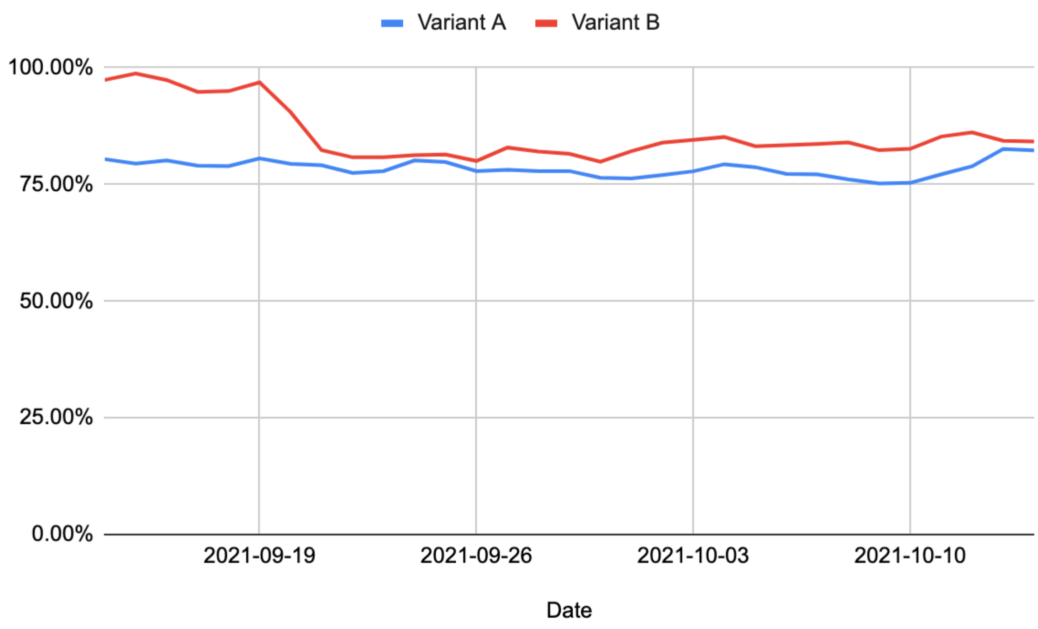

Figure 7.3: Curves showing the development rewarded video completion rate in variants during the testing period.

Although the completion rate for Variant B remains above Variant A throughout the time period, there is a drop starting on the 19th of September. Even though the reason for this is not known, likely there still is one instead of this being a random occurrence. Cause for the drop could be a change in campaigns displayed by networks or a change in how well each network is performing in the auctions. There can small differences in the user experience delivered by individual partners even among the higher-quality networks.

7.4. Retention, Average Revenue per User and Lifetime Value

Despite the differences being very small for variants in retention and estimated total revenue, one should bear in mind that even small fluctuations can have great impact when there are a lot of users. If the retention or total revenue for Variant B remained higher even by >1% than Variant A, the small improvement on a daily basis accumulates on the long run and holds great potential to retain more players and optimize mediation for a better revenue gain. However, as it was explained in results, by the end of the experiment the gap closed and Variant A slightly outperformed Variant B, so this effect probably wouldn't take place in this case.

Looking at the average revenue per user data from Firebase, the figure is slightly higher for Variant A. However, in the figures dug from BigQuery, it seems that the LTV is 8.06% higher in Variant B. Both LTV (lifetime value) and ARPU (average revenue per user) include ads and in-app purchases and measure similar aspects. ARPU is calculated by dividing the total revenue with the number of users, meanwhile LTV per user presents the total value a user has generated to the company. As the results on revenue are contradictory, it is unclear which option would absolutely be the best.

If in reality the Firebase data is correct, the revenue trends would be in line with the increase of bidding traffic with lower CPMs. If, on the other hand, the LTV data fetched from BigQuery is accurate, it would probably mean that the users in Variant B have possibly seen more ads or they've made more purchases in the game. Either way, it is quite surprising that such conflict exists in the result as both pieces of information are from the same source.

Moving on to some event data from BigQuery, as explained in the results, there weren't any differences among targeted groups. It was expected that perhaps having less ad networks in the mix could decrease the number of crashes happening for users during game play. It was speculated that the odds would be in favour of less crashes and other ad-related issues and the networks removed were the ones most commonly responsible for problems. Against the expectations there was no change detected in the number of issues arising for users in Variant B. It can be worth considering that perhaps the SDKs integrated with the application could still be causing issues for the application even when their ads aren't displayed. Therefore, there would be no change between variants regarding this as both targeted groups played the same versions of the game with the same set of SDKs integrated.

7.5. Findings

Based on the results and the analysis discussed above, it is not very easy to draw a direct conclusion. During the A/B testing, an observation of how retention and revenue progressed in the beginning looked like the Variant B would beat baseline by a small difference. However, as the end of experiment was closing in, the positions switched, and Variant B was no longer performing better in neither aspect. The changes made between variants were very small and therefore, it can be stated that the changes made in this test did not impact retention positively, but neither can it be confirmed that the result would be negative in the long run.

In order to figure this out, there would have to be another test run for a longer period of time. Looking at the development towards which the retention was going, it would be likely that having Variant B replace the baseline would be harmful for retention in the long run. Variant A outperforming Variant B in retention means that the retention rate was decreasing slower for the better-performing variant. It would be interesting to see where the difference would settle – when would Variant B stop decreasing faster than Variant A?

Considering other aspects potentially contributing to user experience, the outcome from Variant B is looking slightly more promising. Starting with rewarded video completion rate, it probably had the most consistent and clear change out of all – the increase of this rate is undeniably a positive outcome and potentially has to do with the updated set of networks. Nevertheless, the reason for the development remains a bit unclear.

As bidding experienced growth in the impression share of voice, the importance of filling and matching ad requests increased. Considering this effect combined with the boost in match and fill abilities, it could by chance be a positive impact on the user experience. However, it does seem like the increase in these percentages comes with a cost of a decreasing CPM level, which could actually be a root cause. When the bidding sources are able to participate auctions with lower bids as there is less competition on the waterfall side, the networks' algorithms could allow them to fill more ad requests. In other words, bidding ad sources are allowed a better access to their targeted audiences at lower prices.

When there are two factors – CPMs and fill- and match rates – weighing, the important question to ask oneself is *which one weighs more?* The answer might depend on who is asked. Some professionals operating in the field are very revenue-driven in their mindset, meanwhile others can be extremely concerned with the user's experience. A healthy approach to successful ad monetization is probably somewhere in the middle, balancing between the two. If a user experience is sacrificed at the cost of gaining maximum revenue, the strategy could backfire in form of decreasing retention, and vice versa.

Even if a single network has a lowered ability to deliver ads, the average rate of a mediation group can remain the same. It's probable that this would be the case as if none of the networks are removed from the waterfall after all, there's a greater chance that one of the many networks will fill the requests. What comes to the rewarded video completion rate, it definitely is a metric worth investigating further. The increase of this rate is a positive change, but what should be asked next is *why does it happen?* The reasoning behind the difference is not clear, and a good way to start exploring the subject would be to review the metrics on ad network level and see if there are any differences between partners.

Coming into a conclusion on the financial side is quite tricky as the results differ from each other depending on where the data originates from, and which metrics are taken into consideration. The average revenue per user seemed to be almost equal between groups with a variation range of a few percentages, meanwhile the LTV revenue showed a difference of 8.06%. There is a possibility that the formulas for calculating figures crucially differ from each other, or that the data or estimations used for calculations are different. Either way, this is potentially a good topic to research more.

As stated in the results-section, the event data from BigQuery did not yield any results in changes apart from the LTV revenue and there wasn't difference between variants on how often ad errors and crashes happened. Nevertheless, the option that some ad networks cause them more often can't be completely excluded based on the test. The applications used by both groups of users were the same and had the same sets of ad network SDKs integrated, and it is indeed possible that an SDK causes problems even when their ads are not shown. To run the same test without the entire set of SDKs of excluded networks would be extremely difficult to implement as it would require two different application packages, which practically is impossible with the tools accessible.

To conclude the outcome of this thesis in one sentence, the research question of '*What is the impact of ads' quality on user experience?'* cannot be directly answered. It probably wouldn't be entirely safe out with Variant B from A/B testing, at least not without further research and experimenting. However, it doesn't mean that the time and effort put in hasn't been valuable – the process has helped in raising further questions that can assist the ads team at Fingersoft to go into the right direction in developing the operations.

Some highly practical takeaways yielded from the process as well. The next logical step would be doing some extensive exploring of metrics such as rewarded video completion rate to find out what causes the positive change and if there are other ways to improve it. As there could possibly be more testing related to user experience and ads in the future, the valuable experience gained from this research will most likely be beneficial. Having already done an A/B test of this kind for ads, it's possible that the threshold for running one again in the future is lowered.

8. CONCLUSION

The mobile gaming being relatively new as an industry is constantly growing and evolving, and companies operating in the field are putting effort into finding the best practices when it comes to monetizing their content. Publishers utilizing ads in their monetization strategies might struggle in finding the sweet spot whilst balancing between providing a good user experience and maximizing ad revenues. Prioritizing income over customer relationships can come with a great cost in form of increasing retention - that's why it's important to explore the relationship between the two.

The ideas presented above led into the following research question: 'What is the impact of ads' quality on user experience?', which was attempted to answer by running an A/B test in the target game. The A/B testing process consisted of geographical segmentation, analysing third-party networks partnering with the commissioner, creating new a waterfall setup in mediation platform, running the A/B test itself and querying data from a data warehouse platform. The main point of interest was user retention, although one metric is not sufficient to make changes into the current practices as there are other factors affecting important aspects such as revenue.

The experiment did indeed yield plenty of results to go over and help with reaching a conclusion even though not all of the results are advocating for the same option of the two variants. Retention did not differ significantly between the groups, yet a couple of other metrics related to user experience did improve for the variant with higher-quality networks only. On the other hand, this variant experienced a slight drop in CPM-levels. However, the overall lifetime value still seemed to increase from the baseline.

The results and analysis proved that there are some potentially meaningful differences in mediation groups' performance depending on what types of ad networks are included. As some of the metrics seem to turn these two variants against each other, user experience being on the other and revenue on the other side, to reach an absolute conclusion, some extensive questioning on the matters would be in place. Even an unclear answer is surprisingly enlightening - in this case it's likely that the research question would be better broken down into smaller, more detailed questions in order to proceed with the topic.

A positive user experience goes hand in hand with success when discussing about mobile games, due to which it is vital to explore options outside the practice currently implemented in ad monetization operations. Only by attempting to continuously seek new and better solutions can a company keep on developing in this department. To find the relevant areas to focus on, asking and answering questions is necessary which makes the output of this thesis a valuable for the commissioner. Had this topic not been researched, for now it would've been left undiscovered that there could be room for improvement in user experience metrics dealt with.

9. DISCUSSION

I found writing this thesis very rewarding, although the process was not entirely unproblematic. As mentioned, gaming in mobile platforms is quite a new industry and so is monetizing the games with ads in comparison to some other forms of digital marketing. Therefore, finding literature resources for the theoretical part of the work was fairly challenging. As the branch is not very mainstream, explaining the concepts related required a plenty of thought from someone more familiar with the processes and practices.

Even though I have already worked with ad monetization for over a year, the research offered plenty of new challenges and learnings. For instance, before starting the thesis I had never planned or created an A/B test in Firebase before, and as a result I feel more confident running experiments in the future. Along the process, I also got more familiar with tools such as remote config and BigQuery and what they are used for. Using BigQuery as one source for test results was an excellent opportunity to learn a bit about SQL and benefit from the learnings.

In addition to practical experience with tools and platforms, I learned to see ads from other perspectives as well. As someone working from the publisher's side, I'm very used to focusing on observing performance through metrics expressing revenue. Researching more about third-party ad networks, user experience and discussing user experience with the quality assurance team are among the factors that have helped me expand my personal point of view.

Expectations I had regarding the A/B test and its results differed from the reality as I was anticipating a clearer outcome. Frankly, the slight ambiguousness of the findings did make it a bit difficult to draw a conclusion. However, I believe that the topic was very important to look into. As the question has now been looked into with this particular way of experimenting, it will be easier to continue down the same path with a different methodology or a different set of questions.

REFERENCES

Adjust GmbH 2021. The Adjust Mobile Measurement Glossary. Cited 21.08.2021. https://www.adjust.com/glossary/

Admonsters. Mobile App Ad Quality Playbook. Cited 13.9.2021. https://www.admonsters.com/downloadpdf/playbook/admonsters-playbook-mobile-app-ad-quality

Applovin 2021. Mobile App Glossary. Cited 21.08.2021. https://www.applovin.com/glossary/

Baremetrics. Glossary: Daily Active User (DAU). Cited 23.09.2021. https://baremetrics.com/academy/daily-active-user

Braze 10.10.2016. The Top 3 Ways to Calculate User Retention Rate with Formulas. Cited 14.09.2021. https://www.braze.com/resources/articles/calculate-retention-rate

Borgdon, Andrew 26.05.2016. Mobile Ads Garage: Episode 4 – Interstitials. Google AdMob. Cited 04.08.2021. https://blog.google/products/admob/mobile-ads-garage-4-interstitials/

Borgdon, Andrew 09.06.2016. Mobile Ads Garage Episode 5: Interstitial Best Practices. Google AdMob. Cited 04.08.2021. https://blog.google/products/admob/the-mobile-ads-garage-episode-5-interstitial-best-practices/

Cohen, Jasmine 08.04.2018. What is ad fill rate and how can app developers increase it? ironSource. Cited 22.09.2021. https://www.is.com/community/blog/what-is-fill-rate/

Cooper, Jackelyn 01.08.2017. Succeed with rewarded video, part 1: Designing your rewarded video placement. MoPub. Cited 10.11.2021. https://www.mopub.com/en/blog/rewarded-video-placements

Digital Limbo 21.08.2019. Fill Rate in Ad Monetization. Cited 23.11.2021. https://www.digitalimbo.com/fill-rate-in-ad-monetization/

Djamasbi, Soussan, McAuliffe, Dan, Gomez, Wilmann, Kardzhaliyski, Georgi, Liu, Wan & Oglesby, Frank. 2014. Designing for Success: Creating Business Value with Mobile User Experience (UX). Cited 15.09.2021. https://link.springer.com/content/pdf/10.1007%2F978-3-319-07293-7_29.pdf

Fingersoft 2021. About Us. Cited 05.07.2021. https://fingersoft.com/about-us/

Fingersoft 2021. Hill Climb Racing 2. Cited 21.08.2021. https://fingersoft.com/games/hill-climb-racing-2/

Firebase Remote Config 2021. Google Firebase Documentation. Cited 09.08.2021. https://firebase.google.com/docs/remote-config

Firebase A/B Testing 2021. Google Firebase Documentation. Cited 09.08.2021. https://firebase.-google.com/docs/ab-testing

Fletcher, Jon 16.07.2020. What Factors Affect CPM? Marfeel. Cited 07.08.2021. https://www.marfeel.com/resources/blog/what-factors-affect-cpm

Google AdMob Help Center. Find you app IDs & ad unit IDs. Cited 14.09.2021. https://support.-google.com/admob/answer/7356431?hl=en

Google AdMob Help Center. Get Started: Match Rate. Cited 22.11.2021. https://support.google.-com/admob/answer/9654749?hl=en

Google AdMob Help Center. Overview of bidding. Cited 14.09.2021. https://support.google.com/admob/answer/9234488?hl=en#zippy=

Guanziroli, Jacopo. 29.06.2020. Digital Pill: Completion Rate for Rewarded Videos. Digitalimbo. Cited 22.11.2021. https://www.digitalimbo.com/completion-rate-for-rewarded-videos/

Gui, Jiaping, Nagappan, Meiyappan & Halfond, William G. J. 22.02.2017. What Aspects of Mobile Ads Do Users Care About? An Empirical Study of Mobile In-app Ad Reviews. Cited 15.09.2021. https://arxiv.org/pdf/1702.07681.pdf

ElHady, Hady 05.11.2020. The Mobile Game Monetization Guide. Instabug Blog. Cited 03.08.2021. https://instabug.com/blog/mobile-game-monetization/

Karnes, KC 31.05.2019. What are Interstitials and How To Use Them Without Penalty. Clevertap. Cited 10.11.2021. https://clevertap.com/blog/interstitial-ads/

Kiiski, Erno 16.04.2019. Overview of the Current Mobile Sports & Racing Games Market. Game Refinery By Vungle. Cited 21.08.2021. https://www.gamerefinery.com/overview-current-mobile-sports-racing-games-market/

Knezovic, Andrea 05.05.2021. 141 Mobile Gaming Statistics for 2021 That Will Blow Your Mind. Udonis Blog. Cited 05.07.2021. https://www.blog.udonis.co/mobile-marketing/mobile-games/mobile-gaming-statistics

Manjrekar, Ishan 22.06.2018. Designing a User Friendly In-game Ad Experience. UX Collective. Cited 02.10.2021. https://uxdesign.cc/designing-a-user-friendly-in-game-ad-experience-a1da5d9dda0a

Morrisroe, Ben 23.09.2020. What is CPM? Publift. Cited 02.09.2021. https://www.publift.com/blog/what-is-cpm

Nations, Daniel 02.12.2020. What are Freemium Games? Lifewire. Cited 04.08.2021. https://www.lifewire.com/what-is-freemium-1994347

Olker, Seyhmus 23.05.2017. Best Practices Of Video Ad Placement In Mobile Games. AppSamurai. Cited 03.10.2021. https://appsamurai.com/best-practices-of-video-ad-placement-in-mobile-games/

Perez, Sarah 11.06.2019. Mobile games now account for 33% of installs, 10% of time and 74% of consumer spend. TechCrunch+. Cited 05.07.2021. https://techcrunch.com/2019/06/11/mobile-games-now-account-for-33-of-installs-10-of-time-and-74-of-consumer-spend/

ProductPlan 2021. User Experience. Cited 02.10.2021. https://www.productplan.com/glossary/user-experience/

The AdMob Team 20.03.2017. Rewarded 101: Important things you need to know about rewarded video. Google AdMob. Cited 04.08.2021. https://blog.google/products/admob/rewarded-101-things-you-need-to-know/

Wolf, Adam 21.02.2015. Lowdown: How to Optimize Mediation for your App Part 2. Google AdMob. Cited 14.09.2021. https://www.blog.google/products/admob/lowdown-how-to-optimize-mediation-for/