Empir Software Eng (2019) 24:170–207
https://doi.org/10.1007/s10664-018-9627-4

An empirical study of game reviews on the Steam platform

Dayi Lin[1] · Cor-Paul Bezemer[1] · Ying Zou[2] ·
Ahmed E. Hassan[1]

Published online: 15 June 2018

Abstract The steadily increasing popularity of computer games has led to the rise of a multi-billion dollar industry. Due to the scale of the computer game industry, developing a successful game is challenging. In addition, prior studies show that gamers are extremely hard to please, making the quality of games an important issue. Most online game stores allow users to review a game that they bought. Such reviews can make or break a game, as other potential buyers often base their purchasing decisions on the reviews of a game. Hence, studying game reviews can help game developers better understand user concerns, and further improve the user-perceived quality of games. In this paper, we perform an empirical study of the reviews of 6224 games on the Steam platform, one of the most popular digital game delivery platforms, to better understand if game reviews share similar characteristics with mobile app reviews, and thereby understand whether the conclusions and tools from mobile app review studies can be leveraged by game developers. In addition, new insights from game reviews could possibly open up new research directions for research of mobile app reviews. We first conduct a preliminary study to understand the number of game reviews and the complexity to read through them. In addition, we study the relation between several game-specific characteristics and the fluctuations of the number of reviews that are

Communicated by: Massimiliano Di Penta

✉ Dayi Lin
dayi.lin@cs.queensu.ca

Cor-Paul Bezemer
bezemer@cs.queensu.ca

Ying Zou
ying.zou@queensu.ca

Ahmed E. Hassan
ahmed@cs.queensu.ca

[1] Software Analysis and Intelligence Lab (SAIL), Queen’s University, Kingston, ON, Canada

[2] Department of Electrical and Computer Engineering, Queen’s University, Kingston, ON, Canada

received on a daily basis. We then focus on the useful information that can be acquired from reviews by studying the major concerns that users express in their reviews, and the amount of play time before players post a review. We find that game reviews are different from mobile app reviews along several aspects. Additionally, the number of playing hours before posting a review is a unique and helpful attribute for developers that is not found in mobile app reviews. Future longitudinal studies should be conducted to help developers and researchers leverage this information. Although negative reviews contain more valuable information about the negative aspects of the game, such as mentioned complaints and bug reports, developers and researchers should also not ignore the potentially useful information in positive reviews. Our study on game reviews serves as a starting point for other game review researchers, and suggests that prior studies on mobile app reviews may need to be revisited.

Keywords Game reviews · Computer games · Steam

1 Introduction

Computer games are a rapidly growing application genre. With the revenue of the game industry reaching $91 billion in 2016 (SuperData Research 2016), PC gaming is expected to grow at a rate of 6.3% annually through 2020 (Takahashi 2016).

However, due to the scale of the computer game industry, developing a successful game is challenging. In addition, prior work has shown that gamers are a group of users that is extremely difficult to satisfy (Chambers et al. 2005), making the quality of games an important issue. In order to improve the user-perceived quality of games, a better understanding of the concerns of gamers is essential for game developers. However, the majority of recent research on the quality of games has focused on quality issues from the perspective of *developers* (Washburn et al. 2016; Lewis et al. 2010; Lin et al. 2017a), while few studies are related to the particular issues that *users* face when playing games (Washburn et al. 2016; Lewis et al. 2010).

Similar to mobile app distribution platforms, such as the Apple App Store and Google Play, many online game distribution platforms allow users to post reviews of a game. These game reviews provide a rich data source that can be leveraged to better understand user-reported issues. Prior work on mobile app reviews has shown the value of studying reviews (Vasa et al. 2012; Hoon et al. 2012; Khalid et al. 2015; Pagano and Maalej 2013).

To get a deeper insight on the user-reported issues of games, in this paper we study the reviews of 6224 games on the Steam platform, one of the most popular digital game distribution platforms. As the first work that studies game reviews from a software engineering perspective, our goal is to understand if game reviews share similar characteristics with mobile app reviews. This understanding will allow us to reason about whether the conclusions and tools from prior mobile app review studies can be leveraged by game developers, thereby helping game developers understand better how to leverage user reviews for improving the user-perceived quality of their games. In addition, our study could serve as a starting point for more longitudinal studies of game reviews, and possibly open up new research directions for research of mobile app reviews.

In the first part of this paper, we conduct a preliminary study on the number, length, language and readability of game reviews.

In addition, we study whether there are game-specific characteristics that have a relation with the number of daily reviews. Our preliminary study shows that most games receive a limited number of reviews each day, with a relatively short length and high readability. There are several different aspects between game reviews and mobile app reviews . In addition, we observe that developers should be prepared to get a peak in the number of received reviews after a sales event.

In the second part of this paper, we first study what gamers talk about in their reviews, to understand if gamers address different things in their reviews than mobile app users. Second, we study how long players play a game before they post a review. This information is unique compared to mobile app reviews, and may provide interesting insights for researchers to help developers design the storyline and levels of a game. In particular, we address the following two research questions (RQs):

RQ1: What are gamers talking about in reviews? We manually identify six categories of reviews. Although negative reviews contain more valuable information for developers, the portion of useful information in positive reviews, such as suggestions for further improving the game design, also should not be ignored by developers. Players appear to value game design over software quality (i.e., the number of bugs in a game).

RQ2: How long do players play a game before posting a review? Gamers play a game for a median of 13.5 hours before posting a review. The first hour playing experience is more important for free-to-play games, as we observe a peak in the number of received reviews for free-to-play games after approximately one hour of playing. Developers should pay particular attention to the design of the first 7 hours of gameplay, as the majority of negative reviews are posted within that period.

Paper Organization The rest of this paper is organized as follows. Section 2 provides a brief description of the Steam platform. Section 3 presents the methodology that we used during our empirical study. Section 4 presents the results of our preliminary study. Section 5 presents the results of our empirical study. Section 6 discusses related work. Section 7 discusses the threats to validity of our study. Finally, Section 8 concludes the paper.

2 The Steam Platform

Steam is a digital game distribution platform, developed by Valve Corporation. Steam is considered to be one of the largest digital distribution platforms for PC gaming, with over 8,000 games available and over 184 million active users (Sergey 2016). Steam offers digital rights management (DRM), multiplayer gaming, and social networking services, through two major components of the Steam platform: the Steam Store (Valve 2016b), and the Steam Community (Valve 2016a). Table 1 shows a comparison between the number of games on Steam and on several other PC gaming distribution platforms.

Users can purchase games from the Steam Store. The games that are purchased from the Steam Store, along with the games that are purchased from third-party vendors and then activated through the Steam platform, are playable for a user after logging in on Steam using the Steam client. The Steam client will verify ownership of the game and automatically install any available updates. It is mandatory to install the latest update in order to play a game through Steam.

In addition, users can enjoy social network-like features such as friends lists through the Steam Community. Game developers and journalists can publish news updates for games on

Table 1 Comparison between the number of games on Steam and on other PC gaming distribution platforms (as of Dec 19, 2017)

Platform	Number of PC games
Steam	18,711
Green Man Gaming[1]	5,978
GamersGate[2]	5,921
Good Old Games (GOG)[3]	2,232
Direct2Drive[4]	1,552
GameStop[5]	1,103
Origin[6]	318

[1] https://www.greenmangaming.com

[2] https://www.gamersgate.com/

[3] https://www.gog.com/

[4] https://www.direct2drive.com/

[5] http://www.gamestop.com/

[6] https://www.origin.com/

so-called channels. In general, although it is not mandatory for developers to post announcements about game updates to one or more channels (e.g., to the *Product Update* channel), developers often do post news updates about their games to keep users informed about the latest news about their games.

The Steam Community also permits users to post reviews of games once they played them. Different from other popular application distribution platforms which use a 5-star rating system for reviews, players are asked to provide their overall feeling about the game: “Recommended” (i.e., a positive review), or “Not Recommended” (i.e., a negative review). The number of playing hours of the reviewed game, the number of played games, and the number of previously posted reviews by the reviewer at this moment are shown alongside the review.

The positive review rate ($\frac{\#\ of\ recommended\ reviews}{\#\ of\ all\ reviews}$) is displayed on the Steam Store page of the game, to advise potential customers. A user can only provide one review of a game, across all versions of the game. The user is allowed to update the review at a later time.

In order to publish a game in the Steam Store, developers need to undergo a tax and identity verification process and pay a product submission fee of $100 for each game. In addition, the game must go through review periods where Steam personnel play each game to check that it is configured correctly, matches the description that is provided on the store page, and does not contain malicious content (Alden 2017). The strict process of publishing a game on the Steam Platform ensures the quality of the games that are available on the Steam Store.

Mobile app stores, such as the Apple App store and the Google Play store, have similar review processes. However, compared to the Apple App Store, which requires an annual developer membership fee of $99 (Apple Inc. 2017), or Google Play which has a one-time membership fee of $25 (Google 2017), Steam requires a submission fee for each product submission.

3 Methodology

This section introduces the methodology of our empirical study of game reviews. We detail how we extracted and processed data. Table 2 presents the description of our collected dataset. Figure 1 gives an overview of our methodology.

3.1 Collecting Basic Game Information

We took a snapshot of all the 8025 games that were available in the Steam Store on March 7th, 2016 using a customized crawler. It is important to select high-quality subjects when conducting software engineering research (Nagappan et al. 2013). As a result, prior studies on digital distribution platforms, such as mobile app stores, removed apps that do not have enough downloads as these apps are likely to be toy or personal projects. We removed games that had less than 25 reviews from our study, to avoid a possible bias in our results due to a small number of reviews. In total, we studied 6224 games. We collected the title, developer, publisher, tags, genres, and current early access status (i.e. whether a game is in the early access stage or not) of games. The tags of a game are specified by its players, while the genres of a game are specified by its developer.

3.2 Collecting Release Notes and User Reviews

In order to obtain the update dates of games, we used the accompanying release notes that were posted on channels in the Steam Community. We used the process described in our prior work (Lin et al. 2017a) to extract release notes from the channels. We briefly describe the process below.

We developed a custom-written crawler to extract all 98,823 news updates for all studied games on March 7th, 2016. The earliest news update that was available on March 7th, 2016 was published on June 18th, 2008. We performed the following steps to extract release notes from all news updates.

1. We kept all news updates that were posted on the *Product Release* or *Product Update* channel.
2. We removed all news updates of which the title does not contain the words *update*, *release*, *patch*, *hotfix*, *change log* **OR** a version number.
3. The news updates that were left, together with the news updates from step 1 were considered as release notes.

Table 2 Dataset description

# of studied games	6,224
# of news updates	98,823
# of release notes	37,613
# of reviews	10,954,956
# of positive reviews	9,393,546
# of negative reviews	1,561,410
# of English reviews	6,768,768
# of reviews with accurate playing hours data	28,159

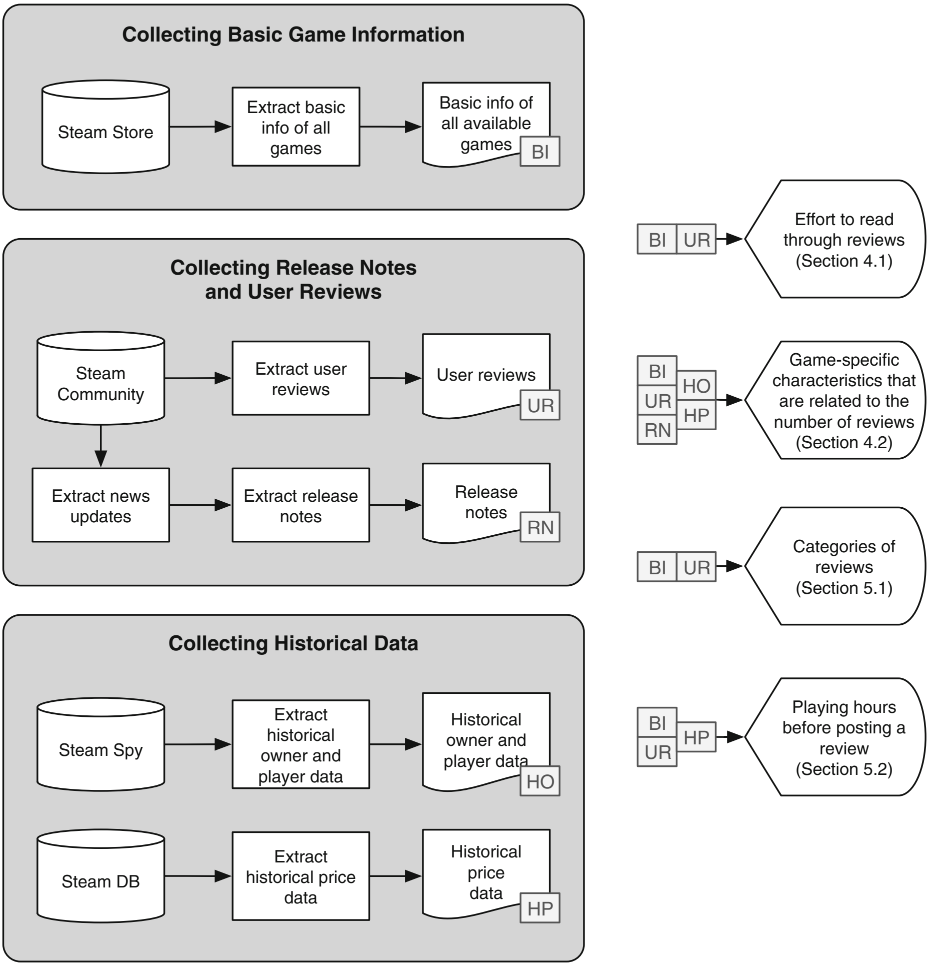

Collecting Basic Game Information
Steam Store
Extract basic info of all games
Basic info of all available games
BI
Collecting Release Notes and User Reviews
Steam Community
Extract user reviews
User reviews
UR
Extract news updates
Extract release notes
Release notes
RN
Collecting Historical Data
Steam Spy
Extract historical owner and player data
Historical owner and player data
HO
Steam DB
Extract historical price data
Historical price data
HP
BI UR
Effort to read through reviews (Section 4.1)
BI UR RN HO HP
Game-specific characteristics that are related to the number of reviews (Section 4.2)
BI UR
Categories of reviews (Section 5.1)
BI UR HP
Playing hours before posting a review (Section 5.2)

Fig. 1 Overview of our study

We identified 37,613 release notes for the studied 6224 games. We validated the precision and recall of our extraction steps in our prior work (Lin et al. 2017a). Our extraction steps have a precision of 89% and a recall of 87%.

We extracted all the reviews for each game from the Steam Community, and filtered-out reviews that contain no words, but only random characters such as smiley faces (e.g., ":)"), as they are non-informative and can be easily filtered-out by developers. There were in total 10,954,956 reviews across all supported natural languages. Steam provides a filter for the language of reviews for a game. We crawled the reviews in each language separately using this filter, to identify the language of each review. However, the number of playing hours (i.e., the number of hours that the reviewer played the game) that is shown with each review is not the number of playing hours at the time of posting the review, but the number of playing hours until now. Hence, in order to study the timing of gamers posting reviews, we developed another real-time crawler which only crawls reviews that are received within the last 6 minutes of the time of crawling, to collect reviews that have an accurate number of playing hours. Therefore, we were able to collect the dataset with an error margin of 6 minutes. We ran the real-time crawler for a month and collected 28,159 reviews with an accurate number of playing hours.

3.3 Collecting Historical Data

We collected the history of the number of owners and the number of players since March 20th, 2015 for all games from Steam Spy (Sergey 2016). Steam Spy is a third-party project which continuously monitors the Steam platform. The owners of a game are people who buy the game on Steam or in retail, then activate the game on Steam; or ones who receive the game through a promotion or as a gift (Sergey 2016). Different from owners, the players of a game are people who play the game during a specific time range. Hence, the number of owners does not necessarily equal the number of players in any given day.

The user profile pages on the Steam Community show the games that a user owns. Theoretically, by going through the profile pages for all users, we can calculate the accurate number of owners for every game. However, with over 184 million users on Steam, it is not practical to churn through all profile pages in a timely manner. Therefore, Steam Spy randomly crawls a representative sample of user profile pages to estimate the number of owners (Orland 2014). To be more accurate, Steam Spy uses a three-day rolling sample to generate the reported numbers of owners, i.e., every day, the data from three days prior are replaced by newly-crawled data. About 1,700,000 randomly-selected profiles are crawled every three days.

We also extracted the price history since November 27th, 2014 for all games from the Steam DB project (Pavel Djundik 2016), another third-party project that monitors the Steam platform. We used the price of a game in U.S. Dollar in our study. We used the differences in the prices of games over time to identify sales events.

3.4 Types of Studied Reviews

We use the developer-provided game genres to distinguish two types of games. We considered games that are tagged with the “Indie” genre as indie games, and games that are tagged with the “Free-to-play” genre as free-to-play games. In addition, we distinguish early-access games using the crawled data (see Section 3.1). In our study, we compared all the studied reviews along the following four dimensions:

1. **Positive reviews and negative reviews.** Prior work has shown that positive reviews and negative mobile app reviews may provide different information (Pagano and Maalej 2013). We study whether positive and negative game reviews are different from each other as well.
2. **Indie game reviews and non-indie game reviews.** Because of the rise of digital distribution platforms such as the Steam platform, indie games have become an important part of the gaming industry after 2004, as these platforms offer a convenient way of distributing games from studios with a smaller budget (Cobbett 2017). To the best of our knowledge, there is no official definition of “indie” games. We use the universal definition as proposed by Stern (2012): *“A game that is both (a) developed to completion without any publisher or licensor interference, and (b) created by a single developer or a small team”*. We assume that the “indie” genre on Steam follows this definition. As the team size and available development resources are very different between indie games and non-indie games, we study indie games and non-indie games separately from each other.

3. **Early access reviews and non-early access reviews.** Our prior work has shown that players of early access games interact differently with the Steam platform during the early access stage (Lin et al. 2017b). Hence, we study if early access reviews are different from non-early access reviews.
4. **Free-to-play game reviews and non-free-to-play game reviews.** We explore if paying for a game has an impact on a user's review behavior.

For each dimension, we compared the total number of reviews, and we manually studied exceptional cases or extraordinary findings. Note that a review can fall into several dimensions (e.g., a review of an indie game can also be a review of a free-to-play game).

4 Preliminary Study of the Characteristics of Game Reviews

In this section, we present our preliminary study of the characteristics of game reviews. As shown in prior work (Vasa et al. 2012; Hoon et al. 2012; Pagano and Maalej 2013; Khalid et al. 2015; McIlroy et al. 2017), mobile app reviews contain useful information for developers to improve the quality of the apps. Similarly, we expect that game reviews will contain valuable information for game developers. It is obvious that the best solution for understanding the issues that users raise in reviews, is to manually read through all the reviews. However, popular games may receive a large number of reviews each day, making it time-consuming for developers to read through all of them. Moreover, the number of reviews that are received each day is under constant fluctuation. For example, Fig. 2 shows the daily number of reviews of the *Dota 2* game. The figure suggests that the number of received reviews each day is unstable, making it hard for developers to assign resources to read through reviews.

In this preliminary study, we first study the number of reviews that games receive each day, the length of the reviews, and the readability of reviews, to understand if game reviews share similar characteristics with mobile app reviews. We then investigate the impact of different game-specific characteristics on the number of reviews that are received each day, to understand what drives this number, and whether the phenomenon is consistent with that of mobile app reviews. The result will answer the question of whether game developers can directly adopt conclusions from prior work on mobile app reviews, and whether prior work on automatically extracting useful information from mobile app reviews can be directly applied to game reviews. Table 3 shows the description of the dataset that is used in this preliminary study along four of the studied dimensions from Section 3.4.

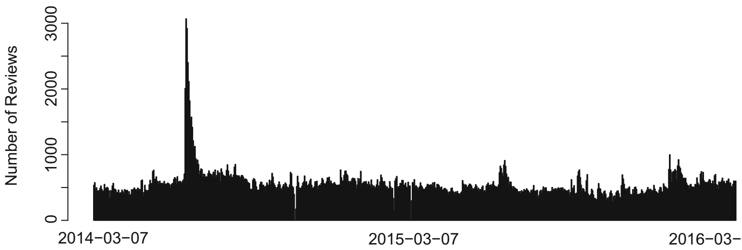


Fig. 2 The number of reviews that the *Dota 2* game received each day. On July 19, 2014 (the peak in the figure), Steam ran a large scale sales event named "Steam Summer Sale 2014", during which players of the game could win free premium game items

Table 3 Preliminary study dataset description

	Indie	Non-indie	Early access	Non-early access	Free-to-play	Non-free-to-play
# of games	3,628	2,596	552	5,672	384	5,840
# of positive reviews	3,664,191(86%)	5,729,355(86%)	973,191(81%)	8,420,355(86%)	1,784,118(84%)	7,609,428(86%)
# of negative reviews	601,376(14%)	960,034(14%)	229,125(19%)	1,332,285(14%)	338,729(16%)	1,222,681(14%)
# of all reviews	4,265,567	6,689,389	1,202,316	9,752,640	2,122,847	8,832,109

4.1 How Many Reviews are Posted and What is their Complexity?

Approach We studied the number and the complexity of reviews from three perspectives: the number of reviews to read each day, the length of the reviews, and the readability of reviews. We studied the readability of all 6,768,768 English reviews, and the number and length of all 10,954,956 collected reviews.

In order to compare the scale and the complexity of different types of reviews, we used the Wilcoxon signed-rank test (Wilcoxon 1945) to compare the distributions for the metrics of different groups of reviews. We grouped the reviews by several different aspects including positive versus negative, early access versus non-early access, free to play versus non-free to play, and different genres. The Wilcoxon signed-rank test is a paired, non-parametric statistical test of which the null hypothesis is that two input distributions are identical. If the p-value computed by the Wilcoxon signed-rank test is smaller than 0.05, we conclude that the two input distributions are significantly different. On the other hand, if the p-value is larger than 0.05, the difference between the two input distributions is not significant. For example, we calculated the medium length of positive reviews and negative reviews of the *Counter-Strike* game, which is 18 characters and 45 characters respectively. We considered the medium length of positive and negative reviews of a game as a pair as the reviews come from the same group of players. We repeated the process for all the studied games, then applied the Wilcoxon signed-rank test to all the pairs. We used a paired test in this section because players of different games may have different review habits, hence by using a paired test we ensured that we were comparing different types of reviews for the same game.

In addition, we calculated Cliff's delta d (Long et al. 2003) effect size to quantify the difference in the distributions of the metrics. We used the following thresholds for interpreting d, as provided by Romano et al. (2006):

$$\text{Effect size} = \left\{ \begin{array}{ll} negligible(N), & \text{if } |d| \leq 0.147. \\ small(S), & \text{if } 0.147 < |d| \leq 0.33. \\ medium(M), & \text{if } 0.33 < |d| \leq 0.474. \\ large(L), & \text{if } 0.474 < |d| \leq 1. \end{array} \right\}$$

We quantified the readability of reviews using the Coleman-Liau index (Coleman and Liau 1975). The Coleman-Liau index is a readability test that is designed to gage the understandability of a piece of text. The index approximates the U.S. grade level thought necessary to comprehend the text. Unlike other readability tests (e.g., the Flesch reading index (Kincaid 1975)), The Coleman-Liau index (CLI) avoids the problem of inaccurately counting syllables (Coleman and Liau 1975). The CLI is calculated using the following formula:

$$CLI = 0.0588L - 0.296S - 15.8$$

where L is the average number of letters per 100 words, and S is the average number of sentences per 100 words. A higher CLI indicates that the text is harder to understand, while a lower CLI indicates that the text is easier to understand. Hence, reviews with a lower CLI should be easier to read through. Table 4 shows examples of reviews with a low and a high CLI respectively.

Findings **96% of the games receive a median of less than 10 reviews per day.** Figure 3 shows the empirical cumulative distribution function of the median number of reviews that are received by each game per day. We removed 20 games with a median number of reviews

Table 4 Examples of reviews with low and high CLI

Example	review content	CLI
A review with a low CLI	*"Very good game, but it was not as good as the first one. It's a fun little game to pass your time, and it's FREE."*	3.7
A review with a high CLI	*"Of course, ironically in the exact same way as robocraft met it's downfall, it was ruined by Greedy developers trying to force their playerbase to spend money on microtransactions with anti-consumer methods of getting the weapon parts they actually want."*	14.4

per day that is greater than 100 from the figure for better demonstration. As shown in Fig. 3, the distribution is extremely skewed. On average, a game receives a median of 2 reviews per day, and 96% of the games receive a median of less than 10 reviews per day. As a result, it should be practical for developers of most games to manually go through all received reviews. However, even developers of games with a relatively low number of reviews per day may not be able to go through all reviews. For example, developers of indie games may only have limited time each week to spend on a game (e.g. because they have an additional full-time job). Hence, the number of reviews that needs to be read for those games could still add up fairly quickly.

It is worth noting that the number is lower than the number of reviews received by mobile apps, which is a median of 22 reviews per day per mobile app. In particular, mobile app users published a median of 31 reviews per app per day for apps in the Games category (Pagano and Maalej 2013).

We manually examined the games that received a median of more than 100 reviews each day, and found that these games are either recently released games, or very popular games.

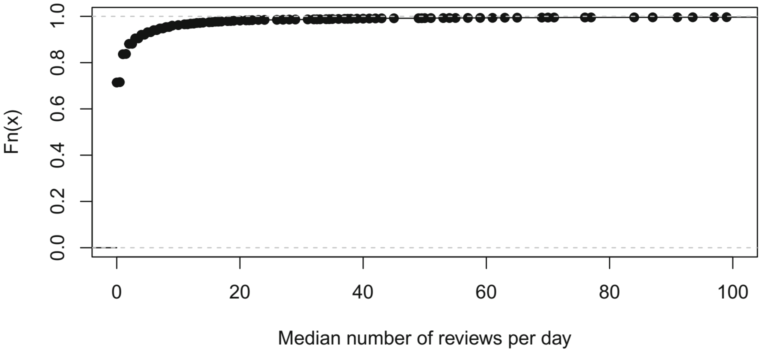


Fig. 3 The empirical cumulative distribution function (Fn(x)) of the median number of reviews that are received by each game per day

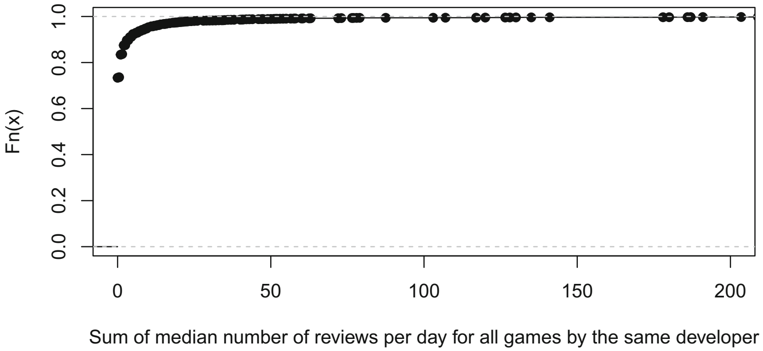


Fig. 4 The empirical cumulative distribution function (Fn(x)) of the sum of median numbers of reviews per day for all games developed by the same developer

For instance, the *Stardew Valley* game[1] was released on February 26, 2016, less than two weeks before our data collection, and received a median number of 586 reviews per day, which is the highest in our dataset. On the other hand, after being released more than 3 years ago, the *Counter-Strike: Global Offensive* game[2], which has more than 25 million owners, receives a median of 514.5 reviews per day. The observation suggests that there may exist game-specific characteristics that have a relation with the number of reviews received by games, such as the lifetime and the number of owners of the games. We further study the relation of such game-specific characteristics with the number of reviews that are received by games per day in Section 4.2.

In addition, we grouped the games by developer, to study how many reviews per day a developer of multiple games would potentially need to read. For this calculation, we included games with less than 25 reviews as well, to get a more accurate overview of the total number of reviews for a developer. Figure 4 shows the empirical cumulative distribution function of the sum of median numbers of reviews that were received by all games from the same developer per day. As shown in the figure, 99% of the developers receive less than 50 reviews in total from all their games.

Most games receive reviews with a median length of 205 characters, or 30 words. Figure 5 shows the distribution of the median length of reviews. We calculated that the median value of the median number of words in reviews per game is 30 words. The lowest median length of reviews is 15 characters for reviews of the *Karos Returns* game[3] (245 reviews in total), and the highest median length of reviews is 1,684 characters for reviews of the *Drizzlepath: Genie* game[4] (29 reviews in total).

We calculated that the median length of reviews across all the games is 93 characters. Our findings show that game reviews are longer than mobile app reviews, which have a median of 61 characters (Pagano and Maalej 2013).

Negative reviews are slightly longer than positive reviews, but the difference is negligible. Figure 5 shows the distribution of the median length of negative and positive reviews. The Wilcoxon signed-rank test shows that the two distributions are significantly different,

[1] http://store.steampowered.com/app/413150

[2] http://store.steampowered.com/app/730

[3] http://store.steampowered.com/app/371310

[4] http://store.steampowered.com/app/438340

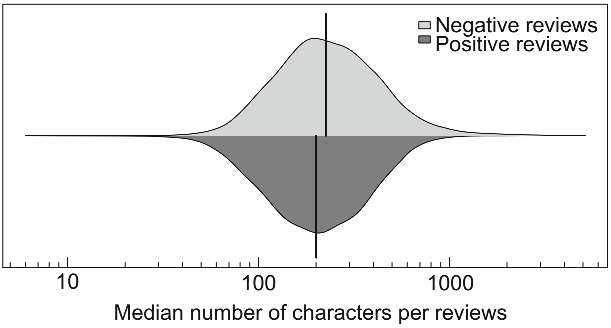


Fig. 5 The distribution of the median length of positive and negative reviews per game. The vertical lines represent the median. The distributions are significantly different ($p < 0.05$), with negligible effect size

however with a negligible Cliff's delta effect size. The negligible effect size indicates that although negative reviews are slightly longer in general, the difference is negligible.

Early access reviews are slightly longer than non-early access reviews. Early access reviews are reviews that are received in the early access stage of an early access game[5]. Early access games allow players to purchase the game during its public beta period while developers continue working on the game. Developers of early access games can receive crucial feedback and bug reports directly from their target community in an earlier development phase. Hence, players may provide more detailed feedback in early access reviews. Our prior work (Lin et al. 2017b) showed that the average rating of reviews is higher during the early access stage. Figure 6 shows the distribution of the median length of early access and non-early access reviews. The Wilcoxon signed-rank test shows that the two distributions are significantly different, with a negligible effect size, indicating that early access reviews are slightly longer than non-early access reviews.

Players write longer reviews for games for which they paid. Figure 7 shows the distribution of the median length of reviews for free-to-play and non-free-to-play games. Free-to-play game reviews have a median length of 105 characters per game, while non-free-to-play games have a median length of 215 characters per game. The Wilcoxon signed-rank test shows that the two distributions are significantly different, with a large Cliff's delta effect size, indicating that non-free-to-play games receive longer reviews than free-to-play games. One possible explanation is that paying for a game makes players feel more strongly about that game.

Reviews for indie games are longer than reviews for non-indie games. Figure 8 shows the distribution of the median length of reviews for indie and non-indie games. The Wilcoxon signed-rank test shows that the two distributions are significantly different, with a small Cliff's delta effect size. The difference is similar to the length of reviews for early access games and non-early access games. The similarity could possibly be explained by the fact that early access games are mostly indie games, as shown in our prior study (Lin et al. 2017b).

Games receive a median of 36% non-English reviews. Figure 9 shows the distribution of the portion of reviews in the top 10 languages. We studied the games with a low portion of English reviews, and observed that most of them were developed by studios from non-English speaking countries. Although some of these games have an English interface, the

[5] http://store.steampowered.com/earlyaccessfaq/

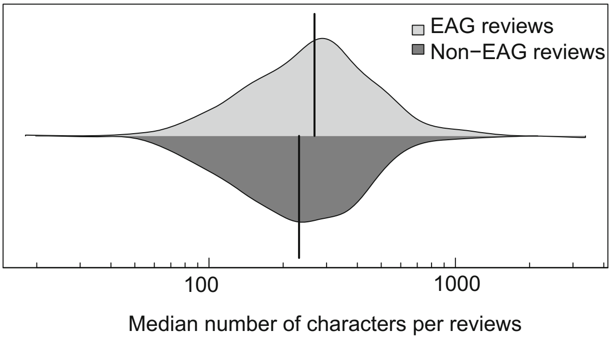


Fig. 6 The distribution of the median length of early access reviews and non-early access reviews per game. The vertical lines represent the median. The distributions are significantly different ($p < 0.05$), with a small effect size

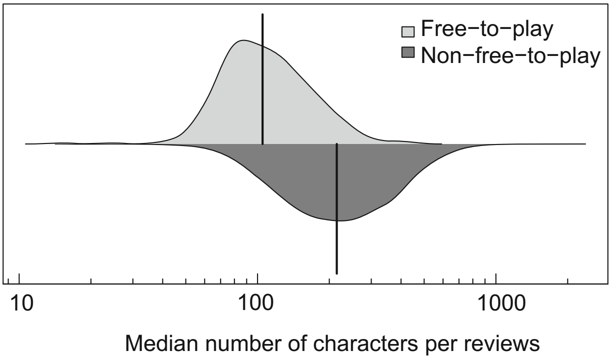


Fig. 7 The distribution of the median length of reviews for free-to-play and non-free-to-play games. The vertical lines represent the median. The distributions are significantly different ($p < 0.05$), with a large effect size

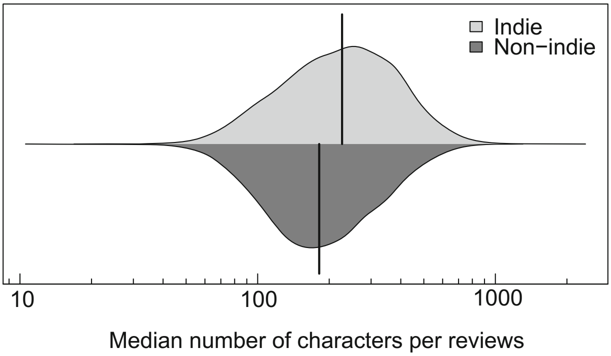


Fig. 8 The distribution of the median length of reviews for indie and non-indie games. The vertical lines represent the median. The distributions are significantly different ($p < 0.05$), with a small effect size

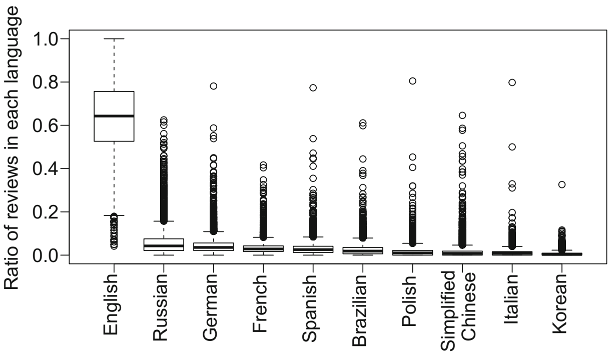


Fig. 9 The distribution of the portion of reviews in the top 10 languages per game

majority of their customers may not speak English. In comparison with mobile app store research, which is usually done on the U.S. version of a store, review language poses a larger threat on Steam, as there is only a single global Steam store. Hence, future studies need to be aware of the considerably large portion of non-English reviews.

Reviews have a median readability level of grade 8. Figure 10 shows the distribution of the median Coleman-Liau index (CLI) of reviews in English. The median value of the distribution is 7.83, and the first and the third quartiles of the distribution are 7.20 and 8.43 respectively, indicating that most game reviews have a median readability level of around US grade 8. We did not observe significant differences in the CLI distribution across different genres of games.

We calculated the median Coleman-Liau index for the reviews of each mobile app in the dataset provided by Grano et al. (2017). The reviews of mobile apps in the dataset have a median CLI of 5.69, which is lower than the median CLI calculated for game reviews. The Wilcoxon rank sum test confirms the significant difference between the readability of game reviews and mobile app reviews, with a large effect size. Game reviews have a significant lower readability than mobile app reviews.

Summary: *Most games receive a limited number of reviews each day, with a relatively short length and high readability. Reviews of early access games are slightly longer. More advanced review selection and summarization techniques are needed for developers of the top 4% games with the most reviews, or for developers who cannot go through their daily reviews for other reasons.*

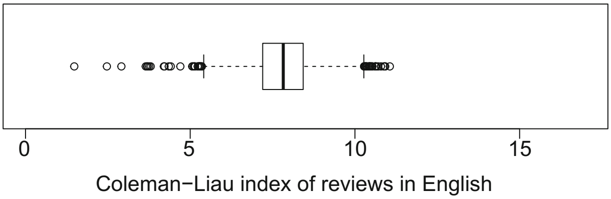


Fig. 10 The distribution of the median Coleman-liau index of reviews in English

4.2 Which Game-Specific Characteristics are Related to the Number of Reviews that are Received Each Day?

Approach We investigated what drives the number of reviews, and whether the phenomenon is consistent with that of mobile app reviews, so that developers can better assign resources to deal with a sudden growth in the number of reviews. We investigated the impact of different game-specific characteristics on the number of reviews that are received each day, including the age of the game, the number of players and the number of owners[6], the developer, the size of the developer studio, information about discounts, and the number of updates. We studied the impact of the aforementioned game-specific characteristics on the number of reviews by building a linear mixed-effect model (Bates et al. 2015), using all 6,224 studied games as training dataset. In a traditional linear regression model, all the independent variables have the same relation with the dependent variable, hence such a model cannot express differences for independent variables at different hierarchical levels (e.g., different games). Unlike traditional linear regression models, linear mixed-effect models have two types of variables, i.e., random effect variables (game-level variables) and fixed effect variables (review-level variables). A mixed-effect model expresses the relationship between the dependent variable (i.e., the number of reviews that are received in one day for a game) and the review-level variables (e.g., the number of players of the game on that day), while taking into consideration the different game-level metrics (e.g., the Steam game id). Table 5 shows the independent variables that were used in the model.

Data Scaling Prior to building our model, we centered and scaled the data, so that we can interpret the coefficients in the model. We applied the `scale` function in R to the numeric variables in Table 5.

Correlation Analysis We checked for fixed effect variables that are highly correlated with one another using Pearson correlation. We used a variable clustering analysis to construct a hierarchical overview of the correlation among the fixed effect variables. We selected only one variable from the sub-hierarchy with correlation $|\rho| > 0.7$ (McIntosh et al. 2016) for inclusion in our models. Figure 11 shows the hierarchical overview of the correlation. There is no fixed effect variable over the threshold. Hence, all variables are kept for further analysis.

Redundancy Analysis Redundant variables (i.e., variables can be explained using other explanatory variables) in an explanatory model will distort the modeled relationship between the explanatory and dependent variables. We used the `redun` function in the `rms` package to detect redundant variables. With a threshold for $R^2 = 0.9$, all variables survived the redundancy analysis, and hence were kept for building the model.

Model Building We used the `lmer` function in the `lme4` package to build the linear mixed-effect model. We also used the `lmerTest` package to calculate the p-value for each fixed effect variable. Tables 6 and 7 shows the result of the model.

[6] Anyone who purchased the game is the owner of the game, but only the people who played the game on that day are counted as the player of the game.

Table 5 A description of the variables of the mixed-effect model

Dependent variables	Effect type	Type	Description
game_id	Random	Categorical	The Steam game id.
developer	Random	Categorical	The developer of the game.
studio_size	Random	Numeric	The number of games that are developed by the developer.
owners	Fixed	Numeric	The number of owners of the game on that day.
players	Fixed	Numeric	The number of players of the game on that day.
eag	Fixed	Boolean	Whether the game is in the early access stage.
age	Fixed	Numeric	The number of days since the initial release of the game.
last_update	Fixed	Numeric	The number of days since the last update of the game.
last_discount	Fixed	Numeric	The discount percentage of the last sale.
last_discount_life	Fixed	Numeric	The number of days after the last sale.

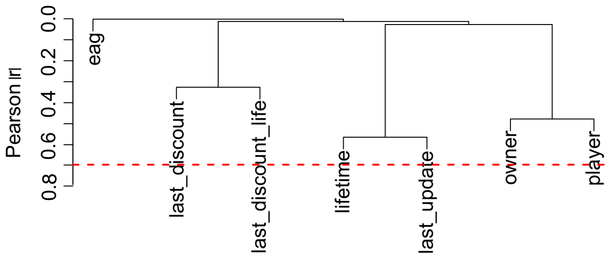


Fig. 11 Hierarchical overview of the correlation among the fixed effect variables. The dotted line shows the threshold ($|\rho| = 0.7$)

Findings The number of players has the strongest relation with the number of reviews. Table 6 shows the mixed-effect model results for fixed effects. The "players" variable has the highest estimated coefficient, while the "owners" variable's estimated coefficient is not high, indicating that the number of active players has a stronger relation with the number of reviews received each day than the owner base.

Although the finding that the number of players has the strongest relation with the number of reviews may look trivial, it actually yields us new information compared to prior studies, such as those of mobile app reviews. Such studies only had access to the number of owners of an app, while our study has access to both the number of owners and the number of active users (players).

A sale event has a stronger relation with an increase in the number of received reviews than releasing an update. It is also worth noting that in Table 6, the "last_discount" variable has the third highest estimated coefficient, while the "last_update" variable has the second lowest absolute estimated coefficient, indicating that a sale event has a greater impact on increasing the number of reviews than releasing an update. A possible explanation is that a sale event can increase the number of players, leading to a higher number of reviews.

The finding is not consistent with prior studies of mobile app reviews. Prior studies on mobile apps tend to ignore paid apps altogether (although there are exceptions Maalej and Nabil 2015; Fu et al. 2013). Our finding shows that discounts are an important factor when studying reviews. In addition, prior studies of mobile apps have shown that mobile app reviews are generally triggered by new releases (Pagano and Maalej 2013). However, our study shows that of all seven fixed effects that we considered, the number of days since the last update of the game has the second lowest impact on the number of received reviews.

Table 6 Model result for fixed effects

	Coefficient	Standard error	p-value	Significant
players	72.77	0.30	< 2e-16	✓
lifetime	−4.27	0.33	< 2e-16	✓
owners	3.91	0.16	< 2e-16	✓
last_discount	2.62	0.07	< 2e-16	✓
eag	2.37	5.22	0.650163	
last_update	0.83	0.22	0.000231	✓
last_discount_life	0.29	0.08	0.000496	✓

Table 7 Model result for random effects

Groups	Variance
game_id	132.030
studio_size	14.102
developer	0.426

Hence, our results are an indication that prior mobile app studies may need to be revisited, thereby taking discounts in the app store into account as well.

> ***Summary:*** *A sale event has a stronger relation with an increased number of reviews than releasing an update. Developers should be prepared to get a surge in the number of reviews after a sales event.*

5 Game Reviews on Steam Platform

In this section, we present the results of our empirical study of game reviews on the Steam platform. First, we discuss the categories of game reviews, and compare those to the taxonomy of mobile app reviews from prior work, to understand if gamers address different things in their reviews than mobile app users. Then, we study the number of playing hours before posting a review. As the number of playing hours before posting a review is a very unique attribute of game reviews compared to mobile app reviews, we study whether this attribute provides interesting insights for researchers, e.g., to provide developers advice for designing the storyline and levels of a game. Our findings can also demonstrate the value of collecting and analyzing app usage times for mobile app developers and researchers.

5.1 RQ1: What are Gamers Talking About in Reviews?

Motivation In our preliminary study, we studied the reviews from several quantitative views. In this RQ, we complete our study of reviews by providing a qualitative viewpoint. The goal is to understand what are the differences between the content of game reviews and the content of mobile app reviews, and among different types of games. We classified reviews into high-level categories, and compared our findings to mobile app review studies.

Approach We manually categorized a statistically representative random sample of English reviews. To obtain the sample, we followed the following steps:

1. To select a representative sample with a confidence level of 95% and a confidence interval of 10%, we need to randomly select at least 96 reviews for each studied dimension (based on the total number of reviews in that dimension). Hence, we randomly selected reviews from the population of all English reviews, and we counted the number of selected reviews from each dimension, until we selected at least 96 reviews from each dimension.
2. We randomly selected 96 reviews from each dimension from the reviews that were selected in the first step, to create a sample of equal size for each dimension.
3. We ended up with a representative sample that contains 472 reviews in total, and 96 reviews across each studied dimension (note that a review may appear in multiple

```
Inputs = All reviews, a list of categories of reviews (which is initially
     empty)

For each review:
   Manually examine the content of this review.

   If the review matches an existing category:
      Label the review with that/those category(-ies).

   Else:
      Add a new category to the list of categories of reviews.
      Restart labelling with new list of categories.

Outputs = All reviews (labelled with appropriate categories), and a list
     of categories of reviews
```

Listing 1 Coding process

dimensions). When considering the representativeness of the sample of 472 reviews, we can draw conclusions with a 95% confidence level and 5% confidence interval. We can also draw conclusions at the dimension-level with a 95% confidence level and a 10% confidence interval.

We performed an iterative process that is similar to *Open Coding* for classifying reviews, as suggested by Seaman et al. (1999, 2008). The procedure is shown in Listing 1.

The procedure starts with an empty list of categories of reviews. For each of the reviews in the sample set, we manually examine the content of the review. If the review matches one or more existing categories in the list, we label the review with those categories. Otherwise, we add a new category to the list and restart labeling with the new list of categories. Note that a review can be categorized into more than one category. For example, if a review contains a suggestion for the game as well as reports a bug, the review would be categorized into both the "Suggestion" and "Bug" categories. Both the first and the second author performed the process individually, and compared the results. The two authors had a (partially) different categorization for 85 out of 472 reviews. The vast majority of the disagreements were cases in which one of the coders assigned an additional label to a review. The conflicts were easily resolved by discussing and coming to an agreement.

During our analysis, we extracted 6 categories from the reviews. Table 8 shows all categories with their description and an example taken from an examined review.

Several studies of mobile app reviews have proposed taxonomies for mobile app review contents (Pagano and Maalej 2013; Gu and Kim 2015; Di Sorbo et al. 2016; Ciurumelea et al. 2017). However, some of these studies focused on the intention instead of the content of reviews (Di Sorbo et al. 2016), while others were only applicable to mobile apps (Ciurumelea et al. 2017). As shown in the previous sections, there are differences between mobile app reviews and game reviews. Therefore, we did not follow the taxonomies proposed for mobile app reviews in this section. We compare our extracted categories to the high level mobile app categories that were proposed by Gu and Kim (2015) in Table 9.

Findings 42% of the reviews provide valuable information to developers. Table 10 shows the percentage of each category. We calculated that the categories that provide valuable feedback for improving the games (i.e., "Pro", "Con", "Suggestion", "Bug") cover 42% of the reviews, suggesting that it is important for developers to read through reviews.

Table 8 Identified categories of reviews

Category	Description	Example
Not helpful	The review contains information that is not helpful for a developer, such as stating the emotion without giving specific reasons.	*"Good game!"*
Pro	The review contains a pro of the game.	*"This game does atmosphere well and the story presented a mystery that seemed worth exploring ..."*
Con	The review contains a con of the game (excluding a bug).	*"... a very, VERY, sharp learning curve ..."*
Video	The review contains an URL to a video review.	*"For my full review please vist: [YouTube link], And watch my video!"*
Suggestion	The review contains a suggestion on how to improve the game.	*"... The game would have been more interesting if you could play with 4 players ..."*
Bug	The review contains a description of a bug that occurs in the game.	*"My game crashed, not once, not twice, but three times in five miniutes..."*

The percentage is slightly higher than the 35% "informative" reviews (i.e., reviews that are potentially useful for developers to improve the quality of the user experience of apps) in mobile app reviews (Chen et al. 2014).

It is worth noting that, although some categories may not be valuable for developers (e.g., "Not helpful"), they may be helpful to other players or potential customers.

Players complain more about game design than bugs. Table 10 shows that only 8% of the reviews mention bugs in games, while 34% of the reviews mention the cons related to game design. Moreover, in negative reviews, 17% mention the bugs while 57% mention the cons related to game design. The percentage of reported bugs in reviews is surprisingly low compared to the cons of game design, suggesting that players value a well-designed gameplay over software quality (i.e., the number of bugs in a game).

Moreover, 42% of the reviews that mention bugs in a game are positive reviews, suggesting that having bugs in a game does not necessarily lead to negative reviews. We examined the negative reviews with bugs, and found that most of the reported bugs in negative reviews

Table 9 Mapping between Gu and Kim's categories (Gu and Kim 2015) and the categories of game reviews that are identified in this paper

This paper	Gu and Kim (2015)
Not helpful	Praise, Others
Pro	Aspect Evaluation
Con	Aspect Evaluation
Video	Others
Suggestion	Feature Request
Bug	Bug Report

Table 10 Categories of reviews (ordered by % of all reviews)

Category	% of all reviews	% of positive reviews	% of negative reviews	% of early access reviews	% of non-early access reviews	% of reviews for indie games	% of reviews for non-indie games	% of reviews for free-to-play games	% of reviews for non-free-to-play games
Not helpful	71	71	55	68	72	66	73	66	71
Pro	38	46	18	41	32	41	30	25	36
Con	34	29	57	27	33	40	31	30	33
Bug	8	7	17	13	8	7	9	14	7
Suggestion	4	4	2	9	2	3	1	3	4
Video	1	1	1	4	1	2	1	2	1

Note that these percentages do not add up to 100% as a single review can be labeled to multiple categories can block players from playing or finishing the game, i.e., they are severe bugs. The most common reported bugs are:

1. Incompatibility (e.g., “Works very very badly on windows 8... Unplayable.”)
2. Crashes (e.g., “works well until a large battle then it crashes. Want to play it but this makes it impossible”, “game keeps crashing”)
3. Bugs blocking users from playing (e.g. “Well it’s been months of trying to get this game to work, but it still doesn’t.”, “...you get back to the map and it just freezes and won’t accept user input...”)

We found the following most common types of bugs in positive reviews:

1. Performance issues (e.g., “Laggy....”)
2. Audio or visual issues (e.g., “Audio tends to fade in and out sporadically and frames drop at specific sequences.”)
3. Crashes (often accompanied by a compliment about game design, e.g., “It is a very interesting game ... The game crashes sometimes.”)

Our prior work (Lin et al. 2017a) found that 64% of the urgent updates of games address feature malfunctions of games (e.g., “Fixed save game does not save your minibike”), most of which are not bugs that block users from playing or finishing the game. As urgent updates cause unnecessary stress on the development team, this finding suggests that developers can re-consider the priority of non-gameplay-blocking bugs, and reduce the number of urgent updates for non-gameplay-blocking bugs by delaying them and bundling them with regular updates.

Negative reviews contain more valuable information about the negative aspects of a game for developers. Table 10 shows that negative reviews have a higher portion of both “Con” and “Bug”, and a lower portion of “Not helpful” reviews, indicating that negative reviews may provide developers with more valuable information about the negative aspects of the current game design. The finding agrees with reported results for mobile app reviews (Hoon et al. 2012), that low-star ratings provide more valuable information to developers.

Positive reviews also provide useful information. Table 10 shows that 29% of the positive reviews discuss cons of the games, and 7% of the positive reviews report bugs in the games. Moreover, positive reviews contain a higher portion of pros of the games, and a slightly higher portion of suggestions, than negative reviews. Knowing what players appreciate about a game is important for developers, as they can ensure that these pros remain or are further improved in future updates. For example, knowing what users consider the pros of a game can help developers to decide whether a feature can be removed. Hence, developers and researchers should not dismiss the information that can be extracted from positive reviews.

Early access games receive more bug reports and suggestions. Table 10 shows that early access reviews have a higher percentage of bug reports, and almost five times the percentage of suggestions of non-early access reviews. These numbers are reasonable considering that the purpose for developers of using the early access model is to gather more early feedback. The finding complements our prior work, in which we show that games have a much more active discussion forum in their early access stage (Lin et al. 2017b).

Indie games receive more suggestions than non-indie games. Table 10 shows that indie games receive a higher percentage of suggestions in reviews, as well as a higher percentage for both pros and cons of the games. A possible explanation is that the player community of indie games is more engaged than the community of non-indie games.

Summary: *We identify 6 categories of reviews. Although negative reviews contain more valuable information for developers, the portion of useful information in positive reviews should not be ignored by developers and researchers. Players appear to value game design over software quality (i.e., the number of bugs in a game).*

5.2 RQ2: How Long do Players Play a Game Before Posting a Review?

Motivation We studied the number of playing hours before posting a review, as this number is a unique attribute that is not found in mobile app reviews. Prior work has shown that the first sustained play session is important for players' engagement (Cheung et al. 2014). With the number of playing hours associated with each review, researchers can quantify and study the importance in depth, and provide developers with suggestions for designing the storylines and levels of a game accordingly. In addition, some online distribution platforms (e.g., Nintendo Game Store[7]) have a minimum requirement for the playing time before allowing a user to post a review (Machkovech 2018), suggesting that reviews with the same rating but different usage times may have different values. The findings of this research question can demonstrate the value of studying app usage time to mobile app developers and researchers.

Approach We use the Wilcoxon rank sum test to compare the distributions of playing hours across different types of reviews and reviews from different types of games, as explained in Section 3.4. The Wilcoxon rank sum test is the unpaired version of the Wilcoxon signed-rank test that we used in Section 4.1. As the reviews with playing hours that are used in this RQ were crawled across all the studied games, we use an unpaired test to compare the distributions. We use Cliff's delta effect size to quantify the difference in the distributions. Table 11 shows the description of the dataset that is used in this RQ.

Findings Gamers play a game for a median of 13.5 hours before posting a review. Figure 12 shows the distribution of the playing hours that are associated with each review. The distribution has a median of 13.5 hours, indicating that half of the reviews are posted within the first 13.5 hours of playing.

Negative reviews are posted after significantly less playing hours than positive reviews. Figure 13 shows the distributions of playing hours for positive reviews and negative reviews. There are 21,874 positive reviews and 6285 negative reviews in the dataset that is used in this RQ. The playing hours for negative reviews are significantly less than the positive reviews, with a small effect size. The median number of playing hours for positive reviews is 15.5 hours, while the median number of playing hours for negative reviews is 6.6 hours. Hence, we suggest that developers should be extremely cautious about the design of the gameplay for the first 6.6 hours, as more than half of the negative reviews are made within the first 6.6 hours of playing.

We also observe a higher peak in the distribution of playing hours for positive reviews, and a more flat distribution of playing hours for negative reviews. One possible explanation is that there may be different reasons that lead to negative impressions which occur at different period of gameplay of a game. To understand more about why people complain

[7] https://www.nintendo.com/games/

Table 11 RQ2 dataset description

	Indie	Non-indie	Early access	Non-early access	Free-to-play	Non-free-to-play
# of games	4,721	3,304	786	7,239	386	7,639
# of positive reviews	9,329	12,624	3,685	18,268	3,567	18,386
# of negative reviews	2,419	3,919	1,333	5,005	1,016	5,322
# of all reviews	11,748	16,543	5,018	23,273	4,583	23,708

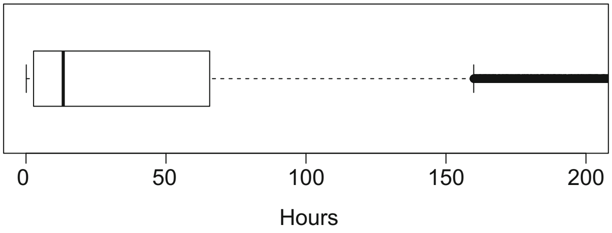


Fig. 12 The distribution of playing hours that are associated with each review

about a game even though they played it for a long time, we manually examined the 63 negative reviews with the longest 1% playing hours. We observed that many of the players who posted such reviews are actually satisfied with the general idea of the game. However, the gaming community (e.g., players who ruin the gameplay), the quality of the latest updates, or the pricing of Downloadable Contents (DLC) of the games disappointed these loyal players, indicating that a badly maintained gaming community, or a poorly planned update may ruin the loyalty of the player base.

We also manually examined the 210 negative reviews with equal to or less than 0.1 playing hour, as 0.1 hour is the smallest granularity at which we monitor playing hours. We identified two major reasons for users to give a negative review after such a short playing time:

1. Severe bugs (e.g., “Doesn’t even log into the game”, “Game crashes every time I try to start it”).
2. Bad design of the game. (e.g., “Gameplay is so boring”, “Not particularly engaging”).

A peak in the number of reviews of free-to-play games is observed after approximately one hour of playing. Figure 14 shows the distributions of playing hours for reviews of free-to-play games and non-free-to-play games. The Wilcoxon rank sum test shows a significant difference between the two distributions, with a negligible effect size. It is worth noting that Fig. 14 shows a density peak at around one hour for free-to-play games, indicating that many free-to-play game players make their judgment of a game after one hour of playing. Possible explanations are that (1) free-to-play games are shorter, or (2) players

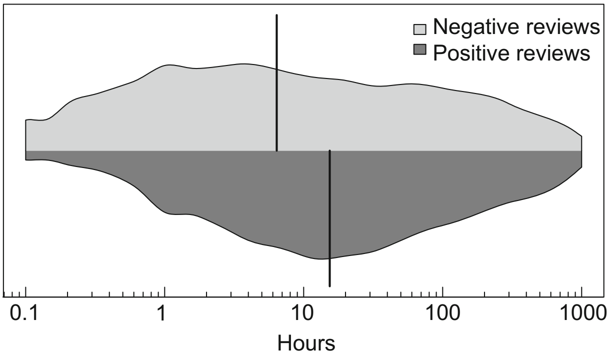


Fig. 13 The distributions of playing hours for positive and negative reviews. The vertical lines represent the median. The distributions are significantly different ($p < 0.05$), with a small effect size

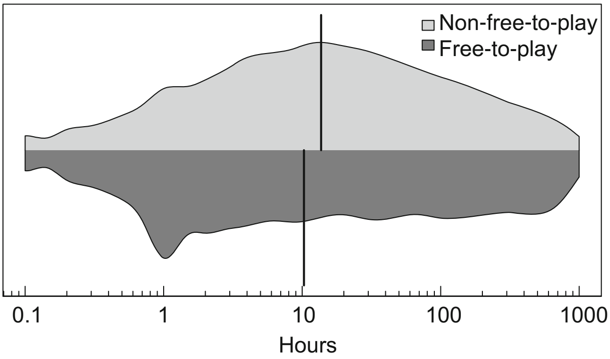


Fig. 14 The distributions of playing hours that are associated with reviews for free-to-play and non-free-to-play games. The vertical lines represent the median. The distributions are significantly different ($p < 0.05$), with a negligible effect size

give up sooner as they did not invest money to buy the game. Moreover, we also observed a different intensity of the peaks around one hour across other dimensions in Figs. 13 and 15, suggesting that the first hour of game design is very important.

We calculated that the median length of the reviews with 1±0.5 playing hours is 39 characters, while the median length of the reviews with 10.3±0.5 playing hours (the median playing hours of free-to-play game reviews) is 38 characters. For non-free-to-play games, the length is 64 and 69 respectively, indicating that reviews that are posted around the first playing hour do not necessary contain less information. The median lengths are shorter than our finding in Section 4.1 because we did not group the reviews by games in this RQ. The median review length (without grouping the reviews per game) for all the reviews studied in Section 4.1 is 80.

Our finding on the importance of the first playing hour agrees with the work by Cheung et al. (2014), which hypothesized that the “first hour experience” is critical for players’ engagement. While Cheung et al.’s work uses “first hour experience” to refer to the first sustained play session, our finding confirms from actual empirical data that the first few

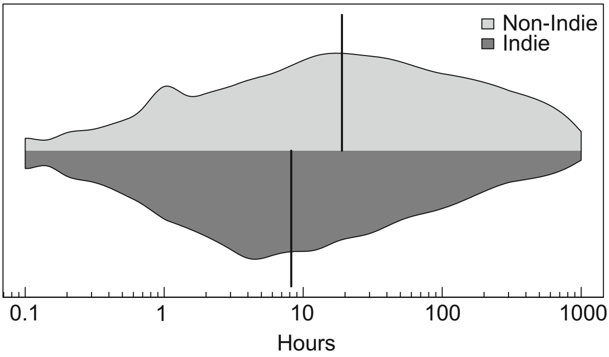


Fig. 15 The distributions of playing hours that are associated with reviews for indie and non-indie games. The vertical lines represent the median. The distributions are significantly different ($p < 0.05$), with a small effect size

hours are important for user experience but also suggests that the first hour is even more important for free-to-play games.

Indie game developers have a shorter time to satisfy players in their games than non-indie game developers. Figure 15 shows the distributions of playing hours for reviews of both indie games and non-indie games. The Wilcoxon rank sum test shows that the playing hours that are associated with reviews for indie games are significantly shorter than non-indie games, with a small effect size. Hence, indie game developers have less time to satisfy players than non-indie game developers. A possible explanation is that indie games may have a shorter storyline. It is worth noting that among the 3628 studied indie games, only 183 (5%) of them are free to play.

Players of games in the early access stage spend more time playing a game before posting a review. Figure 16 shows the distributions of playing hours for early access reviews and non-early access reviews. The Wilcoxon rank sum test shows that the playing hours that are associated with early access reviews are significantly longer than non-early access reviews. However, the effect size is negligible. This finding agrees with our prior work (Lin et al. 2017b), which suggests that players of early access games tend to be more tolerant of the quality of a game during its early access stage.

Players of casual games spend the least time playing a game before posting a review. Figure 17 shows the distributions of playing hours for each game genre. Casual games have the lowest range of playing hours, indicating that casual game players make their judgment about a game faster than other genres. The genre with the highest median playing hours is the Massively Multiplayer game. In addition, the median number of playing hours for negative reviews is lower than for positive reviews for all game genres.

Summary: *Gamers play a game for significantly less time before posting a negative review than before posting a positive review. The first hour playing experience is more important for free-to-play games. Developers should pay particular attention to the design of the first 6.6 hours of gameplay, as the majority of negative reviews are posted within that period.*

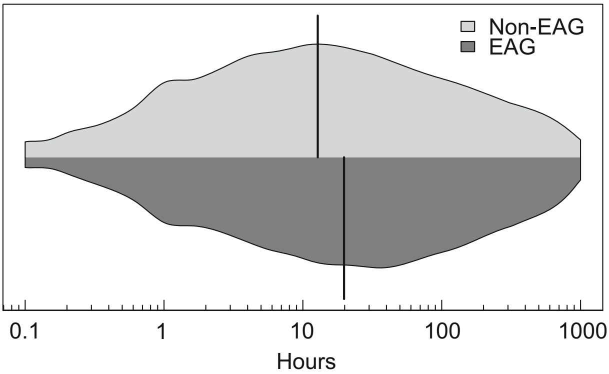


Fig. 16 The distributions of playing hours that are associated with reviews for EAG and non-EAG. The vertical lines represent the median. The distributions are significantly different ($p < 0.05$), with a negligible effect size

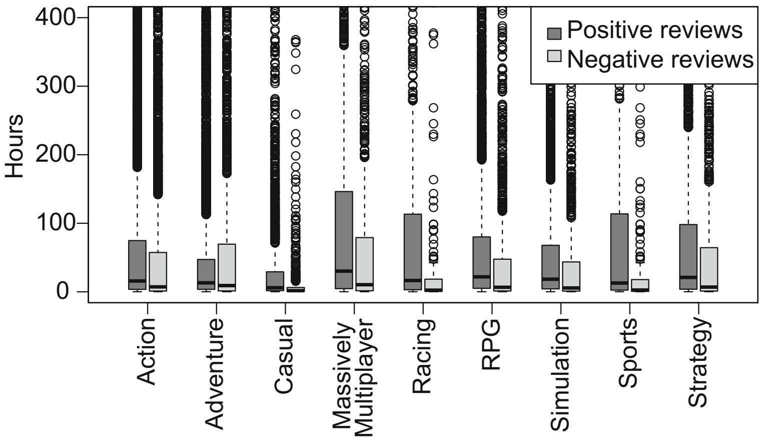


Fig. 17 The distributions of playing hours that are associated with reviews for games in each genre. Outliers greater than 400 are removed for better demonstration

5.3 Implications of our Findings

In this section, we discuss the implications of our findings for researchers and future studies that focus on reviews of online distribution platforms (e.g., mobile app stores and the Steam platform).

Reviews for games are different from reviews for mobile apps. Throughout our study of reviews for games on the Steam platform, we found that in several aspects, game reviews are different from mobile app reviews. Firstly, the median number of reviews received by games per day (2) is much lower than the number for mobile apps (22). However, game reviews are longer than mobile reviews. Secondly, prior research on mobile app reviews only had access to the number of owners of an app, while our study analyzed both the number of owners and the number of active users (players) of games, and found that the number of players has a stronger relation with the number of reviews received per day. In addition, prior mobile app studies that do study paid apps, tend to ignore the impact of discounts. These studies of mobile app reviews should be revisited, as we found that a sales event has the strongest relation with an increase in the number of received game reviews. Finally, games receive a higher percentage of reviews that contain useful information for a developer than mobile apps. Future studies should investigate further whether existing methods for analyzing mobile app reviews can be applied to game reviews. Two observations that we made during our manual analysis of game reviews, which limit the applicability of the automated analysis tools that are currently popular in mobile app review research, are that (1) game reviews tend to contain sarcastic language and that (2) game reviews tend to contain game-specific terminology. As a result, many automated techniques for analyzing natural language do not achieve a high accuracy on game reviews.

Different types of games have different reviews. In our study, we compared all the studied reviews along four dimensions: positive and negative; indie and non-indie; early access and non-early access; free-to-play and non-free-to-play. We found that the median length of reviews is significantly different along every studied dimension. We also noticed that reviews provide different types of information to developers along every studied dimension.

For example, indie games receive more suggestions than non-indie games. Therefore, future studies of game reviews should consider the impact of different types of games.

Information that can be extracted from positive reviews should not be ignored by future studies. Previous studies in mobile app reviews often focus on the negative side of the reviews (Khalid et al. 2015; McIlroy et al. 2016), and rarely consider positive reviews, or the praise in reviews (Panichella et al. 2015). However, we found that 29% of the positive reviews discuss cons of the games, and 7% of the positive reviews report bugs in the games despite their positivity. Positive reviews also contain a higher portion of suggestions than negative reviews. In addition, knowing about what players appreciate in a game can help with making important decisions about the evolution of a game. Hence, the helpful information in positive reviews should not be ignored by future studies.

The number of playing hours before posting a review provides a unique and helpful insight for developers. In Section 5.2, we showed that the number of playing hours associated with game reviews is a unique attribute that is not found in mobile app reviews. This attribute yields new information that can be leveraged by game developers for game design. We know from prior work (Cheung et al. 2014) that the design of the initial gameplay is important. Using the data from the Steam platform, we can quantify and study this importance in depth for all types of games. In this paper, we showed that the number of playing hours provides useful information. Future longitudinal studies should be done to draw definitive conclusions about how gameplay is correlated with the playing time.

In addition, our study sheds light on the fact that reviews that have the same rating, but are posted after different usage times, may provide different information. For example, negative reviews on the Steam platform that are posted after many playing hours, are usually associated with a bad community or update, while negative reviews with few playing hours are usually caused by severe bugs or bad design. Unfortunately, current mobile app stores do not provide the usage time with reviews. As a result, prior studies on mobile app reviews treat all reviews with the same rating equally. However, our findings show that it could be beneficial for researchers and developers to collect and analyze the usage time for mobile apps as well. Hence, mobile app stores and developers should consider integrating the usage time into mobile app reviews. For example, mobile app stores could identify “early” and “late” reviews, to give developers and other mobile app users more context about a user’s opinion.

6 Related Work

In this section, we discuss prior research related to our work. The contribution of our work in comparison to prior work is that we are the first, to the best of our knowledge, to study user reviews from digital game distribution platforms.

6.1 Mining Digital Distribution Platforms

Most of the work about mining digital distribution platforms focuses on mining mobile app stores. Martin et al. (2017) surveyed the field of app store analysis for software engineering. They observed an increasing scale of app samples and a diverse set of techniques and applications in app store analysis, highlighting the health and future potential of the field.

Mining data from digital gaming platforms is an area that has been gaining attention recently. Chambers et al. (2005) analyzed two years of game-related data, such as server traffic and player numbers, from several sources, including Steam. They demonstrated the difficulty of providing enough resources at launch time of a game and showed that gamers are extremely difficult to satisfy. In our prior work (Lin et al. 2017a), we studied urgent updates of popular games on the Steam platform. One of our major findings is that the chosen update strategy by a game developer affects the number of urgent updates that are released. We also studied the early access games on the Steam Platform (Lin et al. 2017b). We suggested game developers to use the early access model as a method for eliciting early feedback and more positive reviews to attract additional new players. In our prior work on the Steam platform, we never studied the contents and the playing hours of game reviews.

Several empirical studies examined the social network of the Steam Community. Blackburn et al. (2014) studied cheaters in the Steam Community. They analyzed more than 12 million player profiles of which 700,000 were flagged as cheaters and showed that the social network of a player (e.g., whether a player has cheating friends) plays an important role in whether a player becomes a cheater. Becker et al. (2012) analyzed the evolution of the Steam Community social network and examined user groups in the Steam community. Sifa et al. (2015) studied cross-game behavior of players in the Steam Community. They analyzed how players that play multiple games on Steam divide their playtime and which games are played by them.

6.2 Empirical Studies on Reviews on Digital Distribution Platforms

Many studies have focused on reviews from mobile app stores. Vasa et al. (2012) and Hoon et al. (2012) analyzed user reviews of mobile apps and found that when users give a negative review to an app, the length of the feedback is greater. Pagano and Maalej (2013) stated that the quality and constructiveness of mobile app reviews vary widely, from helpful advices and innovative ideas to insulting offenses.

Many studies focused on automatically extracting useful information from mobile app reviews. Fu et al. (2013) proposed a system called WisCom to analyze mobile app reviews at the market, app, and review level. Iacob and Harrison (2013) proposed MARA, a prototype for automatic retrieval of mobile app feature requests from online reviews. Di Sorbo et al. (2016) introduced an approach to summarize recommended software changes from user reviews. Gu and Kim (2015) presented a framework to classify reviews into five categories and used a pattern-based parser to extract software aspects. Panichella et al. (2015) presented a taxonomy to classify app reviews as well as an approach to automatically classify app reviews to the proposed categories. Palomba et al. (2017) introduced an approach to not only extract feedback from reviews, but also link it back to software artifacts. Similarly, Ciurumelea et al. (2017) also defined a high and low level taxonomy containing mobile specific categories, and an automatic approach to classify the reviews and link them back to source code files. Man et al. (2016) proposed a framework to analyze app issues across different mobile app distribution platforms. Villarroel et al. (Villarroel et al. 2016) introduced a solution to categorize, cluster, and automatically prioritize reviews. Genc-Nayebi and Abran (2017) systematically reviewed literature about opinion mining from mobile app store user reviews. Martens and Johann (2017) performed an exploratory study of the emotional sentiment of seven million reviews from the Apple App Store.

Several studies focused on the categorization of mobile app reviews. Khalid et al. (2015) manually identified 12 types of complaints that users complain about in the reviews. McIlroy et al. (2016) studied the multi-labeled nature of reviews from 20 mobile apps, and proposed an automatic multi-labeling approach for mobile app reviews.

Some studies examined the mobile app review mechanism. Ruiz et al. (2016) examined more than 10,000 unique mobile apps in the Google Play store and stated that due to the evolving nature of mobile apps, the current displayed score generated from reviews is not dynamic enough to show the changing user satisfaction level. McIlroy et al. (2017) studied the value of developers responding to mobile app reviews, and found that there are positive effects to responding to reviews, with a median increase of 20% in the rating. Hassan et al. (2018) studied the dialog between users and developers, and found that developers and users use the response mechanism as a rudimentary user support tool. Noei et al. (2017) study the relation of both device attributes and app attributes with the user-perceived quality of Android apps, and found that the code size has the strongest relationship with the user-perceived quality.

As discussed in Section 5.3, we show in this paper that game reviews are different from mobile app reviews in several aspects. In addition, computer gamers have a very unique culture compared to users of other types of software. During the study, we observed a considerable amount of reviews that use sarcastic language or game-specific terminologies, making it difficult to apply existing generic tools for mobile app review analysis on game reviews in our study.

7 Threats to Validity

This section presents the threats to the validity of our findings.

7.1 Internal Validity

A threat to the validity of our findings is that we only studied reviews that were written in English for the research questions that involve the contents of reviews. However, there is an obvious limitation in reading reviews in all languages. Future studies should validate whether our observations hold for non-English reviews.

Although on the platform level, there are no incentives for gamers to write reviews, there may exist games that provide an in-game incentive (making it hard for us to find out without actually playing the game). The incentives could possibly bias our results.

To understand what are gamers talking about in reviews, we manually categorized a statistically representative random sample of English reviews. Our sample size is 472 reviews for English reviews, which has a confidence level of 95% and a confidence interval of 5; and 96 reviews for each categories of reviews that we studied, which has a confidence level of 95% and a confidence interval of 10. Although the sample size is relatively small compared to the population of 6,768,768 English reviews, our sample is statistically representative of the whole population of game reviews on Steam.

Another threat to the validity of our findings in Section 5.2 is that we only collected one month of reviews that have an accurate number of playing hours. Future studies are needed on a larger dataset to verify our findings. In addition, as there may be other factors influencing player's playing hours, such as the player's expectation based on the hype that was created for a game, our findings do not suggest causations.

We conducted manual analysis to understand the content of reviews in Section 5.1. We have attempted to apply latent Dirichlet allocation (LDA) (Blei et al. 2003) to automatically extract topics from reviews. However, we did not get meaningful topics. A possible explanation is that players use a large amount of game-specific terminology in their reviews, which limits the applicability of LDA. In addition, although the two authors had a partially different categorization for 85 out of 472 reviews, the vast majority of the disagreements were cases in which one of the authors assigned an additional label to a review (hence they were no major conflicts). The conflicts were resolved by discussing and coming to an agreement.

Other threats to the validity of our study concern the metrics that were used in Section 4.2 in our model for the game-specific characteristics that are related to the number of reviews that are received each day. The number of owners used in our study were estimated from a representative sample by Steam Spy. Although a three-day rolling sample was used to increase the accuracy, there can still exist a deviation from the actual number of owners. However, because the sales data is confidential in the game industry, this is the most accurate method to our knowledge to estimate the number of owners of a game. In addition, we estimated the lifetime of games using the release date as advertised on the Steam Store page. This number is an estimation because developers are allowed to change that release date. We observed that for some games that already existed before they were released on Steam, developers changed the release date to the real release date. We do not have data (reviews, price, etc.) between the real release date and the date that the game was released on Steam. However, we expect it is sufficiently accurate to be used to give a good estimation of the lifetime of games.

7.2 External Validity

In our empirical study, we studied the reviews on Steam. The findings of our study may not generalize to other reviews on different distribution platforms. However, as stated in Section 2, Steam is the largest digital distribution platform for PC gaming. Hence, the reviews on Steam are representative for a large number of reviews. We also compared our findings on game reviews to mobile app reviews where possible.

8 Conclusion

The competition within the game industry, and the hard-to-please user base has made the quality of games an increasingly important issue. As game reviews are a direct reflection of user concerns, a better understanding of reviews can help developers produce games with higher user-perceived quality.

In this paper, we performed an empirical study on the reviews of 6224 games on the Steam platform, a popular platform for digital game distribution. We studied the number and the complexity of reviews, the type of information that is provided in the reviews, and the number of playing hours before posting a review.

The most important findings of our study are:

1. Negative reviews are often posted after only half of the playing hours of positive reviews.
2. A large number of reviews for free-to-play games are posted after approximately one hour of playing.
3. Players complain more about game design than bugs in their reviews.
4. Although negative reviews contain more valuable information for developers, the useful information in positive reviews should not be dismissed.
5. Game reviews are different from mobile app reviews along several aspects.

Based on our findings, we provide the following suggestions for future studies:

1. Due to the difference we discovered between game reviews and mobile app reviews, future studies should investigate further how to adjust existing methods for analyzing mobile app reviews to apply them to game reviews.
2. When studying game reviews, the impact of different types of games should be considered.

3. Information that can be extracted from positive reviews should not be ignored by future studies.
4. The number of playing hours before posting a review should be included in future studies, as the attribute provides a unique insight into how a player's opinion is related to their playing time. Mobile app stores and developers should consider integrating the usage time into mobile app reviews.

We believe that our findings and suggestions can help researchers conduct future studies in game reviews, and in turn help developers improve the user-perceived quality of their games. Future studies should investigate advanced methods that help developers who are not able to read all reviews of their games each day, such as developers of the top 4% games, or developers who have an additional full-time job. Furthermore, our findings show that game reviews are different from mobile app reviews in many aspects, and reveal several important factors that are often ignored by mobile app researchers (e.g., the impact of discounts on the number of reviews). Prior work on mobile app reviews needs to be revisited to take such factors into account.

Acknowledgements We are grateful to Sergey Galyonkin, the owner of Steam Spy, who generously gave us access to all the historical data of Steam collected by Steam Spy for this research.

References

Alden (2017) Steamworks Partner Program. https://partner.steamgames.com/steamdirect, (last visited: Mar 29, 2018)

Apple Inc. (2017) Choosing a Membership. https://developer.apple.com/support/compare-memberships/, (last visited: Mar 29, 2018)

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Softw 67(1):1–48

Becker R, Chernihov Y, Shavitt Y, Zilberman N (2012) An analysis of the Steam community network evolution. In: Proceedings of the 27th Convention of Electrical & Electronics Engineers in Israel (IEEEI). IEEE, pp 1–5

Blackburn J, Kourtellis N, Skvoretz J, Ripeanu M, Iamnitchi A (2014) Cheating in online games: a social network perspective. ACM Trans Internet Technol (TOIT) 13(3):9

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet Allocation. J Mach Learn Res 3(Jan):993–1022

Chambers C, Feng Wc, Sahu S, Saha D (2005) Measurement-based characterization of a collection of online games. In: Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement, USENIX Association, pp 1–1

Chen N, Lin J, Hoi SCH, Xiao X, Zhang B (2014) Ar-miner: Mining Informative reviews for developers from mobile app marketplace. In: Proceedings of the 36th International Conference on Software Engineering (ICSE). ACM, New York, pp 767–778

Cheung GK, Zimmermann T, Nagappan N (2014) The first hour experience: How the initial play can engage (or lose) new players. In: Proceedings of the First ACM SIGCHI Annual Symposium on Computer-human Interaction in Play (CHI PLAY). ACM, New York, pp 57–66

Ciurumelea A, Schaufelbühl A, Panichella S, Gall HC (2017) Analyzing reviews and code of mobile apps for better release planning. In: Proceedings of the 24th International Conference on Software Analysis, Evolution and Reengineering (SANER), IEEE, pp 91–102

Cobbett R (2017) From Shareware superstars to the Steam gold rush: How indie conquered the PC. http://www.pcgamer.com/from-shareware-superstars-to-the-steam-gold-rush-how-indie-conquered-the-pc/ (last visited: Mar 29 2018)

Coleman M, Liau TL (1975) A computer readability formula designed for machine scoring. J Appl Psychol 60(2):283

Di Sorbo A, Panichella S, Alexandru CV, Shimagaki J, Visaggio CA, Canfora G, Gall HC (2016) What would users change in my app? Summarizing app reviews for recommending software changes. In: Proceedings of the 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering. ACM, pp 499–510

Fu B, Lin J, Li L, Faloutsos C, Hong J, Sadeh N (2013) Why people hate your app: Making sense of user feedback in a mobile app store. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp 1276–1284

Genc-Nayebi N, Abran A (2017) A systematic literature review: Opinion mining studies from mobile app store user reviews. J Syst Softw (JSS) 125:207–219

Google (2017) How to use the Play Console. https://support.google.com/googleplay/android-developer/answer/6112435?hl=en, (last visited: Mar 29, 2018)

Grano G, Di Sorbo A, Mercaldo F, Visaggio CA, Canfora G, Panichella S (2017) Android apps and user feedback: a dataset for software evolution and quality improvement. In: Proceedings of the 2nd ACM SIGSOFT International Workshop on App Market Analytics. ACM, pp 8–11

Gu X, Kim S (2015) What parts of your apps are loved by users?. In: 30Th international conference on automated software engineering (ASE). IEEE, pp 760–770

Hassan S, Tantithamthavorn C, Bezemer CP, Hassan AE (2018) Studying the dialogue between users and developers of free apps in the google play store. Empir Softw Eng 23(3):1275–1312

Hoon L, Vasa R, Schneider JG, Mouzakis K (2012) A preliminary analysis of vocabulary in mobile app user reviews. In: Proceedings of the 24th Australian Computer-Human Interaction Conference. ACM, pp 245–248

Iacob C, Harrison R (2013) Retrieving and analyzing mobile apps feature requests from online reviews. In: 10Th working conference on mining software repositories (MSR). IEEE, pp 41–44

Khalid H, Shihab E, Nagappan M, Hassan AE (2015) What do mobile app users complain about? IEEE Softw 32(3):70–77

Kincaid J (1975) Derivation of New Readability Formulas: (automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel Research Branch report 8–75, Chief of Naval Technical Training, Naval Air Station Memphis

Lewis C, Whitehead J, Wardrip-fruin N (2010) What went wrong: a taxonomy of video game bugs. In: Proceedings of the 5th International Conference on the Foundations of Digital Games (FDG). ACM, pp 108–115

Lin D, Bezemer CP, Hassan AE (2017a) Studying the urgent updates of popular games on the Steam platform. Empir Softw Eng 22(4):2095–2126

Lin D, Bezemer CP, Hassan AE (2017b) An empirical study of early access games on the Steam platform. Empirical Software Engineering 23(2):771–799

Long JD, Feng d, Cliff N (2003) Ordinal Analysis of Behavioral Data. Wiley, Hoboken

Maalej W, Nabil H (2015) Bug report, feature request, or simply praise? on automatically classifying app reviews.In: 23Rd international requirements engineering conference (RE). IEEE, pp 116–125

Machkovech S (2018) Nintendo: letting our fans review video games might not be a good idea. https://arstechnica.com/gaming/2018/02/nintendo-has-opened-the-doors-to -fans-game-reviews-on-nintendo-com/, (last visited: Mar 29, 2018)

Man Y, Gao C, Lyu MR, Jiang J (2016) Experience Report: Understanding cross-platform app issues from user reviews. In: 27Th international symposium on software reliability engineering (ISSRE). IEEE, pp 138–149

Martens D, Johann T (2017) On the emotion of users in app reviews. In: Proceedings of the 2nd International Workshop on Emotion Awareness in Software Engineering. IEEE Press, pp 8–14

Martin W, Sarro F, Jia Y, Zhang Y, Harman M (2017) A survey of app store analysis for software engineering. IEEE Trans Softw Eng (TSE) PP(99):1–32

McIlroy S, Ali N, Khalid H, Hassan AE (2016) Analyzing and automatically labelling the types of user issues that are raised in mobile app reviews. Empir Softw Eng 21(3):1067–1106

McIlroy S, Shang W, Ali N, Hassan A (2017) Is it worth responding to reviews? a case study of the top free apps in the Google Play store. IEEE Softw 34(3):64–71

McIntosh S, Kamei Y, Adams B, Hassan AE (2016) An empirical study of the impact of modern code review practices on software quality. Empir Softw Eng 21(5):2146–2189

Nagappan M, Zimmermann T, Bird C (2013) Diversity in software engineering research. In: Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering (ESEC/FSE). ACM, New York, pp 466–476

Noei E, Syer MD, Zou Y, Hassan AE, Keivanloo I (2017) A study of the relation of mobile device attributes with the user-perceived quality of android apps. Empir Softw Eng 22(6):3088–3116

Orland K (2014) Introducing Steam Gauge: Ars Reveals Steam's most popular games. http://arstechnica.com/gaming/2014/04/introducing-steam-gauge-ars-reveals-steams-most-popular-games/ (last visited: Mar 29 2018)

Pagano D, Maalej W (2013) User feedback in the appstore: An empirical study. In: 21St international requirements engineering conference (RE). IEEE, pp 125–134

Palomba F, Salza P, Ciurumelea A, Panichella S, Gall H, Ferrucci F, De Lucia A (2017) Recommending and localizing change requests for mobile apps based on user reviews. In: Proceedings of the 39th International Conference on Software Engineering (ICSE). IEEE Press, pp 106–117

Panichella S, Di Sorbo A, Guzman E, Visaggio CA, Canfora G, Gall HC (2015) How can i improve my app? classifying user reviews for software maintenance and evolution. In: International conference on software maintenance and evolution (ICSME). IEEE, pp 281–290

Pavel Djundik MB (2016) Steam DB - Steam Database. https://steamdb.info/, (last visited: Mar 29, 2018)

SuperData Research (2016) Market Brief — Year in Review 2016. https://web.archive.org/web/20170702202939/https://www.superdataresearch.com/market-data/market-brief-year-in-review/, (last visited: Mar 29, 2018)

Romano J, Kromrey JD, Coraggio J, Skowronek J, Devine L (2006) Exploring methods for evaluating group differences on the NSSE and other surveys: Are the t-test and Cohen's d indices the most appropriate choices. In; Annual meeting of the Southern Association for Institutional Research

Ruiz IJM, Nagappan M, Adams B, Berger T, Dienst S, Hassan AE (2016) Examining the rating system used in mobile-app stores. IEEE Softw 33(6):86–92

Seaman CB (1999) Qualitative methods in empirical studies of software engineering. IEEE Trans Softw Eng (TSE) 25(4):557–572

Seaman CB, Shull F, Regardie M, Elbert D, Feldmann RL, Guo Y, Godfrey S (2008) Defect categorization: making use of a decade of widely varying historical data. In: Proceedings of the 2nd International Symposium on Empirical Software Engineering and Measurement (ESEM). ACM, pp 149–157

Sergey G (2016) SteamSpy - All the data and stats about Steam games. http://steamspy.com/, (last visited: Mar 29, 2018)

Sifa R, Drachen A, Bauckhage C (2015) Large-scale cross-game player behavior analysis on steam. In: Artificial Intelligence and Interactive Digital Entertainment International Conference. AAAI Press, pp 198–204

Stern C (2012) what makes a game indie: a universal definition. http://sinisterdesign.net/what-makes-a-game-indie-a-universal-definitio n/, (last visited: Mar 29, 2018)

Takahashi D (2016) PwC: Game Industry to grow nearly 5% annually through 2020. https://venturebeat.com/2016/06/08/the-u-s-and-global-game-industries-will-grow-a-healthy-amount-by-2020-pwc-forecasts/ (last visited: Mar 29 2018)

Valve (2016a) Steam Community. http://steamcommunity.com/, (last visited: Mar 29, 2018)

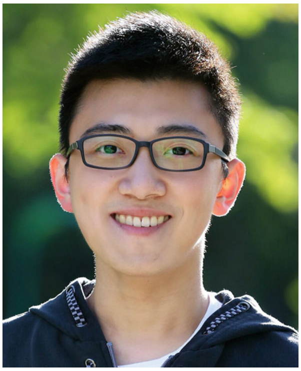
Valve (2016b) Steam Store. http://store.steampowered.com/, (last visited: Mar 29, 2018)

Vasa R, Hoon L, Mouzakis K, Noguchi A (2012) A preliminary analysis of mobile app user reviews. In: Proceedings of the 24th Australian Computer-Human Interaction Conference. ACM, pp 241–244

Villarroel L, Bavota G, Russo B, Oliveto R, Di Penta M (2016) Release planning of mobile apps based on user reviews. In: Proceedings of the 38th International Conference on Software Engineering (ICSE). ACM, pp 14–24

Washburn Jr, M, Sathiyanarayanan P, Nagappan M, Zimmermann T, Bird C (2016) What went right and what went wrong: An analysis of 155 postmortems from game development. In: Proceedings of the 38th International Conference on Software Engineering (ICSE). IEEE/ACM, pp 280–289

Wilcoxon F (1945) Individual comparisons by ranking methods. Biom Bullet 1(6):80–83



Dayi Lin is a Ph.D. student in the Software Analysis and Intelligence Lab (SAIL) at Queen's University, Canada. His research interests include mining software repositories and empirical software engineering, in particular, Software Engineering related aspects on PC games. His research aims at providing a better understanding of Software Engineering aspects on PC games, to help game developers produce games with better quality and player satisfaction. Contact him at dayi.lin@cs.queensu.ca. More information about Dayi is available on his website: http://lindayi.me.

Cor-Paul Bezemer currently works as a postdoctoral research fellow in the Software Analysis and Intelligence Lab (SAIL) at Queen's University in Kingston, Canada. His research interests cover a wide variety of software engineering and performance engineering-related topics. His work has been published at premier software engineering venues such as the ESEC-FSE, ICSME and ICPE conferences and the TSE and EMSE journals. He is one of the vice-chairs of the SPEC research group on DevOps Performance. Before moving to Canada, he studied at Delft University of Technology in the Netherlands, where he received his BSc (2007), MSc (2009) and PhD (2014) degree in Computer Science. More about Cor-Paul can be read on his website: http://sailhome.cs.queensu.ca/~corpaul.

Ying (Jenny) Zou is the Canada Research Chair in Software Evolution. She is a professor in the Department of Electrical and Computer Engineering, and cross-appointed to the School of Computing at Queen's University in Canada. She is a visiting scientist of IBM Centers for Advanced Studies, IBM Canada. Her research interests include software engineering, software reengineering, software reverse engineering, software maintenance, and service-oriented architecture. More about Ying and her work is available online at http://post.queensu.ca/~zouy.

Ahmed E. Hassan is the Canada Research Chair (CRC) in Software Analytics, and the NSERC/BlackBerry Software Engineering Chair at the School of Computing at Queen's University, Canada. His research interests include mining software repositories, empirical software engineering, load testing, and log mining. Hassan received a PhD in Computer Science from the University of Waterloo. He spearheaded the creation of the Mining Software Repositories (MSR) conference and its research community. Hassan also serves on the editorial boards of IEEE Transactions on Software Engineering, Springer Journal of Empirical Software Engineering, Springer Journal of Computing, and PeerJ Computer Science. Contact ahmed@cs.queensu.ca. More information at: http://sail.cs.queensu.ca/.