What's in a Review: Discrepancies Between Expert and Amateur Reviews of Video Games on Metacritic

TIAGO SANTOS, Graz University of Technology, Austria

FLORIAN LEMMERICH, RWTH Aachen University, Germany

MARKUS STROHMAIER, RWTH Aachen University, Germany

DENIS HELIC, Graz University of Technology, Austria

As video game press (“experts”) and casual gamers (“amateurs”) have different motivations when writing video game reviews, discrepancies in their reviews may arise. To study such potential discrepancies, we conduct a large-scale investigation of more than 1 million reviews on the Metacritic review platform. In particular, we assess the existence and nature of discrepancies in video game appraisal by experts and amateurs, and how they manifest in ratings, over time, and in review language. Leveraging these insights, we explore the predictive power of early expert vs. amateur reviews in forecasting video game reputation in the short- and long-term. We find that amateurs, in contrast to experts, give more polarized ratings of video games, rate games surprisingly long after game release, and are positively biased towards older games. On a textual level, we observe that experts write rather complex, less readable texts than amateurs, whose reviews are more emotionally charged. While in the short-term amateur reviews are remarkably predictive of game reputation among other amateurs (achieving 91% ROC AUC in a binary classification), both expert and amateur reviews are equally well suited for long-term predictions. Overall, our work is the first large-scale comparative study of video game reviewing behavior, with practical implications for amateurs when deciding which games to play, and for game developers when planning which games to design, develop, or continuously support. More broadly, our work contributes to the discussion of wisdom of the few vs. wisdom of the crowds, as we uncover the limits of experts in capturing the views of amateurs in the particular context of video game reviews.

140

CCS Concepts: • **Human-centered computing** → *Empirical studies in collaborative and social computing*;

Keywords: video game reviews; review aggregator websites; wisdom of the crowds

ACM Reference Format:

Tiago Santos, Florian Lemmerich, Markus Strohmaier, and Denis Helic. 2019. What's in a Review: Discrepancies Between Expert and Amateur Reviews of Video Games on Metacritic. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3, CSCW, Article 140 (November 2019). ACM, New York, NY. 22 pages. <https://doi.org/10.1145/3359242>

1 INTRODUCTION

Understanding the many facets of how expert knowledge compares to the wisdom of the crowds is a fundamental research question [6, 19, 44, 66, 69]. Is expert-produced knowledge more reliable and trustworthy than that by crowds? In particular, in subjective domains, such as experience goods [54], are expert opinions representative of or even needed at all by the crowds?

Authors' addresses: Tiago Santos, Graz University of Technology, Graz, Austria, tsantos@iicm.edu; Florian Lemmerich, RWTH Aachen University, Aachen, Germany, florian.lemmerich@cssh.rwth-aachen.de; Markus Strohmaier, RWTH Aachen University, Aachen, Germany, markus.strohmaier@cssh.rwth-aachen.de; Denis Helic, Graz University of Technology, Graz, Austria, dhelic@tugraz.at.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 2019 Copyright held by the owner/author(s).

2573-0142/2019/11-ART140

<https://doi.org/10.1145/3359242>

Fig. 1. These sample Metacritic webpages display stark contrasts (Fig. 1a) as well as noteworthy agreement (Fig. 1b) in the appraisal of video games by experts and amateurs. This work analyzes differences and similarities in the review ratings and text depicted in the lower part of the Figures, namely the “critic (or expert) reviews” (orange) and “user (or amateur) reviews” (blue) section.

A prominent example of frequent clashes between expert and crowdsourced views on experience goods are video game reviews. As the video game press (henceforth *experts*) and the crowd of gamers (henceforth *amateurs*) produce video game reviews separately while following different motivations, remarkable evaluative discrepancies in their reviews may arise (e.g., see Figure 1a). Such discrepancies influence reputation and ultimately sales of a video game, frequently resulting in controversy surrounding experts and game developers, and occasionally even causing amateur-called boycotts¹. However, as depicted in Figure 1b, there are examples highlighting the opposite may also occur, as experts and amateurs also express agreeing views on a video game’s reputation.

Recently, only a few initial studies [31, 57] analyzed this phenomenon by investigating differences and similarities in video game rating and review language of experts and amateurs, and we still lack a large-scale analysis of video game reviews. The facts that video game reviews impact (i) sales as estimated via numerous regression models [11, 21, 74, 75], (ii) player experience as surveyed by multiple authors [28, 42, 43] and potentially even (iii) developer plans as implicated by the game piracy study of Drachen et al. [17] further support our view that video game reviews pose a relevant and fertile testbed for comparing expert with crowdsourced inputs.

Research Questions. Hence, in this work, we address the following research questions:

- RQ1 (Existence):** Is there substantial disparity between experts and amateurs in their video game appraisal?
- RQ2 (Characterization):** How can we characterize potential disparities, in particular along temporal and textual dimensions?
- RQ3 (Impact):** What are predictive strengths and weaknesses of expert and amateur reviews for short-term and long-term video game reputation?

¹The video game “Star Wars Battlefront II” is a prominent example of a severe disagreement between experts and amateurs, as indicated by mixed press coverage and overwhelmingly negative amateur reactions, in particular calling for a game boycott (e.g. https://www.metacritic.com/game/pc/star-wars-battlefront-ii/user-reviews#user_review_7754293).

Approach. To address our research questions, we build a large-scale dataset of more than a million video game reviews from the video game portion of Metacritic, a popular online review aggregator. Within the frame our research questions provide, we operationalize video game reputation as amateur review ratings. We empirically analyze reviewing behavior of experts and amateurs, in particular along temporal and textual dimensions. Leveraging this analysis, we predict future short-term and long-term reputation of video games.

Findings. We find a more balanced rating behavior by experts in comparison to amateurs, whose ratings are more polarized and often indicate disagreement with experts. Experts review games shortly after release, on the contrary to amateurs, who still review games years after release and tend to give increasingly positive ratings in those cases. On the textual level, experts write more complex texts and strike a rather detached tone, whereas amateur review language is easier to understand and more emotional. Interestingly, the topic nostalgia arises in reviews by both. Based on these insights, we show video game reputation among amateurs can be accurately predicted with up to 91% and 80% ROC AUC based on amateur and respectively expert reviews in the short-term, and 72% and respectively 71% ROC AUC in the long-term. Prediction models based on reviews from both yield even better performances, suggesting experts and amateurs complement each other.

Contributions. Our extensive investigation of ratings and language in video game reviews by experts and amateurs is of unprecedentedly large scale, and offers actionable insights into similarities and discrepancies between expert and amateur video game appraisal. In particular, by bridging reviews of experts to those of amateurs and vice-versa, we help unify views on overall video game reputation, thereby supporting other video game industry stakeholders (and in particular producers and developers) attain a clearer grasp of the market. Further, we derive suggestions for online commerce platforms to weigh expert and amateur opinions in e.g. recommender systems. For example, video game recommender systems could, when computing recommendations, more appropriately quantify the impact of ratings awarded long ago or more recently, and provide higher weights for expert ratings of older games.

Overall, our work thus contributes to the overarching discussion on wisdom of the few vs. wisdom of the crowds from the particular, nuanced viewpoint of video game reviews.

2 RELATED WORK

Online Reviews. Previous work on online reviews focused on their rating and textual valence, as well as economical impact in sales of books [9, 12, 26], movies [18, 24, 26, 29, 33, 61, 70], restaurants [30, 45] and video games [11, 21, 74, 75]. As these works establish the existence and strength of a relationship between review characteristics and market performance of experience goods [54], we turn our attention to another indicator of public reception of video games in particular: the interplay between expert and amateur review ratings and texts. Multiple authors [9, 12, 26, 70] found that amateur review ratings of books and movies have a bimodal distribution, which may be indicative of a selection bias in items that amateurs review [35]. In the review helpfulness study of Danescu-Niculescu-Mizil et al. [12], the authors discuss how increasingly large variance of amateur ratings translates into a bimodal distribution of review helpfulness as evaluated by amateurs. We replicate these findings in our analysis of the amateur review rating distribution, and extend them via our comparison with the expert review rating distribution.

More closely related to our work are comparative studies of expert and amateur reviews of movies [14, 59, 70] and books [16]. In a survey of 169 college students with varying degrees of exposure to movies, Plucker et al. [59] find a continuum of rating behavior, where more experienced students rate movies more like experts. De Jong et al. [14] perform a textual comparison of 72 expert and amateur reviews of movies, which reveals amateurs evaluate movies from a personal standpoint and experts employ a more informative, contextualizing style. Estimating revenue and

sales of 136 movies and 179 digital cameras via expert and amateur reviews, Wang et al. [70] uncover significant interaction effects between expert and amateur reviews when product ratings have high variances. In their study of 100 book reviews, Dobrescu et al. [16] assess which book characteristics experts and amateurs appreciate more strongly. We go beyond these previous analyses by studying a large-scale dataset of video game reviews, and by exploring temporal rating behavior by experts and amateurs and how both textually manifest (dis-)appreciation for games.

On the topic of video game reviews, previous studies surveyed how players experience playing a game, and how this experience depends on their past exposure to video games [10] and relates to Metacritic critic and user scores [28]. In particular, Johnson et al. [28] call for a textual analysis of amateur reviews to explore further dimensions of player experience. This is a research gap which (i) Thominet [68] explores in a manual study of 180 amateur reviews on the online game shop Steam², (ii) Lin et al. [40] inspect in reviews of over 6224 games also on Steam, and (iii) we also address. Past correlational [57] and textual [31] studies of expert and amateur reviews on Metacritic cover the most similar kind of dataset to the one we handle, but those studies do not go beyond a static, small-scale analysis. Again, we expand on these studies with temporal and textual insights from a large-scale dataset of expert and amateur video game reviews on Metacritic. Also, we leverage these insights for short- and long-term prediction of game reputation among amateurs.

In research directions orthogonal to our work, multiple authors [12, 39, 46, 52, 53, 55] focus on fake review detection in a plethora of online marketplaces and review websites. Other authors uncovered the existence of herding effects in the language [20, 51] and rating [35, 36] of online reviews. In this work, we estimate the potential impact of fake reviews as rather negligible. Furthermore, we alleviate the short-term impact herding effects might have by focusing on the long-term horizon as well, as herding effects weaken over time [35].

Recommender Systems. In the literature on recommender systems, multiple authors [1, 7, 8, 15, 37, 41, 48, 49, 62, 73] leverage review text to predict user ratings per item and thereby improve item recommendations in various settings. Previous work [1, 8, 15, 41] utilizes reviews to address, in collaborative filtering and related approaches, common problems such as cold start, data sparsity and noise. Beyond addressing such data issues, Cheng et al. [8] and other authors [7, 37, 48, 49, 62, 73] extract topics [7, 8, 48], user sentiment [37, 73] and viewpoints [62], and categories and attributes [8, 49] of product reviews to provide interpretable recommendations and improve upon baseline recommender systems. Notably, Lei et al. [37] show jointly estimating item reputation and review sentiment of a user and her friends leads to improved recommender system performance. Note we predict game reception among *a group of users* (amateurs), while the previously mentioned recommender system and other natural language processing [60] and deep learning [67] methods predict ratings by *single users*. In particular, Tang et al. [67] extract user-specific word sentiment from movie and restaurant reviews to predict ratings, whereas we contextualize review text sentiment by user group. Overall, as our results advance temporal and textual understanding of video game reviews, our work may contribute to improve recommender systems leveraging review text.

Wisdom of the Few vs. Wisdom of the Crowds. Our work also contributes to the overarching discussion on the wisdom of individual experts vs. that of the crowds. Previous work [6, 19, 66] identified numerous instances of the performance of crowds surpassing that of experts. However, there is also evidence [5, 44, 69] for the deterioration and (e.g. cognitive) biases of crowds as well. With our comparison of video game review production by experts and amateurs, we extend the experts vs. crowds literature to the setting of experience goods [54]. In this setting, subjective views dominate discussion and opinion discrepancies may thus arise. Interestingly, recent work [64] suggests disagreements of the kind we uncover may positively influence (discussion) outcomes.

²store.steampowered.com

3 MATERIALS AND METHODS

Background on Metacritic. In this work, we study a dataset covering the video game reviews on the Metacritic platform. Metacritic is a popular online review aggregator which provides review ratings and text snippets by critics of many different experience goods [54], such as movies, TV shows, music records and in particular also video games. Beyond the review information provided by external critics we term *experts*, Metacritic also offers a platform for website users, which we term *amateurs*, to rate and review aforementioned experience goods. As illustrated in Figure 1, a Metacritic video game review website consists of (i) an upper part with aggregated review ratings by experts (“Metascore”) and amateurs (“User Score”), and (ii) a lower part with individual expert and amateur reviews. In this work, we focus on this lower part of Metacritic video game review webpages, i.e. on expert and amateur reviews. These reviews consist of a rating on differently graded scales and a review text. We convert all expert ratings, originally mapped by Metacritic from diverse scales to a 0 – 100 scale [50], to the amateurs’ 0 – 10 scale by dividing expert ratings by ten and odd-even rounding the result. Note that the expert review text provided by Metacritic is a brief summary of the actual review written by the expert. Metacritic typically provides a link to the expert’s external website (if it exists, which is not always the case).

Data Collection. We collected all Metacritic video game reviews by experts and amateurs as of December 2018. Each video game review comprises, besides text and a rating, a review date, the name of the review author and, for amateur reviews only, the number of up- and downvotes. Note that Metacritic does not allow voting on reviews of experts. Besides reviewing data, we gathered video game metadata also available at Metacritic, namely the game platform, genre, developer and release date. In this work, we study rating and reviewing behavior on Metacritic; we thus do not parse experts’ full reviews from their external websites.

Pre-Processing and Resulting Dataset. Game release dates in Metacritic may consist only of a quarter and a year. We accommodate such cases (less than 300 in total) by postulating the release date as the first day of the first month in that quarter. Furthermore, we prune all reviews written in languages other than English via Lui and Baldwin’s language identification tool [47]. We list the descriptive statistics of our pre-processed dataset in Table 1. Our dataset features more than 400k games released over the course of 34 years and more than a million reviews. Despite the large imbalance in the number of experts (in total 482) compared to the number of amateurs (234,853), we observe a roughly 40-60% split of reviews authored by experts and amateurs.

Preliminary Descriptive Analysis. As expert reviews on Metacritic are only synopses of their full review, their median review length (in words) is considerably shorter than the amateurs’ (c.f.

Table 1. Dataset statistics.

Dataset Statistics	
# games	417,391
# game genres	174
# platforms	21
# developers	115,662
Game release years	1984–2018
# experts	482
# amateurs	234,853
# reviews	1,088,731
... thereof # expert reviews	457,228
... thereof # amateur reviews	631,503
Median length of expert reviews (in words)	32
Median length of amateur reviews (in words)	76
# helpful votes in amateur reviews	4,717,255

Fig. 2. Dataset overview. We visualize key characteristics of our dataset (from left to right) and assess their impact on our empirical analyses: lengths of review texts, number of games released per year and number of reviews produced. In Figure 2a, we show the distribution of review length in words (amateur review length distribution truncated at the 90th percentile). Metacritic provides only expert review summaries and allows amateurs to produce longer reviews. We thus control for review length in our subsequent analyses. In Figure 2b, we depict the number of games released per year with a minimum of three expert and three amateur reviews on Metacritic. This number increases over time but stabilizes at an average of around 800 games since the year 2003. This period thus allows for isolating rating and reviewing biases with respect to older games. We illustrate the distribution of the number of reviews of experts and amateurs in Figure 2c. The number of reviews by amateurs follows a heavy-tailed distribution, whereas experts have few commonalities in the number of reviews they produce. We thus expect different levels of experience reviewing video games to correspond to complementary (rather than overlapping) viewpoints.

Figure 2a for a summary of these distributional differences). Opposite to the experts, there is a considerable amount of long and detailed amateur reviews, which are substantially different from short, evaluative summaries. Hence, to ensure a fair comparison basis, we control for length in our subsequent analyses of review text. Our control consists of an upper limit in the word count by reviews of both experts and amateurs. To determine a threshold for the word count upper limit, we compute the Standardized Mean Difference (SMD) on the following covariates of the text length: counts of characters, syllables and sentences. SMD is the difference between the means of a covariate in amateur and expert reviews divided by the standard deviation of that covariate in amateur reviews. We set the word count upper limit to the 95th percentile (i.e. 72 words) of the word count distribution of expert reviews, as this value minimizes the median absolute SMD of the three covariates. Changing the word count upper limit to values in the range of balanced covariates (i.e. those with absolute SMD smaller than 0.25 [65]) did not significantly alter our results.

Further, any trends in amateur rating over the years could be confounded by the fact Metacritic features fewer older games (cf. Figure 2b), so the available ones are the more popular ones. However, between 2003 and the end of 2018, the number of games released per year and reviewed at least three times by both experts and amateurs on Metacritic stabilizes at around 800 games. Hence, in our analyses of long-term rating behavior, we focus on this time period, where trends in amateur rating over the years should not be (strongly) confounded by growing numbers of games available.

Finally, we inspect how many experts and amateurs produce how many reviews in Figure 2c. We find a heavy-tailed curve of amateur review counts, as most amateurs write a few reviews only and few of them write more than 1000 reviews. Experts, however, have much less commonalities in the amount of reviews they write. Interpreting review counts per reviewer as an experience indicator, we expect this discrepancy may result in opinion divergences between experts and amateurs.

Fig. 3. Rating distribution. The review rating density of amateur and expert reviews differs in the number of peaks and their locations, as amateurs rate games notably often a zero or a ten, while experts an eight. This indicates higher polarization of ratings by amateurs in contrast to experts.

Methodology. To address our research questions, we first resort to a descriptive statistical analysis of rating by experts and amateurs. Specifically, to address our first research question, we examine the probability distribution of ratings by experts and amateurs.

For our second research question, we first perform descriptive statistical, comparative analyses of ratings by experts and amateurs along multiple dimensions, such as game metadata, helpfulness voting behavior by amateurs, and time (in the short- and long-term). Following this numerical study, we inspect the linguistic characteristics of the reviews themselves. To that end, we focus on textual form and style via analyses of textual complexity, readability and part of speech, and we assess review text content via sentiment analysis, subgroup discovery and Latent Dirichlet Allocation [4], an unsupervised topic model.

This empirical analysis underlies our answer to the third research question, i.e. our design of prediction experiments to forecast game reputation among amateurs from expert and amateur rating and reviewing data. We present the standard machine learning techniques we employ alongside our exposition of these prediction tasks.

4 EMPIRICAL RESULTS

We structure our analyses of expert and amateur video game reviewing discrepancies in two components corresponding to research questions RQ1 (Section 4.1) and RQ2 (Section 4.2). In the former, we inspect review rating distributions, and, in the latter, we characterize them across game metadata, temporal, and textual dimensions. Then, we leverage those insights to address RQ3 via prediction models for video game reputation among amateurs (Section 4.3).

4.1 RQ1 (*Existence*): Is there substantial disparity between experts and amateurs in their video game appraisal?

We answer **RQ1** positively, i.e. we find substantial disparity in video game appraisal by experts and amateurs. Operationalizing video game appraisal as review rating, we observe, in Figure 3, the rating distribution of expert ratings is unimodal and right-skewed, as experts most often award the grade eight. In contrast, reviews by amateurs form a bimodal rating distribution, with a peak at grade zero and another at grade ten. Distributional differences are significant at $p < 0.001$ according to a Kolmogorov-Smirnov two-sample test. Further, although the median rating of both experts and amateurs is equal to eight, the top quartile of expert ratings is eight and the bottom quartile six, and respectively ten and five in amateur ratings. Differences in mean and standard deviation of ratings underscore the rating discrepancies: The mean and respectively standard deviation of expert

Fig. 4. Mean rating of genres, developers and platforms of games with most reviews by both experts and amateurs. In this and all subsequent Figures (where applicable), error bars indicate bootstrapped 95% confidence intervals, many of which are too small to be visible. Although genre preferences appear more universal than platform and especially developer ones, we observe significant differences in all of them, signaling further discrepancies in expert and amateur tastes.

ratings is 7.34 and 1.62, whereas that of amateur ratings is 6.97 and 3.4 (bootstrapped differences in mean and in standard deviation are both significant at $p < 0.001$). To frame these results, out of 14,111 games with at least three expert and three amateur reviews, expert rating differs in absolute values from amateurs' by two or more grades in 4,945 games, i.e. in more than 35% of cases.

Combined, these findings indicate a strong disparity between expert and amateur video game ratings, and thus their appraisal. Moreover, we observe a more balanced rating behavior by experts, when compared to rather polarized ratings given by amateurs.

4.2 RQ2 (Characterization): How can we characterize potential disparities, in particular along temporal and textual dimensions?

In our answer of RQ2, we uncover rating discrepancies related to game metadata, amateur review helpfulness, time, and review text. Specifically, amateurs signal clearly different appreciation for top game developers, genres and platforms, when compared to experts. By upvoting critical amateur reviews more highly when disagreeing with experts than when both agree on a rating, amateurs appear to signal the strength of expert-vs.-amateur discrepancies in video game appraisal. Experts award comparable ratings over the years and review games shortly after release. On the other hand, amateurs give lower and more polarized ratings over time, and not only react more slowly to a game release, but also review games years after it and under an increasingly positive light. Finally, experts write rather formal, complex and detached reviews in comparison to amateurs, who exhibit more emotions in their reviews. We also find evidence of nostalgic reminiscence in reviews by both.

Experts Are More Attuned to Game Meta-Properties. We turn our attention to relationships between expert and amateur ratings of game meta properties. Figure 4 highlights differences in mean rating associated with genres, developers and platforms of most reviewed games by both experts and amateurs. Specifically, we plot top ten genres, developers and platforms which gathered the highest number of reviews by both experts and amateurs. All differences between expert and amateur mean ratings across genres, developers and platforms are statistically significant ($p < 0.001$, Welch t-test, except for the Sci-Fi genre ($p = 28.44$)). We see experts rate on average almost all genres, developers and newer platforms notably higher than amateurs. This indicates experts are more attuned to and amateurs more critical of certain meta-properties of games they review.

Amateur Review Helpfulness Signals Discrepancy. Next, we investigate how helpfulness ratios (i.e., the number of “helpful” votes divided by total votes) of amateur reviews relate to

(a) Distribution of helpfulness ratio per absolute deviation of rating from the mean
 (b) Distribution of helpfulness ratio per signed deviation from the mean rating of games on which experts and amateurs agree
 (c) Distribution of helpfulness ratio per signed deviation from the mean rating of games on which experts and amateurs disagree

Fig. 5. Helpfulness of amateur review rating relates to its deviation from mean game rating. In Figure 5a, we see helpfulness of amateur reviews declines with higher absolute deviations of review rating from mean game rating. In Figures 5b and 5c, negative x-axis values indicate when a review awards grades lower than a game's average. Median review helpfulness is higher for reviews with above-average ratings when amateurs agree with experts (Fig. 5b), and notably higher for below-average ratings in the case of disagreement (Fig. 5c). This suggests the use of review helpfulness as a metric for the strength of expert-amateur discrepancies in rating, and it also indicates amateurs coordinate to signal distaste for games experts and other amateurs did not sufficiently condone.

discrepancies between expert and amateur reviews. In that direction, Figure 5a shows boxplots with the distribution of helpfulness ratios conditioned on the absolute deviation of the review rating from the overall mean rating of a game. We observe that helpfulness ratios of reviews are in general considerably lower if their rating differs from the overall mean (see [12] for comparable results in another domain). For a more detailed analysis, Figure 5b and 5c show distributions of signed deviations restricted to the 500 games with the lowest and respectively highest differences between expert and amateur ratings³. For these two groups, we observe opposite effects. For games where experts and amateurs agree, more positive reviews are generally considered more helpful: Note median helpfulness is overall higher for positive deviations from mean rating than for negative deviations (cf. Fig. 5b). In contrast, when experts and amateurs disagree, moderately more critical reviews are seen as decisively more helpful: The median helpfulness of reviews with ratings one to three grades below the mean is higher than those of reviews with above-average grades (see Fig. 5c). Recalling from Figure 3 that rating discrepancies often occur when experts award a high grade and amateurs a low one, we deduce amateurs give more helpfulness votes to more punishing reviews. The analogous holds for reviews with high grades, as amateurs upvoted slightly more positive reviews. Overall, we thus find review helpfulness signals expert-amateur rating disagreement.

Experts Review Games Swiftly After Release; Amateurs (Still Review) Years After. We illustrate the reaction times of experts and amateurs to game release⁴, in a time scale of days in Figure 6a and months in Figure 6b. Overall, experts swiftly review games upon their release, posting their reviews in a median of nine days after game release (cf. Fig. 6a). Amateurs, on the other hand, review games in a median of 65 days after release. Further, amateurs review games for remarkably longer periods of time after release than experts, as signaled by the gap in the amount of games amateurs, rather than experts, reviewed as late as ten to 100 months after release (cf. Fig. 6b).

³We consider only games with at least three reviews.

⁴Here, we include only reviews from after game release; experts and amateurs may also review beforehand (via e.g. previews).

(a) Distribution of Days Passed between Game Release Date and Reviews (b) Count of Games With Reviews per Month Since Game Release (c) Mean Daily Expert Rating Frequency Relative to the Grade (d) Mean Weekly Amateur Rating Frequency Relative to the Grade

Fig. 6. Timing and rating of reactions to game release. Fig. 6a depicts distributions of days passed between game release dates and reviews produced by experts and amateurs; Fig. 6b the number of games with reviews per month after game release date. Experts review games rather shortly after release, while amateurs produce more reviews for extended periods of time. We illustrate, in Fig. 6c and 6d, daily expert and respectively weekly amateur rating frequency relative to all reviews with a given grade. We choose a daily resolution for expert ratings since experts review games earlier than amateurs. Expert ratings converge faster, and are not as polarized as amateurs': Most frequent expert ratings early on are of rather similar colors (hence overall similar), while high frequencies of grades zero and ten by amateurs reveal polarization shortly after game release.

Amateurs, Moreso Than Experts, Award Polarized Ratings Shortly After Release. We visualize expert rating behavior in the first two weeks after game release in Figure 6c, and amateur ratings in first eight weeks in Figure 6d. Specifically, we plot, over time, the frequency of review ratings relative to all reviews of a given grade. In the plot of expert ratings given shortly after game release (cf. Fig. 6c), we observe not only rather balanced ratings (as grades of similar color are next to each other even early on), but also a rapid convergence of rating frequencies after six days post game release. As depicted in Figure 6d, the highest proportion of amateur grades are given in the week of game release. In contrast to experts, grade zero and grade ten are among the top three grades amateurs most frequently award in the release week. Further, more than 62% of reviews with grade zero and more than 50% of those with grade ten are written in these first eight weeks after release. By week four, grades awarded by amateurs all have roughly the same frequency. Overall, this rating behavior may indicate rather polarized reactions of amateurs shortly after game release.

For Amateurs, Games Were “Better In the Good Old Days”. We observe, in Figures 7a and 7d, that, for games released in or after 2003, mean rating of amateur reviews increases, while standard deviation decreases, as the years pass after game release. We remark, again, this trend could be confounded by a relative lack of older games to review in comparison to newer ones, but Figure 2b dispels this concern for the time period spanning 2003 until 2018. Hence, we see this finding as an indicator for a positive rating bias, by amateurs, towards older games. We do not replicate the visualization for expert ratings, since experts review games shortly after release (cf. Fig. 6a).

Amateurs Become Stricter and More Polarized Over Time In Comparison to Experts. In Figures 7b and 7e, we observe (i) constant mean and slightly declining standard deviation of ratings by experts, in contrast to (ii) declining mean and strongly growing standard deviation of ratings by amateurs. Recall expert reviews from before 2010 are not dated on Metacritic. However, as experts review a game rather shortly after release (cf. Fig. 6 and discussion thereof), we estimate review year as the year of game release, and we represent mean estimated expert rating for the years before 2010 as a dotted line, which is in line with the development after 2010. These findings indicate experts seem to rate rather evenly over time, suggesting the existence of expert consensus. We

Fig. 7. Long-term temporal evolution of ratings by experts (solid lines) and amateurs (dashed lines). The mean of review rating by amateurs increases, while the standard deviation decreases, with time passed since game release (cf. Figs. 7a and 7d). This indicates a positive rating bias of amateurs with respect to older games. Note, in Figs. 7b and 7e, the mean and standard deviation of review rating of experts does not change much over the course of calendar years (expert reviews of games from before 2010 are not dated – the dotted line segments indicate estimated expert review years). Amateurs, however, give lower mean ratings over time, and the standard deviation of their ratings increases. This may indicate amateurs are getting stricter and more polarized over time. As far as monthly ratings are concerned (cf. Figs. 7c, 7f), there is a noticeable peak in expert ratings from August until November and a corresponding trough in amateur ratings around November, potentially signaling strong reactions to a surge in game releases before the holiday season.

interpret the declining mean ratings by amateurs as a sign of increasing strictness with games they review, rather than more recent games having lower quality (which expert reviews would have also reflected). Growing standard deviation in amateur ratings suggests increasing polarization among amateurs. Plotting mean and standard deviation of ratings per month (cf. Figs. 7c and 7f), we also observe seasonal differences in mean ratings by experts and amateurs, in particular in the months leading up to the holiday season.

Experts Write More Complex and Detached Reviews, While Amateurs More Accessible and Emotional Ones. After controlling for review text word count (cf. Fig. 2a and discussion thereof), we observe experts use longer words and sentences than amateurs, and the expert reviews are less readable according to the Flesch and Flesch-Kincaid readability scores⁵ (differences significant at p-values < 0.001, Welch t-test). The samples of reviews of highly rated games we list in the upper part of Table 2 exemplify the more complex textual formulations employed by experts.

⁵We caution that reading ease scores may not be meaningful measures in this context of short texts and that pre-processing may impact readability score [34, 56].

A part of speech (POS) analysis of expert and amateur reviews reveals how experts and amateurs structure their reviews. Experts use more nouns (POS tag NN) and adjectives (JJ), as well as more conjunctions (IN and CC) and the 3rd person singular (VBZ): They write more like Academic English [3] and employ a more well-structured and distanced style. In contrast, amateurs notably use more personal pronouns like “me” and “myself” (PRP), suggesting they employ a more personally involved narrative in their video game reviews (differences in mean POS tag frequency are significant at $p < 0.001$, Welch t-test). The exemplary amateur review of a lowly rated game at the bottom of Table 2 gives a personal account of the amateur’s experience with the game.

We further explore differences in emotional involvement via a sentiment analysis of review text. First, we observe the Spearman-rank correlation between review rating and compound review text sentiment, as estimated by VADER [27], values 0.31 for expert reviews and 0.55 for amateur reviews (both significantly different from zero and from each other, bootstrapped $p < 0.001$). Further, positive sentiment of expert reviews is on average 0.18 while that of amateurs 0.23, and negative sentiment of expert reviews is also on average lower than that of amateurs, at 0.08 vs. 0.09 (all differences significant, bootstrapped $p < 0.001$). This supports our view amateurs write more emotionally charged reviews than experts (cf. examples of Table 2).

Additionally, we employ a subgroup discovery technique with a binomial test interestingness measure [2] to identify words that are characteristic for either expert or amateur reviews, i.e., words that appear significantly more often in one of the review sets⁶. Overall, experts address certain game-related aspects more verbosely and less bombastically, as they mention “narrative” and “presentation” and employ substantives such as “charm” while avoiding adjectives like “horrible”. On the other hand, amateurs use simpler, more emotionally charged vocabulary (such as “love” or “suck”) in their reviews, which thus appear more exaggerated. Further, both example review comparisons we list in Table 2 support that review vocabulary is in line with our previous findings.

In Their Reviews, Experts Focus on Factual Descriptions of Gaming, While Amateurs Describe Gameplay and Experiences. To assess topic-level differences in expert and amateur reviews we apply a commonly used topic modeling technique, Latent Dirichlet Allocation (LDA) [4]. We adapt the setup outlined by Zhang et al. [72] and also use the Stanford Topic Modeling Toolbox [22] in our LDA application. In Appendix Section A.1, we provide details on our pre-processing procedure for LDA, a procedure which leads us to report on 15 topics. We list, in Table 3, the top ten

⁶As a result, the top 10 words most characteristic to expert reviews are “single player”, “likely”, “charm”, “fails”, “narrative”, “offering”, “certainly”, “presentation”, “might” and “manages”. For amateur reviews, the most characteristic top 10 words are “say”, “money”, “bad”, “think”, “love”, “people”, “bought”, “horrible”, “play” and “suck”.

Table 2. Exemplary reviews [sic] of highly and lowly rated games by experts and amateurs illustrate structural, lexical and emotional discrepancies between their review texts.

Game	Author	Grade	Text
MLB The Show 17	Expert	10	MLB The Show 17 builds on last year’s version, which many cited as the best in the series, by adding some key features. [...] All of these additions make the game feel more like real baseball. This authenticity makes games more fun to play, as it’s easier than ever to become invested in every pitch.
	Amateur	9	Freaking awesome. Any true ballplayer would appreciate the game they’ve put together with The Show. Didn’t play another game for a full 7 months after it’s release... it’s THAT good.
Hannah Montana: The Movie	Expert	3	Hannah Montana: The Movie isn’t a total flop. There are some good points to the game such as the array of songs, fairground games and the gorgeous Miley Cyrus, while this game isn’t worth the price tag it’s definitely worth a rent.
	Amateur	0	Worst game on the Wii ever. My little sister played this game and said loudly “I hate this game”. So I took a look at the game to fin the reason why my lil’ sis hates it. I totally agree with her hating it. The horrible graphic, tedious gameplay, and dumb minigames made this game suck BIG TIME. 0/10

topics (i.e. with greatest weight) we extract from reviews by experts and separately amateurs. The table includes a name for each topic, which we annotate manually, top ten most frequent words per topic and average topic weight per review. We include, in Appendix Section A.2, the bottom five LDA topics. Those topics reflect only video game genres, and we thus do not analyze them further.

These results indicate experts and amateurs discuss general game categories and their feelings about games using different vocabulary registers. Beyond obvious similarities in words used to describe common video game genres, this LDA topic analysis may also suggest game aspects experts and amateurs rather focus on: Experts appear to describe game modalities (words “online”, “single-player” in the topic “Shooter Games”) and context (topic “Sequels” which is not present in amateurs’ topics), while amateurs seemingly rather discuss gameplay itself (words “mission”, “combat” in the topic “Shooter Games”) and appealing gaming concepts (words “experience”, “art” in the topic “Gaming Experience” which is not present in experts’ topics).

Finally, we highlight a clear-cut topic containing words related to gaming nostalgia (“Nostalgia” topic) in reviews by both experts and amateurs. Inspecting less frequent words in that topic and selecting those more closely related to nostalgia from the topic’s top ten words (by experts and amateurs), we get a clearer picture of gaming nostalgia-related words, namely “old”, “classic”, “retro”, “origin”, “collection”, “remastered” and the word “nostalgia” itself. As nostalgia appears inherently related with time, we study the temporal dynamics of those nostalgia keywords, as depicted in Fig. 8. Increasing usage of nostalgia-related vocabulary may point towards its growing influence over time, which may be key to grasping the nature and timing of long-term gamer preferences.

Summary of RQ2 Findings. Discrepancies in video game reviews by experts and amateurs manifest themselves on multiple levels. Experts are less critical of top game genres, developers

Table 3. We list, for experts and separately for amateurs, a label we propose for each of the top ten LDA topics, their top ten most frequent words, and average topic weight per review. The LDA analysis reveals not only how experts and amateurs evaluate games, but also which game aspects they discuss about them, such as genre and in particular also nostalgia for older games.

Experts			Amateurs		
Topic	Top Words	Weight	Topic	Top Words	Weight
Negative Reviews	bad, graphic, lack, repetitive, isn’t, frustrating, design, idea, problem, poor	0.1426	Mixed Reviews	control, addictive, puzzle, worth, simple, little, music, boring, challenge, nice	0.1462
Positive Reviews	puzzle, challenge, platform, unique, adventure, level, design, charm, beautiful, visuals	0.1011	Action Games	fight, awesome, fan, mode, control, action, cool, boss, buy, hard	0.1013
Multiplayer Games	multiplayer, online, price, friend, arcade, worth, party, simple, addictive, lack	0.0876	Racing Games	race, car, mode, year, online, drive, buy, sport, control, awesome	0.1009
Horror Games	adventure, episode, character, horror, dark, end, atmosphere, narrow, survival, shadow	0.0718	Negative Reviews	buy, money, worst, bug, terrible, review, release, horrible, boring, waste	0.1007
Sequels	improves, world, content, origin, expansion, franchise, predecessor, sequel, addition, change	0.0685	Shooter Games	shooter, repetition, mission, weapon, bore, action, combat, nice, awesome, pretty	0.0824
Movie Games	movie, character, adventure, kid, humor, entertaining, gamer, ninja, repetitive, film	0.0630	Nostalgia	amazing, platform, control, origin, music, version, fan, classic, adventure, perfect	0.0723
Nostalgia	platform, challenge, classic, level, genre, design, man, old, legend, difficulty	0.0606	Gaming Experience	end, amazing, beautiful, effects, experience, episode, mass, series, choice, art	0.0690
Strategy Games	strategic, war, star, tactic, space, battle, genre, combat, campaign, depth	0.0570	Strategy Games	war, strategy, total, simulation, bug, battle, campaign, unit, space, patch	0.0527
Shooter Games	shooter, multiplay, campaign, online, weapon, co-op, zombie, dead, solid, single-player	0.0568	Adventure Games	rpg, world, combat, quest, fallout, amazing, dragon, mod, explore, fantasy	0.0510
Sports Games	sport, football, improvement, featuring, baseball, soccer, graphics, simulation, online, basketball	0.0564	Positive Reviews	amazing, perfect, year, awesome, life, think, masterpiece, greatest, simply, excellent	0.0503

Fig. 8. Relative frequency of usage of nostalgia-related vocabulary in amateur reviews. The usage, by amateurs, of vocabulary related to nostalgia clearly increases over the years, hinting at its importance long after games get released.

and platforms than amateurs, who upvote (very) critical reviews in cases of large disagreement with experts. Experts swiftly review games after their release, whereas amateurs review games years after release and are positively biased towards older games. We observe amateurs (in contrast to experts) award lower, more polarized ratings over the years. Review language reflects a more professional and detached tone by experts, while the amateurs' tone is more casual and emotional.

4.3 RQ3 (*Impact*): What are predictive strengths and weaknesses of expert and amateur reviews for short-term and long-term video game reputation?

To assess the extent to which early expert and amateur reviews capture amateur video game reputation, we design two prediction experiments leveraging and validating our previous findings. Our results show early amateur review data is very predictive of both short-term and long-term video game reputation also among amateurs. Expert review data is on par with amateur review data in long-term predictions, and both combine to produce best forecasts in both prediction tasks.

4.3.1 Task Description. We conduct two binary prediction tasks, where we forecast video game reputation. We operationalize video game reputation as amateur review rating. For our binary prediction tasks, we define class separating thresholds. We exhaustively experiment with variations of these thresholds and while predictive performance of all prediction models decreases slightly with a weaker contrast between classes⁷, our results and interpretations remain mostly the same.

Short-Term Prediction Task. Given all review data up to one week after release and video game metadata, we predict if amateur ratings per game, starting one month after game release, will be among the top quartile of amateur ratings (i.e. rating ten) or among the bottom quartile (corresponding to rating five or lower). As the amateur rating distribution is bimodal, a two class prediction target with well separated classes is in line with amateur rating behavior⁸. Figure 6d motivates our choice of the first week after game release as the time window for training prediction models. As amateur review ratings vary the most in this time window, prediction models with good performance would be most useful during this time period.

Long-Term Prediction Task. Given all review data up to one week after release and video game metadata, we predict if the number of reviews a game receives in the time period starting three

⁷We observe that the availability of more data for prediction partially offsets performance decreases due to less contrast between two classes.

⁸Another reason why we do not predict rating grades directly is data sparsity: Amateur grades of e.g. two are very underrepresented (cf. Fig. 3), and effectively predicting such grades would thus require over-sampling or merging into other classes (which would be similar to our setup).

years after release will be in the top or bottom quartile of the distribution of amateur review counts per game. We focus on review counts rather than ratings in the long-term due to the fact most review ratings per game are given in the first two months after game release, and thus average rating per game does not change much in the time period after that (cf. Fig. 6d). Amateur review counts of many games, however, still increase long after game release, as we observed in Figure 6b.

4.3.2 Experimental Setup. The pre-processing routines we use in our prediction tasks involve, as motivated by our discussion of Figure 2a, placing an upper bound on review length in words at 72 words. Further, we impose a minimum amount of reviews per game to include it in a train set. We again set this minimum value at three reviews, to balance between the amount of training instances (i.e. games) available for prediction, and unnecessary noise in the training dataset in the form of games with only two often very disparate (expert or amateur) reviews.

For both of our prediction tasks, we employ standard machine learning configurations provided by Pedregosa et al. [58], i.e., we do not perform hyperparameter search. We measure predictive performance via the area under the receiver operating characteristic curve (ROC AUC), the probability a binary classifier ranks a randomly chosen instance of a given class higher than another randomly chosen instance of the other class. We use a five-fold train-test split and we report on ROC AUC averaged over the folds. For both prediction tasks, we evaluate the classifiers logistic regression, support vector machines and gradient boosted trees. We present only the predictive performance of the latter, as they give the best results.

Prediction Models. Our empirical analysis of rating and reviewing behavior of experts and amateurs informs our choice of features for the two predictive tasks. We create a series of models with these features to assess their individual importance per prediction task. We define the following groups of features:

- (i) Rating: Expert/amateur grade (Fig. 3), number of days elapsed between game release and review dates (Fig. 6).
- (ii) Metadata: Gaming platform, game genre and developers (Fig. 4). We encode these metadata features as three binary variables indicating if a game's platform, genre or developer is among the top ten most frequently reviewed of its kind (or not).
- (iii) Release Time: Year and month of game release date (Fig. 7).
- (iv) Review Form: Number of syllables, number of words, number of sentences, readability features (Flesch and Flesch-Kincaid) (Sec. 4.2 "Experts Write More Complex Reviews").
- (v) Review Content: Review sentiment, binary variables for the presence of words experts and amateurs use in reviews of top and bottom quartile ratings⁹ and, for the long-term task, a binary variable for the presence of nostalgia-related words¹⁰ (Sec. 4.2 "Experts Write More Complex Reviews", Table 3 and Fig. 8).
- (vi) All: All the above mentioned features.

Furthermore, per prediction task, we devise two separate sets of models, with one using only review data from experts and the other only from amateurs. With this choice, we assess not only the extent to which expert ratings and reviews are predictive of amateurs', but we also compare their predictive performance with respect to an amateur data-only model. Also, we assess if both models combined lead to better predictive performance, as this would indicate experts and amateurs complement each other to form a fuller picture about future video game reputation. To further contextualize results, we also compare ROC AUC values of our models with a simple rule-based baseline: For a given game and its ratings in the first week, this baseline predicts the game belongs

⁹These word lists include the top ten most frequent words we uncovered in the subgroup discovery analysis and LDA topics. We use separate word lists for words employed by experts and for those by amateurs, i.e. we keep a total of four word lists.

¹⁰This word list includes words of the LDA topic "nostalgia" of both experts and amateurs.

Fig. 9. We predict game reputation among amateurs in the short-term (one month after game release) and long-term (three years after). We distinguish our models by their training data: reviews by experts (Figs. 9a, 9d), amateurs (Figs. 9b, 9e) and both (Figs. 9c, 9f). Vertical dashed lines indicate baseline model performance. Our features capture the first week of review data (feature groups “Review Content”, “Review Form” and “Rating”), as well as timing (“Release Time”) and other metadata (“Metadata”) information available at game release. In both tasks, a combination of all feature groups and reviews by both yields highest ROC AUC values. Review rating is among the most predictive features, and the importance of other features varies with prediction horizon.

to the top rating quartile in the short-term (and, in the long-term prediction task, to the top review count quartile) if it was in the top quartile in the first week (i.e. in the training dataset) already.

Recall that we include, in pre-processing for prediction experiments, only games with at least three reviews. As the set of games with at least three expert reviews is different from the set of games with at least three amateur reviews, the expert review-based and amateur review-based models include different games and different quantities of games. However, the dataset sizes are comparable: We have a total of 728 games for the expert review-based and 809 games for the amateur review-based short-term prediction, and 904 games for the expert review-based and 1164 for the amateur review-based long-term prediction. Further, the prediction classes are roughly evenly represented in all experiments, as the class splits range from 39% / 61% to 48% / 52%.

4.3.3 Results. Figure 9 summarizes our results in both prediction tasks. At 0.91 ROC AUC, the model trained with all feature groups and on amateur review data achieves high classification performance values in the short-term prediction task (cf. Fig. 9b). In the long-term and using the amateur review-based models, we get 0.73 ROC AUC (Fig. 9e). The performance of the expert review-based models is lower in the short-term (Fig. 9a; 0.8 ROC AUC) and slightly lower in the long-term (Fig. 9d; 0.72 ROC AUC) prediction tasks. These ROC AUC values are significantly better than random or stratified majority baselines, which reach values of only 0.5. The baseline rule-based model (cf. dashed lines throughout Fig. 9) achieves only 0.57 ROC AUC in the expert review-based

and 0.75 in the amateur review-based short-term prediction, and 0.51 in both variations of the long-term prediction. Therefore, our models strongly outperform such a baseline as well.

We achieve maximal classification performance, at 0.92 ROC AUC in the short-term and 0.74 ROC AUC in the long-term prediction tasks, with models trained on both kinds of reviews (cf. Figs. 9c and 9f).

We now analyze the influence of our feature groups on the prediction performance. In all prediction settings, the combination of all features achieves the best performance. The “Rating” feature group appears to be the most predictive set of features in both prediction tasks and with all variants of the training data, with the exception of long-term prediction from amateur reviews. There, a model using only the “Release Time” feature group attains the second highest performance. However, review text feature groups also carry predictive importance, in particular “Review Content” in the short-term prediction, and “Review Form” in the long-term¹¹. Finally, models combining expert with amateur reviews boast classification performance gains in the feature groups “Rating” and “Review Content” ranging from 0.01 to 0.04 ROC AUC.

5 DISCUSSION

In our study of an online video game review aggregator, we uncovered remarkable discrepancies (and few similarities) in video game appraisal by experts and amateurs along temporal and textual dimensions. We leveraged these characteristics for a series of prediction experiments for video game reputation among amateurs, which achieved high predictive accuracy. We now reflect on the implications of these findings for video game market players and the experts vs. crowds discussion.

Existence of Video Game Appraisal Discrepancies. On the level of video game review rating, we found substantial discrepancies, as the rating distribution of experts is unimodal and that of amateurs bimodal. Extrapolating this finding suggests (very) different perspectives of the few vs. those of the crowd on the same experience good may co-exist. Thus, we believe this poses a significant challenge when consolidating expert and amateur opinions of experience goods in general. Consensus finding in such settings requires appropriately understanding and weighing expert and amateur views, a task which benefits recommender systems for online marketplaces.

Characterization of Video Game Appraisal Discrepancies. We observed experts, often video game journalists¹², employ a clinical, detached writing style, which reflects their experience and is reflected in their unimodal rating distribution. Amateurs, on the other hand, often reviewing just a few games, using emotionally charged language and awarding rather extreme ratings, appear dedicated to either glorify games or express their grievances with them, and to at times signal their disagreement with experts. We thus believe gauging overall public opinion of a video game requires adequately discounting the emotion in often extreme video game ratings and reviews by amateurs to distill actionable (or more detached) feedback. This could in turn help video game developers and publishers with video game production. More broadly, our findings suggest reconciling subjective views of the crowd with those of the few entails learning to interpret their signaling mechanisms.

Beyond short-term signals, we also observed how amateurs tend to give lower ratings over the years and better ratings as time passes after game release, a development which may correlate with nostalgia for older games. We believe these trends reflect growing competition between developers in the video game market and increasingly stringent consumer demands, as well as a positive bias towards older classic games and leniency towards their outdated technical prowess.

¹¹Adding the nostalgia-based feature to the “Review Content” group in long-term predictions improves expert review-based models by 0.02 (and amateur review-based ones by 0.01) ROC AUC in comparison to “Review Content” without that feature.

¹²The full list of expert publication venues, available at <https://www.metacritic.com/faq#item20> as of December 2018, includes many well-known websites in the video game milieu, such as gamespot.com or ign.com, for which teams of video game professionals produce content.

Specifically, our study advances current understanding of nostalgia effects [13, 25, 63, 71] in the popular sense, i.e. as a kind of yearning for one's childhood [13, 25, 63], rather than the clinical and psychological sense [71]. With our large-scale empirical analysis, we show video game reviews can be leveraged to quantify strength and timing of nostalgia-related evaluations of video games. We believe these findings may potentially empower not only video game developers but also experienced good producers to assess how and when to recycle elements from older developments.

Prediction of Video Game Reputation. We begin by briefly discussing the “Release Time” feature group: The fact this feature group does not perfectly predict long-term video game reputation simply shows there are older games which fail (and respectively newer games which succeed) at growing long-term reputation. This contradicts the assumption that, as older games have more time to amass more interest and hence reviews, this feature group alone should have great predictive accuracy. Not including this feature group impairs the prediction performance of the expert review-based model by -0.01 and the amateur review-based model by -0.06 ROC AUC. This underscores (i) the strength of expert reviews for assessing long term video game reputation and (ii) the growing importance of the temporal dimension in long-term amateur evaluation of video games.

As experts typically react within a few days of game release, our prediction models facilitate estimations of game reputation shortly after release date. Our prediction experiments also showed early amateur reactions are highly indicative of future reputation. This underlines the potential for game developers to engage with gamers early on and potentially address feedback with updates. Moreover, experts may want to reflect on better capturing short-term concerns by amateurs, as well as focus on their ability to write reviews that stand the test of time. Further, video game recommender systems leveraging review text may obtain performance gains by assigning greater weight to expert opinion of older games. We reason the wisdom of the few may lie herein: Their greater experience, when compared to the crowd, translates into valuable long-term views.

Recall we assess, with models trained on both expert and amateur reviews, if expert reviews can be regarded as complementary to amateur reviews. Using such models, we obtain performance gains in both prediction tasks of 0.01 ROC AUC on the full model and up to 0.04 ROC AUC on the feature groups related to review ratings and content. We find these performance improvements, albeit minor, remarkable, as ROC AUC values were already high with amateur review data only. Hence, we conclude experts (i.e. the few) provide value on their own and in addition to amateurs (i.e. the crowd), as both combine to form a more comprehensive view.

Limitations. In this work, we exclude game ratings of amateurs that do not contribute an accompanying review to their rating. This data on amateur ratings represents, however, another large user population, which silently provides feedback on games. We also expect the full reviews by experts, as written on their own websites and publication venues, to provide further insights into their rating and reviewing behavior. Furthermore, in this work, we characterize the group of experts and amateurs as a whole. We know, however, from Figure 2c and previous work [10, 59], that there are subpopulations of both experts and amateurs with varying degrees of expertise and thus reviewing behavior. Hence, though beyond the scope of the present study, we believe tailoring prediction models to individual amateurs is a promising direction for future work. We also stress our aim in this work is comparing experts and amateurs, rather than optimizing prediction models.

Another issue in review datasets is fake and duplicate review production and identification [12, 39, 46, 52, 53, 55]. As Mukherjee et al. [53] found, behavioral cues, rather than linguistic ones, are most effective in characterizing fake review production. As we do not investigate individual behavior, we also do not control for such fake reviews. However, as noted by Danescu-Niculescu-Mizil et al. [12], fake reviews often feature very similar text. In our dataset and using usual text similarity thresholds of 70% [12, 32] and the Levenshtein distance [38] to match review text (excluding reviews of same game releases in multiple platforms), we estimate a mere 0.1% of all expert reviews and 0.2% of all

amateur reviews are near matches. Both values are well below the usual thresholds of 14% to 16% other studies [46, 53] reported. Therefore, this facet of fake reviews is negligible in our dataset.

Finally, we remark our work is a large-scale study of the video game portion of the website Metacritic. Thus, there is an opportunity to extend our work to other platforms for experts and amateurs, such as MobyGames¹³ or even Steam, which recently introduced curators, an expert type of user role. However, Metacritic is one of the largest websites of its kind, as it aggregates reviews from well-known video game experts (which other websites such as Steam do not), and it attracts activity from many amateurs (in particular more than the comparable website MobyGames).

6 CONCLUSION

Summary. In this work, we presented the first large-scale study of video game reviewing behavior by experts and amateurs. We uncovered substantial differences (and some similarities) in short- and long-term rating and reviewing by experts and amateurs. Amateur reviews and ratings are highly polarized as they often rate video games as zero or ten, whereas expert behavior is more balanced, both in rating and over time. The polarization of amateur reviews is corroborated by amateurs often using emotionally charged vocabulary that exhibits stronger sentiment compared to experts. We leveraged our empirical observations to design a series of prediction experiments that demonstrate the feasibility of predicting video game reputation among amateurs in the short- and long-term and that test the predictive qualities of expert reviews.

Implications. Our work thus highlights challenges in bridging subjective views from experts (the few) vs. amateurs (the crowd) and how to overcome such challenges, as we exemplify in the context of video game appraisal. Knowing which mechanisms experts and amateurs utilize to signal (dis)approval and (dis)agreement facilitates the improvement of review aggregators and recommender systems in online marketplaces. We believe combining such signals and capitalizing on the strengths of both the few and the crowd, in particular the former's experience and the latter's emotions, results in the most comprehensive view of experience goods in general.

Future Work. Extending our textual analyses to the temporal dimension [23] would complement our understanding of rating trends. It would also be fruitful to assess, beyond studies of reviewer experience, if there are influential reviewers whose ratings affect others'. We also believe investigating why there appears to be more consensus among experts than among amateurs, ideally via surveys of their motivations, is promising future work. Finally, our approach can be applied to other domains (e.g. book or movie reviews), as our methods are generic and can be easily extended.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable feedback on the manuscript. Tiago Santos is a recipient of a DOC Fellowship of the Austrian Academy of Sciences at the Institute of Interactive Systems and Data Science of the Graz University of Technology.

REFERENCES

- [1] Xavier Amatriain, Neal Lathia, Josep Pujol, Haewoon Kwak, and Nuria Oliver. 2009. The Wisdom of the Few: A Collaborative Filtering Approach Based on Expert Opinions from the Web. In *SIGIR*.
- [2] Martin Atzmueller. 2015. Subgroup discovery. *WIREs Data Mining and Knowledge Discovery* (2015).
- [3] Douglas Biber. 2010. *Longman Grammar of spoken and written English*. Longman.
- [4] David Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet Allocation. *JMLR* (2003).
- [5] Keith Burghardt, Emanuel Alsina, Michelle Girvan, William Rand, and Kristina Lerman. 2017. The myopia of crowds: Cognitive load and collective evaluation of answers on Stack Exchange. *PLoS ONE* (2017).
- [6] Hailiang Chen, Prabuddha De, Yu Jeffrey Hu, and Byoung-Hyoun Hwang. 2014. Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies* (2014).

¹³mobygames.com

- [7] Zhiyong Cheng, Ying Ding, Xiangnan He, Lei Zhu, Xuemeng Song, and Mohan Kankanhalli. 2018. A³ NCF: An Adaptive Aspect Attention Model for Rating Prediction. In *IJCAI*.
- [8] Zhiyong Cheng, Ying Ding, Lei Zhu, and Mohan Kankanhalli. 2018. Aspect-Aware Latent Factor Model: Rating Prediction with Ratings and Reviews. In *WWW*.
- [9] Judith Chevalier and Dina Mayzlin. 2006. The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research* (2006).
- [10] Georgios Christou. 2013. A comparison between experienced and inexperienced video game players' perceptions. *Human-centric Computing and Information Sciences* (2013).
- [11] Joe Cox. 2014. What makes a blockbuster video game? An empirical analysis of US sales data. *Managerial and Decision Economics* (2014).
- [12] Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian Lee. 2009. How Opinions are Received by Online Communities: A Case Study on Amazon.com Helpfulness Votes. In *WWW*.
- [13] Fred Davis. 1979. *Yearning for Yesterday: A Sociology of Nostalgia*. Free Press.
- [14] Ilona De Jong and Christian Burgers. 2013. Do consumer critics write differently from professional critics? A genre analysis of online film reviews. *Discourse, Context & Media* (2013).
- [15] Qiming Diao, Minghui Qiu, Chao-Yuan Wu, Alexander Smola, Jing Jiang, and Chong Wang. 2014. Jointly Modeling Aspects, Ratings and Sentiments for Movie Recommendation (JMARS). In *KDD*.
- [16] Loretta Dobrescu, Michael Luca, and Alberto Motta. 2013. What makes a critic tick? Connected authors and the determinants of book reviews. *Journal of Economic Behavior & Organization* (2013).
- [17] Anders Drachen, Kevin Bauer, and Robert Veitch. 2011. Only the Good... Get Pirated: Game Piracy Activity vs. Metacritic Score. In *FDG*.
- [18] Wenjing Duan, Bin Gu, and Andrew Whinston. 2008. Do online reviews matter?—An empirical investigation of panel data. *Decision Support Systems* (2008).
- [19] Francis Galton. 1907. Vox Populi. *Nature* (1907).
- [20] Eric Gilbert and Karrie Karahalios. 2010. Understanding Deja Reviewers. In *CSCW*.
- [21] Adams Greenwood-Ericksen, Scott Poorman, and Roy Papp. 2013. On the Validity of Metacritic in Assessing Game Value. *Eludamos. Journal for Computer Game Culture* (2013).
- [22] The Stanford Natural Language Processing Group. 2010. Stanford Topic Modeling Toolbox. <https://nlp.stanford.edu/software/tmt/tmt-0.4/>. The Stanford Natural Language Processing Group (December 18, 2018).
- [23] Will Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *ACL*.
- [24] Thorsten Hennig-Thurau, André Marchand, and Barbara Hiller. 2012. The relationship between reviewer judgments and motion picture success: re-analysis and extension. *Journal of Cultural Economics* (2012).
- [25] Morris Holbrook and Robert Schindler. 1991. Echoes of the Dear Departed Past: Some Work in Progress on Nostalgia. *ACR North American Advances* (1991).
- [26] Nan Hu, Paul Pavlou, and Jennifer Zhang. 2006. Can Online Reviews Reveal a Product's True Quality? Empirical Findings and Analytical Modeling of Online Word-of-Mouth Communication. In *EC*.
- [27] Clayton Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *ICWSM*.
- [28] Daniel Johnson, Christopher Watling, John Gardner, and Lennart Nacke. 2014. The Edge of Glory: The Relationship between Metacritic Scores and Player Experience. In *CHI PLAY*.
- [29] Mahesh Joshi, Dipanjan Das, Kevin Gimpel, and Noah Smith. 2010. Movie Reviews and Revenues: An Experiment in Text Regression. In *NAACL HLT*.
- [30] Dan Jurafsky, Victor Chahuneau, Bryan Routledge, and Noah Smith. 2014. Narrative framing of consumer sentiment in online restaurant reviews. *First Monday* (2014).
- [31] Andrew Kehoe and Matt Gee. 2017. "Please please don't buy this game like I did. I feel terrible and wish I could return it!": A corpus-based study of professional and consumer reviews of video games. In *CL*.
- [32] Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. 2006. Automatically assessing review helpfulness. In *EMNLP*.
- [33] Timothy King. 2007. Does film criticism affect box office earnings? Evidence from movies released in the US in 2003. *Journal of Cultural Economics* (2007).
- [34] George Klare. 1976. A Second Look at the Validity of Readability Formulas. *Journal of Reading Behavior* (1976).
- [35] Gael Lederrey and Robert West. 2018. When Sheep Shop: Measuring Herding Effects in Product Ratings with Natural Experiments. In *WWW*.
- [36] Young-Jin Lee, Kartik Hosanagar, and Yong Tan. 2015. Do I Follow My Friends or the Crowd? Information Cascades in Online Movie Ratings. *Management Science* (2015).

- [37] Xiaojiang Lei, Xueming Qian, and Guoshuai Zhao. 2016. Rating Prediction based on Social Sentiment from Textual Reviews. *IEEE Transactions on Multimedia* (2016).
- [38] Vladimir Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* (1966).
- [39] Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Lauw. 2010. Detecting Product Review Spammers Using Rating Behaviors. In *CIKM*.
- [40] Dayi Lin, Cor-Paul Bezemer, Ying Zou, and Ahmed Hassan. 2019. An empirical study of game reviews on the Steam platform. *Empirical Software Engineering* (2019).
- [41] Guang Ling, Mike Lyu, and Irwin King. 2014. Ratings Meet Reviews, a Combined Approach to Recommend. In *RecSys*.
- [42] Ian Livingston, Lennart Nacke, and Regan Mandryk. 2011. The Impact of Negative Game Reviews and User Comments on Player Experience. In *ACM SIGGRAPH Game Papers*.
- [43] Ian Livingston, Lennart Nacke, and Regan Mandryk. 2011. Influencing Experience: The Effects of Reading Game Reviews on Player Experience. In *International Conference on Entertainment Computing*.
- [44] Jan Lorenz, Heiko Rahut, Frank Schweitzer, and Dirk Helbing. 2011. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences* (2011).
- [45] Michael Luca. 2016. Reviews, Reputation, and Revenue: The Case of Yelp.com. *Harvard Business School NOM Unit Working Paper* (2016).
- [46] Michael Luca and Georgios Zervas. 2016. Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science* (2016).
- [47] Marco Lui and Timothy Baldwin. 2012. langid.py: An Off-the-shelf Language Identification Tool. In *ACL System Demonstrations*.
- [48] Julian McAuley and Jure Leskovec. 2013. Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text. In *RecSys*.
- [49] Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning Attitudes and Attributes from Multi-Aspect Reviews. In *ICDM*.
- [50] Metacritic. December 18, 2018. How We Create the Metascore Magic. <https://www.metacritic.com/about-metascores>.
- [51] Loizos Michael and Jana Otterbacher. 2014. Write Like I Write: Herding in the Language of Online Reviews. In *ICWSM*.
- [52] Arjun Mukherjee, Bing Liu, and Nat Glance. 2012. Spotting Fake Reviewer Groups in Consumer Reviews. In *WWW*.
- [53] Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Nat Glance. 2013. What Yelp Fake Review Filter Might Be Doing?. In *ICWSM*.
- [54] Phillip Nelson. 1970. Information and consumer behavior. *Journal of Political Economy* (1970).
- [55] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey Hancock. 2011. Finding Deceptive Opinion Spam by any Stretch of the Imagination. In *ACL HLT*.
- [56] João Palotti, Guido Zuccon, and Allan Hanbury. 2015. The Influence of Pre-Processing on the Estimation of Readability of Web Documents. In *CIKM*.
- [57] Hyemin Park and Haewon Byun. 2016. Correlation Analysis: Game Professional Score and User Score on Steam. *International Journal of Multimedia and Ubiquitous Engineering* (2016).
- [58] Fabian Pedregosa et al. 2011. Scikit-learn: Machine Learning in Python. *JMLR* (2011).
- [59] Jonathan Plucker, James Kaufman, Jason Temple, and Meihua Qian. 2009. Do Experts and Novices Evaluate Movies the Same Way? *Psychology & Marketing* (2009).
- [60] Lizhen Qu, Georgiana Ifrim, and Gerhard Weikum. 2010. The Bag-of-Opinions Method for Review Rating Prediction from Sparse Text Patterns. In *ACL*.
- [61] David Reinstein and Christopher Snyder. 2005. The Influence of Expert Reviews on Consumer Demand for Experience Goods: A Case Study of Movie Critics. *The Journal of Industrial Economics* (2005).
- [62] Zhaochun Ren, Shangsong Liang, Piji Li, Shuaiqiang Wang, and Maarten de Rijke. 2017. Social Collaborative Viewpoint Regression with Explainable Recommendations. In *WSDM*.
- [63] Robert Schindler and Morris Holbrook. 2003. Nostalgia for early experience as a determinant of consumer preferences. *Psychology & Marketing* (2003).
- [64] Feng Shi, Misha Teplitskiy, Eamon Duede, and James Evans. 2019. The wisdom of polarized crowds. *Nature Human Behaviour* (2019).
- [65] Elizabeth Stuart. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* (2010).
- [66] James Surowiecki. 2005. The Wisdom of Crowds. *Anchor Books* (2005).
- [67] Duyu Tang, Bing Qin, Ting Liu, and Yuekui Yang. 2015. User Modeling with Neural Network for Review Rating Prediction. In *IJCAI*.
- [68] Luke Thominet. 2016. 10/10 Would Review Again: Variation in the Player Game Review Genre. *Technical Communication Quarterly* (2016).

[69] Christian Wagner, Sesia Zhao, Chris Schneider, and Huaping Chen. 2010. The Wisdom of Reluctant Crowds. In *HICSS*.
[70] Feng Wang, Xuefeng Liu, and Eric Fang. 2015. User Reviews Variance, Critic Reviews Variance, and Product Sales: An Exploration of Customer Breadth and Depth Effects. *Journal of Retailing* (2015).
[71] Tim Wildschut, Constantine Sedikides, Jamie Arndt, and Clay Routledge. 2006. Nostalgia: Content, Triggers, Functions. *Journal of Personality and Social Psychology* (2006).
[72] Jason Shuo Zhang, Chenhao Tan, and Qin Lv. 2018. “This is why we play”: Characterizing Online Fan Communities of the NBA Teams. In *CSCW*.
[73] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit Factor Models for Explainable Recommendation based on Phrase-level Sentiment Analysis. In *SIGIR*.
[74] Feng Zhu and Xiaoquan Zhang. 2006. The Influence of Online Consumer Reviews on the Demand for Experience Goods: The Case of Video Games. In *ICIS*.
[75] Feng Zhu and Xiaoquan Zhang. 2010. Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics. *Journal of Marketing* (2010).

A APPENDIX

A.1 Pre-Processing for LDA

Before running LDA on expert and amateur reviews, we (i) concatenate all expert (and separately amateur) length-controlled reviews per game (with a minimum of 3 reviews), (ii) tokenize the review text, (iii) convert it to lower-case, (iv) remove common stop words, (v) apply the Porter stemmer, and (vi) remove the 30 most common and very rare words (i.e. those occurring in less than 10 reviews), as well as video game-specific words such as Nintendo, Zelda and others (to learn about reviewing practices rather than specific video games). We sanity-check our pre-processing routine for LDA by calculating perplexity scores for numbers of topics ranging from 5 to 50. As expected, perplexity scores decline with more topics: They decrease from 1,960 at 5 topics to 1,722 at 15 topics, and stabilize between 1,722 and 1,640 for numbers of topics between 15 and 50. We also inspect the word distribution of LDA topic models trained on 20 and 25 topics, and we conclude the core topics we present are robust to increasing the number of topics, as they remain even in LDA models with more topics. Manually inspecting LDA models trained with between 5 and 10 topics, we find such models amalgamate many finer concepts. Hence, we set the number of topics at 15.

A.2 Remaining LDA Topics

We list the bottom five LDA topics with a label we manually annotate, their top 10 words, and average weights, for expert and amateur reviews in Table 4.

Table 4. LDA topics eleven through fifteen.

Experts			Amateurs		
Topic	Top Words	Weight	Topic	Top Words	Weight
Fighting Games	fight, definite, console, fighter, port, graphics, edit, visual, evil, character	0.0563	Multiplayer Games	multiplayer, campaign, map, single, mode, duty, online, shooter, battlefield, buy	0.0453
Adventure Games	rpg, combat, character, fantasy, quest, dragon, battle, final, warrior, dungeon	0.0552	Free2Play Games	dlc, free, pay, community, content, update, friend, money, team, map	0.0386
Racing Games	race, racer, drive, car, speed, track, arcade, graphic, need, simulation	0.0473	Horror Games	horror, evil, zombie, survival, resident, atmosphere, scary, dead, fan, scare	0.0346
Mobile Games	city, super, mobile, level, platform, mission, app, run, world, iphone	0.0430	Open World Games	amazing, world, uncharted, batman, mission, city, open, awesome, gear, everything	0.0294
Music Games	hero, music, rock, song, guitar, band, dance, rhythm, batman, park	0.0242	Online Games	server, online, city, mmo, drm, expansion, content, single, buy, money	0.0263

Received April 2019; revised June 2019; accepted August 2019