Crowds and Critics: How Professional and Amateur Reviews Differ on Metacritic

ABSTRACT

Both professional and amateur reviews can substantially shape how products fare in the marketplace. Yet, we know little about how these two viewpoints compare. In this paper, we bridge this gap by contrasting professional and amateur reviews within the domain of videogames. Examining nearly 200K reviews from the popular site Metacritic, we find professionals favor phrases describing the broader market, while amateurs emphasize social aspects of playing games. Amateurs skew towards higher scores, but are the primary source of low scores; that is, they show considerably more variance. However, both groups express the same average levels of sentiment. Finally, regressing future videogame sales on reviews, we find that a combination of professional scores, volume of professional reviews, and volume of amateur reviews provides a moderately powerful predictive model for future game sales. This is the first work we know of to systematically contrast professional and amateur reviewers, addressing a question at the heart of many contemporary debates: How do crowds and domain specialists compare?

INTRODUCTION

> Do not buy this game. Do not rent this game, do not look at this game on the shelf, don't even think about this game lest someone at Ubisoft find out and they prep a Just Dance 2. Such would be the end of all things, mark my words.
>
> — *Just Dance* professional review; score: 20/100

> Screw professional reviews. My wife and I have played DDR, Dancing with the Stars, We Cheer, that other Wii game with the microphone—all terrible. This game is great because it is fun. End of story. All you pro reviewers are wrong.
>
> — *Just Dance* amateur review; score: 10/10

The Wii game *Just Dance* released to near-universal critical derision. With a "metascore" of 49/100 based on 21 professional reviews, the review site Metacritic gives the game its worst possible rating category. Yet amateur reviewers loved it—yielding an average 8/10 score over 107 ratings. And at the time of this writing, *Just Dance* has gone on to sell 6.9M units and spawn a franchise with three sequels and versions on all major videogame platforms. While the case of *Just Dance* is in many ways exceptional, it does suggest a question: Just how divergent are professional product reviews from amateur ones?

Particularly in the entertainment industry, the opinions of elite critics and journalists hold sway. Professional film critics such as The New York Times's A. O. Scott are regarded as important for recognizing hidden gems and over-hyped flops. Yet despite this widespread regard, relatively little science exists on how indicative these professionals' opinions are for market success. Nor do we know how they differ from the opinions of ordinary consumers.

In this paper, we bridge this gap by comparing professional critics and amateur reviewers within the context of videogame reviews. Who knows the market best, professionals or amateurs? Who will best predict a product's eventual success or failure? Do professionals and amateurs differ in their expressions of sentiment? How do professionals and amateurs describe products when reviewing? While previous work has investigated the question of predicting sales from reviews [3,6,8,10,13,17,23], the differences between professionals and amateurs have received little attention. With this work, we believe we address a question at the heart of many contemporary debates: How do crowds compare to domain specialists and who is more effective for predicting future outcomes?

We bring together a corpus of 197K reviews from the review aggregator Metacritic with sales data on 7K videogames. In summary, we find that a combination of professional score, volume of professional reviews, and volume of amateur reviews effectively predicts game sales. Amateurs skew toward higher review scores than professionals, but also provide the main source of low review scores. That is, they show more variation. Amateurs and professionals also show very similar levels of average review subjectivity and polarity; however, amateurs tend toward more extreme and more varied expressions of sentiment. Finally, professionals and amateurs systematically differ in their language, with experts focusing on market factors and amateurs focusing on social aspects of playing games.

First, we present an overview of our corpora, characterizing the distribution of professional and amateur review scores and sentiment expressions. Next, we model total sales-to-date as a function of professional and amateur review scores before a game's release. Finally, we train classifiers on review text to find key words differentiating professional and amateur reviews. We conclude this paper by discussing implications this work has for CMC theory and for practical issues, such as combining reviews from professionals and amateurs.

RELATED WORK

Technologies like wikis, blogs, discussion forums, and online review venues have enabled new distributed forms of knowledge production, information sharing, social support, and opinion aggregation [2]. People read movie critics' reviews of new films; now Netflix and Amazon recommend movies based on what fellow moviegoers like. Subject matter experts wrote encyclopedias; now there is Wikipedia. People with health conditions went to doctors; now there's PatientsLikeMe.

In addition to these commercial examples, amateurs have proven valuable across many domains, providing insight and creative energy to fields as diverse as biology, astronomy, mathematics and animation [4, 5, 15, 18]. However, professionals have traditionally provided many of these services, prompting the question: what value are professionals in these systems? Efforts have varied in how they integrate professionals with amateurs. Several prominent Citizen Science projects feature distinct roles for professionals (scientists) and amateurs (citizens/laypeople). For example, birders participating in the eBird project enter their observations of birds. Then, "automated data quality filters developed by regional bird experts review all submissions" and "local experts review unusual records that are flagged by the filters" [21]. On the Encyclopedia of Life anyone can contribute but materials from unvetted contributors are initially marked as unreviewed. EOL curators are "professional scientists" and "experienced citizen scientists" who have applied and been approved [1]. As an example non-scientific case, Metacritic aggregates review scores from a selected set of professional reviewers for movies, games, television, and music. These review scores are compiled into an aggregate "metascore" that measures the overall quality of a product. Amateurs may also post reviews to the site that are displayed alongside these professional reviews, but are not counted toward the overall metascore of a title.

Quantifying the value of amateur activities is a challenging task. Predicting books sales [3, 13], film box office sales [6, 23], and (to a lesser extent) videogame sales [9, 19] from online reviews or blog chatter have all attracted attention as ways of grounding the value of amateur reviewers. In early work in this area, Gruhl et al. demonstrated blog chatter can predict spikes in book sales rank on Amazon [13]. Chevalier and Mayzlin found improvements in review scores on book sales sites precede increases in book sales [3]. Forman, Ghose, and Wiesenfeld found that increases in product reviews including identity-descriptive information associated with subsequent increases in online book sales [11].

Predicting box office sales has found volume of mentions on blogs or Yahoo!'s movie pre-release discussion forums can predict film box office sales [8, 17]. Results conflict, however, on whether review score averages are [6, 14] or are not [8, 17] predictive of sales, beyond the power of pure volume of reviews. Sentiment has also often proven useful in predicting sales. Yu et al. found sentiment expressed in blog posts can improve an autoregressive model through a carefully chosen lexicon of sentiment-laden words [23]. Specialized classifiers that differentiate product description text from opinion related text were shown to improve sales prediction by Ghose and Ipeirotis [11]. We address the question of review subjectivity and its relation to sales prediction and reviewer expertise, finding no strong predictive powers in the domain of videogame sales.

Few of these early works explicitly examine whether and how different types of reviewers produce varying review scores. Dellarocas et al. note a weak correlation between amateur and professional film reviews and find that removing professional review information (average score only) is slightly less detrimental to their model's performance than removing amateur review information (a combination of average score, number of reviews, entropy of review gender distribution, entropy of reviewer age distribution) [6]. Kim et al. find frequency of online word-of-mouth mentions and expert review valence are predictive of box office sales [16]. Plucker et al. compare student, online novice, and professional film reviewers, finding greater experience watching films correlates with increased similarity to professional reviews [20].

Earlier studies of online reviewing examined amateur reviewers in terms of their motivations for reviewing [7, 12, 22] and characteristic language used in videogame reviews in particular [24]. We complement this work through an assessment of the differences between professional and amateur reviewers through examining review volume, score, and sentiment rating in a linear model that allows direct comparison of predictive power. Further, our analysis of review text and score distributions provides in-depth insight into how professional and amateur reviews differ. Rather than characterize amateur reviewer traits in isolation we seek to understand how they compare against professionals.

Our work broadens previous analyses of videogame sales that have been limited to smaller subsets of games or sources for review data. Ehrenfeld's thesis found volume of game mentions on the NeoGAF online game discussion forum is predictive of game weekly sales in a support vector regression model [9]. Marcoux and Selouani employed an autoregressive neural network model to predict games sales from review scores, volume, and related features using data from the IGN videogame news and reviews website after first performing nonlinear transformations of the model data [19]. Compared to these approaches we examine a much larger set of games[1], compare professional and amateur reviews, and draw from a longer-term and larger set of reviews that aggregate over websites. Our results are directly interpretable in terms of relative feature impact in linear models.

METHOD

We crawled every amateur and professional review for videogames on the Metacritic website[2] through December 7, 2012. We chose videogames both because it is a relatively understudied product domain (movies and books have received far

[1] Ehrenfeld does not report the number of games used in his analysis, but limits to data over the course of 42 weeks of releases, which is a subset of our total corpus spanning 425 weeks. Marcoux and Selouani examine 74 games, while we examine 860 or 600, depending on the model.

[2] http://www.metacritic.com

more research attention), and also because videogames fit into a broader research project. For every game we collected the following information: console (the hardware the game software was made for), title, publisher (company responsible for distributing the game), developer (studio responsible for making the game), release date, current metascore, current average "user" (amateur) score, genre (according to Metacritic), and ESRB rating (an age-appropriateness rating). Note that some titles may appear on multiple consoles. We treat these as separate games as they reach potentially different audiences and may vary in their implementation.

From every game we also collected all amateur and professional reviews, including their text, review score, time of review, and a flag indicating whether the review came from a amateur or professional. Metacritic converts professional review scores from many formats (such as letter grade, 0-10 range, 0-100 range) into a 0-100 score. Amateur reviews are limited to a 0-10 score range. For comparison, we divide all professional scores by 10 to have all reviews on a 0-10 scale. Metacritic only provides summary excerpts from professional reviews and we limited ourselves to this text to make review length more comparable between amateurs and professionals. Our final corpus consists of 197,383 reviews: 138,843 from professionals and 58,540 from amateurs. Of these, 4,331 professional and 729 amateur reviews covering 958 games were in the first 10 weeks prior to the release of the game—these were used in our future sales prediction tasks.

VGChartz sales data

For sales data we scraped information from the VGChartz website[3]. VGChartz tracks weekly, annual, and lifetime game sales data from a variety of outlets and is primarily targeted toward sales of games in the United States. Their data is most accurate for console games (rather than computer or mobile phone), so we limit ourselves to examining the following consoles using only US sales data: Sony's PlayStation 3 (PS3), PlayStation Portable (PSP) and Vita; Nintendo's Wii, DS, and 3DS; and Microsoft's Xbox360. These consoles are considered to make up the current generation of videogame hardware and are the primary game console distribution platforms. VGChartz typically only records the first 10 weeks of game sales at weekly granularity, limiting our analysis to these weeks. We collected total lifetime sales-to-date for 7,467 games and weekly sales for the first ten weeks of game sales for 4,902 games.

Statistical Methods

We apply three different methods to compare and contrast professional and amateur reviews: regressing sales on reviews, extracting sentiment from text, and searching for keywords distinguishing professionals from amateurs.

Sales prediction regression

Our sales predictions investigated:

1. predicting total lifetime sales using reviews in the 10 weeks prior to the release of a game (total vs pre-10)
2. predicting net first 10-week sales using the same pre-release review subset (10-week vs pre-10)
3. predicting lifetime sales comparing metascore, professional, and amateur review scores using all available reviews (total vs meta)

Predictive accuracy was assessed through mean squared error (MSE) computed with leave-one-out cross-validation via bootstrapped resampling with R's boot package[4].

After matching game reviews to sales data using game titles and consoles, we had lifetime sales data for 6,809 games, and weekly sales for 600 games. Of the 6,809 games with lifetime sales 2,902 had both professional and amateur reviews, while 839 additionally had metascores. Metascores are only assigned when at least four professional critics approved by Metacritic have reviewed the game.

We took a conservative approach of only keeping exact game title matches between Metacritic and VGChartz data without manipulating titles (e.g., using lowercase versions or removing punctuation). We believe it is better to have a slightly smaller dataset than misattribute review scores to different games. All games were matched by a concatenation of title and game console, as games on different consoles vary in implementation, release date, and audience demographics.

Reviews and sales data were aligned based on weeks since release. Reviews were binned into weekly periods based on time since the release of the game. For example, all reviews within 7 days of the game's stated release date (for that console) were binned into the first week. Sales values were logged to get closer to normality. From reviews we constructed features for mean and median review scores, number of reviews, and review length.

Review sentiment extraction

We employed the Pattern toolkit[5] to analyze review text sentiment. For every review, we parsed the text into sentences, computed per-word sentiment (both polarity and subjectivity) and computed the per-review mean, median, minimum, maximum, and variance of sentiment values. Subjectivity values measure the extent to which a text conveys an opinion. Polarity values estimate that opinion as positive or negative. Below we refer to these per-sentence aggregation values as the review "mean", "median", "minimum", "maximum", or "variance" of polarity or subjectivity. "Overall" subjectivity and polarity measure the review text as a whole, instead of computing per-sentence values and aggregating them. Considering these metrics allowed us to study expressions of average review sentiment (median, overall), skew in review sentiment (mean vs median, variance), or extreme review sentiment (minimum, maximum). All features (including score features above) were centered and scaled by subtracting their mean and dividing by their standard deviation.

Reviewer text analysis

We examined the power of review text words to classify reviewers as professionals or amateurs (labeled 0 and 1, respectively) using penalized binomial regression with a lasso penalty to favor the use of a small set of terms. We prepared

[3] http://www.vgchartz.com

[4] http://cran.r-project.org/web/packages/boot

[5] http://www.clips.ua.ac.be/pages/pattern

predicted variable	review set	total games	R^2	MSE	null deviance	control deviance	model deviance
lifetime sales-to-date	pre-10	600	0.360	0.669	562	500	360
net first 10 week sales	pre-10	600	0.421	0.642	599	502	347
lifetime sales-to-date	meta	839	0.446	0.586	839	736	464
lifetime sales-to-date	meta sub	839	0.430	0.599	839	736	478

Table 1: Linear regression models. Review set indicates which subset of reviews were used: "pre-10" indicates only reviews from the 10 weeks prior to the launch of the game, "meta" indicates the full review database and metascore information, and "meta sub" is the same model without metascore information. Mean squared error (MSE) values are from leave-one-out cross-validation of models. Null deviance reports null model deviance; control deviance includes only console, game genre, and maturity (ESRB) rating as predictors; model deviance uses full set of predictive variables including controls. All deviance differences were significant (χ^2 test $p < 0.001$).

our review texts using several standard methods from text analysis using the R programming language package tm.[6] We removed whitespace, punctuation, numbers not part of words, and common English words. All text was lowercased and stemmed to group similar words. We tokenized the reviews into single words, requiring words be at least 3 letters long and appear in at least 10 documents. After constructing a full corpus with 86,376 terms we removed the most sparse terms to produce a set of 2,581 terms.

For binomial regression we used R's glmnet package[7] for generalized linear models. These models account for collinearity of terms (their frequent appearance together altering relative importance) and can control for sparsity (relative infrequency of terms). Controlling collinearity is important to prevent overweighting words that often appear together. Accounting for sparsity enables the model to exclude terms with little predictive power, yielding a smaller and more interpretable set of results. We used a lasso penalty to encourage the model to use the smallest set of terms possible. The parsed text corpus was used to examine review text but not for sales prediction—future work should explore which text features are most predictive of game sales and how they relate to reviewer types.

RESULTS

Before describing results on our prediction tasks it is important to understand the characteristics of the review and sales corpora we collected. It is not apparent *a priori* how Metacritic amateurs behave compared to the professionals whose reviews are featured on the site. Below we explore how professionals and amateurs differ in aggregate scoring and review sentiment expressions.

Review scores

Amateurs and professionals show clear differences in review score assignment (Figure 1). Professionals tend to provide reviews distributed more tightly around a mean 7.2 and median 7.5 score, while amateurs show greater variation with a mean 7.4 and median 9.0 score (Table 2). A Wilcoxon test found these differences to be significant (p $<$ 0.001). Amateurs favor providing higher scores than professionals, but are the only ones likely to provide low review scores, seen by a larger portion of the cumulative review distribution on lower scores (Figure 1a). Amateurs tend to provide few reviews on any single game, while professionals tend to more frequently provide many reviews (Figure 1b). Thus, amateurs tend to focus on a few high-profile games, but provide a wider range of scores than professionals.

Factors relating to professionals' role and amateurs' motivations may explain these differences. Professionals are often under threat of blacklisting for providing low review scores and have a reputation to preserve by not consistently giving high review scores. Game distributors are also unlikely to provide reviewers with free game copies for review if they anticipate low ratings, while professionals are unlikely to review low profile and low quality titles that will not drive traffic to their websites. These factors could combine to skew professionals towards reviewing games generally favorably without providing overly positive reviews.

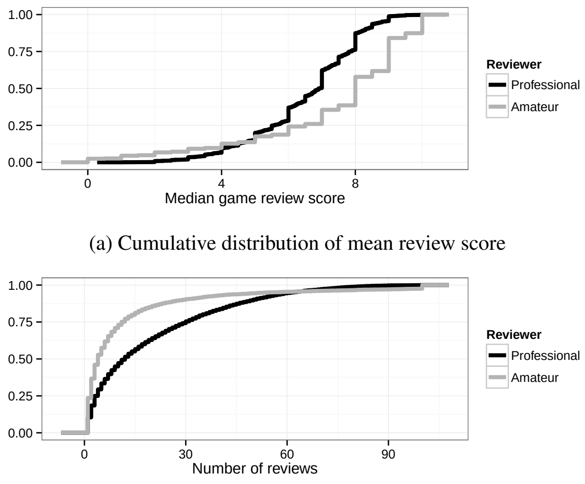
Amateurs, by contrast, are most likely to review games that provided a great or terrible experience [12]. As Metacritic is a major game review outlet, reviewers are likely to provide high scores to games they enjoyed, while attacking games they found poor quality or a waste of money. "Fanboy" culture also likely plays a role in driving reviewing behavior. Our text analysis results corroborate these findings, showing amateurs often describe product value, while professionals focus on

	professionals		**amateurs**	
	mean	median	mean	median
score	7.18	7.5	7.44	9.0
mean polarity	0.14	0.12	0.15	0.13
median polarity	0.14	0.10	0.12	0.00
max polarity	0.25	0.22	0.51	0.50
min polarity	0.04	0.00	-0.17	-0.06
variance polarity	0.03	0.00	0.08	0.06
overall polarity	0.10	0.09	0.07	0.06
mean subjectivity	0.50	0.50	0.48	0.48
median subjectivity	0.50	0.50	0.47	0.50
max subjectivity	0.65	0.69	0.86	1.00
min subjectivity	0.36	0.37	0.15	0.00
variance subjectivity	0.04	0.00	0.10	0.11
overall subjectivity	0.59	0.60	0.61	0.60

Table 2: Comparison of professional and amateur review text sentiment.

[6] http://cran.r-project.org/web/packages/tm

[7] http://cran.r-project.org/web/packages/glmnet



(a) Cumulative distribution of mean review score

(b) Cumulative distribution of number of reviews per game

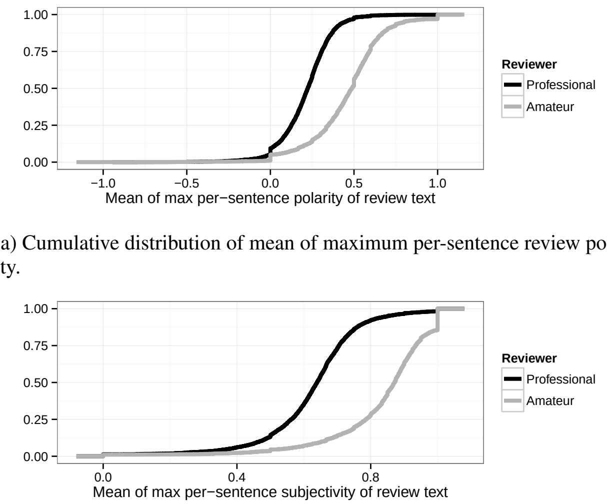
Figure 1: Game review corpus characteristics.

potential purchaser game feature interests and demographics (see below). Both amateur and professional reviewers acknowledge the strongly split gaming demographics, shown through the prevalence of related terms (Table 5). Our results suggest amateurs and professionals may have different habits and practices, suggesting different underlying drives and purposes for reviewing that merit additional investigation.

Review sentiment

Reviewers and professionals show similar average levels of review sentiment, but differ in expressions of extreme sentiment (Figure 2 and Table 2). For our analysis we considered both mean and median aggregations of polarity and subjectivity features over all reviews; all results reported were significant subject to a Wilcoxon test at $p < 0.001$, although these differences may not appear in Table 2 when rounding to two decimal places. While differences in mean and median per-review polarity and subjectivity were significant, their magnitude was small. Both professionals' and amateurs' review distribution are slightly skewed toward more positive polarity and mild subjectivity (compare means and medians in Table 2).

Professionals and amateurs barely differ in mean polarity over a review, both centering on a mildly positive typical review. However, amateurs vary their sentence-level polarity more within reviews and show a strong skew towards at least one strongly positive and/or negative review sentence when compared to professionals (Figure 2a). Similarly, while professionals and amateurs show nearly the same level of moderate mean subjectivity, amateurs tend to use at least one highly subjective sentence (Figure 2b). Interestingly, amateurs are also more likely to employ at least one highly *objective* sentence, relating to their overall larger variance in expressions of subjectivity. Contrary to our expectations, professionals are not particularly objective when compared to amateurs, except in terms of the most polarizing and subjective sentences used in their reviews.



(a) Cumulative distribution of mean of maximum per-sentence review polarity.

(b) Cumulative distribution of mean of maximum per-sentence review subjectivity.

Figure 2: Cumulative review sentiment (polarity and subjectivity) distributions comparing reviewer type

One limitation of this interpretation, however, is that professional text was limited to summaries and thus may not reflect the intended valence of the review, but a revised summary meant to convey factual information for featuring on Metacritic. These results may also reflect limitations of our sentiment analysis and merit further investigation using full review text from the original review sites.

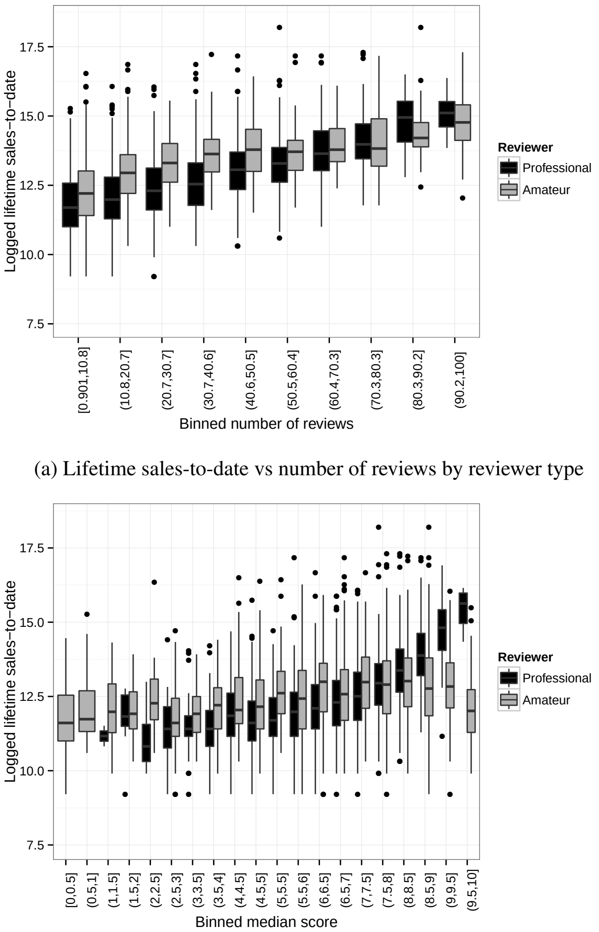
Sales Prediction

We used linear models to predict sales for each game, controlling for game console, genre, and Entertainment Software Rating Board (ESRB) maturity rating. While linear models may have limited predictive power for nonlinear relationships they afford direct interpretation useful to analysts employing the results.

We constructed four models: predicting total sales from reviews in the 10 weeks prior to the game's release (total vs pre-10), predicting net first 10 week sales from the same pre-release review subset (10 week vs pre-10), and predicting total sales using all reviews available with metascore, professional, and amateur score information and including or excluding metascore information (total vs meta and total vs meta sub). Table 1 summarizes the models. Tables 3 and 4 summarize the predictive variables, standardized β coefficients, and model fitting results for each model.

Review scores and volume both show clear relationships to the lifetime sales of a game (Figures 3a and 3b). Anecdotally, many have claimed that a metascore over 90 (9 on our rescaled range) leads to a substantial boost to game sales. Our data supports this claim when considering aggregated professional scores, although the range of relevance is broader.[8] Amateur scores do not show a clear relationship to sales:

[8] We found similar trends when examining Metascores, although the dataset is limited to fewer games.



(a) Lifetime sales-to-date vs number of reviews by reviewer type

(b) Lifetime sales-to-date vs mean review score by reviewer type

Figure 3: Mean lifetime sales-to-date vs review scores and volume.

increasing mean amateur review score does not appear to relate to any substantial gain in sales. By contrast, professional review scores show a clear increasing relationship with sales, starting around scores above 60 but becoming most prominent for scores above 80. While the [95-100] category has only 25 professional reviews and thus is of dubious predictive value, the [85-90] and [90-95] categories consist of 371 and 56 reviews and thus are of greater predictive merit.

Prediction Models

Our models show moderately strong predictive power (R^2 between 0.360 and 0.446, MSE between 0.669 and 0.586) for logged sales using standardized variables (Tables 3 and 4). Regression results report only standardized β coefficients significantly different from 0 ($p < 0.001$); "n.s." indicates non-significant coefficients. In all cases test models had a statistically significantly (χ^2 test, $p < 0.001$) smaller deviance than control models that accounted for game console, genre, and ESRB rating. Thus, all models explain variance in game sales not accounted for game market factors, showing the predictive value of Metacritic amateur and professional reviews.

For each predictive task we computed all possible linear models formed from subsets of review score and text features and scored them according to the Bayesian Information Criterion (BIC) using R's leaps package.[9] Using the BIC allows our model selection process to penalize overly complex models to reduce the likelihood of overfitting. Relative importance of model features was measured using the lmg metric in R's relaimpo package.[10] The lmg metric calculates the average R^2 contribution of a feature averaged over orderings among predictive variables—we use relative values among predictors to compare their importance.

Total and net 10 week sales vs pre-release reviews

Marketers often seek to predict sales for unreleased games relying solely on information available prior to a game's release. For this application, we examined two models that used reviews only in the 10 weeks prior to the release of a game to predict either lifetime sales-to-date or net sales over the first 10 weeks after a game releases.

The BIC identified the following predictive features: median review score, number of reviews, type of review (amateur or professional), and interactions between type of reviewer and median score, number of reviews, mean of mean per-sentence review polarity and mean of mean per-sentence review subjectivity. With an adjusted R^2 of 0.360 and MSE of 0.669 the pre-release review model is moderately effective at predicting lifetime sales. Table 1 shows the slightly better performance of the pre-release review model when predicting first 10 week sales. This is not surprising as most game sales occur during the initial launch and marketing period, with relatively few games having more long-term strong sales or shifts in sales. In this dataset, net first 10 week sales and total lifetime sales-to-date have a Spearman's rank correlation coefficient of 0.849, confirming this relationship.

We compared these models with a third model that employed all reviews (rather than solely pre-release reviews) for predicting total sales ("total vs all"); this model covered 2902 games, rather than the 600 in the previous models. Using all reviews gives an upper bound on performance for sales prediction given the information in our reviews, at least as we have processed it. All three models for sales prediction considering review features and their interactions with reviewer types show qualitatively similar effects (Table 3).

In general, number of reviews is most important for sales prediction, followed by median review score, and the interaction between median review score and reviewer type. Review sentiment factors are less important and have smaller coefficients in most cases. These three aspects account for most of these models' predictive power. Median review scores have the most positive effect on sales with the interaction of these scores with reviewers as amateurs being equally powerful but *negative*. Higher scores suggest strong sales, but not when coming from amateurs, who tend to provide the same scores across games of varying commercial success. Number of reviews has a weaker positive effect than review scores.

[9] http://cran.r-project.org/web/packages/leaps

[10] http://cran.r-project.org/web/packages/relaimpo

Together these results show professional scores and review volume indicate positive critical reception, while amateur reviews are only useful as a gauge of popular interest.

Volume of professional and amateur reviews both reflect relative attention to a game and the relative publishing and marketing resources backing the game's developer. More amateur reviews reflect more people have played a game and thus likely already purchased it. More professional reviews likely reflect publisher budget and expectations. Review copies are typically distributed to professionals when publishers expect games to be reviewed well, bolstering their review numbers. Further, only more well-financed and established publishers are able to push their games to be reviewed by venues that Metacritic would index and report from.

While direct causality cannot be read off from these results it is clear the amount of attention amateurs or professionals devote to a game relates to its sales success. Future work should explore more sophisticated features for predicting sales such as game previews, media campaigning, and interest expressed on Twitter, Facebook, Youtube and other media outlets. These sources may provide complimentary information on broader interest in a product compared to the narrower base of Metacritic reviewers.

Comparing Metascores, professionals, and amateurs

To better understand the efficacy of different scoring systems we compared the predictive power of metascores, combined professional review scores, and combined amateur scores for predicting lifetime sales-to-date. From the total dataset, 840 games have data on lifetime sales-to-date, metascores, and reviews from both professionals and reviews. We predict lifetime sales as metascores are running values that are updated without a historical trace—thus we cannot know historical values of metascores prior to a game's launch. Professional and amateur reviews were aggregated across the full set of data for a fair comparison with metascores.

According to the BIC criteria the best linear model (adjusted $R^2 = 0.4457$, MSE $= 0.586$) uses (standardized β in parenthesis): metascores (0.449), median professional scores (-0.093), median amateur scores (-0.190), and number of professional reviews (0.282) (Table 4). Removing metascore information shifts the model to positively weight amateur review scores (0.223) and add number of amateur reviews (0.235) as a significant predictive factor. The model without metascores is a somewhat worse fit ($R^2 = 0.4298$, and MSE $= 0.599$) according to a χ^2 test ($p < 0.001$).

As in the previous models, volume of amateur and professional reviews captures a large portion of the model's predictive power, using metascores for score information. The small differences in model fit suggest metascores are no more predictive than amateur or professional scores combined by simple weighting schemes (mean or median).

That amateur scores were not predictive of sales when coupled to metascores surprised us. To understand this effect we examined the Spearman's rank correlation between metascores, amateur scores, and professional scores. Surprisingly, metascores are very strongly correlated with median amateur scores ($\rho = 0.978$) and more weakly correlated with median professional reviews ($\rho = 0.545$). The comparable overall model power between using only amateur and professional data for lifetime sales and adding metascores suggests the metascore weighting scheme may not be particularly powerful compared to simpler methods.

	with metascores		**no metascores**	
	β	imp.	β	imp.
metascore	0.449	11.49%	-	-
median professional score	-0.093	1.87%	-0.071	1.90%
median amateur score	-0.190	8.51%	0.223	15.68%
number reviews in metascore	n.s.	12.51%	-	-
number professional reviews	0.282	21.60%	0.280	28.28%
number amateur reviews	n.s.	12.62%	0.235	22.48%

Table 4: Regression model for lifetime sales-to-date comparing metascores, amateurs, and professionals. β are standardized regression coefficients when predicting log-scaled and standardized lifetime sales. All values not listed as not significant ("n.s.") are significant at $p < 0.05$. With metascores, the model has $R^2 = 0.4457$ and $MSE = 0.586$; without metascores has $R^2 = 0.4298$ and $MSE = 0.599$. The metascore model is significantly different from the null model (residual deviance 465 on 816 degrees of freedom, null deviance 839 on 838 degrees of freedom, χ^2 test $p < 0.001$). The model without metascores is significantly different from both the null model (residual deviance 478 on 818 degrees of freedom, null deviance 838 on 838 degrees of freedom, χ^2 test $p < 0.001$) and the model with metascores (χ^2 test $p < 0.001$).

REVIEW TEXT CLASSIFICATION

Our previous analyses show reviews are predictive of sales. But what aspects of reviews distinguish them, particularly in terms of the words used? We explored this question by using review text in a bag of words model to classify reviewers as professionals or amateurs using a penalized binomial regression model. Understanding these text-level differences can elucidate what characteristics distinguish professional review content from amateur review content.

The classification model achieved an F1 score of 0.9359 (precision 90.96% and recall 96.38%) over the set of 197,383 reviews for which we had data. 96.38% of professionals were correctly labeled, and 77.28% of amateurs were correctly labeled, reflecting the imbalanced proportions of these categories in our data set—70.34% of reviews come from professionals, 29.66% from amateurs. 10-fold cross-validation found the model had a MSE of 0.4980. The penalized regression model employed 2,570 terms, assigning 146 a weight of zero, 1,235 positive weights (being an amateur), and 1,190 negative weights (being a professional).

Below we examine several of the terms from these groups to understand how amateurs and professionals differentiate themselves in text descriptions with term coefficients in parenthesis. As words were stemmed we included completions of the stems in parentheses to help interpretation. Table 5 illustrates a selection of words with strong predictive power: β coefficients are reported in parentheses with positive values predicting a reviewer to be a amateur. Word categories were

	10 week vs pre-10		total vs pre-10		total vs all	
	β	rel. importance	β	rel. importance	β	rel. importance
median review score	0.64691	18.63%	0.553424	19.88%	0.30145	15.80%
number of reviews	0.23578	21.06%	0.235556	20.88%	0.53365	53.32%
is amateur	0.43711	2.60%	0.606100	10.76%	0.37623	4.79%
is amateur x median score	-0.61708	12.93%	-0.505708	11.01%	-0.21179	2.60%
is amateur x number of reviews	0.51815	1.98%	0.431584	1.76%	n.s.	n.s.
is amateur x mean review polarity	0.25741	4.70%	0.213716	4.50%	0.03635	2.31%
is amateur x mean review subjectivity	-0.20898	2.77%	-0.170682	2.33%	-0.02785	0.17%

Table 3: Regression models for lifetime sales-to-date and first 10 weeks net sales. β are standardized regression coefficients when predicting log-scaled and standardized lifetime sales. All values are significant at $p < 0.001$.

derived through inspection and are not part of the predictive model, but shown to highlight trends in how amateurs and professionals word their reviews.

Professionals were most recognizable by references to major game genres and the games market. Compared to amateurs, professionals are more likely to mention words linking a game to marketing and game industry competition, using terms such as “tie(-)in” (-1.53) and “competitor” (-1.44). References to game genres included descriptions of gameplay perspective—“firstperson” (-0.86) and “thirdperson” (-1.01)—as well as gameplay style–“action” (0.36), “roleplay” (-1.34), and “brawler” (-0.78). Professional reviews often characterize a game within the broader industry sales cycle and specific genres of interest. This makes intuitive sense—professional reviews are often a form of marketing for companies, acting to define the intended audience of a game through describing how it relates to a broader marketing strategy and genre. Professionals more often refer to game features and the demographics of players who may be interested for these reasons.

Amateurs were recognizable by slang terms, references to game purchasing and reviewing, family, multiplayer, and aspects of the running game. Slang terms included phrases such as “wtf” (2.32), “imo” (4.59), “lol” (2.77), and “meh” (1.20). Reviews often mentioned aspects of reviewing games through references to review websites and aspects of “bias” (2.45) and “opinion” (1.46). Game purchasing appeared through game pricing, purchase options, and places to purchase games (e.g., the “Gamestop” (3.31) retail chain). Amateurs often reference the social dynamics of play, mentioning family and multiplayer as well as problems with actual game bugs or flaws. Mentioning live gameplay features and family members reflects amateurs’ stronger emphasis on the experience of playing with others, rather than assessing a potential product for (implicitly individual) consumption as done in professional reviews.

Combined, these results paint a picture of professionals taking a role of identifying purchase products and describing their features, while amateurs relate their games to broader consumption practices and gaming culture. Amateurs are free to reference particular game distributors or review biases in ways professionals cannot. Compared to professionals, amateurs typically relate games back to playing behavior, experiences, and social interactions. In part these results help clarify the power of professional review scores to predict sales better than amateur review scores. Professionals describe games in a way to guide purchasing decisions, while amateurs are more likely to reflect on a game in their play practices and purchasing experiences while giving relatively little information on the game’s features.

DISCUSSION

Having woven much of our discussion throughout our results, we will now summarize our findings and bring them to bear on practical problems faced by reviewing communities. Our analyses of game sales and reviews uncovered aspects of the different roles amateurs and professionals play in promoting and consuming games. Professionals employ a narrow range of review scores, describe games in terms of product purchase decisions, and are predictive of lifetime sales both in terms of scores assigned and volume of reviews. In sum, professionals act as dispassionate expert sources of information and indicators of overall investment in game production values and marketing push. In contrast, amateurs are more likely to give low review scores, skew towards higher review scores, make reference to gaming culture and experiences around playing a game, and are only predictive of sales through the volume of reviews provided. Together these results portray amateur reviewers as consumers evaluating the quality of a game experience and acting as indicators of overall market uptake and attention to the game.

Practical Implications

The most straightforward application for our work is predicting game sales from Metacritic reviews. Publishing firms can use this as a guide to predict sales based on trends in pre-release reviewing from professionals and amateurs. Such predictions can guide decisions to vary the resources devoted to marketing, developing, or supporting a game (particularly those with online or ongoing components). Marketers might attempt to drive amateur review volume and word-of-mouth in an effort to improve early and lifetime sales. Although we have not demonstrated a causal connection, this strategy presents itself as one worth exploring.

Review sites such as Metacritic develop ad-hoc schemes for weighting the scores of multiple reviewers using different review systems. Employing predictive models based on score or features could potentially provide a unifying metric for game “quality” that combines scores onto a single scale. Review sites could then feature the most predictive reviews.

category	word	β
game genre	roleplay	-1.34
	thirdperson	-1.01
	openworld	-0.68
	singleplay(er)	-0.67
	oldschool	-0.47
	sidescrol(ler)	-0.32
game market	tie(-)in	-1.53
	competitor	-1.44
	holiday	-1.09
	followup	-0.64
	brand	-0.59
	publish	-0.41
	budget	-0.40
player demographics	diehard	-1.09
	gamer	-0.32
	fanboy	1.30
slang	awsom(e)	5.45
	imo	4.59
	lol	2.77
	wtf	2.32
game reviewing	metacrit(ic)	3.03
	bias	2.45
	opinion	1.46
	critic	1.27
game purchasing	gamestop	3.31
	dlc	1.16
	preorder	1.06
family	wife	2.74
	son	2.06
running game	server	1.40
	lagg(y)	1.27
	glitch(y)	0.75
multiplayer	teamwork	0.57
	team	0.46
	splitscreen	0.43

Table 5: Major categories of review words distinguishing reviewers and professionals. Values in parentheses report β coefficients from binomial regression model where positive weights predict amateurs (as opposed to professionals).

Game marketers and producers could select potential reviewers most likely to bolster sales. Predictive amateur reviewers could even charge for their review services. Further, reviewers could develop novel scoring systems that are automatically scaled against other systems. Losing familiar amateur guides to scoring may hurt interpretability for amateurs, but may also allow both amateurs and professionals to focus more on aspects of game features or experience without filtering these results through an arbitrary number.

Limitations

Our work has several important limitations both in terms of the models employed and generalization of our results. We chose to use linear predictive models for sales. Linear models allow for direct interpretation but lose the nuances of these almost certainly non-linear relationships. Further, while clear relationships exist in our data they do not necessarily indicate causality and likely reflect a host of intervening factors. Game unit pricing and sales, marketing efforts, developer and publisher renown, and other external factors around sales regulation and economic circumstances all affect game sales.

The data we used potentially limits generalization based on our model. Our review corpus is derived from a single website and professional reviews were limited to summary text. It is unclear whether the amateur reviewing patterns on other websites would show similar relationships, as Metacritic is considered a prominent source for game review information. Perhaps the number of amateur reviews only matters from major review outlets, while other venues have weaker relationships with sales.

Our text-based models are limited to unigrams in a bag-of-words model. We thus lose information on linguistic qualifications, negations, and context that may provide additional nuance to reviews. The terms we found are thus an initial foray into the space of how reviews differ between professionals and amateurs. Even capturing simple adjective-noun relations would add depth and potential insight into what aspects of products reviewers describe and how they describe them [24].

Future Work

Beyond improving the sophistication of the natural language processing techniques used to examine review texts we might compare amateurs and professionals from other websites or sources. Do amateurs of different websites provide different perspectives on games? Metacritic is known as a major review center and thus likely reflects the opinions of more devoted game players. Other review sources such as Amazon may have a wider audience that is indicative of general opinions or perspectives on games. Combining these sources may uncover how amateurs of these sites differ and have varying predictive powers. For example, more “casual” games’ sales may be better predicted by Amazon reviews while Metacritic reviews may better capture sales of more niche games. However, establishing who is a professional and who is an amateur may be more challenging in these communities.

Conclusions

We analyzed amateur and professional reviews and examined their power to predict game sales. Pre-release reviews can predict both early and long-term sales, with volume of response from both amateurs and professionals as the most important predictive features, followed by professional (but not amateur) scores. Metascores strongly correlate with average amateur scores and provide roughly equal predictive power. While professional reviewers emphasize game marketing features, amateurs focus on the atmosphere of playing a live game with others. This is the first work we are aware of to quantitatively contrast professional and amateur reviewers, speaking to the broader question of the crowd and domain specialists compare with one another.

REFERENCES

1. Encyclopedia of life. http://www.eol.org. Accessed 29 May 2013.
2. Y. Benkler. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press, May 2006.
3. J. A. Chevalier and D. Mayzlin. The effect of word of mouth on sales: Online book reviews. Technical report, National Bureau of Economic Research, 2003.
4. S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović, et al. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307):756–760, 2010.
5. J. Cranshaw and A. Kittur. The polymath project: lessons from a successful online collaboration in mathematics. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 1865–1874. ACM, 2011.
6. C. Dellarocas, X. Zhang, and N. Awad. Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing*, 21(4):23–45, 2007.
7. N. Diakopoulos and M. Naaman. Towards quality discourse in online news comments. In *Proc. ACM 2011 Conf. on Computer Supported Cooperative Work*, pages 133–142. ACM, 2011.
8. W. Duan, B. Gu, and A. Whinston. Do online reviews matter? an empirical investigation of panel data. *Decision Support Systems*, 45(4):1007–1016, 2008.
9. S. Ehrenfeld. Predicting video game sales using an analysis of internet message board discussions. Master's thesis, San Diego State University, 2011.
10. C. Forman, A. Ghose, and B. Wiesenfeld. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research*, 19(3):291–313, 2008.
11. A. Ghose and P. Ipeirotis. Designing novel review ranking systems: predicting the usefulness and impact of reviews. In *Proceedings of the ninth international conference on Electronic commerce*, pages 303–310. ACM, 2007.
12. E. Gilbert and K. Karahalios. Understanding deja reviewers. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 225–228. ACM, 2010.
13. D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 78–87. ACM, 2005.
14. T. Hennig-Thurau, A. Marchand, and B. Hiller. The relationship between reviewer judgments and motion picture success: re-analysis and extension. *Journal of Cultural Economics*, 36(3):249–283, 2012.
15. B. Kanefsky, N. G. Barlow, and V. C. Gulick. Can distributed volunteers accomplish massive data analysis tasks. *Lunar and Planetary Science*, 1, 2001.
16. S. H. Kim, N. Park, and S. H. Park. Exploring the effects of online word of mouth and expert reviews on theatrical movies' box office success. *Journal of Media Economics*, 26(2):98–114, 2013.
17. Y. Liu. Word-of-mouth for movies: Its dynamics and impact on box office revenue. *Journal of marketing*, 70(3):74–89, 2006.
18. K. Luther, C. Fiesler, and A. Bruckman. Redistributing leadership in online creative collaboration. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1007–1022. ACM, 2013.
19. J. Marcoux and S. Selouani. A hybrid subspace-connectionist data mining approach for sales forecasting in the video game industry. In *2009 WRI World Congress on Computer Science and Information Engineering*, volume 5, pages 666–670. IEEE, 2009.
20. J. A. Plucker, J. C. Kaufman, J. S. Temple, and M. Qian. Do experts and novices evaluate movies the same way? *Psychology & Marketing*, 26(5):470–478, 2009.
21. B. L. Sullivan, C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling. ebird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142(10):2282–2292, 2009.
22. A. Xu and B. Bailey. What do you think? a case study of benefit, expectation, and interaction in a large online critique community. In *Proc. ACM 2012 Conf. on Computer Supported Cooperative Work*, pages 295–304. ACM, 2012.
23. X. Yu, Y. Liu, X. Huang, and A. An. Mining online reviews for predicting sales performance: A case study in the movie domain. *IEEE Transactions on Knowledge and Data Engineering*, 24(4):720–734, 2012.
24. J. P. Zagal, N. Tomuro, and A. Shepitsen. Natural language processing for games studies research. *Journal of Simulation & Gaming (S&G), Special Issue on Games Research Methods*, 43(3):353–370, 2011.