This manuscript is a late draft of an article that was accepted for publication in *Simulation & Gaming*. The version of record can be accessed at https://doi.org/10.1177/1046878117703680

A Taxonomy Approach to Studying How Gamers Review Games

Matthew J. Koehler

Brian Arnold

Spencer P. Greenhalgh

Liz Owens Boltz

Michigan State University

Author Note

Correspondence concerning this article should be addressed to Matthew J. Koehler, College of Education, 620 Farm Lane, East Lansing, MI 48824

Email: mkoehler@msu.edu

Abstract

Background. Player-generated reviews of **video games** represent a large, rich, and under-explored source of data for exploring what makes for an effective game.

Aim. We explore whether existing theory, in the form of the Bedwell, Pavlas, Heyne, Lazzara, and Salas (2012) **taxonomy**, suitably captures the issues that players raise when they review games.

Method. User-submitted **game reviews** were coded along the dimensions of the Bedwell et al. (2012) **taxonomy** to test the frequency of usage of each dimension.

Results. We found some support for the use of **taxonomy**, as four of nine **taxonomy** features were frequently present in **game reviews**. We also found support for other features of reviews not encapsulated by the Bedwell et al. (2012) **taxonomy**. Specifically, we found that players often reviewed **video games**: a) holistically; b) by comparing them to other games, game franchises and other reviews; and c) by judging the value of games in terms of time, money, and effort. These results have implications for using **game reviews** for future research.

Keywords: game features, video games, game reviews, taxonomy

A Taxonomy Approach to Studying How Gamers Review Games

> "If you make art, people will talk about it. Some of the things they say will be nice, some won't. You'll already have made that art, and when they're talking about the last thing you did, you should already be making the next thing."
> – Neil Gaiman

Games are often touted as an engaging and enjoyable medium—155 million Americans played computer games in 2015, and the video game industry has, in terms of revenue, surpassed the movie, music, and television industries (Entertainment Software Association, 2015). This growth and enthusiasm is mirrored in the growing community surrounding these games, which now consists of people of all ages and from all continents who are collectively investing millions of hours and billions of dollars in this hobby (McGonigal, 2011).

As in any field of research, some topics, questions, and methodologies are over-represented while others remain under-researched. To date, gaming research has traditionally employed experimental and quasi-experimental designs that use data stemming from surveys, assessments, observations, and user tracking as data sources. These approaches have helped game researchers, educational researchers, and psychologists to establish the power of games to motivate and immerse (Brown & Cairns, 2004) and, in some cases, perform as powerful learning tools (Clark, Tanner-Smith, & Killingsworth, 2016; Wouters, van Nimwegen, van Oostendorp, & van der Spek, 2013). The field as a whole has made important progress in understanding the broad outlines of what players get from games.

In this paper, however, we concern ourselves with one prominent, yet under researched, approach to understanding gaming. That is, we attend to the research on commercial games (as

opposed to games developed by researchers) and to the potential of using player reviews of commercial games as data to address cognitive, motivational, and affective gaming outcomes. Increasingly, professional and amateur reviewers are taking to the Internet to share their opinions of and experiences with games. For example, by 2012, Gamespot.com had accumulated over 29 million reviews on over 12,000 video games (Zhu & Fang, 2015). Review data on sites such as *GameSpot*, *IGN*, and *VideoGameGeek* represent very large corpora of ecologically-valid experiences generated by gamers, who have considerable experience playing and interacting with games.

The few studies that have used game review data come from studies of commercial gaming, and those studies were conducted to understand disparate topics, including the readability of reviews (Zagal, Tomuro, Shepitsen, 2012), salient features of game reviews (Zagal, Ladd, & Johnson, 2009), usability (Pinelle, Wong, & Stach, 2008; Zhu & Fang, 2015), and core factors (Calvillo Gámez, 2009). These efforts have predominantly employed automated and bottom-up approaches such as grounded theory or Natural Language Processing (NLP; i.e., approaches that use computers to find patterns in text).

For game reviews to be useful as data to explore cognitive, motivational, and affective outcomes of commercial games, however, researchers must be able to analyze these data through a theoretical lens that is germane to the field. Therefore—and unlike most previous work using player review data—we apply a top-down approach using existing theory: a taxonomy specifically created to describe the most salient features of video games; the differences between them; and their cognitive, motivational, and affective implications. In doing so, we hope to better understand if current theories of game design and game features, as represented in taxonomies, are rich enough to capture the experiences and reviews of actual gamers.

Purpose and Research Questions

The purpose of this study is to explore the suitability of an existing taxonomy of game design features for characterizing reviews of video games. Although several taxonomies to describe games have been proposed over the years, we chose to focus on Bedwell, Pavlas, Heyne, Lazzara, and Salas's (2012) taxonomy because its creators explicitly worked to synthesize and integrate previous taxonomies, drew on the knowledge of expert gamers, and set out to map game elements to a wide variety of salient outcomes. Specifically, the Bedwell at al. (2012) taxonomy explores the cognitive, motivational, and affective outcomes of gaming. It has nine overarching dimensions (*Action Language*, *Assessment*, *Conflict/Challenge*, *Control*, *Environment*, *Game Fiction*, *Human Interaction*, *Immersion*, and *Rules/Goals*).

In the present study, we looked for references to each of the nine dimensions of this taxonomy in 200 amateur reviews of commercial video games from the *VideoGameGeek* website (http://videogamegeek.com). We evaluated how suitable this approach is for characterizing game reviews; in particular, we used the criteria described in the following research questions to evaluate how *useful*, *complete*, and *predictive* this approach is:

1. *Useful*: How frequently do each of the dimensions appear in reviews of digital games?
2. *Complete:* How often do other (non-taxonomy) features appear in reviews of digital games?
3. *Predictive:* Are there differences between the features that appear in reviews of highly-rated and poorly-rated digital games?

If the taxonomy is shown to suitably represent users' game reviews, then a large source of game review data becomes easily accessible to researchers in addition to existing approaches that study cognitive, motivational, affective, and behavioral outcomes. This data could be used in

addition to, or instead of, other data sources that may be more difficult to obtain. However, if the taxonomy does not suitably characterize game reviews, more work will be necessary to determine whether game review data is well-suited to testing theories of salient game features.

Literature Review

In this section, we review previous research related to game reviews as data sources and the use of taxonomies for studying gaming.

Research Using Game Reviews as Data

Natural Language Processing approaches. Many of the studies using game reviews as data have employed Natural Language Processing (NLP) approaches that use computers to find patterns and themes in text. Researchers using Natural Language Processing approaches on game reviews have investigated such issues as the readability of reviews (Zagal, Tomuro, Shepitsen, 2012), the positive and negative sentiments contained in reviews (Drake, Ringger, & Ventura, 2008; Zagal, Tomuro, Shepitsen, 2012), the aesthetics of game play (Zagal & Tomoru, 2010), the similarity between games (Raison, Tomuro, Lytinen, & Zagal, 2012), cultural differences in game appraisal (Zagal & Tomuro, 2013), and the use of adjectives in game reviews (Zhu & Fang, 2015).

Two NLP studies are of particular interest because they directly explore specific features that users find salient when reviewing games. Zagal and Tomuro (2010) investigated the aesthetic language used in reviews, and Zhu and Fang (2015) examined typical adjective groupings in game reviews. Although the studies used different analytical approaches, both extracted groupings of adjectives used to describe games in user-submitted reviews. Clusters (Zagal & Tomuro, 2010) and factors (Zhu & Fang, 2015) of these adjectives appear to suggest groups of features that reviewers used to comment on games. Zagal and Tomoru (2010)

identified six clusters that spoke broadly about aesthetic features of gameplay: *Pacing, Complexity, Cognitive Accessibility, Scope, Demands*, and *Impact*. Zhu & Fang (2015) identified six factors that spoke to usability: *Playability, Creativity, Usability, Competition, Sensation,* and *Strategy*.

Inductive and deductive coding approaches. There are some notable exceptions to the NLP approach to analyzing game reviews. For example, Pinelle et al. (2008) used content analysis (bottom-up coding) to study 108 game reviews published to *GameSpot* by professional reviewers. Focusing on issues of usability, the authors identified 12 categories of usability problems, with the five most frequently-appearing categories accounting for over 58% of the issues reported by reviewers: a) *unpredictable or inconsistent response to user's actions*; b) *doesn't provide adequate training or help*; c) *difficult to control actions in the game;* d) *response to user's action not timely enough;* and e) *does not provide enough information on game status*.

Calvillo Gámez (2009) used grounded theory (i.e., bottom-up coding) of 116 professional game reviews and interviews collected in 2006 to establish the "core elements" of the gaming experience. Five main elements were identified to inform the development of Calvillo Gámez's *Core Elements of Game Experience* theory: *Gameplay, Environment, Ownership, Control,* and *Facilitators*. Using a similar approach and goal, Bond & Beale (2009) examined 33 professional reviews from GameSpot UK. They focused on what factors led to making a game "good" or "bad" as reflected in overall player ratings of games. They found 13 factors predicting game success: *Gameplay, Environment, Storytelling, User Interaction, Customization, Social, Variety, Technical, Cohesion, Maintenance, Price, Franchise*, and *Quantity*.

Wang, Shen, and Ritterfeld (2008) used a combination of bottom-up coding and *a priori* codes derived from the literature on game enjoyment to generate 30 "commonsense fun factor

categories" (p. 30). They then used these categories as codes for performing content analysis of 60 professionally-conducted game reviews of 30 games from *GamePro* and *IGN*. Over 50% of the coded instances fell into the top five most popular categories: *Overall Game Design*, *Visual Presentation*, *Control*, *Audio Presentation*, and *Complexity and Diversity*.

Interviews and focus groups. Interviews and focus groups are two related methodologies for studying gamers' motivation to play games. They are similar to player reviews in that responses come from participants in their own words but differ in that the use of interviews and focus groups allows researchers to specifically target important aspects of motivation, cognition, and emotion that players might not consider on their own. For example, Brown & Cairns (2004) interviewed seven gamers about issues of immersion. Using grounded theory, they found three progressive levels of engagement in the data—initial *engagement* moving on to greater involvement in the form of *engrossment*, which is followed by *total immersion*. Similarly, based upon the responses of 16 participants in focus groups, Poels, de Kort, and Ijsselsteijn (2007) found nine categories that emerged from the data to describe players' experiences with games*: Enjoyment, Flow, Imaginative Immersion, Sensory Immersion, Suspense, Competence, Negative Affect, Control*, and *Social Presence*.

Summary and limitations of existing research. The existing research using game reviews, focus groups, and interviews has explored various approaches to understanding important features of game play. This research to date, however, has some important limitations. First, most of the research has employed NLP approaches or bottom-up approaches. While useful for building theory, these approaches do not directly address (or code for) any existing theory—repeated use of the NLP approach risks creating an overabundance of theories, few of which are actually used. Second, studies employing video game reviews as data have typically used

professional reviews. There are, of course, many reasons to prefer professional reviews: They are detailed, structured, analytic, and suited to the task of decomposing a game into elements. However, the present study is interested in what can be learned, if anything, from the large corpus of reviews left by everyday gamers. A third limitation of the current research is the age of the prior studies using game reviews, most of them having been published prior to 2007—a lot has changed since 2007 in gaming (e.g., games have become increasingly more online, cooperative, and social) and in the way games are reviewed (increasingly more crowdsourced).

Taxonomies of Game Attributes

Researchers have long struggled to reach a consensus about the elements that make up a good game (Klabbers, 2009). Efforts to definitively organize game features, called taxonomies (Tobias, Fletcher, & Wind, 2014), began in earnest with the work of Malone (1981; Malone & Lepper, 1987). Malone proposed four individual motivations (*Challenge*, *Curiosity*, *Control*, and *Fantasy*) and three interpersonal motivations (*Cooperation*, *Competition*, and *Recognition*) as the common core for gameplay (see Figure 1). Each individual motivation in this approach can be further deconstructed. For example, the motivation to be challenged can be expressed in terms of *Goals, Uncertain Outcomes, Performance Feedback,* and *Self-esteem.*

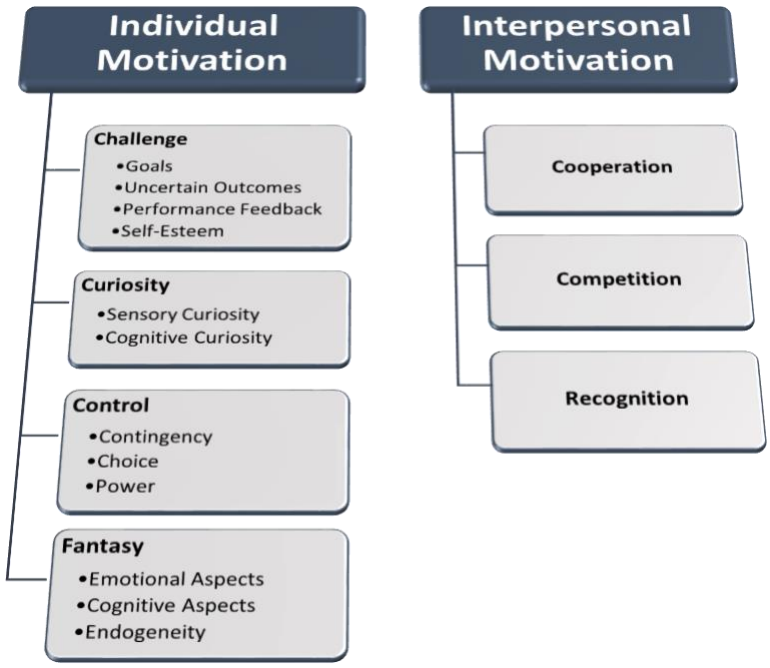

Figure 1. Heuristics for making learning fun, according to Malone and Lepper (1987)

Since 1987, others have made incremental modifications to this proposed framework. For example, Juul (2003) proposed a taxonomy that also considers player's *Effort Level* and the *Game Outcomes*. Leemkuil, de Jong, and Ootes (2000) included *Rules* and *Constraints*; Baranauskus, Neto, and Borges (2001) added *Risk*; and Thiagarajan (1999) included *Closure* and *Contrivance*.

Concerned with the growing number of different approaches to categorizing games, some researchers have attempted to define a universal taxonomy of game attributes so that game design features might be described and analyzed more consistently. For example, Wilson et al. (2009) reviewed previous taxonomies in order to develop a list of eighteen game features:

Adaptation, *Assessment*, *Challenge*, *Conflict*, *Control*, *Fantasy*, *Interaction (equipment)*, *Interaction (interpersonal)*, *Interaction (social)*, *Language/Communication*, *Location*, *Mystery*, *Pieces or Players*, *Progress and Surprise*, *Representation*, *Rules/Goals*, *Safety*, and *Sensory Stimuli*. The goal in doing so was to help researchers connect each feature to specific cognitive, skills-based, and affective outcomes using varied sources of data and methodologies.

Bedwell et al.'s (2012) Taxonomy

Following up on Wilson et al.'s (2009) study, Bedwell et al. (2012) used surveys and card sorting activities with developers and players as participants to refine the 18 categories into a more concise, orthogonal taxonomy avoiding conceptual overlap (see Table 1). This follow-up taxonomy was born out of a concern that there was too much overlap between elements of the Wilson et al. (2009) taxonomy. For example, the authors describe how the *Pieces or Players* category overlaps with the *Representation* category in the Wilson et al. (2009) taxonomy, proposing that it is difficult to change the *Pieces or Players* (the "objects or people being included in the game narrative," p.31) without also changing the *Representation* ("the player's perceptions of the game … that make the game appear psychologically real," p.31). That is, changing the objects or people in the game changes players' very perception of how real the game is.

Bedwell et al. (2012) were also responding to the lack of common terminology and the difficulty researchers have experienced in disentangling categories of game features both theoretically and practically. This study uses the Bedwell et al. (2012) framework as the taxonomy for data analysis because this taxonomy: a) arose from the need to develop a *universal* taxonomy, b) includes many core features of prior taxonomies, c) significantly simplifies and

refines the Wilson et al. (2009) framework, and d) is of particular value in theoretical and applied settings.

Table 1

Taxonomy of Game Features (Bedwell et al., 2012)

Category	**Description**
Action Language	the ways a player makes their intent known to the system
Assessment	the type and means of feedback that is given to the player within the game
Conflict/Challenge	the nature and level of difficulty of various aspects of the game
Control	the amount of agency afforded to the player
Environment	the way the setting or location of the game is represented
Game Fiction	the game's story or plot
Human Interaction	the amount or type of human-to-human contact within the game
Immersion	the player's affective response to the game world and its story, including its sensory stimuli (graphics, sound, music, etc.)
Rules/Goals	the clarity of a game's rules and objectives

We initially tested the utility of the Bedwell et al. (2012) taxonomy in a pilot study (Boltz, Arnold, & Greenhalgh, 2015) by applying the taxonomy to reviews on the website *VideoGameGeek* (VGG; http://videogamegeek.com). We found that players tended to perceive the *Conflict/Challenge, Immersion,* and *Control* attributes as having a stronger influence on enjoyment than *Action Language* and *Assessment* but also noted that Bedwell et al.'s (2012) categories were not sufficient to account for all of the reasons players provided for their enjoyment of video games. While this pilot study demonstrated that the Bedwell et al. (2012) taxonomy showed promise for capturing important features of game reviews, it also had

significant limitations—for example, it was based on a small number of games and relied largely on descriptive statistics to make claims.

The current study seeks to extend the findings of our earlier work through the incorporation of more data, the inclusion of inferential statistics, and a test of the taxonomy-based approach to analyzing player-generated game reviews.

Method

This study employs an observational, descriptive design to analyze secondary data retrieved from online game reviews. As people use the Internet, they leave "digital traces" (Lazer et al., 2009, p. 721) of their activity that can subsequently be collected and analyzed. In this study, we examine the digital traces left behind by video game players in the form of reviews they have published online.

Data Sources

Data for this study came from game reviews retrieved from *VideoGameGeek* (VGG), an extensive and detailed website for the gaming community to gather and display information about games. Within the VGG community, game players—and not a professional staff—populate the site with game information, ratings (scores from 1 to 10), and short game reviews (as text). Even though these reviews are short and colloquial, our study focused on these as data because they represent a large, ecologically-valid set of data that more directly reflects current approaches to gaming communities and game reviews.

We used the VGG Application Programming Interface (API) to collect a sample of reviews for qualitative coding. From a practical standpoint, we decided that coding 200 reviews was feasible while still representing a very large number of reviews compared with previous studies. In deciding which 200 reviews to sample, we purposefully sampled 4 reviews each from

the 25 highest- and 25 lowest-rated games (according to average game rating) published between 2010 and the present that had at least 4 reviews coming from the top quintile of reviews in terms of word count. The key choices and reasoning for these decisions are summarized below:

- High- and low-rated games were used to test the application of the taxonomy on both ends of the quality spectrum.
- Concentrating on 2010 to the present allowed us to focus on understanding contemporary players rather than have our results representing several generations of video game.
- Four reviews per game were sampled to ensure that reviews were representative of each game (rather than biased by one single review).
- The top quintile of game reviews by length contained the most verbose reviews on the site and resulted in reviews 61 words or longer (mean of 160.13 words). This was the mostly likely sample from which specific taxonomy attributes were likely to be identified and coded.

Measures

Each of the 200 reviews in our sample was coded for the presence of 11 different measures, which included the nine dimensions of Bedwell et al.'s (2012) taxonomy (Table 1) and the following additional categories suggested by a pilot study (Boltz, Arnold, & Greenhalgh, 2015):

- *Value Judgment*: Players' perceptions of what they gained or lost from playing in terms of time, money, replay, experience, or return on investment.
- *Comparisons*: Players' comparisons of a game to other reviews, games, and external factors that influenced players' expectations and opinions.

Each review was coded with a "1" for each measure that was present in the review and with a "0" for each measure that was absent from the review. It is important to note that these measures are not mutually exclusive: A review could potentially be coded positively for all 11 categories or even for none of them. Table 2 provides a sample review of *Journey*, a 2012 single-player and multi-player adventure game for the PlayStation platform. The color of the highlighted text in the review indicates the coding category assigned.

Table 2

Sample Reviews and Assigned Coding Categories

Text of the Game Review	**Coding Key**
ps3. Beat. Gives me a similar feeling to Ico, which I love. Very beautiful. The music and atmosphere are fantastic. There's a story buried here too- kind of feels like playing through some abstract SF/Fantasy short story. Very short but I think you'll want to play it again and the trophies add some replay value. I find the experience of it sticking in my mind, much like Ico and Shadow of the Colossus, and many people obviously get the same strong feeling from it- hence the awards and accolades. With that said, I can see the "not a game" argument here- there is no challenge. You can take your time exploring the world or can go through it very quickly and experience it kind of like a movie but that is it. So it's not a traditional game, but it is pretty awesome entertainment if experienced on its own terms.	*Comparison* – Players' comparison of a game to other reviews, games, and external factors *Immersion* – the player's affective response to the game world and its story, including its sensory stimuli (graphics, sound, music, etc.) *Game Fiction* – the game's story or plot *Value Judgment* – Players' perceptions of what they gained or lost from playing in terms of time, money, replay, experience or return on investment *Conflict/Challenge* – the nature and level of difficulty of various aspects of the game

Coding Procedures

Having developed these measures, we took several steps to establish their reliability. Prior to coding, two raters developed shared, consistent understandings of the taxonomy, a

coding manual, and procedures about how to apply the coding categories to the data. Both raters then coded 10 reviews together to further establish their shared understanding, after which they coded 20 reviews separately and met to discuss and resolve differences. Inter-rater reliability was calculated for the remaining 170 reviews, which were coded separately by both raters. All differences in coding were resolved by consensus. Raters showed high levels of inter-rater reliability for each of the 11 categories—percent agreement ranged between 90.6% and 98.2% and Cohen's kappa ranged between .70 and .93 across the 11 categories.

Results

Using the criteria of *useful*, *complete*, and *predictive*, we examine how well the Bedwell et al. (2012) taxonomy represents the data in the 200 player-generated game reviews.

Useful: How Often Taxonomy Features Appeared in Reviews

Our first research question explored how *useful* the Bedwell et al. (2012) taxonomy was for characterizing user reviews by considering how often each category in the taxonomy was used. It is important to note that from an absolute perspective, features ranged considerably in how frequently they were used in game reviews (see Table 3). On average, a feature appeared in a review 24.6% of the time.

Using tests of equal proportions, we compared the proportion of reviews associated with each feature with the mean overall rate of usage (24.6%). *Rules/Goals*, *Immersion*, *Conflict/Challenge*, and *Game Fiction*—the four most commonly referenced elements of game design—all appeared significantly more often than average, suggesting that players find these design attributes more important when reviewing games. Conversely, *Human Interaction*, *Assessment*, *Environment*, and *Control*—the four least commonly referenced elements of game

design—appeared significantly less often than average, suggesting that reviewers find these attributes less important.

Complete: How Often Other Features Appeared in Reviews

Our second research question studied how frequently the two other (i.e., non-taxonomy) codes were used to describe the data in order to better understand how *complete* the Bedwell et al. (2012) taxonomy is for describing user review data. These two other codes, *Value Judgment* and *Comparisons*, arose from an earlier pilot study (Boltz, Arnold, & Greenhalgh, 2015).

Table 3

Proportion of Game Features by Type of Game Review, by High- and Low-rated Games

	Type of Game Review		
Game Feature	Low-Rated Games (N =100)	High-Rated Games (N=100)	Combined (N = 200)
Control	.060	.110	.085 **
Environment	.060	.110	.085 **
Human Interaction	.070	.150	.110 **
Assessment	.130	.090	.110 **
Action Language	.180	.260	.220
Rules/Goals	.360	.340	.350 *
Conflict/Challenge	.480	.310 †	.395 *
Immersion	.520	.310 †	.415 **
Game Fiction	.390	.490 †	.440 **
Average Feature Usage	.250	.241	.246

† difference of proportion between high- and low-rated games is statistically significant, $p < .05$

* feature is used significantly more or less than the overall mean usage of .246, $p < .05$

** feature is used significantly more or less than the overall mean usage of .246, $p < .001$

These two non-taxonomy codes are notable in that both were used quite often (see Table 4). *Value Judgments*—perceptions about cost or benefits in terms of time, money, experience or value—appeared in 30.5% of the reviews we considered, more than five out of the nine original Bedwell et al. (2012) features. These judgments refer to features of games related to how long a game is and how much someone can play (or replay) it while continuing to feel that it is valuable. For example, some reviews included these phrases: "I'll be coming back to play both games again," "the worst three hours of my life," and "Well worth the couple of bucks I paid for it".

Table 4

Proportion of Other Features by Type of Game Review, by High- and Low-rated Games

	Type of Game Review		
Review Category	Low-Rated Games (N =100)	High-Rated Games (N=100)	Combined (N = 200)
Comparisons	.720	.810	.765 **
Value Judgment	.320	.290	.305

** significantly more comparisons were used than value judgments, $p < .001$

The *Comparison* code was even more common, appearing in the majority of reviews (76.5%) and more than any of the Bedwell et al. (2012) features. Players often compared one game to other titles (e.g., "Gives me a similar feeling to Ico, which I love"), to other installments

in the franchise (e.g., “RB3 is a step forward for the franchise in some ways, but a decided step back in others”), or to other reviews (e.g., “After hearing all the negative feedback on the internet about this game at the time of release, I was actually quite pleasantly surprised”).

Predictive: Differences Between High-Rated and Low-Rated Games

Our third question examined how effectively the taxonomy *predicts* outcomes, specifically game enjoyment. Our approach centered on comparing how often taxonomy categories appeared in high- and low-rated games. In terms of the overall average number of taxonomy reviews in high- and low-rated games, we found no significant differences, suggesting that reviewers comment on about the same number of game features whether they are describing a well-received or poorly-received game.

Subsequently, we investigated whether particular taxonomy features appeared more often in high- or low-rated games. To do this, we carried out tests of equal proportions comparing how frequently each taxonomy feature appeared between high- and low-rated games (see Table 3). Reviewers of high-rated games (compared to low-rated games) were significantly more likely to include *Game Fiction* in their review and significantly less likely to include *Conflict*/*Challenge* and *Immersion.* This suggests that—on average—reviewers were more likely to comment on the story or world of a game when they enjoyed the game, and more likely to focus on the difficulty or immersion of a game when a game received low ratings.

Discussion

The purpose of this paper was to determine the suitability of a taxonomy-based approach for analyzing game review data in terms of player enjoyment. We explored the suitability of the Bedwell et al. (2012) taxonomy for characterizing user review data using three criteria: *useful,*

complete, and *predictive*. In this section, we evaluate this taxonomy before commenting on other implications of our findings.

Taxonomy Evaluation

Our three tests produced mixed results with regard to our purpose. With our first test, we found that *some*—but not all—features in the Bedwell et al. (2012) taxonomy were *useful* for capturing what reviewers found salient in their reviews. Specifically, *Rules/Goals*, *Immersion*, *Conflict/Challenge*, and *Game Fiction* were each present in more than one-third of reviews. In contrast, four features were rarely used (less than 15% of reviews): *Human Interaction*, *Assessment*, *Environment*, and *Control.*

In our second test, we found that the taxonomy was not *complete*, in that the *ad hoc* categories of *comparisons* (used in over 76.5% of reviews) and *value judgments* (used in 30.5%) seemed to be highly salient to reviewers. That *comparisons* appeared more often than any of the Bedwell et al. (2012) features and that *value judgments* appeared more than over half of them demonstrate that this taxonomy does not account for everything that players may comment on when reviewing games.

In our third test, we found that the taxonomy showed *predictive* power in distinguishing between high- and low-rated games. We found that *Game Fiction* was more important in distinguishing a high-rated game, with reviews commenting on the in-game story and presumably praising it. In contrast, reviews of low-rated games had increased commentary about *Immersion* and *Conflict/Challenge*, implying that the presentation and difficulty of a game were often among the reasons that it was poorly-received.

In summary, we find that the Bedwell et al. (2012) taxonomy has some use in characterizing game reviews but that it may not be the best way of framing player considerations

of games. In the sample of 200 reviews, the majority of text doesn't correspond to any particular feature of the Bedwell et al. (2012) taxonomy. Instead these portions of text invoke vague concepts such as gameplay and comparisons to other games, thereby providing an overall evaluation rather than a detailed critique. Rather than talking about specific features of games, large portions of reviews focused on more holistic evaluations, employing ill-structured concepts like how much play they get out of a game and how a game compares to other games. For example, consider the following example review excerpts:

> "It's possible to hit a wall and then going back to it doesn't appeal. I hit the wall."
>
> "I finally beat it out of sheer determination but it wasn't worth the effort."
>
> "While still being a very good game, it is a letdown of the series."
>
> "No other game besides some Final Fantasy games made me more emotionally invested in the characters. "

Connections to Prior Research

For the features of the Bedwell et al. (2012) taxonomy that were used frequently—*Rules/Goals*, *Immersion*, *Conflict/Challenge*, and *Game Fiction*—this study adds support to the salience of these design features as raised in previous studies (Bedwell et al., 2012, Bond & Beale, 2009; Zagal & Tomuro, 2010; 2013; Zhu & Fang, 2015). Beyond the Bedwell et al. (2012) taxonomy, the present study also identified *Comparison* (e.g., players' comparison of a game to other reviews, games, and external factors) as a highly-salient feature of many reviews. This concurs with the Bond & Beale (2009) study showing that *Franchise* consideration (how a game fits in within a series) was important to gamers. This also resonates with the findings of Ryan, Kaltman, Hong, Mateas, and Wadrip-Fruin (2015), which suggest that people naturally compare games, tending to seek out games that are related to the ones they like. Zagal, Ladd, and

Johnson (2009) also used the *Game Context* category to capture the high number of times players contextualize games with respect to other games.

Also frequent in the present study was the use of *value* judgments (e.g., players' perceptions of what they gained or lost from playing in terms of time, money, experience or return on investment). This directly corroborates the findings of Bond and Beale (2009), which showed that "Price (value for money)" (p. 419) is an important consideration for gamers. Value is also examined in the Zagal and Tumoro (2013) study of cross-cultural differences in games reviews—they found that a game's *replay value* was an especially important consideration for American reviewers. Although demographics are not available for VGG review data we collected, it is likely that the vast majority of VGG reviewers are American.

More generally, most of the bottom-up approaches to studying game reviews have established categories that seem to describe the large number of holistic comments present in our sample but were not coded for in the Bedwell et al. (2009) framework. For example, factors such as *game design, visual presentation,* and *control* (Wang et al., 2008) as well as *pacing, complexity,* and *impact* (Zagal & Tomoru, 2010) address topics raised in the holistic comments in our sample that broadly describe the game in terms of fun, difficulty (or lack thereof), and how the game has or has not impacted the player.

Methodological Considerations

We found that the Bedwell et al. (2012) taxonomy was somewhat useful for characterizing and describing player-generated reviews on the VGG website. The most straightforward conclusion is that some features of the taxonomy seem salient in the data at hand, whereas others seem to have limited application. There are, however, methodological limitations that complicate this interpretation.

First, consider the length and purpose of the reviews we examined. VGG—like the Internet Movie Database, iTunes, various "app stores," and other crowdsourced sites—represents a different kind of "review." These services seek large numbers of numerical ratings in order to establish the wisdom of the crowd. Comments (e.g., reviews) often accompany these ratings, but the expectation (or perhaps convention) is that these are short. In the case of VGG, the length of the reviews we examined averaged 160 words. It is clear that reviewers on the VGG site are talking to other gamers (and not academics). As researchers, we are attracted to understanding what qualities reside in these different types of reviews, but, as results show, doing so presents some challenges when that was not their intended purpose.

Furthermore, it is difficult to establish whether reviewers do not think about the Bedwell et al. (2012) criteria or, instead, just do not write about them as they speak to peers. That is, because players implicitly and explicitly compare games to other games, they may not find all features included in this kind of taxonomy relevant to discussion because other players already know and understand some of them. For example, when reviewing a game in a well-known genre, reviewers may only bring up features that deviated from the expected ones, preferring to emphasize other aspects of a game.

The taxonomy approach we tried here might well be more successful if applied to a different set of reviews. For example, professional reviews tend to be much longer and more systematic. The Bedwell et al. (2012) categories may well prove to be more salient in the context of professional reviews. Researchers may also consider whether giving players scaffolds (e.g., section headers or sub-ratings) along dimensions of interest might also yield better data. For example, players might specifically be given the *Human Interaction* category as an element to include in their review. Given such prompts, it may (or may not) emerge that players have a lot

to say about the topic.

Conclusion

We set out to discover the affordances and constraints of an existing game taxonomy for analyzing players' reviews of games. We found *limited* support for using Bedwell et al.'s (2012) taxonomy to characterize the short, player-generated reviews that we examined in this study—some elements of the taxonomy were used frequently, whereas others were not. We did, however, find evidence of frequent reference to player *comparisons* (to other games, other reviews, and individual or community expectations) and *value* (cost in terms of time, money, or opportunity), a finding consistent with prior research. That is, although this taxonomy did not prove as useful as we had anticipated, this study does suggest that there are rich themes that may be explored in these short, crowd-sourced reviews that players make for each other. Employing other methodological approaches to examine themes in these reviews might add additional support for prior research that has focused on *comparisons* and *value*, aesthetic features (Zagal & Tomoru, 2010), usability (Zhu & Fang, 2015), or "fun factors" (Wang et al., 2008).

Another direction for future work may explore the use of taxonomies like Bedwell et al.'s (2012) for examining longer, comprehensive, and more detailed accounts of gamers' experiences. That is, interviews, focus groups, and professional game reviews may all be better targets for exploring the potential of taxonomy-guided coding of gamers' experiences, conceptions, and interaction with games.

References

Baranauskas, C. C., Neto, N. G. G., & Borges, M. A. F. (2001). Learning at work through a multi-user synchronous simulation game. *International Journal of Continuing Engineering Education and Life Long Learning*, *11*, 251–260. doi:10.1504/IJCEELL.2001.000397

Bedwell, W. L., Pavlas, D., Heyne, K., Lazzara, E. H., & Salas, E. (2012). Toward a taxonomy linking game attributes to learning: An empirical study. *Simulation & Gaming*, *43*, 729-760. doi:10.1177/1046878112439444

Boltz, L. O., Arnold, B., & Greenhalgh, S. P. (2015). Players, features, and enjoyment: Entertainment games as models for educational games. In D. Slykhuis & G. Marks (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference 2015* (pp. 822–829). Chesapeake, VA: Association for the Advancement of Computing in Education (AACE).

Bond, M., & Beale, R. (2009). What makes a good game?: Using reviews to inform design. In *Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology* (pp. 418–422). London, England, UK: British Computer Society.

Brown, E., & Cairns, P. (2004). A grounded investigation of game immersion. In *CHI'04 Extended Abstracts on Human Factors in Computing Systems* (pp. 1297–1300). New York, NY: Association for Computing Machinery. doi:10.1145/985921.986048

Calvillo Gámez, H. E. (2009). *On the core elements of the experience of playing video games* (Doctoral dissertation). Retrieved from UCL Discovery (18510)

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research, 86,* 79–122. doi:10.3102/0034654315582065

Drake, A., Ringger, E., & Ventura, D. (2008). Sentiment regression: Using real-valued scores to summarize overall document sentiment. In *Proceedings of the 2008 IEEE International Conference on Semantic Computing* (pp. 152-157). Washington, D. C.: IEEE Computer Society.

Entertainment Software Association. (2009). *Industry facts*. Retrieved from http://www.org.id.tue.nl/IFIP-TC14/documents/ESA-Essential-Facts-2009.pdf

Juul, J. (2003). The game, the player, the world: Looking for a heart of gameness. In M. Copier & J. Raessens (Eds.), *Proceedings of the Level Up: Digital Games Research Conference* (pp. 30-45). Utrecht, Netherlands: Utrecht University.

Klabbers, J. H. G. (2009). Terminological ambiguity: Game and simulation. *Simulation & Gaming*, *40*(4), 446–463. https://doi.org/10.1177/1046878108325500

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., … Van Alstyne, M. (2009). Computational social science. *Science*, *323*(5915), 721–723. doi:10.1126/science.1167742

Leemkuil, H., De Jong, T., & Ootes, S. (2000). *Review of educational use of games and simulations*. Retrieved from http://doc.utwente.nl/28235/1/review_of_educational.pdf

Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive Science, 5,* 333–369. doi:10.1016/S0364-0213(81)80017-1

Malone, T. W., & Lepper, M. R. (1987). Making learning fun: A taxonomy of intrinsic motivations for learning. In R. E. Snow & M. J. Farr (Eds.), *Aptitude, learning, and*

instruction: Cognitive and affective process analysis (Vol. 3, pp. 223–253). doi:10.1016/S0037-6337(09)70509-1

McGonigal, J. (2011). *Reality is broken*. New York, NY: The Penguin Press.

Pinelle, D., Wong, N., & Stach, T. (2008). *Heuristic evaluation for games: Usability principles for video game design*. Paper presented at the CHI Conference, Florence, Italy.

Poels, K., De Kort, Y., & Ijsselsteijn, W. (2007). "It is always a lot of fun!" Exploring dimensions of digital game experience using focus group methodology. In *Proceedings of the 2007 Conference on Future Play* (pp. 83–89). doi: 10.1145/1328202.1328218

Raison, K., Tomuro, N., Lytinen, S., & Zagal, J. P. (2012). Extraction of user opinions by adjective-context co-clustering for game review texts. In H. Isahara & K. Kanzaki (Eds.), *Advances in Natural Language Processing* (pp. 289-299). Heidelberg, Germany: Springer. doi:10.1007/978-3-642-33983-7_29

Ryan, J. O., Kaltman, E., Hong, T., Mateas, M., & Wardrip-Fruin, N. (2015). People tend to like related games. In *Proceedings of the 10th International Conference on the Foundations of Digital Games*. Santa Cruz, CA: Society for the Advancement of the Science of Digital Games.

Thiagarajan, S. (1999). Team activities for learning and performance. In H. D. Stolovitch & E. J. Keeps (Eds.), *Handbook of human performance technology* (pp. 518–544). San Francisco, CA: Jossey-Bass.

Tobias, S., Fletcher, J. D., & Wind, A. P. (2014). Game-based learning. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology*, (4th ed., pp. 485-503). doi:10.1007/978-1-4614-3185-5_38

Wang, H., Shen, C., & Ritterfeld, U. (2009). Enjoyment of digital games: What makes them "seriously" fun? In U. Ritterfeld, M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects*. New York, NY: Routledge. doi:10.4324/9780203891650

Wilson, K. A., Bedwell, W. L., Lazzara, E. H., Salas, E., Burke, C. S., Estock, J. L., . . . Conkey, C. (2009). Relationships between game attributes and learning outcomes: Review and research proposals. *Simulation & Gaming, 40*, 217-266. doi:10.1177/1046878108321866

Wouters, P., van Nimwegen, C., van Oostendorp, H., & van der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology, 105*(2), 249–265. doi:10.1037/a0031311

Zagal, J. P., Ladd, A., & Johnson, T. (2009). Characterizing and understanding game reviews. *In Proceedings of the 4th International Conference on Foundations of Digital Games* (pp. 215-222). doi:10.1145/1536513.1536553

Zagal, J. P., & Tomuro, N. (2010). The aesthetics of gameplay: a lexical approach. In *Proceedings of the 14th International Academic MindTrek Conference: Envisioning Future Media Environments* (pp. 9-16). doi:10.1145/1930488.1930492

Zagal, J. P., & Tomuro, N. (2013). Cultural differences in game appreciation: A study of player game reviews. In *Proceedings of the 8th International Conference on the Foundations of Digital Games* (pp. 86-93). Santa Cruz, CA: Society for the Advancement of the Science of Digital Games.

Zagal J. P., Tomuro N., & Shepitsen A. (2012). Natural Language Processing in game studies research: An overview. *Simulation & Gaming, 43*(3), 356–373. doi:10.1177/1046878111422560

Zhu, M., & Fang, X. (2015). A lexical approach to study computer games and game play experience via online reviews. *International Journal of Human-Computer Interaction, 31*(6), 413-426. doi:10.1080/10447318.2015.1036228