# Assessing User Experiences with ZORQ: A Gamification Framework for Computer Science Education

Sherri Weitl Harms
University of Nebraska at Kearney
harmssk@unk.edu

Adam Spanier
University of Nebraska at Omaha
aspanier@unomaha.edu

John Hastings
University of Nebraska at Kearney
hastingsjd@unk.edu

Matthew Rokusek
University of Nebraska Lincoln
mrokusek4@huskers.unl.edu

## Abstract

ZORQ is a gamification software framework designed to increase student engagement within undergraduate Computer Science (CS) education. ZORQ is an attractive learning method that (1) utilizes numerous gamification elements, (2) provides a collaborative, game-development based learning approach, (3) offers an opportunity for students to explore a complex, real-world software development implementation, and (4) provides students with a high level of engagement with the system and a high level of social engagement in its collaborative customization.

The usage of ZORQ was assessed using quantitative, qualitative and sentiment analyses in a Data Structures and Algorithms course over five years. The overwhelmingly positive results show that students were satisfied with their user experience and ZORQ was beneficial to their educational experience. By triangulating results from multiple analyses, this study adds to a deeper understanding of how gamification can improve learning and retention and provides a novel, robust, holistic methodology for evaluating user experiences.

Keywords: Gamification, CS Education, User Experience, Sentiment Analysis, Software Usability

## 1. Introduction

Gamification is "the use of design elements characteristic for games in non-game contexts" (Deterding et al., 2011). As found in Ahmad (2020), gamification exhibits effective outcomes when used as a tool to teach complex computer science concepts in high-level educational courses. By implementing game elements into non-game environments, gamification seeks to recreate the 'fun' atmosphere provided by computer games (Deterding et al., 2011). By leveraging the inherent 'fun-ness' of computer games, gamification allows educators to better engage and motivate students to participate in complex and difficult course material (Ahmad et al., 2020). In this way, gamification provides the much needed component to education by which fun can be used to transform students' attitudes towards learning (Alsawaier, 2018).

While gamification presents beneficial outcomes like engagement and motivation when applied to high-level learning in CS education (Alsawaier, 2018; Deterding et al., 2011), mixed results regarding improvement in student learning have been reported (Bai et al., 2020; Hamari et al., 2014; Toda et al., 2019). Some studies indicate that, should a well-designed application not be followed, gamification has the potential to yield negative outcomes (Oliveira et al., 2022). Furthermore, recent studies undertaken to investigate the influence of gamification on student experience, indicate that a non-personalized, "one size fits all" gamification design can achieve less motivation and engagement than a gamification application (GA) designed specifically for the user it aims to influence (Klock et al., 2018; Maheu-Cadotte et al., 2018; Oliveira et al., 2022).

To best understand GAs and their corresponding outcomes when applied to undergraduate CS courses, functional and characteristic classifications can be applied to group similar applications within well-defined domains (Spanier et al., 2021). Classifications play an important role in conceptualizing, analyzing, and understanding existing scientific data (Simpson, 1961, Bailey, 1994, Bailey, 2005). Once classified, broader patterns can be observed among GAs in a larger sense. Spanier (2021) presents a classification system with two broad classifications for

URI: https://hdl.handle.net/10125/102771
978-0-9981331-6-4

GAs: static and dynamic. Static gamification presents an application where every participant interacts with a well-defined gamification interface (Spanier et al., 2021). Dynamic gamification adds an extra layer of flexibility by providing a gamification development framework by which instructors can mold and shape the gamification experience based on student input (Spanier et al., 2021).

While research concerning dynamic gamification is limited, dynamic gamification presents potential benefits to CS education by allowing students to participate not only in the gamified application, but also the development of its implementation (Spanier et al., 2021). At the present, empirical dynamic GAs have received limited quantitative analyses in terms of observed outcomes when applied to undergraduate CS courses. This lack of analyses necessitates organized research effort aimed at applying and understanding quantifiable outcomes of existing dynamic GAs.

One such application that presents notable potential to meet the flexibility and personalization requirements is the ZORQ gamification framework (Hastings et al., 2022). The ZORQ gamification framework's primary goal to increase student engagement, motivation, and success in undergraduate CS education courses. ZORQ is a unique dynamic gamification framework in which students design and program autonomous ships that navigate within a two-dimensional game universe filled with obstacles. Furthermore, ZORQ implements a dynamic pre-phase wherein educators and students can modify the gamification framework source code and update game settings on the fly. This allows users to more accurately cater the application to the needs of a course. While ZORQ meets the requirements for a dynamic gamification framework-based application, and was briefly assessed in Hastings et al. (2022), the sentiment of user satisfaction while using ZORQ in an undergraduate CS class has yet to be analyzed.

### 1.1. Research Questions

The following research questions guide the study:

RQ.1 What key terms do students associate with the ZORQ Gamification Framework?

RQ.2 How satisfied are students with the ZORQ Gamification Framework?

RQ.3 What percentage of student-selected key terms are positive, negative, or neutral?

RQ.4 What sentiments did students express after participating in the ZORQ intervention?

This research aims to analyze student sentiment and satisfaction based on data collected during ZORQ implementations over the course of five years in an undergraduate Data Structures and Algorithms (DSA) class at a regional university. The study makes use of an open-ended Microsoft Product Desirability Toolkit (PDT) survey and unigram lexical sentiment analysis tools. While single instrument quantitative analysis provides adequate resolution for outcome analysis, this research aims to create higher analysis resolution through several analyses, including: (1) an overall analysis of the PDT word choices, (2) a quantitative sentiment analysis of the PDT word choices, (3) a qualitative sentiment analysis of the PDT word choice comments and open-ended questions in the study, and (4) a quantitative sentiment analysis of the PDT word choice comments and open-ended questions in the study. The combination of using the qualitative data provided by the PDT method with existing sentiment analyses tools is a novel quantitative approach for measuring software desirability.

## 2. ZORQ: A Gamification Framework for CS Education

ZORQ is a gamification framework in which space ships navigate a 2D game universe filled with objects which affect a ship either positively or negatively (Hastings et al., 2022). Examples of game objects that have been used include: fuel, shields, mines, black holes, ship-jump portals, electromagnetic pulses, bullets and lasers, as shown in the Figure 1 screenshot. Gamified elements include constant feedback, clear rules, goal-oriented challenges, freedom to fail, points and a leaderboard, as shown in the upper left corner. In the current version of ZORQ, ships earn points for each frame in which they remain active/alive, and also by gathering resources which award bonus points. Remaining active is the primary means by which ships maximize points. Negative encounters cause a ship to be deactivated for five seconds, after which they respawn in a different location. Examples of negative encounters include running into an obstacle, getting sucked into a black hole, or being successfully targeted by another ship after which the attacking ship steals a percentage of the attacked ship's score. Engagements between ships generally favor the ship with more resources (e.g., fuel, bullets, shield energy, etc.) or a superior strategy.

### 2.1. ZORQ Usage in a DSA course

ZORQ was introduced one month before the end of the semester to give sufficient time for students to interact with the framework. The class collaboratively

Figure 1. ZORQ Screenshot

designs and implements framework adjustments as desired to create a unique instance of ZORQ. These changes included adding or removing game objects, or changing the behavior of the game or existing objects. For example, over time, new game bonuses have been added, such as shields which allow a ship to block any sort of attacks for a period of time. With perfective changes, each new semester used the most recent version of ZORQ.

Once the configuration of the framework was completed, students next design and implement code to create and automatically control their own ships. ZORQ is noteworthy in that it not only allows for student-directed general customization upfront, through its design, it also adapts to the needs of individual students by allowing them to creatively implement ship controllers entirely of their own design in ways that best motivate them. Students were instructed to develop a controller that implements a philosophy of their own design. When designing controllers, students generally focus on maximizing their scores, and surviving as long as possible, although some students might pursue other goals such as path finding or resource gathering. Grading on the assignment was not tied to how high ships score when the game runs, but rather how well students: put effort into the assignment, implement their philosophy, describe their philosophy to the class, document their code, and test their code. A focus on implementing a philosophy rather than on "competition" has allowed for creative solutions. For example, one student implemented a 'copy cat' controller that would mirror the actions of the closest opposing controller. This focus also addresses the issue of some students not enjoying an atmosphere of competition.

The students were provided small sample skeleton controllers that demonstrate, in a basic way, what a controller can do by analyzing the state of the game (e.g., the game objects, their locations, and what they are doing) and then selecting an action to take. The example controllers also demonstrate how to utilize some of the built in utilities, e.g., the function to compute angles to other game objects and demonstrate that capable controllers can be built without using AI concepts, which students will see in a later course. In addition, in order to have something to test against, students were provided a library of precompiled, obfuscated controllers from previous semesters.

A single class period at the end of the semester was dedicated to running the students' controllers. This was intended partly so students don't get caught up in the need to repeatedly compete with each other. During the demonstration day, each student discussed their creation. This activity served the additional benefit of helping students get comfortable talking in front of peers, as well as talking about their creation. Students submitted their code to their own private repository which they shared with the instructor for grading.

## 3. Background Methodologies

### 3.1. Software Desirability

The Microsoft Product Desirability Toolkit (PDT) (Benedek and Miner, 2002) is a well-known qualitative analysis tool used to evaluate user experience and to conduct usability testing of software (C. Barnum, 2020). This approach was originally designed to ask the user to complete a usability test for a product, then pick the five "cards" (terms) from a group of 118 "product reaction cards" that best match their reaction to the system (Benedek and Miner, 2002). In the second phase of the PDT, the selected cards become the basis of a guided interview aimed at soliciting feedback and comments regarding the users' experiences and rationale for the terms they selected. As described in Barnum (2020), the PDT it was created to "understand the illusive, intangible aspect of desirability resulting from a user's experience with a product". The product reaction cards have been used extensively as a qualitative tool for assessing desirability resulting from a user's experience and satisfaction with a product (C. Barnum, 2020; C. M. Barnum and Palmer, 2010; Booth and Stumpf, 2013; Hastings et al., 2010; Li and Wang, 2014; Tullis and Stetson, 2006; Veral and Macías, 2019). The PDT is described as the closest tool that uses psychometric theory to create a user experience (UX)-relevant measure of product or service desirability (Lewis and Sauro, 2020).

The advantages of using the PDT are "1) it aims

to avoid a bias toward the positive found in typical questionnaires (e.g., it has been found that if a respondent thinks that a survey intends to assess the quality of a product, they are likely to provide more positive answers about quality) and 2) it is able to more effectively uncover constructive negative criticisms in the guided interview" (Hastings et al., 2010). In the second phase of the PDT, a rich and revealing story of user experience is constructed as users comment on their word choice. Triangulating these findings with post-test questionnaire data and direct observation strengthens the understanding of the desirability factor (C. M. Barnum and Palmer, 2010). Additionally, the text from the user comments are ripe for sentiment analysis.

The PDT provides a way to triangulate findings from other feedback mechanisms, with potential to produce more meaningful and substantive results of user experiences (C. M. Barnum and Palmer, 2010). However, Veral (2019) noted that the high number of available cards makes it necessary to think of improvements on the original method.

### 3.2. Lexical Sentiment Analysis

Sentiment reflects the opinions or views of a person, event, or service. Sentiment analysis utilizes computation to mine emotions and sentiment from text to determine general opinions and sentiments that occur in online textual data repositories (Asghar et al., 2014). Sentiment analysis is often used in social media and microblogging research to determine public sentiment concerning specific topics, services, and/or events.

Before sentiment analysis can occur, data must be preprocessed and cleaned through the removal of: (1) numbers, (2) URLs, (3) HTML Tags, (4) null entries, and by (5) negation, and (6) slang removal (Jianqiang and Xiaolin, 2017). Sentiment analysis relies on tokenization, stemming, lemmatization, and the removal stopwords in the textual data (Asghar et al., 2014; Jeong et al., 2011). Tokenization is the process by which a natural language is broken down into components (Ahuja and Dubey, 2017). For example, the English sentence "The cat is 7.89 lbs." results in the following tokens: "The" "cat" "is" "7.89" "lbs" ".". Stemming is the process by which words in the text are reduced to their resulting stems (Asghar et al., 2014). For example, 'automatic', 'automaton', and 'automate' would result in the stem 'automat'. Lemmatization is the reduction of word inflections to a common root (Asghar et al., 2014). The words, 'fire,' 'fires,' 'fired,' and 'firing' all reduce to a common lemma, 'fire'. Stopwords consist of any non-discriminative word that does not provide useful sentiment data (Saif et al., 2014).

To carry out unigram lexical sentiment analysis in the cleaned and preprocessed data set, the following process is utilized: (1) each response word and it's correlated responses are collected into individual csv files, (2) the response texts are tokenized, (3) stopwords are removed, (4) the response text is lemmatized, (5) words with a length less than 2 are removed, (5) double spaces are eliminated, (6) lemmatized data is rejoined in alphabetical sentences, (7) words are tagged by part of speech, (8) each word receives an individual positive and negative sentiment score, (9) and finally, the overall average positive to negative sentiment is calculated per each statement (Jianqiang et al., 2018).

Total sentiment values for each word are calculated by summing both the positive and negative PMI (point-wise mutual information) measurement derived from the Senti-WordNet lexical dictionary (Jianqiang et al., 2018). The WordSentiment function used to provide word specific values in each response is:

$$WordSent(w) = PMI(w, pos) - PMI(w, neg)$$

To determine total response sentiment, all WordSent scores contained within a response $T$ are summed and the resulting output determines total sentiment ResponseSent of the tweet (Jianqiang et al., 2018). The ResponseSent function is as follows:

$$ResponseSent(T) = \sum_{i=1}^{|T|} n_i$$

where each response $T$ is a set of words $ST$ $n_i \in T$ and $n$ is the number of words in each response $T$. The ResponseSent score is then normalized (Jianqiang et al., 2018).

After calculating the ResponseSent score for each response, the scores are averaged by taking the mean of all sentiment scores in the product reaction term csv. Mean sentiment (MS) in the set of responses D is determined by using the function:

$$MS(D) = \frac{\sum_{i}^{n} x_i}{n}$$

where $x_i$ is the current, ith, ResponseSent score.

## 4. ZORQ User Study Methods and Data

### 4.1. Data Collection

An anonymous online exit survey was created to assess students who had used ZORQ in a DSA course at a regional university. The study received ethical approval from the university, and students were given informed consent to opt-in to completing the survey, but

no incentive was provided. Since the DSA course is a required course that all students must pass to continue in the program, only students who passed the course were surveyed, to ensure each student was only surveyed one time. The first survey was conducted shortly after the completion of the fall 2019 semester, for students who had completed the course in 2017-2019. The exit survey was also given to students who completed the course in fall 2020 and then again in fall 2021 within a few weeks after the end of the semester. Of the 98 students in the population of students who have completed the course in the semesters surveyed, there were 49 completed responses, as reported in Hastings (2022). Nine responses were from female students, (18.3%) which matches the male/female distribution in the overall population of this course. One student did not indicate gender, and 39 respondents identified as male.

To evaluate the research questions, a usability feedback survey based on the PDT was used. To deal with the problem of the high number of available cards noted by Veral (2019), research often uses a subset of the 118 PDT cards (C. M. Barnum and Palmer, 2010; Booth and Stumpf, 2013; Hastings et al., 2010; Li and Wang, 2014; Veral and Macías, 2019). For this study, we used the same set of 55 words as Barnum (2010) and shown in the original article (Benedek and Miner, 2002). As in Hastings (2010), the participant was not given advance warning as to how a selected term would be used or that a follow-up comment would be required in an attempt to reduce the effect of bias on the selection of any term.

The set of PDT words were statistically evaluated and categorized (positive, neutral, and negative) by Veral (2019). These categories were used in this study. With consideration for the potential bias of positive feedback, of the 55 words selected, the same percentage of positive terms from the original PDT (60%) was kept. Negative and neutral terms each accounted for 20% of the remaining words selected.

The terms that the students were provided to select from in best describing their experience with the game environment system, were the following positive terms Accessible, Appealing, Attractive, Collaborative, Comprehensive, Consistent, Customizable, Desirable, Easy To Use, Efficient, Empowering, Exciting, Familiar, Fast, Flexible, Fresh, Fun, High Quality, Inviting, Motivating, Organized, Personal, Relevant, Reliable, Sophisticated, Stimulating, Straight Forward, Time-saving, Trustworthy, Usable, Useful, Valuable; neutral terms Complex, Connected, Overbearing, Overwhelming, Patronizing, Predictable, Rigid, Simplistic, Time Consuming, Too technical, Unconventional, Unpredictable; and negative terms Busy, Confusing, Frustrating, GetsInTheWay, HardToUse, Inconsistent, Intimidating, Notvaluable, Slow, Stressful, Uncontrollable.

Because this was an anonymously online survey, to implement the second phase of the PDT method, each participant was asked to comment on each of the five terms selected in the online survey itself, rather than conducting a face-to-face follow-up interview session.

### 4.2. Analysis Methods

To answer RQ.1, students were asked to select five of the 55 PDT words provided, using the prompt: "Pick 5 words from the following group which best describe your experience with the game environment system." The terms selected indicate what terms the students associate with the ZORQ gamification framework and are used to gain insight into the student's reaction to ZORQ.

Quantitatively evaluating the words selected provided responses to RQ.2, RQ.3, and RQ.4, and helped to better understand the students' experiences. Each of the 49 respondents selected five terms (245 selections total). As a means to quantitatively summarize the user's satisfaction with experience, the number of times each term was selected was tallied, and totaled by category.

The collection of the respondents' five-word PDT response term groups (PRTG) were evaluated for the number of positive, neutral, and negative terms in each. The number of PRTGs with a majority of positive terms and similarly, negative terms, were tallied. Each PRTG set was also evaluated for sentiment using lexical sentiment analysis.

A qualitative analysis of the PDT word choice comments was used to triangulate the findings from the other feedback mechanisms in response to RQ.2 and RQ.4.

Existing unigram sentiment analysis techniques as described above were used on the text from the user provided survey comments for the selected PDT words and the text from the open-ended questions. The sentiment analysis provides a new feedback mechanism to understand the sentiment expressed after the user experience, in response to RQ.2 and RQ.4.

## 5. Results

### 5.1. Overall PDT Word Selection Results

Of the 55 terms presented to the respondents, 34 of the terms had at least two respondent selections and are included in the resulting word-cloud based on frequency of terms selected, as shown in Figure 2. The top six terms selected by respondents, along with the number of respondents who selected the term are: fun

(26), stimulating (23), valuable (14), and exciting (13), motivating (12), and customizable (12). The word-cloud provides a response to RQ.1 and indicates that students had overall positive engagement experiences when using ZORQ.


Figure 2. ZORQ Word Cloud Analysis

## 5.2. Quantitative PDT Word Selection Sentiment Results

Figure 2 also provides a starting response to RQ.2. Positive terms are shown in blue, negative terms are shown in red, and neutral terms are shown in green. As shown in Figure 2, there were 28 positive terms selected, with the term fun having the highest positive term frequency with 26 respondent selections; six negative terms selected, with the term intimidating having the highest negative term frequency with seven respondent selections; and seven terms selected that were classified as neutral, with the term complex having the highest neutral term frequency with 11 respondent selections.

RQ.3 and RQ.4 are answered several ways. Out of the total 245 total possible word selections, 193 (79%) terms selected were positive terms, 26 (11%) were neutral terms, and 26 (11%) were negative terms, as shown in Figure 3, using the categorization for the words as positive, neutral or negative based on Veral (2019).

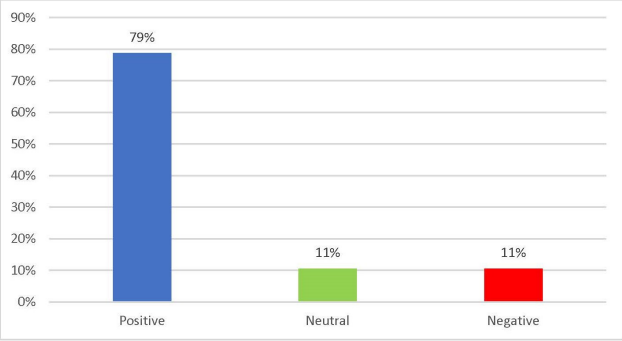


Figure 3. Sentiment of Terms Selected in the PRTGs

Using the same categorization of sentiment, of the 49 respondent PRTGs, 22 PRTGs had five positive terms and 13 had four positive terms for a total of 35 (71%) of the respondents selecting at least four positive terms in their PRTGs as shown in Table 1. Also shown, 32 (65%) respondent PRTGs did not include any negative term, and 12 PRTGs included only one negative term, for a total of 44 (90%) of the respondents selecting at most one negative term.

Overall, 43 (88%) of the respondents selected a majority (3 or more) positive terms; only three (6%) respondents selected a majority of negative terms. Three (6%) of the respondents did not have a majority of either positive or negative terms in their PRTG.

Table 1. Number of Positive, Neutral, and Negative Terms in the PRTGs

| Number of Terms | Positive | Neutral | Negative |
|---|---|---|---|
| 5 | 22 | 0 | 0 |
| 4 | 13 | 0 | 1 |
| 3 | 8 | 0 | 2 |
| 2 | 3 | 0 | 2 |
| 1 | 1 | 16 | 12 |
| 0 | 2 | 28 | 32 |

To triangulate these results, unigram lexical sentiment analysis was used as a method independent of the Veral (2019) categories. Of the 49 PRTGs, 11 exhibited negative sentiment (22%), 10 were neutral (20%), and 28 exhibited positive sentiment (58%). The most negative PRTG sentiment was −0.475 and resulted from the PRTG containing the terms: $PRTG_n$ ={fun, hard to use, slow, uncontrollable, and unconventional}. The most positive PRTG sentiment was 0.5 and resulted from two different PRTGs: $PRTG_{p1}$ ={accessible, appealing, fun, inviting, and stimulating} and $PRTG_{p2}$ ={appealing, attractive, collaborative, exciting, and stimulating}. The average sentiment score for all PRTGs is 0.072.

## 5.3. Qualitative PDT Word Selection Comments Sentiment Results

A review of the respondents' comments provide another means to evaluate their user experience with ZORQ. The respondents' comments were overwhelmingly positive for the terms they selected. For example, comments provided for the most commonly selected term, fun, included: "The autonomy offered by the project and the uncertainty of how your ship would stack up against your classmates' were two aspects that were very enjoyable"; "It was very enjoyable to work on. In fact, I would say it probably ties for my favorite project that I completed throughout

my studies"; "While I was learning more about coding by learning the system and coding my algorithm, it didn't seem like coding as I was enjoying the time I spent working on the project"; "It was fun to problem solve and find why my ship was not doing what it was supposed to"; and "I had alot of fun completing the project. It was perhaps the most fun that I have ever had on a school project."

Students comments on the "neutral" terms tended to be positive in nature. One respondent's explanation of his/her choice of the term "complex" was "Being larger and more complex than other projects I had worked on, it required a fairly thorough understanding of all the skills a student should have acquired up to that point." The "neutral" term unconventional three comments were: "This assignment is unlike many of the other tasks we see in the undergraduate curriculum it almost endeavors to inject fun into the classroom in an unexpected way"; "The assignment itself seemed unconventional for a project in this course nevertheless I think important concepts can be learned from the unconventional approach by assigning a final project that was very enjoyable to students. I think many people put more effort into and learned more from the assignment than they would have otherwise"; and "I chose unconventional because it is not the typical type of project that is given. I have never done anything else quite like this for a school related project".

Students comments on the "negative" terms tended to be helpful or positive, in nature. For example, student comments for the most selected negative term, "intimidating", included "The initial scale of possible things to do is a bit menacing. A basics guide (written or video) would help ease into the project better."; and "On the outside looking in, coding a ship that is able to act correctly for every single instance seems intimidating. In practice it is much simpler than I thought it was going to be."; and "Sometimes I felt overwhelmed due to the fact, that is hard to track the classes because it's so much to learn. I think that should exist a guideline to keep in track everything.". Student comments on the next most selected negative term, "frustrating", included, "Since the code base we started with was bigger than we were used to, it was sometimes frustrating to figure out what to do or where to look - but that frustration also made it more engaging."; and "It was difficult and frustrating at times but it got you to think about what was going on."

Overall, student comments include "Kept me connected and engaged to the system"; "Worked to try and create the best ship that I could. This made it very engaging for me."; "It was one of the most exciting and unique projects I have done"; "It drew me in. I spent time on it because I wanted to, not because I had to"; "I had a lot of fun completing the project"; "I liked how I had the freedom to design my own ship. I also liked how I had to understand a system that was in place already and know how to use it in order to implement my ship."; and "It is set as simply as it can be, so a beginner is able to dive in and understand most of what is going on."

### 5.4. Quantitative PDT Word Selection Comments Sentiment Results

Lexical sentiment analysis of each comment group related to a PDT response term indicates the most positive response word as measured by associated comments is attractive with an average comment sentiment score of $\bar{x} = 0.583$. The most negative response term is slow with an average comment sentiment score of $\bar{x} = -0.0625$.

The top 5 most positive response terms by comment sentiment average were attractive with an $\bar{x}$ of 0.583, desirable with an $\bar{x}$ of 0.359, motivating with an $\bar{x}$ of 0.170, accessible with an $\bar{x}$ of 0.140, and high quality with an $\bar{x}$ of 0.583. The top 5 most negative response terms by comment sentiment average were unconventional with an $\bar{x}$ of $-0.231$, fresh with an $\bar{x}$ of $-0.227$, unpredictable with an $\bar{x}$ of $-0.226$, hard To use with an $\bar{x}$ of $-0.149$, and frustrating with an $\bar{x}$ of $-0.104$. The total average of all response term comment groups is 0.044.

Upon the completion of the survey, all participants were asked to provide what they liked best about the system. Each response was likewise evaluated for sentiment using lexical sentiment analysis. Of the 38 total responses, 3 were negative (8%), the lowest score coming in at $-0.063$, 10 were neutral (26%), and the remaining 25 were positive (66%), the highest being 0.375. The overall average sentiment score for all responses came to 0.108.

## 6. Discussion

### 6.1. User Satisfaction and Sentiment with Using ZORQ

Overall, in analyzing the results from all methods used in this study, students overwhelmingly responded with positive sentiment to their experience with ZORQ (RQ.4), and in terms of their satisfaction (RQ.2) of the ZORQ software. This aligns well with existing research about the value of gamification when applied to CS education (Spanier et al., 2021). Of the words selected, 79% were positive, and 86% selected a majority of positive words in their PRTGs, using the Veral (2019) categories. Using lexical sentiment analysis, 58% of the PRTGs were positive. The qualitative analysis

of the respondent comments indicate overwhelmingly positive sentiment. The sentiment analysis response term comment averages indicated that students generally favored interactions with ZORQ. This indication is corroborated by the existence of only 10 negative sentiment response term averages as opposed to 29 positive averages. Further, the overall average of all comment groups is $\bar{x} = 0.044$ indicating a generalized positive sentiment. While 0.044 is not notably high, positivity conclusively represents the majority of all response term comment groups.

As noted in section 5.2, fun presented the most positive sentiment and intimidating presented the most negative. These findings are somewhat at odds with the comment sentiment analysis averages. The term fun generated an average comment sentiment score of $\bar{x} = 0.094$ and fell into the $9^{th}$ most positive slot during lexical sentiment analysis, while the term intimidating generated an average comment positive sentiment score of $\bar{x} = 0.046769$ and sits at the $17^{th}$ most positive spot.

Interestingly, some response terms with expected positive or negative values generated unexpected comment sentiment averages. Most notably, empowering generated a negative average comment sentiment score of $\bar{x} = -0.078$ while overwhelming generated a positive average comment sentiment score of $\bar{x} = 0.002$. gets in the way also generated a positive average comment sentiment scores of $\bar{x} = 0.093$.

While lexical sentiment analysis yields generally accurate sentiment scores, the utilization of a unigram lexical approach does not take into account the syntactic composition of each response. In this way, some responses can result in erroneous sentiment scores. One such instance occurred when a student chose fun. The comment reported by the student was, "It was fun to problem solve and find why my ship was not doing what it was suppose to." While the response is innately positive in nature, the resulting sentiment score rated the response as negative, scoring it at −0.083. This swap occurred due to the unigram approach taken in this research. The negative words problem find why and not outscored the positive terms fun solve and supposed.

## 6.2. ZORQ as a Teaching Tool

A review of respondent comments, such as those shown above associated with the terms fun and intimidating provide additional support to the benefit of ZORQ as a teaching tool. They also provide helpful input to improve the implementation of ZORQ in the classroom, such as the student suggestion for "a basics guide (written or video) to help ease into the project".

Anecdotally, ZORQ is used in later courses, and faculty frequently hear students fondly discussing and reminiscing about their experiences with ZORQ during the DSA course. Observations of student performance in later courses after using ZORQ suggest better student maturity and comprehension in preparation for proposing and implementing their own independent projects.

## 6.3. Survey Design and Limitations

Even though time had elapsed between the usage of ZORQ in the DSA course and the survey completion for the students who used ZORQ in 2017 and 2018, no significant difference was seen between their responses and those of students who were surveyed right after course completion.

To avoid duplicate student surveys, only students who passed the course were surveyed. This may insert a selection bias as students who did not pass had lower grades and were unable to complete the survey until they passed the course. This situation is challenging for gamification as such a design choice likely pushes findings up (Sanchez et al., 2020).

Because it took five years of implementation in the DSA course to gather the data used in this study, using a control group was not feasible. The small sample size for quantitative analyses and the lack of a control group are noted limitations of this study. However, this study has numerous ways it can be continued in future work as explained below.

## 6.4. Analysis Method Benefits

The combination of using the qualitative data provided by the PDT method with existing sentiment analyses tools is a novel quantitative approach for measuring software desirability. Through this combined use of quantitative and qualitative analyses, this research created higher analysis resolution of user satisfaction of a gamification framework, even though a small sample size and no control group were realities.

Previous educational gamification research indicates that, due to the influence of contextual factors like participant backgrounds, instructor skills, and game element knowledge on gamification implementations, the incorporation of qualitative research in gamification is highly important (Alsawaier, 2019). Qualitative data sources can play a vital role in getting students' views, their criticism, their evaluation, or at least their full reaction to how the gamified events were designed and implemented (Alsawaier, 2019). A number of distinctly qualitative research efforts (Ahmad et al., 2020; Deterding et al., 2011; Luo, 2021; Maheu-Cadotte et al., 2018; Manzano-León et al., 2021;

Oliveira et al., 2022) exist relating to gamification in CS education, but the inherently qualitative nature of gamification has generated significantly less quantitative data. Furthermore, the lack of quantitative research prohibits the combination, analysis, and extraction of outcomes based on qualitative and quantitative data in a mixed-methods manner.

Balancing qualitative and quantitative research elements allows learners' experiences and sentiment concerning gamified events to be taken into consideration while also providing a more grounded conclusion through joining and certification of discoveries. Mixed-methods research is any research in which, "the investigator collects and analyzes data, integrates the findings, and draws inferences using both qualitative and quantitative approaches or methods in a single study or a program of inquiry" (Tashakkori and Creswell, 2007). Though both qualitative and quantitative research each exhibit their own advantages and disadvantages, the combination of the two often creates a broader, more flexible approach. According to Migiro and Magangi (2011), mixed-methods research presents five notable advantages: (1) different methods can be used for different purposes, (2) triangulation can occur, (3) quantitative results can be explained with qualitative analysis, (4) qualitative data can generate a testable theory, and (5) each mixed-method study is enhanced with a supplemental data set. The scope of what can be learned on the effect of gamification in a learning environment can be greatly expanded with balanced qualitative and quantitative elements as it allows learners' experiences and sentiment of the gamified events to be taken into consideration (Alsawaier, 2019).

### 6.5. Overall

As a note on the applicability of the PDT as a survey tool for use in studying the effectiveness of gamification applications, it was quick and easy to construct and distribute and has a strong foundation in software product evaluation (Hastings et al., 2010).

Several metrics should be considered when developing studies on gamification use in education (Luo, 2021). Luo noted that there is a lack of studies that achieved meaningful gamification in the educational domain, engagement is one key measure of gamification's effectiveness, and studies should focus more on why it is effective and what makes it effective, rather than merely assessing whether it is effective. This study shows that meaningful gamification in education can be achieved, when student engagement is through coding within the gamification system itself, as done within the ZORQ framework.

## 7. Future Work and Conclusions

For future utilization of ZORQ in the DSA course, showing ZORQ earlier in the semester should be explored, so students can have some ideas about changes that they want to make. The use of ZORQ in other courses, such as Software Engineering, and Artificial Intelligence needs to be improved.

In terms of the ZORQ system implementation, students have expressed an interest in seeing a view of the action looking out from the front of the ship. There are also future plans to make ZORQ available from the code repository, with a more formalized initialization and installation process for other schools to use and future plans to use deep learning to create controllers by learning from the state of the system and how well existing controllers performed.

Additionally, because the sample size was quite small for quantitative analyses and there was no control group, future work includes studies designed to address these issues, as well as studies to look at the impact ZORQ had on student retention. A longitudinal study that investigates the novelty effects of ZORQ as a gamification tool and that examines individual differences for all students is needed.

A deeper exploration of the analysis approach for evaluating software desirability used in this paper as a generalized methodology is needed. Using this method, software desirability sentiment is quantifiably and qualitatively evaluated by merging the use of qualitative PDT (Benedek and Miner, 2002) with recent machine learning lexical sentiment analysis algorithms (Asghar et al., 2014).

In summary, ZORQ provides students with a high level of engagement with the system and a high level of social engagement in its collaborative customization. The results of this study support a conclusion that the use of ZORQ gamification framework increases student engagement and success within undergraduate CS education. Additionally, by adopting a novel triangulation of several analyses, this study adds to a deeper understanding of how gamification could improve learning and retention (Alsawaier, 2019) and provides a robust, holistic methodology to evaluate software desirability.

## References

Ahmad, A., Zeshan, F., Khan, M. S., Marriam, R., Ali, A., & Samreen, A. (2020). The impact of gamification on learning outcomes of computer science majors. ACM Trans. Comput. Educ., 20(2). https://doi.org/10.1145/3383456

Ahuja, S., & Dubey, G. (2017). Clustering and sentiment analysis on twitter data. 2017 2nd International Conference on Telecom. and Networks (TEL-NET), 1–5.
Alsawaier, R. S. (2018). The effect of gamification on motivation and engagement. The International Journal of Information and Learning Technology, 35(1), 56–79.
Alsawaier, R. S. (2019). Research trends in the study of gamification. The International Journal of Information and Learning Technology, 36(5), 373–380.
Asghar, M. Z., Khan, A., Ahmad, S., & Kundi, F. M. (2014). A review of feature extraction in sentiment analysis. Journal of Basic and Applied Scientific Research, 4(3), 181–186.
Bai, S., Hew, K. F., & Huang, B. (2020). Does gamification improve student learning outcome? evidence from a meta-analysis and synthesis of qualitative data in educational contexts. Edu. Research Review, 30, 100322.
Bailey, K. D. (1994). Typologies and taxonomies: An introduction to classification techniques. Sage Pubs.
Bailey, K. D. (2005). Typology construction, methods and issues.
Barnum, C. (2020). Usability testing essentials. Elsevier.
Barnum, C. M., & Palmer, L. A. (2010). More than a feeling: Understanding the desirability factor in user experience. CHI '10 Extended Abstracts on Human Factors in Computing Systems, 4703–4716.
Benedek, J., & Miner, T. (2002). Measuring desirability: New methods for evaluating desirability in a usability lab setting. Proceedings of the Usability Professionals' Association Conference.
Booth, T., & Stumpf, S. (2013). End-user experiences of visual and textual programming environments for arduino. IS-EUD.
Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining" gamification". Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments, 9–15.
Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamifcation work? - a literature review of empirical studies on gamification. 47th Hawaii International Conference on System Sciences, 3025–3034.
Hastings, J. D., Mirasano, A., Latchininsky, A., & Schell, S. P. (2010). Carma: Assessing usability through a non-biased online survey technique. 2010 43rd Hawaii International Conference on System Sciences, 1–10.
Hastings, J. D., Weitl-Harms, S., Spanier, A., Rokusek, M., & Henszey, R. (2022). Zorq: A gamification framework for computer science education (in submission). 2022 IEEE Frontiers in Education Proceedings, 1–9.
Jeong, H., Shin, D., & Choi, J. (2011). Ferom: Feature extraction and refinement for opinion mining. ETRI Journal, 33(5), 720–730.
Jianqiang, Z., & Xiaolin, G. (2017). Comparison research on text pre-processing methods on twitter sentiment analysis. IEEE Access, 5, 2870–2879.
Jianqiang, Z., Xiaolin, G., & Xuejun, Z. (2018). Deep convolution neural networks for twitter sentiment analysis. IEEE Access, 6, 23253–23260.
Klock, A. C. T., Ogawa, A. N., Gasparini, I., & Pimenta, M. S. (2018). Does gamification matter? a systematic mapping about the evaluation of gamification in educational environments. Proceedings of the 33rd Annual ACM Symposium on Applied Computing, 2006–2012.
Lewis, J., & Sauro, J. (2020). 10 things to know about the microsoft desirability toolkit [Accessed: April 10, 2022]. https://measuringu.com/microsoft-desirability/
Li, Y., & Wang, X.-y. (2014). Mobile interface studies about style description and influential factors. 2014 International Conference on Management Science & Engineering 21th Annual Conference Proceedings, 578–583. https://doi.org/10.1109/ICMSE.2014.6930281
Luo, Z. (2021). Educational gamification from 1995 to 2020: A bibliometric analysis. 2021 the 6th International Conference on Distance Edu. and Learning, 140–145.
Maheu-Cadotte, M.-A., Cossette, S., Dubé, V., Fontaine, G., Mailhot, T., Lavoie, P., Cournoyer, A., Balli, F., & Mathieu-Dupuis, G. (2018). Effectiveness of serious games and impact of design elements on engagement and educational outcomes in healthcare professionals and students: A systematic review and meta-analysis protocol. BMJ Open, 8(3).
Manzano-León, A., Camacho-Lazarraga, P., Guerrero, M. A., Guerrero-Puerta, L., Aguilar-Parra, J. M., Trigueros, R., & Alias, A. (2021). Between level up and game over: A systematic literature review of gamification in education. Sustainability, 13(4), 2247.
Migiro, S., & Magangi, B. (2011). Mixed methods: A review of literature and the future of the new research paradigm. African journal of bus. mgmt., 5(10), 3757–3764.
Oliveira, W., Hamari, J., Joaquim, S., Toda, A. M., Palomino, P. T., Vassileva, J., & Isotani, S. (2022). The effects of personalized gamification on students' flow experience, motivation, and enjoyment. Smart Learning Environments, 9(16), 14–26.
Saif, H., Fernandez, M., He, Y., & Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of twitter. Ninth International Conference on Language Resources and Evaluation.
Sanchez, D. R., Langer, M., & Kaur, R. (2020). Gamification in the classroom: Examining the impact of gamified quizzes on student learning. Computers & Education, 144, 103666. https://doi.org/https://doi.org/10.1016/j.compedu.2019.103666
Simpson, G. G. (1961). Principles of animal taxonomy. Columbia Univeristy Press.
Spanier, A. M., Weitl-Harms, S. K., & Hastings, J. D. (2021). A classification scheme for gamification in computer science education: Discovery of foundational gamification genres in data structures courses. 2021 IEEE Frontiers in Education Conference (FIE), 1–9.
Tashakkori, A., & Creswell, J. W. (2007). The new era of mixed methods.
Toda, A., Klock, A., Oliveira, W., Palomino, P., Rodrigues, L., L. ahd Shi, Bittencourt, I., Gasparini, I., Isotani, S., & Cristea, A. (2019). Analysing gamification elements in educational environments using an existing gamification taxonomy. Smart Learning Environments, 6(1), 1–14.
Tullis, T., & Stetson, J. (2006). A comparison of questionnaires for assessing website usability. Usability professional association conference. Vol. 1. 2004.
Veral, R., & Macıas, J. A. (2019). Supporting user-perceived usability benchmarking through a developed quantitative metric. International Journal of Human-Computer Studies, 122, 184–195. https://doi.org/https://doi.org/10.1016/j.ijhcs.2018.09.012