Game Over? A Review of Gamification in Information Retrieval

Alan Medlar
University of Helsinki
alan.j.medlar@helsinki.fi

Dorota Głowacka
University of Helsinki
dorota.glowacka@helsinki.fi

Abstract

We present a systematic literature review of gamification in information retrieval, identifying 32 peer-reviewed articles spanning a period from 2012 to present day. Our study highlights that, with the exception of crowdsourcing relevance judgments, very few studies have investigated the application of gamification to search systems. Moreover, in articles that we would strictly define as being related to gamification (i.e. not games with a purpose or serious games), a majority did not present any real-world systems, user studies or empirical results. In numerous other fields, gamification has been used to successfully increase user engagement and productivity. We argue that the potential of gamification has been overlooked in information retrieval and, in particular, in interactive search settings. We conclude by discussing experimental obstacles and the potential benefits of evaluating gamification strategies in future interactive search systems.

1 Introduction

Gamification refers to the incorporation of design elements commonly found in video games, such as points, badges and leaderboards, into non-game applications with the goal of increasing user engagement [Deterding et al., 2011]. Gamification increases engagement by promoting gameful experiences [Huotari and Hamari, 2017]. For example, in the language learning app DuoLingo[1], users can become more concerned with maintaining their streak (number of consecutive days of use) than learning the language[2]. Since around 2010, gamification has been studied in a wide variety of application domains [Koivisto and Hamari, 2019], including early work in information retrieval for crowdsourcing relevance judgments [Eickhoff et al., 2012]. Shortly afterwards, it was argued that the engagement potential of gamification should be investigated in several areas of information retrieval, including interactive information retrieval [Čudanov et al., 2014] and collaborative information seeking [Fernández-Luna et al., 2014], however, none of these research directions appear to have been developed further.

Recently, user engagement has become an important topic in interactive information retrieval, where it has been widely studied in relation to different search task characteristics, such as task

[1] https://www.duolingo.com/
[2] Personal experiences of the authors.

interest [Edwards and Kelly, 2016, 2017; O'Brien et al., 2020; Sinnamon et al., 2021] and task complexity [Wildemuth et al., 2014; O'Brien et al., 2020]. We argue that designing search systems that promote user engagement in low interest/high complexity search tasks will necessarily be based on extrinsic motivators, such as gamification. Unfortunately, there is currently no comprehensive overview of how such technologies have already been used in information retrieval.

In this article, we present the first systematic literature review of gamification in information retrieval. In our review, we identified 32 articles published from 2012 to present day. We analyzed bibliographic trends, study characteristics, application domains, gamification affordances and experimental outcomes. We conclude with a discussion focusing on the limited application of gamification outside of crowdsourced relevance judgments and the non-empirical nature of existing gamification research in information retrieval. Aside from a brief survey by Muntean and Nardini [2015], there are no other reviews of gamification in information retrieval or other search-related domains.

2 Background

2.1 Gamification

2.1.1 Definition

Gamification is defined as *"the use of game design elements in non-game contexts"* [Deterding et al., 2011]. These game design elements can be related to achievement (e.g. scores, badges, leaderboards), social connection (cooperation, teams, voting) and immersion (avatars, narrative) [Yee, 2006; Koivisto and Hamari, 2019]. Unlike video games, however, that are primarily designed for entertainment, gamified applications perform an instrumental function while affording gameful experiences [Huotari and Hamari, 2017]. For example, a fitness tracker might record different aspects of physical activity, but a gamified fitness tracker could reward you for reaching your daily exercise goal. While both applications serve the same instrumental purpose, the gamified version attempts to increase engagement in what might otherwise be an unfulfilling activity.

2.1.2 Domains and Affordances

Since being introduced around 2010, gamification has been applied to an exceptionally wide variety of domains. A recent systematic review, however, showed that nearly half of all empirical gamification research is related to education and learning [Koivisto and Hamari, 2019]. Indeed, only three domains: education [Bonde et al., 2014], health [Alahäivälä and Oinas-Kukkonen, 2016] and crowdsourcing [Eickhoff et al., 2012], account for 70% of empirical research in the field [Koivisto and Hamari, 2019]. This concentration in few domains is perhaps inevitable, as the promise of gamification is often said to be behavior change, where a given activity requires long-term commitment to achieve the desired results (see Outcomes, below).

A similar pattern can be seen in which gamification affordances[3] are implemented and studied:

[3]Affordances refer to *"the relationships between the properties of an object and the capabilities of an agent"* [Norman, 2013]. For the purposes of this review, however, we can simply think of affordances as "interface elements", though it is a more general design concept.

where, despite numerous possibilities, there is a heavy skew towards achievement tracking affordances, such as scores/experience points and badges [Koivisto and Hamari, 2019]. A majority of systems will feature multiple affordances, though studies tend not to separate out which affordance is responsible for a given outcome, only evaluating the gamified system as a whole [Hamari et al., 2014].

2.1.3 Outcomes

A large proportion of gamification studies only measure domain-specific outcomes, i.e. the impact on instrumental/productive work done by users. However, many also report on psychological and behavioral outcomes. Psychological outcomes tended to be related to perceptions of the system or user experience (e.g. perception of fun, engagement, affect, etc.). Despite the promise of many gamified systems being behavior change, few studies capture the longitudinal impact of gamification [Koivisto and Hamari, 2019]. Non-domain specific behaviors tend to be instantaneous behaviors, such as system usage, time taken, willingness to use and number of contributions, which are easier to collect, but do not capture the stated purpose of gamification.

2.1.4 Theory

Several theoretical frameworks have been used to explain the motivational potential of gamification, with the most widely cited being self-determination theory (SDT) [Landers et al., 2015]. SDT states that humans are driven by three basic psychological needs: autonomy, competence and relatedness. These needs determine our overall sense of well-being and how we address them is guided by sources of intrinsic and extrinsic motivation [Ryan and Deci, 2000]. Different gamification studies identify with different aspects of SDT, with some arguing that achievements and competition are intrinsically motivating (e.g. [Hamari et al., 2014]), while others view game design elements as a source of extrinsic motivation (e.g. [Mekler et al., 2013]). Other theoretical frameworks, such as operant conditioning and the theory of gamified learning, mostly emphasize the extrinsic motivational effects of gamification [Landers et al., 2015]. These latter approaches suggest that beneficial activities (e.g. exercise) should be rewarded, creating positive associations that ultimately lead to behavior change [Zichermann and Cunningham, 2011; Landers, 2014].

2.1.5 Related Topics

Two closely related topics to gamification are *games with a purpose* and *serious games*. While these terms are frequently used interchangeably with gamification (or referred to as simply "crowdsourcing" and "education" applications of gamification, respectively [Koivisto and Hamari, 2019]), neither involve non-game contexts nor aim to change user behavior. We therefore consider them separately in this review.

2.2 Games with a Purpose

Games with a purpose are where players perform useful work as a by-product of playing a game, e.g. in classification and optimization tasks [Von Ahn, 2006]. In the ESP game, the first game with a purpose, two players are presented with the same random image and both must guess what

label their partner will give to the image. Players are scored on the basis of how many image labels they agree on within a given time limit. The true goal of the ESP game, however, is to produce accurate image labels [Von Ahn and Dabbish, 2004]. Another, more complex example is Foldit, a puzzle game where players attempt to solve the protein folding problem and are scored on the Gibbs free energy of the tertiary structure of a given protein [Cooper et al., 2010]. Games with a purpose are often used together with crowdsourcing to produce data sets efficiently and cheaply. Outside of crowdsourcing, instead of paying participants, the incentive to participate is in playing the game.

2.3 Serious Games

Serious games are games that are played for reasons other than pure entertainment [Abt, 1987]. Serious games differ from games with a purpose in terms of who stands to benefit from the outcome: serious games are intended to benefit the player, whereas games with a purpose benefit whomever is running the game. Serious games are typically used in educational scenarios, including basic numeracy and literacy for those with learning disabilities [Lämsä et al., 2018] and to enhance medical education and surgical skills [Graafland et al., 2012]. Serious games are often used to communicate information in healthcare, in particular to children and adolescents, e.g. about smoking cession [Derksen et al., 2020] and living with chronic disease [Charlier et al., 2016]. Meta-analyses have highlighted numerous factors of successful serious games, such as feedback and adaptability, that contribute to enhanced learning experiences and outcomes [Ravyse et al., 2017].

2.4 Criticism

Previous systematic reviews have highlighted a lack of experimental studies and, where experimental studies exist, a tendency to consider the system holistically, instead of investigating the impact of individual affordances [Koivisto and Hamari, 2019]. Others have highlighted how gamification is used by employers to exploit workers by capturing *"play in the pursuit of neoliberal rationalization and the managerial optimization of working life and labour"* in order to *"adapt behaviour to capital"* [Woodcock and Johnson, 2018]. However, the systems studied in an academic context tend to be ones that are opted into, rather than being mandated by an employer.

3 Review Procedure

The goal of the study was to explore the research literature on how gamification is applied in information retrieval and other search-related domains. The selection of publications was conducted in 3 steps based on the QUORUM statement [Moher et al., 2000], that defines a procedure for meta-analyses. Figure 1 provides an overview of how papers were identified for inclusion in this study.

3.1 Step 1: Literature Search

3.1.1 Source Selection

We wanted to be as inclusive as possible, and include not only articles published in information retrieval venues, but related to the wider concept of information seeking as well. For this reason, we chose two scientific repositories to conduct literature searches: ACM Digital Library (ACMDL) and Web of Science (WOS, formerly Web of Knowledge). ACMDL provides access to almost 3 million articles related to computing through the ACM guide to computing literature. WOS indexes almost 80 million articles across all fields of science.

3.1.2 Search Terms

Following other systematic reviews on gamification, we used the search query "gamif*", but also required there to be another search term from the following list: "information retrieval", "information seeking", "search engine", "web search" and "retrieval". While "gamif*" does not explicitly reference games with a purpose or serious games, it tends to capture both as gamification is often used as an umbrella term for these related topics. Furthermore, our core interest is in gamification in a strict sense, i.e. *in non-game contexts*. We did not limit search results to any particular publication dates, venues or types of article.

3.1.3 Search Procedure

We searched ACMDL and WOS using our search terms. WOS did not allow us to limit search terms to bibliographic data (title, abstract, keywords), so we searched using "All fields". In ACMDL, we similarly searched "Anywhere" as we already needed to filter on bibliographic information in a later step. The search returned a total of 781 articles (ACMDL = 725, WOS = 56). Duplicates articles were removed, i.e. articles retrieved by both ACMDL and WOS (11), leaving 770 articles that were used in the next step.

3.2 Step 2: Selection Criteria

3.2.1 First Exclusion

All research papers from step 1 were exported from ACMDL and WOS into BibTex format. Articles without a title or abstract in its BibTex entry were removed (29). Next, we filtered out articles that did not contain the terms "gamif*" or "game*" in the title, abstract, keywords or publication venue (464). Finally, we filtered out articles that did not contain any of the IR-related search terms (i.e. "information retrieval", "information seeking", "search engine", "web search" and "retrieval") in the title, abstract, keywords or publication venue (195). These exclusion criteria left a total of 82 articles.

We inspected the excluded articles to understand why so many papers were filtered out. A majority of excluded articles did not contain the word "gamification" in the bibliographic data, but were found in search results because it was mentioned in, for example, future work.

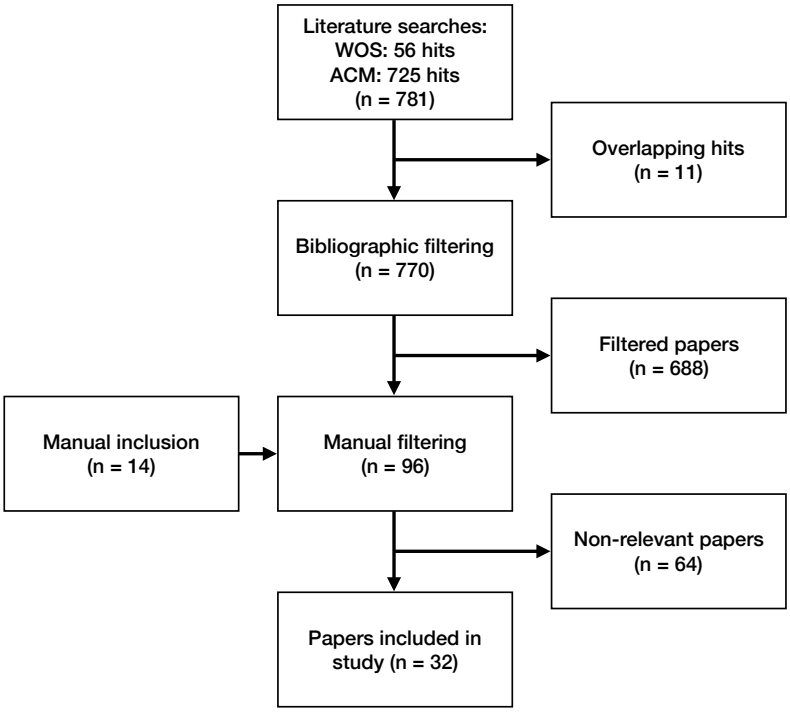

Figure 1: Flowchart summarizing the literature review process.

3.2.2 Manual Inclusion

During this step, we noticed that neither ACMDL nor WOS returned the papers from the GamifIR workshops from 2015 or 2016, so these 14 papers were included manually. This increased the total number of articles to 96 that were used in the next step.

3.3 Step 3: Validity Assessment

3.3.1 Screening Criteria

In this step, we wanted to narrow down the collection of articles to papers that: (i) are related to an area of gamification, (ii) are related to information retrieval or other search-related fields, (iii) are peer-reviewed articles, and (iv) are written in English. These criteria needed to be checked manually as the first exclusion step would (correctly) retain articles that mention "search engine", but also articles that mention "search engine optimization" (i.e. SEO, in the online marketing sense). Similarly, "retrieval" included papers that "information retrieval" missed, but also referred to studies related to human memory.

3.3.2 Second Exclusion

Based on the above criteria, we manually excluded a further 64 articles. The articles excluded during this step included extended abstracts, keynotes and workshop overviews (10). We additionally excluded several reviews/surveys, as these articles could directly reference many of the

articles included in this review (6). One article was not written in English. The remaining excluded articles were not relevant to this review either because they were unrelated to gamification or information retrieval (47). After this step, 32 articles remained in the collection.

3.3.3 Inter-rater Reliability

The above manual assessment was performed by the first author. To understand the reliability of this assessment, the second author independently assessed a subset of articles covering 30% of the papers that were input to the validity assessment step. Inter-rater reliability was assessed using the unweighted Cohen's Kappa coefficient, $\kappa = 0.926$ ($p = 8.88 \times 10^{-7}$). Kappa values greater than 0.9 are regarded as "almost perfect" [McHugh, 2012]. We therefore included all 32 articles in our final analysis (all papers included in this review are cited individually in Table 3).

4 Results

In the following analyses, we report on publication trends, study characteristics and which domains, affordances and outcomes feature in gamification studies.

4.1 Bibliographic Analysis

Table 1 shows a summary of the publication dates of all the articles included in the study. After a brief surge of activity during 2014–2016, coinciding with the Gamification in Information Retrieval (GamifIR) series of workshops, the number of publications decreased until 2020 when no studies were published. We cannot say whether this drop was due to a lack of interest in the field, or other difficulties, such as problems conducting user studies during the COVID-19 pandemic.

Table 2 breaks down the same data by publication type, showing that a majority of articles were published at workshops. Not all papers from the GamifIR workshop were included in this study as they were unrelated to search (the workshop also accepted papers related to, for example, machine learning [Di Nunzio et al., 2016]). While only a quarter of articles were full length papers, i.e. long conference or journal articles, they included papers published at both SIGIR [Eickhoff et al., 2012; Megorskaya et al., 2015] and ECIR [Gligorov et al., 2013], suggesting that ideas related to gamification received wide exposure in the information retrieval research community.

4.2 Study Characteristics

We categorized all papers included in this study along two dimensions: the type of study and the type of gamification. We defined three types of study:

Experimental study: Experimental studies were those where a system or approach was compared against a baseline or ground truth data set. For example, in Jin et al. [2016] the quality of relevance judgments were compared between multiple treatment and control groups.

Empirical study: Empirical studies were where a quantitative analysis was performed, but there was no explicit comparison to a baseline. For example, in Shmelev et al. [2016] a quiz to

Year of publication	Freq.
2012	2
2013	1
2014	10
2015	8
2016	7
2017	2
2018	1
2019	1
2020–2021	0

Table 1: Publication counts by year.

Type of publication	Freq.
Workshop papers	18
Conference papers (short)	5
Journal articles	4
Conference papers (long)	4
Book chapters	1

Table 2: Publication counts by publication type.

develop information-seeking skill was tested with a group of students, but not compared to a baseline.

No study: We categorized articles with no empirical findings as no study. For example, Fernández-Luna et al. [2014] speculates how gamification could be applied to collaborative information-seeking systems, but includes no system implementation nor user studies.

Following the definitions given in Section 2, we defined three categories of article:

Games with a purpose: Games with a purpose are designed to produce useful work (e.g. relevance judgments) as a by-product of users playing the game.

Serious games: Unlike games with a purpose, serious games do not produce data, but are intended to educate or inform players as they are playing the game.

Gamification: Gamified systems have an instrumental purpose other than entertainment (i.e. they are not games), but include game design elements to promote gameful experiences.

The outcome of this exercise is shown in Table 3. The most common kind of article is games with a purpose (18/32 articles), whereas both serious games and gamification had only 7 articles associated with each category. The most common type of study was different for each type of gamification: (i) games with a purpose were most likely to be experimental studies (10/18), (ii) serious games were most likely to be empirical studies (4/7) and (iii) gamification articles were most likely to have no study (6/7). No articles categorized as either serious games or gamification featured any experimental studies.

As noted in the table, there were several articles that used gamification affordances, but were categorized as serious games because they had no instrumental purpose other than education. If these 3 articles were instead categorized as gamification, it would only make the observed trends of which kind of study is most represented by each type of article more pronounced.

	Experimental studies	Empirical studies	No study
Games with a purpose	Eickhoff et al. [2012] Gligorov et al. [2013] Harris [2014] Lux et al. [2014] Brenner et al. [2014] Rosani et al. [2015] Megorskaya et al. [2015] Jin et al. [2016] Ganguly and Jones [2016] Harris [2017]	He et al. [2014] Schlötterer et al. [2015] Radu et al. [2015] Riegler et al. [2015] Xu et al. [2016] Pinto and Viana [2019]	Siebenlist and Knautz [2012] Dalton et al. [2018]
Serious games	None	Rybak et al. [2015] Moazzam et al. [2015] Karatassis and Fuhr [2016] Shmelev et al. [2016]	Azzopardi et al. [2014]* Wilhelm-Stein and Eibl [2015]* Karatassis [2017]*
Gamification	None	Barr et al. [2016]	Agoritsas et al. [2014] Galli et al. [2014] Fernández-Luna et al. [2014] Shovman [2014] Čudanov et al. [2014] Meder et al. [2016]

Table 3: All papers included in review categorized by the type of study (experimental, empirical or no study) and gamification category (games with a purpose, serious games or gamification). $*$ = included gamification affordances, but had no instrumental purpose other than education.

4.3 Sample sizes

We extracted sample sizes from the 20 articles that included user studies (of the 21 articles with experimental or empirical studies, 1 article was based on simulation). Where more than a single study was performed in a given article, we used the total number of users. Table 4 shows the wide range of sample sizes used in studies, ranging from 7-1750. The sample sizes for games with a purpose were generally higher than serious games, which can be attributed to the use of crowdsourcing platforms versus specific user groups, and because most of the studies of serious games appeared to be pilot studies. The one empirical gamification study had the largest sample size of all from an observational study of library users [Barr et al., 2016], but this is not representative of gamification studies in general [Koivisto and Hamari, 2019].

Category	Sample sizes
Games with a purpose (experimental)	10, 10, 82, 83, 96, 274, > 300, 370, 795
Games with a purpose (empirical)	7, 8, 26, 28, 118, 352
Serious games (empirical)	10, 10, 15, 116
Gamification (empirical)	1750

Table 4: Sample size by study category for papers that involved user studies.

4.4 Domains

Table 5 summarizes the domains identified as the broad focus of each article included in this review. In cases where multiple domains applied, we chose what we deemed to be the most appropriate single domain. For example, web search literacy was counted as *education* and not *web search* [Karatassis, 2017]. The *specialist search* domain included scientific literature, library catalogs, video and enterprise search. The most popular domain is relevance judgments/tagging due to the high proportion of games with a purpose identified in this study. As in other reviews, we identified education as an important domain as well [Koivisto and Hamari, 2019]. Taken together, relevance judgments and education account for over half of the articles. The examples with the lowest counts, such as collaborative information seeking and interactive information retrieval, were speculative descriptions of how gamification could be applied to those fields [Galli et al., 2014; Čudanov et al., 2014; Fernández-Luna et al., 2014].

4.5 Affordances

Despite the low number of articles categorized as gamification (of which all but one contained no study), we extracted all affordances that were mentioned to understand their relative importance in information retrieval. Table 6 shows which affordances were mentioned in how many articles. The number of affordances mentioned in each article ranged from 2-10. As shown in more general reviews, the most popular affordances are badges, leaderboards and points [Koivisto and Hamari, 2019]. A majority of these articles were speculative (6/7 articles), and it is therefore unknown which affordances are better suited to the particular search tasks and problems associated with information retrieval.

4.6 Outcomes

Only Barr et al. [2016] reported an empirical study related to gamification (see Table 3). The stated goal of this paper was to increase engagement with the library, however, the only outcomes reported were related to system usage: the number of badges awarded and the aggregate number of points awarded to each college within a university. These outcomes were not compared to a control group and we, therefore, do not know whether gamification actually impacted user engagement. Similarly, the comparisons between colleges were not made in reference to a baseline, so it is unclear whether the differences are attributable to gamification or not.

Domain	Freq.
Relevance judgments/tagging	11
Education	7
Specialist search	5
Web search	2
Other crowdsourcing	2
Information seeking	2
General IR	1
Interactive IR	1
Collaborative information seeking	1

Table 5: Publication counts by main application domain.

Affordance	Freq.
Badges, achievements	9
Leaderboards	9
Points	8
Customization, avatar	5
Levels	3
Progress bars	3
Virtual goods	3
Feedback messages	3
User activities stream	1
Group achievements	1

Table 6: Affordances mentioned in articles categorized as gamification and serious games that used gamification (see Table 3).

5 Discussion

This systematic review shows that the IR research community has generally viewed gamification as a tool to collect gold standard data sets, rather than increase user engagement in search. Indeed, 11/32 papers focused on the gamified crowdsourcing of relevance judgments and tags (Table 5). While this review included articles from core IR conference venues, a majority of papers (17/32) came from the GamifIR series of workshops, including 5/7 articles related to gamification in a strict sense, i.e. using game design elements in *non-game* contexts. These articles argued that gamification should be investigated in fundamental areas of information retrieval, such as web search [Shovman, 2014], interactive information retrieval [Čudanov et al., 2014] and collaborative information seeking [Fernández-Luna et al., 2014], but to date none of these ideas have been followed through on. Indeed, this finding appears to be part of a broader pattern: games with a purpose tend to be evaluated experimentally, serious games empirically (i.e. no clear baseline or comparisons) and gamified applications tend not to feature any study at all (Table 3). We speculate that these differences are related to the relative difficulties of evaluation in each category, with games with a purpose being the easiest (e.g. using crowdsourcing platforms like Amazon Mechanical Turk) and gamification being the hardest (requiring between-subject or longitudinal user studies). Papers related to education, i.e. serious games, did not feature experimental studies either, however, this has previously been observed in reviews of computer science education research in general [Valentine, 2004; Lishinski et al., 2016].

Despite the limited popularity of gamification in information retrieval (Table 1), the number of gamification studies published in other fields continues to grow each year [Koivisto and Hamari, 2019]. This lack of interest could, therefore, be due to real or perceived obstacles specific to information retrieval. Evaluations of gamified search systems will struggle with similar issues to those identified in the Interactive [Dumais and Belkin, 2005] and HARD tracks [Allan, 2005] at

TREC. In particular, that interactions between users, tasks and systems can bias experiments and affect reproducibility [Lagergren and Over, 1998]. The study of gamification would be especially challenging as any increase in user engagement will vary between users (i.e. users may respond differently from one another, with some feeling encouraged by gamification and others not) and for different levels of task interest and complexity (even if gamification increases user engagement, there may be little to gain if task interest is already high). The potential for high variability is hinted at by the increasing sample sizes used in gamification studies, with Koivisto and Hamari [2019] reporting an average sample size of 74.5 study participants compared to an earlier review that found the average to be 20. However, evaluating the impact of gamification in interactive information retrieval has the advantage that the experimental infrastructure already exists to (i) control task interest and task complexity [Kelly et al., 2015; Capra et al., 2017], and (ii) to distinguish between different task types, such as lookup and exploratory search [Athukorala et al., 2016; Medlar et al., 2017]. These techniques would allow for more controlled experiments than are possible in other domains and permit us to target affordances to specific search behaviors.

There are inherent limitations to systematic literature reviews due to their dependence on search terms and human judgment. In our study, we aimed to be inclusive by searching multiple scientific repositories with multiple search terms synonymous with information retrieval. Despite the possibility of subjectively applying exclusion criteria, the inter-rater reliability was very high. Lastly, our study underestimates the extent to which games with a purpose and serious games have featured in information retrieval research. For example, games with a purpose have been used to improve image search [Ma et al., 2009] and investigate search behavior [Purvis and Azzopardi, 2012], but neither of these papers were included in this study as they did not explicitly use the term "gamification" or "gamified". As we were primarily motivated by understanding gamification in non-game contexts, however, we did not seek to expand the scope of the literature search.

Acknowledgements

We would like to thank Yang Liu for insightful comments on an earlier draft of this manuscript.

References

Clark C Abt. *Serious games*. University press of America, 1987.

Thomas Agoritsas, Emma Iserman, Nicholas Hobson, Natasha Cohen, Adam Cohen, Pavel S Roshanov, Miguel Perez, Chris Cotoi, Rick Parrish, Eleanor Pullenayegum, et al. Increasing the quantity and quality of searching for current best evidence to answer clinical questions: protocol and intervention design of the macplus fs factorial randomized controlled trials. *Implementation Science*, 9(1):1–16, 2014.

Tuomas Alahäivälä and Harri Oinas-Kukkonen. Understanding persuasion contexts in health gamification: A systematic analysis of gamified health behavior change support systems literature. *International journal of medical informatics*, 96:62–70, 2016.

James Allan. Hard track overview in TREC 2003 high accuracy retrieval from documents. Technical report, 2005.

Kumaripaba Athukorala, Alan Medlar, Antti Oulasvirta, Giulio Jacucci, and Dorota Glowacka. Beyond relevance: Adapting exploration/exploitation in information retrieval. In *Proceedings of the 21st international conference on intelligent user interfaces*, pages 359–369, 2016.

Leif Azzopardi, Martin Bevc, Andrew Gardner, David Maxwell, and Abdullah Razzouk. Pagefetch 2: Gamification the sequel. In *Proceedings of the First International Workshop on Gamification for Information Retrieval (GamifIR)*, pages 38–41, 2014.

Matthew Barr, Kay Munro, and Frank Hopfgartner. Increasing engagement with the library via gamification. In *Proceedings of the Third International Workshop on Gamification for Information Retrieval (GamifIR)*, 2016.

Mads T Bonde, Guido Makransky, Jakob Wandall, Mette V Larsen, Mikkel Morsing, Hanne Jarmer, and Morten OA Sommer. Improving biotech education through gamified laboratory simulations. *Nature biotechnology*, 32(7):694–697, 2014.

Markus Brenner, Navid Mirza, and Ebroul Izquierdo. People recognition using gamified ambiguous feedback. In *Proceedings of the First International Workshop on Gamification for Information Retrieval (GamifIR)*, pages 22–26, 2014.

Rob Capra, Jaime Arguello, and Yinglong Zhang. The effects of search task determinability on search behavior. In *European Conference on Information Retrieval*, pages 108–121. Springer, 2017.

Nathalie Charlier, Nele Zupancic, Steffen Fieuws, Kris Denhaerynck, Bieke Zaman, and Philip Moons. Serious games for improving knowledge and self-management in young people with chronic conditions: a systematic review and meta-analysis. *Journal of the American Medical Informatics Association*, 23(1):230–239, 2016.

Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, et al. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307):756–760, 2010.

Mladen Čudanov, Diana Parlić, and Adam Sofronijević. Proposed framework for gamifying information retrieval: case of dart-european research theses portal. In *Proceedings of the Second International Conference on Technological Ecosystems for Enhancing Multiculturality*, pages 185–190, 2014.

Jeffrey Dalton, Victor Ajayi, and Richard Main. Vote goat: Conversational movie recommendation. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1285–1288, 2018.

ME Derksen, S van Strijp, AE Kunst, JG Daams, Monique WM Jaspers, and MP Fransen. Serious games for smoking prevention and cessation: A systematic review of game elements and game effects. *Journal of the American Medical Informatics Association*, 27(5):818–833, 2020.

Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. From game design elements to gamefulness: defining" gamification". In *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, pages 9–15, 2011.

Giorgio Maria Di Nunzio, Maria Maistro, and Daniel Zilio. Gamification for machine learning: The classification game. In *GamifIR@ SIGIR*, 2016.

Susan T Dumais and Nicholas J Belkin. The trec interactive tracks: Putting the user into search. *TREC: Experiment and evaluation in information retrieval*, pages 123–152, 2005.

Ashlee Edwards and Diane Kelly. How does interest in a work task impact search behavior and engagement? In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pages 249–252, 2016.

Ashlee Edwards and Diane Kelly. Engaged or frustrated? disambiguating emotional state in search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 125–134, 2017.

Carsten Eickhoff, Christopher G Harris, Arjen P de Vries, and Padmini Srinivasan. Quality through flow and immersion: gamifying crowdsourced relevance assessments. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 871–880, 2012.

Juan M Fernández-Luna, Juan F Huete, Humberto Rodríguez-Avila, and Julio C Rodríguez-Cano. Enhancing collaborative search systems engagement through gamification. In *Proceedings of the First International Workshop on Gamification for Information Retrieval (GamifIR)*, pages 42–45, 2014.

Luca Galli, Piero Fraternali, and Alessandro Bozzon. On the application of game mechanics in information retrieval. In *Proceedings of the First International Workshop on Gamification for Information Retrieval (GamifIR)*, pages 7–11, 2014.

Debasis Ganguly and Gareth JF Jones. A gamified approach to relevance judgement. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 214–220. Springer, 2016.

Riste Gligorov, Michiel Hildebrand, Jacco Van Ossenbruggen, Lora Aroyo, and Guus Schreiber. An evaluation of labelling-game data for video retrieval. In *European Conference on Information Retrieval*, pages 50–61. Springer, 2013.

Maurits Graafland, Jan M Schraagen, and Marlies P Schijven. Systematic review of serious games for medical education and surgical skills training. *Journal of British Surgery*, 99(10):1322–1330, 2012.

Juho Hamari, Jonna Koivisto, and Harri Sarsa. Does gamification work?–a literature review of empirical studies on gamification. In *2014 47th Hawaii international conference on system sciences*, pages 3025–3034. Ieee, 2014.

Christopher G Harris. The beauty contest revisited: Measuring consensus rankings of relevance using a game. In *Proceedings of the First International Workshop on Gamification for Information Retrieval (GamifIR)*, pages 17–21, 2014.

Christopher G Harris. Finding the best job applicants for a job posting: A comparison of human resources search strategies. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 189–194. IEEE, 2017.

Jiyin He, Marc Bron, Leif Azzopardi, and Arjen de Vries. Studying user browsing behavior through gamified search tasks. In *Proceedings of the First International Workshop on Gamification for Information Retrieval (GamifIR)*, pages 49–52, 2014.

Kai Huotari and Juho Hamari. A definition for gamification: anchoring gamification in the service marketing literature. *Electronic Markets*, 27(1):21–31, 2017.

Yuan Jin, Mark James Carman, and Lexing Xie. A little competition never hurt anyone's relevance assessments. In *Proceedings of the Third International Workshop on Gamification for Information Retrieval (GamifIR)*, 2016.

Ioannis Karatassis. Websail: Computer-based methods for enhancing web search literacy. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*, pages 403–405, 2017.

Ioannis Karatassis and Norbert Fuhr. Gamification for websail. In *Proceedings of the Third International Workshop on Gamification for Information Retrieval (GamifIR)*, 2016.

Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-ching Wu. Development and evaluation of search tasks for iir experiments using a cognitive complexity framework. In *Proceedings of the 2015 international conference on the theory of information retrieval*, pages 101–110, 2015.

Jonna Koivisto and Juho Hamari. The rise of motivational information systems: A review of gamification research. *International Journal of Information Management*, 45:191–210, 2019.

Eric Lagergren and Paul Over. Comparing interactive information retrieval systems across sites: The trec-6 interactive track matrix experiment. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 164–172, 1998.

Joni Lämsä, Raija Hämäläinen, Mikko Aro, Raine Koskimaa, and Sanna-Mari Äyrämö. Games for enhancing basic reading and maths skills: A systematic review of educational game design in supporting learning by people with learning disabilities. *British Journal of Educational Technology*, 49(4):596–607, 2018.

Richard N Landers. Developing a theory of gamified learning: Linking serious games and gamification of learning. *Simulation and gaming*, 45(6):752–768, 2014.

Richard N Landers, Kristina N Bauer, Rachel C Callan, and Michael B Armstrong. Psychological theory and the gamification of learning. In *Gamification in education and business*, pages 165–186. Springer, 2015.

Alex Lishinski, Jon Good, Phil Sands, and Aman Yadav. Methodological rigor and theoretical foundations of cs education research. In *Proceedings of the 2016 ACM conference on international computing education research*, pages 161–169, 2016.

Mathias Lux, Mario Guggenberger, and Michael Riegler. Picturesort: gamification of image ranking. In *Proceedings of the First International Workshop on Gamification for Information Retrieval (GamifIR)*, pages 57–60, 2014.

Hao Ma, Raman Chandrasekar, Chris Quirk, and Abhishek Gupta. Improving search engines using human computation games. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 275–284, 2009.

Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

Michael Meder, Till Plumbaum, and Sahin Albayrak. Learning gamification design-an usability first approach for the enterprise infoboard experiment. In *Proceedings of the Third International Workshop on Gamification for Information Retrieval (GamifIR)*, 2016.

Alan Medlar, Joel Pyykkö, and Dorota Glowacka. Towards fine-grained adaptation of exploration/exploitation in information retrieval. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 623–627, 2017.

Olga Megorskaya, Vladimir Kukushkin, and Pavel Serdyukov. On the relation between assessor's agreement and accuracy in gamified relevance assessment. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 605–614, 2015.

Elisa D Mekler, Florian Brühlmann, Klaus Opwis, and Alexandre N Tuch. Do points, levels and leaderboards harm intrinsic motivation? an empirical analysis of common gamification elements. In *Proceedings of the First International Conference on gameful design, research, and applications*, pages 66–73, 2013.

Waqas Moazzam, Michael Riegler, Sagar Sen, and Mari Nygård. Scientific hangman: Gamifying scientific evidence for general public. In *Proceedings of the Second International Workshop on Gamification for Information Retrieval (GamifIR)*, pages 26–33, 2015.

David Moher, Deborah J Cook, Susan Eastwood, Ingram Olkin, Drummond Rennie, and Donna F Stroup. Improving the quality of reports of meta-analyses of randomised controlled trials: the quorom statement. *Oncology Research and Treatment*, 23(6):597–602, 2000.

Cristina Ioana Muntean and Franco Maria Nardini. Gamification in information retrieval: State of the art, challenges and opportunities. In *IIR*, 2015.

Don Norman. *The design of everyday things: Revised and expanded edition.* Basic books, 2013.

Heather L O'Brien, Jaime Arguello, and Rob Capra. An empirical study of interest, task complexity, and search behaviour on user engagement. *Information Processing and Management*, 57(3):102226, 2020.

José Pedro Pinto and Paula Viana. Improving youtube video retrieval by integrating crowdsourced timed metadata. *Journal of Intelligent and Fuzzy Systems*, 37(6):7207–7221, 2019.

James Purvis and Leif Azzopardi. A preliminary study using pagefetch to examine the searching ability of children and adults. In *IIiX*, pages 262–265. Citeseer, 2012.

Anca-Livia Radu, Aliaksandr Autayeu, Bogdan Ionescu, and Fausto Giunchiglia. High quality photo collection via gamification. In *Proceedings of the Second International Workshop on Gamification for Information Retrieval (GamifIR)*, pages 19–25, 2015.

Werner Siegfried Ravyse, A Seugnet Blignaut, Verona Leendertz, and Alex Woolner. Success factors for serious games to enhance learning: A systematic review. *Virtual Reality*, 21(1): 31–58, 2017.

Michael Riegler, Ragnhild Eg, Lilian Calvet, Mathias Lux, Pål Halvorsen, and Carsten Griwodz. Playing around the eye tracker-a serious game based dataset. In *Proceedings of the Second International Workshop on Gamification for Information Retrieval (GamifIR)*, pages 34–40, 2015.

Andrea Rosani, Giulia Boato, and Francesco GB De Natale. Eventmask: A game-based framework for event-saliency identification in images. *IEEE Transactions on Multimedia*, 17(8):1359–1371, 2015.

Richard M Ryan and Edward L Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist*, 55(1):68, 2000.

Jan Rybak, Krisztian Balog, and Kjetil Nørvåg. Ir game: How well do you know information retrieval papers? In *Proceedings of the Second International Workshop on Gamification for Information Retrieval (GamifIR)*, pages 7–12, 2015.

Jörg Schlötterer, Christin Seifert, Lisa Wagner, and Michael Granitzer. A game with a purpose to access europe's cultural treasure. In *Proceedings of the Second International Workshop on Gamification for Information Retrieval (GamifIR)*, pages 13–18, 2015.

Vadim Shmelev, Maria Karpova, Nikita Kogtikov, and Alexey Dukhanov. Students' development of information-seeking skills in a computer-aided quest. In *2016 IEEE Frontiers in Education Conference (FIE)*, pages 1–4. IEEE, 2016.

Mark Shovman. The game of search: What is the fun in that? In *Proceedings of the First International Workshop on Gamification for Information Retrieval (GamifIR)*, pages 46–48, 2014.

Tobias Siebenlist and Kathrin Knautz. The critical role of the cold-start problem and incentive systems in emotional web 2.0 services. In *Indexing and Retrieval of Non-Text Information*, pages 376–405. De Gruyter Saur, 2012.

Luanne Sinnamon, Limor Tamim, Samuel Dodson, and Heather L O'Brien. Rethinking interest in studies of interactive information retrieval. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pages 39–49, 2021.

David W Valentine. CS educational research: A meta-analysis of SIGCSE technical symposium proceedings. *ACM SIGCSE Bulletin*, 36(1):255–259, 2004.

Luis Von Ahn. Games with a purpose. *Computer*, 39(6):92–94, 2006.

Luis Von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, 2004.

Barbara Wildemuth, Luanne Freund, and Elaine G Toms. Untangling search task complexity and difficulty in the context of interactive information retrieval studies. *Journal of Documentation*, 2014.

Thomas Wilhelm-Stein and Maximilian Eibl. Teaching the IR process using real experiments supported by game mechanics. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 312–317. Springer, 2015.

Jamie Woodcock and Mark R Johnson. Gamification: What it is, and how to fight it. *The Sociological Review*, 66(3):542–558, 2018.

Feifei Xu, Feng Tian, Dimitrios Buhalis, Jessika Weber, and Hongmei Zhang. Tourists as mobile gamers: Gamification for tourism marketing. *Journal of Travel and Tourism Marketing*, 33(8): 1124–1142, 2016.

Nick Yee. Motivations for play in online games. *CyberPsychology and behavior*, 9(6):772–775, 2006.

Gabe Zichermann and Christopher Cunningham. *Gamification by design: Implementing game mechanics in web and mobile apps.* " O'Reilly Media, Inc.", 2011.