Measuring how motivation affects information quality assessment: a gamification approach

Marko Poženel, Aljaž Zrnec and Dejan Lavbič

Marko Poženel, Aljaž Zrnec and **Dejan Lavbič**. 2022. **Measuring how motivation affects information quality assessment: a gamification approach**, PLOS ONE, 17(10).

Abstract

Purpose – Existing research on the measurability of information quality (IQ) has delivered poor results and demonstrated low inter-rater agreement measured by Intra-Class Correlation (ICC) in evaluating IQ dimensions. Low ICC could result in a questionable interpretation of IQ. The purpose of this paper is to analyse whether assessors' motivation can facilitate ICC.

Methodology – To acquire the participants' views of IQ, we designed a survey as a gamified process. Additionally, we selected Web study to reach a broader audience. We increased the validity of the research by including a diverse set of participants (i.e. individuals with different education, demographic and social backgrounds).

Findings – The study results indicate that motivation improved the ICC of IQ on average by 0.27, demonstrating an increase in measurability from poor (0.29) to moderate (0.56). The results reveal a positive correlation between motivation level and ICC, with a significant overall increase in ICC relative to previous studies. The research also identified trends in ICC for different dimensions of IQ with the best results achieved for completeness and accuracy.

Practical implications – The work has important practical implications for future IQ research and suggests valuable guidelines. The results of this study imply that considering raters' motivation improves the measurability of IQ substantially.

Originality – Previous studies addressed ICC in IQ dimension evaluation. However, assessors' motivation has been neglected. This study investigates the impact of assessors' motivation on the measurability of IQ. Compared to the results in related work, the level of agreement achieved with the most motivated group of participants was superior.

Keywords

information quality; motivation; information quality assessment; inter-rater agreement; task-at-hand; gameful design

1 Introduction

Making the best possible decisions requires information of the highest quality. As the amount of information available grows, it becomes increasingly difficult to distinguish quality from questionable information (Arazy, Kopak, and Hadar 2017). The problem of poor information quality can weaken our decision processes, so we need more reliable measures and new techniques to assess the quality of information (Arazy and Kopak 2011; Ji-Chuan and Bing 2016; Fidler and Lavbič 2017). Unfortunately, such assessment can itself be very demanding (Arazy and Kopak 2011; Arazy, Kopak, and Hadar 2017; Yaari, Baruchson-Arbib, and Bar-Ilan 2011).

In general, the term *information quality (IQ)* represents the value of information for a given usage. However, IQ often refers to people's subjective judgment of the goodness and usefulness of information in certain information use settings (Yaari, Baruchson-Arbib, and Bar-Ilan 2011; Michnik and Lo 2009). The literature has widely adopted a multidimensional view of IQ (West and Williamson 2009; Arazy, Kopak, and Hadar 2017) to support more effortless management of its complexity.

The measurability of IQ has gained substantial attention in recent years (Yaari, Baruchson-Arbib, and Bar-Ilan 2011; Lian et al. 2018; Zhu et al. 2020). Most research in this field has been limited to measuring the quality of structured data (e.g. data in databases where a scheme is defined in advance) (Helfert 2001; Tilly et al. 2017; Madhikermi et al. 2016). Measuring the IQ of unstructured data (e.g. Wikipedia articles) requires different approaches that include interdisciplinary components (Batini et al. 2009). The research community proposed several determinants of IQ and there is a growing concern regarding how to best identify quality information (Arazy, Kopak, and Hadar 2017). Only a few studies presented inter-rater agreement results using Interclass Correlation Coefficient (ICC) statistics, and multiple guidelines for the interpretation of ICC inter-rater agreement values exist (Landis and Koch 1977; Cicchetti 1994; Koo and Li 2016). Regardless of the ICC interpretation used, the values reported in recent studies are poor or at best moderate (Fidler and Lavbič 2017; Arazy and Kopak 2011). This demonstrates that reaching consensus among various raters is difficult when measuring IQ.

Research problems regarding efficient IQ measurement remain relatively underexplored. Previous papers studied some of the cues that affect IQ assessment on selected sources of data (Arazy and Kopak 2011; Arazy, Kopak, and Hadar 2017). However, the research community needs additional case studies to evaluate the inter-rater reliability of IQ dimensions (a single aspect of data that can be measured and improved) in various settings to help increase the external validity of the cues and factors.

Being motivated means having an incentive to do something (Ryan and Deci 2000a). Intrinsically motivated does something for its own sake, for the sheer enjoyment of a task, while extrinsically motivatied does something in order to attain some external goal or meet some externally imposed constraint (Hennessey et al. 2015). To the best of our knowledge, no previous research has investigated how motivation affects IQ assessment and whether it has a significant impact on inter-rater agreement. In this paper, we propose a new approach that improves the measurability of IQ by considering various IQ dimensions. Specifically, we study the effect of motivation on IQ measurement and inter-rater reliability. Researchers have always seen motivation as an important factor that influences learning performance (Mamaril et al. 2013; DePasque and Tricomi 2015; Tokan and Imakulata 2019). Our goal in this work is to corroborate that motivation also affects the measurability of IQ.

In related work, Arazy and Kopak (2011) studied the measurability of IQ in Wikipedia articles, and Fidler and Lavbič (2017) narrowed the object of a study to individual paragraphs. In this work, we evaluate IQ of hints that (i) correspond to selected IQ dimensions, (ii) have diversified predefined quality, and (iii) help participants in progressing through the gamified process. Specifically, we evaluate the relevance of gamified task hints targeting IQ dimensions of accuracy, objectivity, completeness, and representation. We are interested in the consistency between multiple raters assessing the same set of hints in a hands-on assignment.

This study contributes to the existing literature concerning IQ measurability and inter-rater reliability. It extends the work presented in Arazy and Kopak (2011) and Arazy, Kopak, and Hadar (2017). To support comparison, we use the categorization of IQ dimensions defined by Lee et al. (2002), previously used in similar studies (Arazy and Kopak 2011; Fidler and Lavbič 2017; Arazy, Kopak, and Hadar 2017).

The remainder of the paper is structured as follows. In section 2, we review related work and introduce the problem statement and our proposed solution. We follow this with presentation of the empirical study design and the experiment in section 3. Then we present the results and discuss the implications and limitations in section 4. Finally, in section 5, we present our conclusions, limitations and suggest directions for future work.

2 Related work

2.1 Assessing the quality of information

With the growing amount of information published every day, IQ has gained huge social importance (Tilly et al. 2017; Zha et al. 2018; Dedeoglu 2019; Danniswara, Sandhyaduhita, and Munajat 2020). Several studies stressed the increase in interest in IQ (Arazy and Kopak 2011; Yaari, Baruchson-Arbib, and Bar-Ilan 2011; Fehrenbacher 2016). This body of research has often focused on dimensions of IQ and what factors affect its measurability (Arazy and Kopak 2011; Arazy, Kopak, and Hadar 2017). Fewer studies focused on the measurability and assessment of IQ. However, issues with IQ are becoming growingly prevalent (Lee et al. 2002), especially with the rise of user-generated content (e.g. Wikipedia) and citizen science, where users participate in simultaneously creating and editing information. Poor-quality user-generated content (UGC) can present an issue for information retrieval services (Figueiredo et al. 2013) and individuals.

The community considers IQ assessment demanding because sources of information lack metadata and IQ criteria are often subjective (Yaari, Baruchson-Arbib, and Bar-Ilan 2011); which makes it hard for multiple raters to agree upon an object's IQ (Arazy and Kopak 2011). Assessment of an object's (Wikipedia article, paragraph, hint, etc.) IQ depends on several factors, including object itself, and the assessor's prior knowledge, differences in domain expertise, cognitive or demographic traits. Previous research that has studied IQ assessment agrees that IQ is not a uniform construct and that it consists of multiple dimensions (Lee et al. 2002; Arazy and Kopak 2011; Arazy, Kopak, and Hadar 2017). Thus, we cannot assess IQ as a whole but according to its underlying dimensions. The research community has proposed several frameworks and underlying dimensions for the assessment of IQ. Richard Y. Wang and Strong (1996) defined a set of dimensions and a framework where dimensions are grouped into a hierarchical model of IQ aspects and their criteria. Several authors (Lee et al. 2002; Ballou and Pazer 2003; Richard Y. Wang et al. 2002) later investigated the initial set of dimensions defined by Richard Y. Wang and Strong (1996) and evaluated the degree to which individual dimensions comply with the needs or expectations of users (Fehrenbacher 2016). In this paper, we apply the set of quality dimensions (accuracy, completeness, objectivity and representation) that researchers used in most previous empirical studies (West and Williamson 2009; Arazy, Kopak, and Hadar 2017).

The dimensions are defined as follows: Accuracy indicates factual correctness of the data and absence of errors (incorrect information, references to non-authoritative sources, and spelling errors); Completeness refers to sufficient coverage of information appropriate for an encyclopedic entry and to the lack of omission of relevant facts (e.g., missing introductory and background information that would help explain the topic's relevance, importance, or its history); Objectivity pertains to an impartial view of the topic and to the absence of subjective language, opinions stated as facts, the omission of alternative perspectives or existing controversies, or a deliberate misrepresentation; and Representation refers to clarity and ease of understanding at a readership level accessible to the general public (using diagrams when required), rational organization, consistent presentation using a single "voice", and concise formatting.

Arazy and Kopak (2011) focused on IQ dimensions and the extent of agreement (i.e. inter-rater reliability) that could be achieved when rating the aforementioned four IQ dimensions. They found that some IQ dimensions are more difficult to assess than others and noted that assessors often employ heuristics during IQ assessment. Arazy, Kopak, and Hadar (2017) explored the role of heuristic principles in IQ assessment, investigating how the consistent application of heuristic principles affects inter-rater agreement according to IQ dimensions. Fehrenbacher (2016) investigated the effects of satisfaction and complexity on the perception of IQ dimensions and found that satisfied users place a higher weight on qualitative than quantitative aspects of IQ. Metzger and Flanagin (2013) focused on cognitive heuristics in credibility evaluation, studying the heuristics that information consumers use when deciding what sources and information to trust online.

Several research efforts sought to assess the IQ of content in a collaborative-writing environment, UGC, and citizen science. Yaari, Baruchson-Arbib, and Bar-Ilan (2011) examined how non-expert users evaluate the quality of Hebrew Wikipedia contents with a focus on identifying the cues and criteria that users find helpful to assess the quality of Wikipedia articles. Lukyanenko, Wiggins, and Rosser (2019) proposed data derived from UGC and citizen science be used for studying innovative approaches to IQ management. Fidler and Lavbič (2017) provide informed insights on students' perception of IQ. They proposed an approach to improve the relevance of Wikipedia articles to meet students' needs.

All of these studies highlighted issues with IQ assessment. With these issues in mind, our study focuses on factors that may have positive effects on reliable measurement and facilitate the assessment of IQ.

2.2 Gamification and motivation

In the previous work that studies the measurability of IQ, authors achieved mediocre results (Arazy and Kopak 2011; Arazy, Kopak, and Hadar 2017). However, measuring IQ depends on elusive factors and presents a challenging task (Arazy and Kopak 2011). The measures for evaluating IQ depend on the source and also the criteria may not be viewed equally by the users and researchers (Arazy and Kopak 2011). Assessment of quality depends on the "fitness" of the data to one's specific assessment purposes (Arazy, Kopak, and Hadar 2017). We assume that the assessors' motivation can also be a key factor for quality assessment of IQ. Gilakjani, Lai-Mei, and Sabouri (2012) showed that motivation is a crucial success factor, especially in learning. W. T. Wang and Hou (2015) studied the influence of various types of motivations on employees' knowledge sharing behaviors and found that hard reward is a key motivational factor next to soft reward. Motivated assessors may also contribute to a higher inter-rater reliability. To improve assessors' motivation, we introduce the concept of gamification in the assessment process.

Gamification (Gameful design) (Pelling 2011) is a concept where we use game-like elements in various systems to increase user participation, motivation, improve engagement, or to retain users continue using the system. In the literature, gamification is often defined as the use of game design elements in non-game contexts (Deterding et al. 2011). It is an innovative approach to stimulate motivation (Sailer et al. 2014). Motivation is hardly unitary phenomenon (Ryan and Deci 2000a), and can be studied from different perspectives. In Self-Determination Theory (SDT) (Deci and Ryan 1985), we distinguish between different types of motivation based on the different reasons or goals that initiated an action (Ryan and Deci 2000a). The most basic distinction is between intrinsic motivation and extrinsic motivation (Ryan and Deci 2000a). Intrinsic motivation is defined as the wish or tendency to execute an action for its own sake, for example because of its interesting, challenging or exiting nature (Deci and Ryan 1985). It enables high-quality learning and creativity (Ryan and Deci 2000a). Extrinsic motivation contrasts with intrinsic motivation, and refers to the pursuit of an instrumental goal, i.e to achieve results that are not related to action performed (Ryan and Deci 2000a; Reiss 2012). Addressing an individual's intrinsic motivation to play and have fun, we can also define gamification as the concept of leveraging the psychological predisposition to engage in gaming, using mechanisms that game designers applied in making video games, as a potential means to make real-world activities more engaging (Kappen and Nacke 2013). Gamification proved to be successful in addressing individuals' motivation and increasing the user's engagement. SDT assumes three universal psychological needs for competence, autonomy, and social relatedness (Ryan and Deci 2000b). The fulfilment of these needs is especially relevant for fostering intrinsic motivation (Sailer et al. 2014). Also the integration of extrinsic motivation can be addresses by fulfilling these needs as well (Deci and Ryan 2000; Sailer et al. 2014). According to SDT, players are likely to be motivated if they experience the feeling of competence, authonomy and social relatedness (Sailer et al. 2014). In their study of gamification in the workplace, Mitchell, Schuster, and Jin (2020) found that if extrinsic motivation is internalized, it can support needs satisfaction, intrinsic motivation, and behavioral intention. Intrinsic motivation was positively associated with behavioural intention in workplace gamification use (Mitchell, Schuster, and Jin 2020). Gamification is used in application fields like sports (Fernandez-Rio et al. 2020), health (Floryan et al. 2020), sustainability (Oppong-Tawiah et al. 2020), education (Zainuddin et al. 2020), marketing and business (Hwang and Choi 2020). Hamari et al. (2016) reported improved learning using computer games in training applications. In e-learning applications, we can use gamification to enhance motivation. In companies, we can employ gamification to increase employee engagement and to motivate them to perform their tasks with more enthusiasm (Gupta and Gomathi 2017). Even when the introduction of gamification into training did not prove to increase outcomes, it increased the levels of learner motivation to acquire those skills (Larson 2020). Gamification domain is vast and the research community continue to discover new areas of application (Mitchell, Schuster, and Jin 2020).

The main components of gamification are game elements, which denote specific game components that can be used in gamification (Werbach and Hunter 2012; Sailer et al. 2014). Game elements such as points, levels and leaderboards have become a constant in gamification, especially due to their use in games (Mekler et al. 2017). The relationship between game elements and Self-Determination Theory is

presented in (Werbach and Hunter 2012). Sailer et al. (2014) also analyzed game elements and linked them to motivational mechanisms that they primarily refer to. In the literature, it has been argued that thoughtful implementation of game elements may improve intrinsic motivation by satisfying users' innate psychological needs (Francisco-Aparicio et al. 2013; Pe-Than, Goh, and Lee 2014; Peng et al. 2012; Mekler et al. 2017). Mekler et al. (2017) studied the effects of individual game elements on motivation and performance. They found that gamification did not affect intrinsic motivation, but their results suggest that in the given context game elements acted as extrinsic incentives. However, Landers (2019) stressed that gamification is not only the addition of game elements and game design to non-game processes but rather the development and design processes supported by extant research. Gamification studies how we can motivate users and change the process that it gamifies (Landers 2019; Bovermann and Bastiaens 2020).

To our knowledge, previously, gamification has not been used to improve IQ assessment and inter-rater agreement. The existing literature that investigated IQ assessment builds on the well-established classic non-gamified process. Existing assessment processes build on extrinsic motivation to perform tasks-at-hand. In this study, we introduce a novel IQ assessment process that employs gamification. We analyzed the existing assessment process, shortened the length of the source document under assessment, and created a new gamified assessment application. The final application contains game-like components, but it has a functional non-game purpose and elements, which are not game-like (Kasurinen and Knutas 2018). Game elements include points, levels, choice elements, progress bars, and leaderboards. Our goal was to increase the assessor's intrinsic motivation to play and have fun and to positively impact the assessors' attitude to perform the IQ assessment tasks. Besides, we included extrinsic motivation through rewards to receive acknowledgment in the hall of fame scoreboard. Thus, we hypothesize a positive effect of gamification on motivation and finally on IQ assessment score.

2.3 Problem statement and proposed solution

The existing work primarily focuses on heuristic principles (i.e. cognitive decision-making processes) that help assessors with IQ assessment and the effects of consistent application of heuristic principles on the inter-rater reliability. However, fewer studies focused on the cues that influence IQ assessment and inter-rater reliability. Our study contributes to a better knowledge of cues that affect inter-rater agreement levels when assessing IQ. It primarily focuses on the effects of motivation on inter-rater reliability, while it also investigates differences in the measurability of IQ dimensions and their corresponding inter-rater agreements.

Assessment is a complex and mentally labor-intensive task. We believe that the degree of effort that assessors are willing to put into the IQ assessment process affects its result, and hence, inter-rater reliability. If all raters are highly motivated, the difference between their assessments will decrease, and inter-rater agreement will improve.

To motivate participants to do their best at assessing IQ dimensions, we introduced gamification to the assessment process. We adjusted the gamified content to provide an engaging assessment environment that supports assessors in their assessment process (Bovermann and Bastiaens 2020). Intrinsically motivated activities facilitate assessors to perform tasks without any kind of conditioning (Alsawaier 2018), while elements of extrinsic motivation help to perform work tasks through rewards (Rosas et al. 2003; Mitchell, Schuster, and Jin 2020).

Focusing on motivation, the researchers conducted some research in the field of IQ. However, they have not thoroughly studied the impact of motivation on assessment consistency, nor did they employ the concept of gamification. Previous research has been more interested in how increased motivation affects the more consistent use of heuristics, resulting in a possible higher inter-rater reliability. For instance, Metzger and Flanagin (2013) focused on the use of cognitive heuristics in credibility evaluation in an online environment.

In the literature, Yaari, Baruchson-Arbib, and Bar-Ilan (2011) indirectly acknowledged the importance of motivation when using the Elaboration Likelihood Model (ELM) to explain why article length is often viewed as an indicator of quality. According to the ELM, users approach the problem of evaluation systematically when they are motivated and have the knowledge about the relevant topics, but make use of rules of thumb when their motivation and relevant knowledge are lacking.

Concentrating on the source, previous approaches in the literature mainly focused on assessing Wikipedia articles; to our knowledge, few or no studies focused on blogs and other sources. Arazy, Kopak, and Hadar (2017) indicated that the measurability of IQ depends on media type and task context.

Based on our review of the literature and study design, we formulate the following research questions (RQ):

- RQ1: To what extent can motivation affect the measurability of IQ for short hints used in a gamified process?
- RQ2: How does motivation influence individual IQ dimensions in terms of inter-rater agreement in the assessment of IQ?

3 Method

3.1 Evaluation mechanics

To address the research questions, we conducted an online quantitative case study. Participation was voluntary and user consent was obtained before the start. As presented in detail in section 3.2, we included a diverse set of participants (i.e. individuals with different education, demographic and social backgrounds), which amplified the validity of our research. To reach a broader audience, we selected a Web study, and to acquire the participants' views of IQ, we designed a survey in the form of a gamified process.

Previous studies on measuring IQ dimensions focused primarily on students (Arazy and Kopak 2011) and university librarians (Arazy, Kopak, and Hadar 2017). The main drawback of existing studies is the small group of participants. In the present study, we followed a design that would investigate smaller data sources under study but use existing data dimensions and existing estimation metrics with a much bigger set of participants as further discussed in section 3.2.

We employed a gamification principle to measure the influence of motivation on the evaluation of IQ. For our experiment, we developed a tool in the form of a Web-based gamified software tool. The overall gamified purpose of the assessment application is to save and raise a little bird to adulthood and return it to the wild (see details in section 3.3). The objective is to complete the gamified process with a minimum number of attempts to receive more points, which addresses the motivational aspect. Because of the dynamic conditions under which the participants gain points, more motivated participants collect a higher number of points for their effort. For higher user engagement in the gamified process of measuring IQ, we included the following game elements: leaderboard (visual display of social comparison), levels (player's progressive) and points (virtual rewards against the player effort) as those elements improve the motivation and performance of participants (Mekler et al. 2017).

The research goal of the gamified process is to assess hints' IQ that corresponds to selected IQ dimensions and help participants in progressing through the gamified process. We argue that the participants' success in resolving the gamified task strongly correlates to the consistency of given IQ evaluations of participants in selected dimensions.

3.2 Participants

Our study had a total of 1225 participants who participated from April 2015 to March 2020. Initially, we directly targeted undergraduate University students whom we had direct access and then all potential participants by utilising mail and social media campaigns. In the process of data cleaning, we excluded participants that did not answer all 24 questions (4 game levels with 6 questions each) and spent a total time of less than 5 min (quick random selection of responses) or more than 3 h (multiple breaks while playing) to complete the gamified process.

We were then left with 1062 responses with the median time to complete all 4 game levels of 11 min 50 s. There were 30.7% female and 69.3% male, with ages ranging from 15 to 69 with a median value of 20 years old.

In general, we targeted a population that has finished high school, since they have more experience with poor IQ. A total of 40.5% of participants stated that poor IQ deeply disturbs them and 37.5% stated that they at least bother about poor IQ. We sought to increase the diversity of participants to enhance the external validity of the research; the participants were 57.6% undergraduate students and 42.4% non-students.

Although, the study required no prior knowledge to participate, we included all participants in a pre-training that provided an introduction to the IQ dimensions (completeness, accuracy, representation, and objectivity) that they had to evaluate later in the gamified process. Before performing gamified tasks, they also completed an evaluation task in which they were asked to measure these IQ dimensions in a short paragraph.

3.3 Measuring IQ dimensions

The gamified process consists of four levels. At each game level, we evaluate one of the selected IQ dimensions (completeness, accuracy, representation, and objectivity) highlighted in section 2. Figure 1 depicts an overview of the measurement of IQ dimensions, while Figure 2 shows comprehensive details. For every IQ dimension, we presented evaluating objects in a random order (see activity A2 in Figure 1). Each object is associated with a hint of a different predefined level of correctness (see Equation (5)). The rater then evaluates the IQ of a hint (for finding an object within a gamified process) before employing it (see activity A6 in Figure 1) to find the correct object. The number of points awarded is a function of the number of attempts and the level of correctness of a given hint (see Equation (6)). Once the rater finds the correct object, he evaluates again the IQ of the same hint, where a calibration to prior evaluation is possible (see activity A9 in Figure 1). The gamified proces ends when the rater finds all objects within a given IQ dimension and evaluates IQ dimensions.

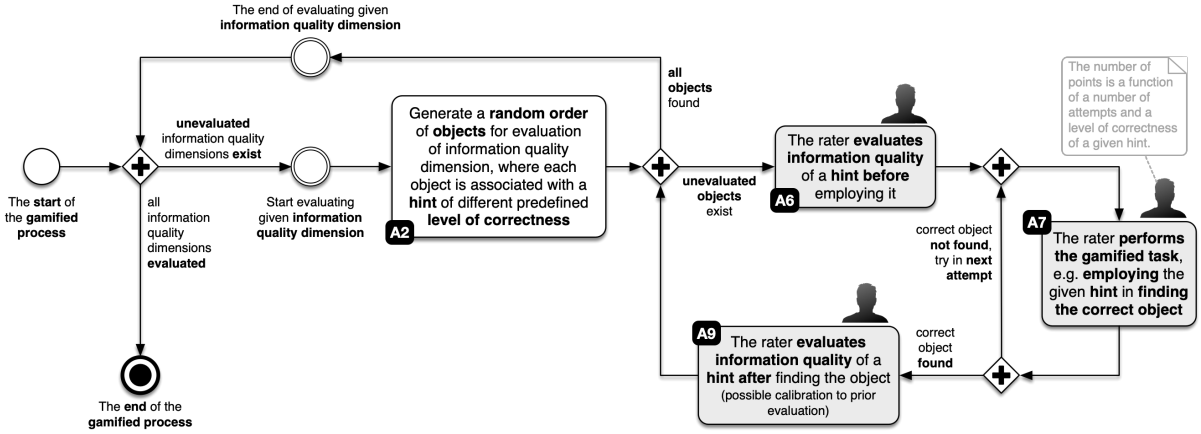

Figure 1: Overview of measuring selected IQ dimensions

Figure 2 depicts that the first step in evaluating IQ dimension is displaying the game rules for the g-th level (see activity A1 in Figure 2), where $g \in [1, 4]$. At the start, each player receives general information about the story of the level and tasks that he must accomplish, along with detailed instructions. At all times there is a progress bar available at the top of the screen that informs the player of his status within the gamified process.

To successfully finish each level of the gamified process, participants must find the correct object $o(g)$, based on the gamified task's context and the hint provided. The objects for each of the game's levels $o(g)$ are selected and displayed to the participant in a random order (see activity A2 in Figure 2).

The **first level** focuses on **completeness**; participants must find a hidden object from the following set

$$o(1) = \{\text{birdie}, \text{worms}, \text{fox}, \text{strawberries}, \text{key}, \text{treasure chest}\} \tag{1}$$

Presented hints are of various completeness levels, where the most complete hint provides information for unique identification of a location of a hidden object, while the least complete hint involves a great level of ambiguity (e.g. there are several possible locations of a hidden object).

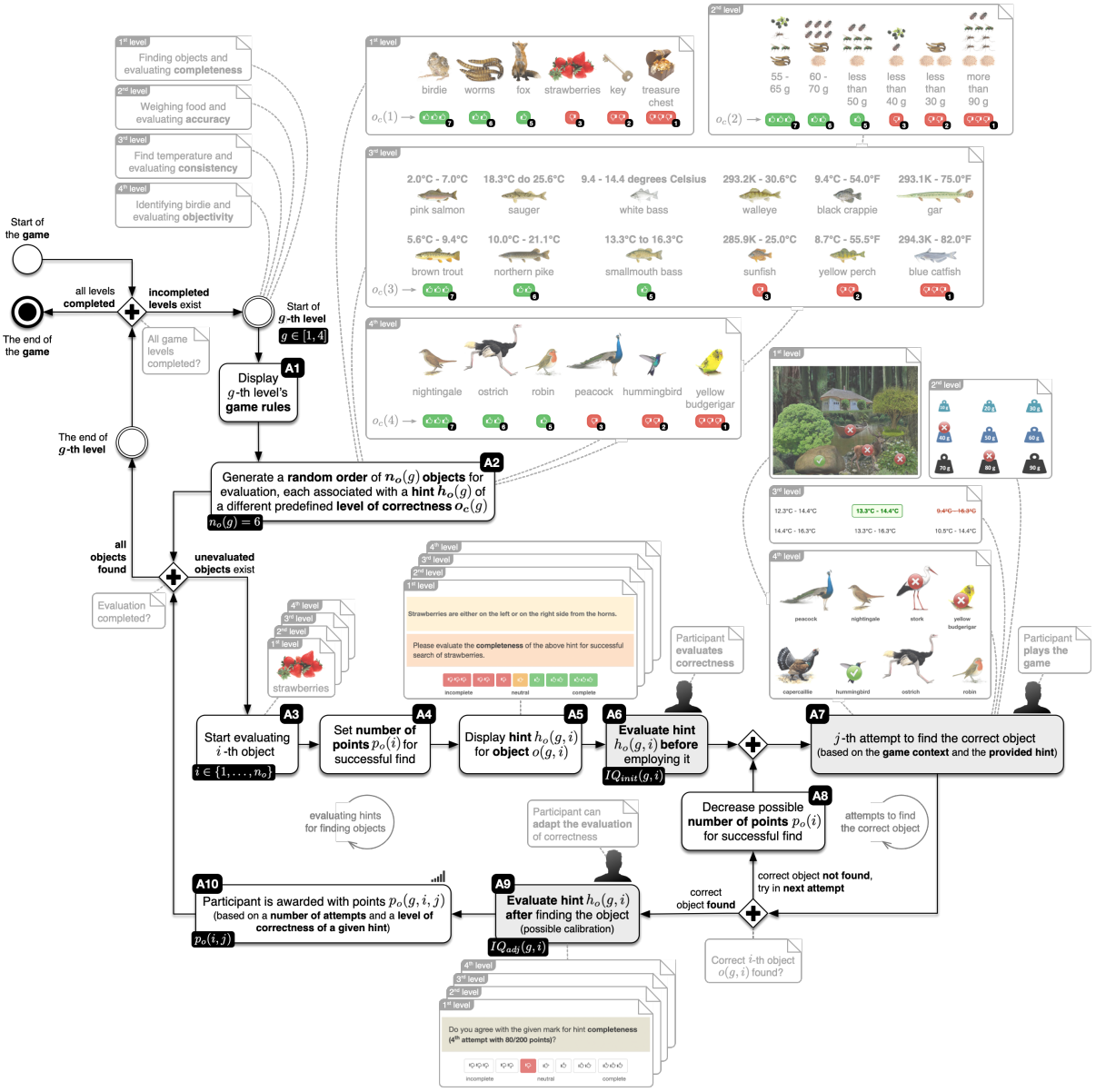

Figure 2: Detailed process of **measuring** selected **IQ dimensions**

The **second level** focuses on **accuracy**; participants must weigh food and select the correct weight of the following objects:

$$
\begin{aligned}
o(2) = \{ & \\
& \{\text{worms + 2 flies + mosquito + blackberries}\}, \\
& \{\text{crumbs + worms + 6 bugs}\}, \\
& \{\text{crumbs + 6 flies + 2 bugs}\}, \\
& \{\text{crumbs + bug + blackberries}\}, \\
& \{\text{2 crumbs + worms}\}, \\
& \{\text{2 crumbs + 4 mosquitos + 3 flies + 2 bugs}\} \\
\} &
\end{aligned}
\tag{2}
$$

Presented hints are associated with scales of various accuracy, from the most accurate with the exact measurement and the least accurate with the false range of measurement.

The **third level** focuses on **representation**; participants must find the correct temperature range for a living habitat of the following fish:

$$
\begin{aligned}
o(3) = \{ & \\
& \{\text{brown trout and pink salmon}\}, \\
& \{\text{sauger and northern pike}\}, \\
& \{\text{white bass and smallmouth bass}\}, \\
& \{\text{walleye and sunfish}\}, \\
& \{\text{black crappie and yellow perch}\}, \\
& \{\text{gar and blue catfish}\} \\
\} &
\end{aligned}
\tag{3}
$$

Presented hints are associated with the representation of temperature ranges that are available in various units (Celsius, Fahrenheit, and Kelvin), where the most consistent representation includes a range with only one unit of measurement, while the least consistent representation presents a mix of various units.

The **fourth level** focuses on **objectivity**; participants must indicate the correct bird

$$
\begin{aligned}
o(4) = \{ & \\
& \text{nightingale, ostrich, robin, peacock,} \\
& \text{hummingbird, yellow budgerigar} \\
\} &
\end{aligned}
\tag{4}
$$

Presented hints include bird's origin, size, and color that are presented by various levels of objectivity with different people, from the most objective ornithologists to the least objective bankers, programmers, and wall painters.

There are $\boldsymbol{n_o}(g) = \mathbf{6}$ **objects** at every level g, each associated with a hint $\boldsymbol{h_o}(g)$ of different **level of correctness** $\boldsymbol{o_c}(g) = \{1, 2, 3, 5, 6, 7\}$, where $o_c(g) = \{1, 2, 3\}$ is associated with correct hints (with value 1 the most correct hint) and $o_c = \{5, 6, 7\}$ are incorrect hints (with value 7 the most incorrect hint). The main part of the task is to evaluate the i-th object based on the provided hint (see activity **A6** in Figure 2).

The function of correctness $o_c(g)$ differs for every level g and represents the selected IQ dimension being measured

$$o_c\,(g) = \begin{cases} o_c = \text{completeness} & ; \quad \text{if } g = 1 \\ o_c = \text{accuracy} & ; \quad \text{if } g = 2 \\ o_c = \text{representation} & ; \quad \text{if } g = 3 \\ o_c = \text{objectivity} & ; \quad \text{if } g = 4 \end{cases} \tag{5}$$

At every level of the game g, participants receive $n_o(g) = 6$ objects in random order for evaluation. When searching for the correct object, participants can try and find the correct object if their previous attempt was incorrect (see activity **A7** in Figure 2).

The **points** awarded for a correct answer $\boldsymbol{p_o(i,j)}$ are decreasing linearly with the **number of attempts** $\boldsymbol{j \geq 1}$ and a predefined **level of correctness** $\boldsymbol{o_c}$ (the less correct the hint is, the more points are awarded if the object is correctly identified) as depicted in Equation (6). We based the quality of hints on a previous analysis and rules that pertain to the dimension under study.

$$p_o\,(i,j) = \begin{cases} 10 \cdot (6-j) \cdot (7-o_c) & ; \quad \text{if } j < 6 \text{ and } o_c(i) \in \{1,2,3\} \\ 10 \cdot (6-j) \cdot (8-o_c) & ; \quad \text{if } j < 6 \text{ and } o_c(i) \in \{5,6,7\} \\ 0 & ; \quad \text{otherwise} \end{cases} \tag{6}$$

Each participant can make multiple attempts to find the object (see activity **A7** in Figure 2) but is, according to Equation (6), motivated to find the correct answer in the minimum number of attempts. Each failed attempt reduces the number of points awarded (see activity **A8** in Figure 2) at a given level g and consequently in the game as a whole.

When a participant at a given game level g starts evaluating the i-th object $o(g,i)$ (see activity **A3** in Figure 2) on a 7-point Likert scale, an associated hint $h_o(g,i)$ with a predefined level of correctness $o_c(g,i)$ is displayed (see activity **A5** in Figure 2). The hint is evaluated with $IQ_{init}(g,i)$ (see activity **A6** in Figure 2) before the participant tries to find the correct object $o(g,i)$ (see activity **A7** in Figure 2) in a minimal number of attempts j, because points $p_o(g,i,j)$ are associated with the number of attempts required for success. After the correct object $o(g,i)$ is found, the previous evaluation $IQ_{init}(g,i)$ can be calibrated with $IQ_{adj}(g,i)$ (see activity **A9** in Figure 2); the participant can alter previously given scores for an IQ dimension, if desired.

The evaluation process of IQ dimensions is complete at the end of the 4-*th* level. At the end, the system gives the participant a score and his overall position in the rankings.

3.4 Evaluation metrics

The interclass correlation coefficient is a reliability index widely used for intra-rater and inter-rater reliability analyses. Since we measured the variation between raters measuring the same group of objects in this work, we focused on inter-rater reliability. According to the guidelines, proposed by Koo and Li (2016), we select ICC(2,1) as a measure of agreement for our inter-rater reliability study (see Figure 3 and Table 1).

The attributes of the ICC(2,1) are:

- the model is two-way random with k raters randomly selected and each hint (total of n hints) measured by the same set of k raters,
- the number of measurements is single measures and reliability is applied to a context where a single measure of a single rater is performed,
- the metric is absolute agreement, where the agreement between raters is of interest, including systematic errors of both raters and random residual errors.

ICC(2,1) is defined as follows

$$\boldsymbol{ICC(2,1)} = \frac{BMS - EMS}{BMS + (k-1) \cdot EMS + \frac{k}{n} \cdot (JMS - EMS)} \tag{7}$$

where

- $\boldsymbol{WMS} = \frac{WSS-BSS}{n\cdot(k-1)}$ is **Within Mean Squares** (from one-way ANOVA),
- $\boldsymbol{BMS} = \frac{BSS}{n-1}$ is **Between Objects Mean Squares** (from one-way ANOVA),
- $\boldsymbol{JMS} = \frac{JSS}{k-1}$ is **Joint (between raters) Mean Squares** (from two-way ANOVA),
- $\boldsymbol{EMS} = \frac{ESS}{(n-1)\cdot(k-1)}$ is **Error (residual) Mean Squares** (from two-way ANOVA).
- $\boldsymbol{ESS} = WSS - BSS - JSS$ is **Error (residual) Sum of Square** (from two-way ANOVA),
- $\boldsymbol{JSS} = n\cdot\sum_{j=1}^{k}\left(\frac{1}{n}\cdot\sum_{i=1}^{n} v_{ij} - m_a\right)^2$ is **Joint (between raters) Sum of Square** (from two-way ANOVA),
- $\boldsymbol{BSS} = k\cdot\sum_{i=1}^{n}\left(\frac{1}{k}\cdot\sum_{j=1}^{k} v_{ij} - m_a\right)^2$ is **Between Objects Sum of Squares** (from one-way ANOVA),
- $\boldsymbol{WSS} = \sum_{i=1}^{n}\sum_{j=1}^{k}\left(v_{ij} - m_a\right)^2$ is **Within Sum of Squares of all raters** (from one-way ANOVA) and
- $\boldsymbol{m_a} = \frac{1}{n\cdot k}\sum_{i=1}^{n}\sum_{j=1}^{k} v_{ij}$.

To measure the reliability of scale we also calculate Cronbach's alpha ICC(3,k), which is defined as

$$\boldsymbol{\alpha} = \frac{BMS - EMS}{BMS} \tag{8}$$

4 Results and discussion

4.1 Results

Participants in our research evaluated the IQ dimensions of hints in a gamified process, where we rewarded their effort with a score. Table 1, Figure 3, and Figure 4 show the results referring to both research questions. The experiment had a two-fold purpose. First, to measure the inter-rater reliability agreement as ICC(2,1) in evaluating various IQ dimensions. Second, to measure the motivation of participants in the gamified process. The involvment of the participants in a form of motivation is measured by game points, related to the performance of players further determined by the number of attempts and predefined level of correctness of a given IQ dimension. The mechanics of points calculation are defined in Equation (6), where each aprticipant can make multiple attempts to find the object but is motivated to find the correct answer in the minimum number of attempts. Each failed attempt reduces the number of points awarded to the player at a given level and consequently in the game as a whole. The aforementioned results ICC(2,1) and groups Q are not related in terms that ICC(2,1) focuses on agreement with other raters, while groups focus on the provided value of IQ dimension by the rater in relation to the predefined level of correctness associated with the dimension.

Table 1 presents detailed inter-class agreement results ICC(2,1) (denoted by ICC) including the measured reliability of scale ICC(3,k) (denoted by α) for various constructs regarding different groups of participants. We divided the participating players' scores into four groups according to the number of points scored in the gamified environment. The groups in Figure 3 are arranged in ascending order by quartile. Group Q4 represents those who achieved mediocre results, while Q1 represents the highest-scoring players.

There are five IQ dimension groups; four for every dimension under investigation, and CIQ, the mean value of all four dimensions.

Table 1: ICC results in our research

IQ dimension	Q	n	k	BMS	WMS	JMS	EMS	ICC(2,1)	ICC(3,k)
Completeness	Q4	6	265	871.10	3.10	6.36	2.45	0.51	1.00
Completeness	Q3	6	254	1,202.42	2.31	3.75	2.02	0.67	1.00
Completeness	Q2	6	264	1,357.12	1.84	3.06	1.60	0.74	1.00

IQ dimension	Q	n	k	BMS	WMS	JMS	EMS	ICC(2,1)	ICC(3,k)
Completeness	Q1	6	279	1,494.92	1.58	2.54	1.39	0.77	1.00
Accuracy	Q4	6	265	442.82	3.77	9.83	2.56	0.31	0.99
Accuracy	Q3	6	254	764.38	2.34	5.05	1.79	0.56	1.00
Accuracy	Q2	6	264	841.57	2.36	5.49	1.74	0.57	1.00
Accuracy	Q1	6	279	956.64	2.20	4.68	1.70	0.61	1.00
Representation	Q4	6	265	137.10	4.39	12.89	2.69	0.10	0.98
Representation	Q3	6	254	161.69	3.46	9.37	2.28	0.15	0.99
Representation	Q2	6	264	229.97	3.02	8.49	1.92	0.22	0.99
Representation	Q1	6	279	227.29	3.15	10.07	1.77	0.20	0.99
Objectivity	Q4	6	265	165.93	3.40	8.66	2.34	0.15	0.99
Objectivity	Q3	6	254	385.43	3.08	6.67	2.36	0.33	0.99
Objectivity	Q2	6	264	453.62	2.79	6.04	2.14	0.38	1.00
Objectivity	Q1	6	279	557.23	2.87	6.63	2.12	0.41	1.00
CIQ	Q4	24	265	403.51	3.66	20.39	2.94	0.29	0.99
CIQ	Q3	24	254	675.34	2.80	11.25	2.43	0.49	1.00
CIQ	Q2	24	264	796.64	2.50	10.60	2.15	0.55	1.00
CIQ	Q1	24	279	882.81	2.45	11.76	2.05	0.56	1.00

Figure 3 depicts ICC results for the selected set of IQ dimensions (completeness, accuracy, representation, and objectivity) for different participants' groups (Q1, Q2, Q3, and Q4).

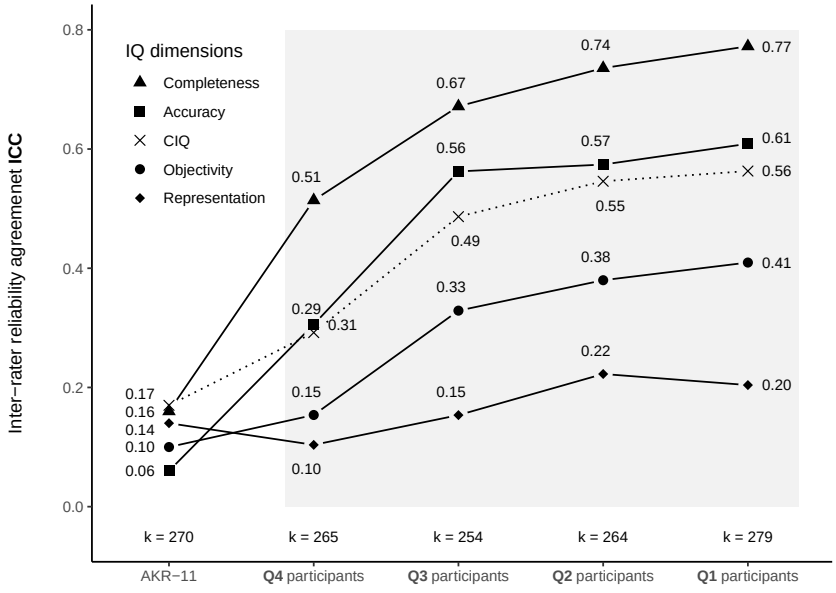

Figure 3: Interclass correlation vs. performance of players. The y-axis represents the ICC, while the x-axis portrays four groups of participants. The groups are divided into quartiles according to the points that participants scored when performing the gamified process. The first column AKR-11 contains results of a similar IQ study, in which Arazy and Kopak (2011) studied the measurability of IQ on the same set of IQ dimensions used in our study. The four highlighted columns (rightmost) exhibit the results of our study, where each column represents one group of players.

There are multiple guidelines for the interpretation of ICC inter-rater agreement values (Landis and Koch 1977; Cicchetti 1994; Koo and Li 2016). Figure 4 summarizes the results of our study by incorporating all aforementioned ICC interpretations. We can observe that highly motivated participants achieved substantially better results compared to the unmotivated ones regardless of the interpretation chosen.

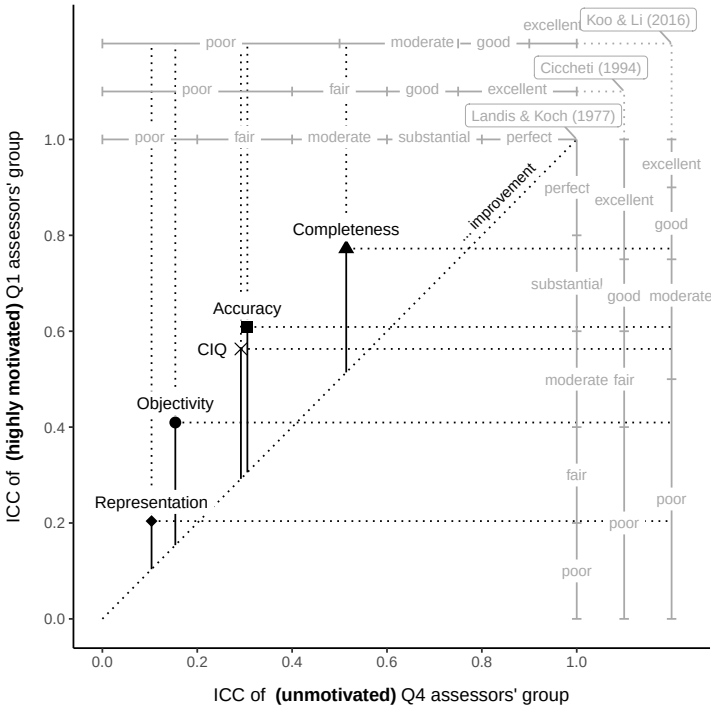

Figure 4: Comparing inter-rater agreement for IQ dimensions in terms of motivation by assessors' groups. The dual-scale data chart depicts the relationship between ICC values and various interpretations of inter-rater agreement. It compares the four IQ dimensions in terms of the extent to which motivation affected the increase in inter-rater agreement. The bottom x-axis denotes ICC for unmotivated users (Q4), while the left y-axis represents ICC values for highly motivated users (Q1). The scales on top of the x-axis and right of the y-axis denote various ICC interpretations for unmotivated (Q4) and motivated (Q1) users, respectively. The figure depicts all IQ dimensions and the mean value of the aforementioned dimensions. ICC values above the identity line (i.e. the dotted diagonal) represent an increase in ICC, while values below represent a decrease in ICC, when comparing unmotivated (Q4) and motivated (Q1) groups.

4.2 Discussion

When analyzing the differences in ICC between the various groups, we can observe that the agreement level increases with players' performance in terms of the final score achieved (see Figure 3). Raters who were more motivated more carefully rated the hints by a given dimension, which lead to more homogenous assessments. Group Q4 with members that attained mediocre results in the gamified process, represents participants who had the poorest motivation and did not focus on the task-at-hand. The results for this group represent a foundation, a basic ICC with which to compare other group' who were more motivated.

When comparing the other three quartiles (Q3, Q2, and Q1), it is evident that inter-rater agreement for all dimensions increases consistently with the increasing game score. We can observe the same for the average score, CIQ. The exception to the rule is the dimension representation; Q1 achieved a slightly lower inter-rater reliability than Q2. However, the ICC for representation is still higher than for Q4 and Q3.

There is a substantial increase in ICC between groups Q4 and Q3 for all dimensions. The results indicate that participants who were even slightly motivated quickly achieved better ICC. Therefore, for RQ1, we can conclude that increased motivation reinforces the measurability of IQ for short hints in the context of a gamified environment.

Based on detailed ICC results for the hands-on task (see Table 1), we can argue in response to RQ2 that ICC has a positive correlation with motivation. The overall results for α reflect a high rate of scale reliability. As shown in Table 1, we reached the highest α for dimension completeness in all four quartile groups (Q4 – Q1), which is also in line with the best results for ICC for that dimension.

In terms of ICC, participants attained the highest agreement levels for the dimension completeness, followed by accuracy, objectivity, and representation. For all four IQ dimensions, we obtained better results than existing studies (Arazy, Kopak, and Hadar 2017; Fidler and Lavbič 2017). However, we must emphasize that our study is not a replication of studies from the literature. This study focuses on other sources under investigation (hints) and other aspects (motivation) that may influence inter-rater reliability.

We observe from our results that raters can consistently identify the quality of a hint if it leads to a problem solution; hence, participants could successfully rate completeness. It is also evident that raters can identify missing information and thus deduce completeness. The latter result is in line with the results of Arazy, Kopak, and Hadar (2017), who reported that the ICC for completeness was substantially higher than for other dimensions, although inter-rater agreement on this dimension was substantially lower in their research. Higher ICC may be obtained for completeness since people have a better understanding of this dimension than the other three dimensions. Participants determined the quality of a hint based on the possible hiding places left when they considered the hint. Since the task was straightforward, the participants succeeded in evaluating the quality of these hints and achieved better inter-rater agreement.

In terms of accuracy, we achieved a moderate (0.61) agreement. Compared to completeness (0.77), the lower result for accuracy may indicate that weight estimation is more challenging than locating an object. However, of all quality dimensions, accuracy gained the biggest increase in inter-rater reliability with a slightly increased motivation (Q4 and Q3).

The measured ICC for the objectivity of motivated participants (Q1) is 0.41, noticeably lower than ICC for completeness (0.77) and accuracy (0.61). However, the result still demonstrates fair agreement. For non-motivated players (Q4), the ICC for objectivity was half that for accuracy and ICC noticeably increased with increased motivation. It is worth noting that the ICC trend for objectivity is very similar to that observed for completeness and accuracy. Not all participants recognized the relevant messages in the hints, and thus did not identify the quality of hints. The meaning of hints remained unclear because of the inclusion of scientific terms, which offered assessors little clarification. As a result, inter-rater reliability was below expectations, though an upward trend was still present.

Participants attained the lowest agreement levels of all four dimensions for the representation dimension. Motivation had a positive impact on representation ICC, but not as much as in the case of other dimensions. Participants struggled with determining the consistency and hence the quality of the hint. As a direct consequence, inter-rater reliability was very low. We believe that this task was the most cognitively challenging of the four, and most participants chose not to invest much effort in solving it.

Based on Landis and Koch (1977), Cicchetti (1994) and Koo and Li (2016) interpretations of ICC, we summarize the results of the study in Figure 4. Focusing on the groups of unmotivated participants (Q4) and the most motivated participants (Q1), the agreement levels were substantially higher for the motivated participants. Motivation increased the CIQ construct by 0.27 with the following interpretations (from Q4 – to Q1): fair – moderate (Landis and Koch 1977), poor – fair (Cicchetti 1994), and poor – moderate (Koo and Li 2016). Comparing our results to findings in the literature (Fidler and Lavbič 2017; Arazy, Kopak, and Hadar 2017) further confirms that this study has achieved results with much greater ICC.

We found an increase in an agreement between the two groups of participants on all four IQ dimensions. We achieved the highest ICC agreement improvement for completeness, which increased from moderate (0.51) to good (0.77), according to the interpretation of Koo and Li (2016). The increase in ICC for completeness (0.26) was slightly below average CIQ (0.27), which still resulted in a superior result in an agreement for completeness. Interestingly, we observed the biggest improvement in ICC for the dimension of accuracy, with ICC increasing by as much as 0.30, improving from poor (0.31) to moderate (0.61) agreement, according to the interpretation of Koo and Li (2016). Motivation proved to be the key factor contributing to better accuracy assessment. The level of agreement of non-motivated participants for the dimensions objectivity (0.15) and representation (0.10) was poor. The introduction of motivation did not improve the quality of assessment enough to reach a moderate agreement for both dimensions (0.41 and 0.20 respectively). However, objectivity yielded an ICC increase slightly below average (0.26),

indicating that motivation is a driving factor for this dimension. Nonetheless, objectivity remains difficult to evaluate consistently, even with motivated assessors. For representation, we observed the lowest agreement and ICC increase (0.10) between the four dimensions. Non-motivated participants achieved poor agreement levels (0.10). Motivation contributed to the rise in inter-rater agreement, but the result remained in the zone of poor agreement (0.20) according to the interpretation of Koo and Li (2016).

4.3 Implications for research and practice

Our study supports the theoretical underpinning of IQ studies and confirms previous findings that IQ is a multidimensional construct that is difficult to measure. We also confirm that inter-rater agreement for different IQ dimensions can vary significantly.

Second, existing research has performed very poorly in assessing the measurability of IQ. According to most ICC interpretations, such results have very low measurability, so their interpretation is questionable. Our study builds on previous IQ-related research and extends it to alternative settings to demonstrate the significance of motivation in IQ assessment. Using gamified tasks to motivate assessors we were able to significantly improve the measurability of IQ (ICC). The correlation between points awarded in the gamified process and the inter-rater reliability agreement was positive for all four IQ dimensions and composite IQ (CIQ). The level of agreement achieved with the most motivated group of participants (Q1) was superior in comparison with results from related work. Future IQ measurement studies should take these results into account if they want interpretable results.

Third, the study extends previous research by introducing gamification to IQ assessment domain. It attempts to avoid rhetorical gamification by creating a renewed assessment process. The evidence reveals that gamification produced increased assessors' motivation leading to a better inter-rater agreement, consequently improving IQ assessment. That also confirms previous findings from (Kasurinen and Knutas 2018; Treiblmaier and Putz 2020) stating that the gamification domain is immense and that researchers discover new application ares continuously.

Finally, the study attempts to motivate other researchers to replicate it in alternative settings to validate or complement our findings. Further studies should investigate additional factors that influence inter-rater reliability, such as heuristic principles used by participants, different sources of information, the size of the source under investigation, and what attributes of assessors affect results.

Information workers and researchers can benefit from our findings by creating IQ assessments in ways that take advantage of increased motivation. This study shows that we can achieve increased motivation by employing the concept of gamification, by including elements such as points, badges, and leaderboards.

Researchers studying the assessment of IQ should recognize motivation as a vital cue affecting IQ assessment. If applicable, they should consider including gamification in their studies. We conducted this study also with student participants. Our results demonstrate that gamification can be used successfully with students. Teachers creating gamified IQ tasks should consider improving the assessment process instead of adding gamification features to look like a gamified process.

Finally, the research provides insight into IQ and its dimensions for consumers of short online news. Many users are not aware of IQ dimensions and might start to consume contents in a more educated way.

4.4 Study validity

We performed activities both in the design phase and later in the data collection and analysis phases intended to increase the validity of our research.

To support internal validity, all participants involved in the experiment participated in the same gamified process, with equivalent study materials, questions, and the same method of obtaining data. The tool used in the experiment was intuitive and easy to use, so no special pre-test training was required for participation, although we performed an initial introduction to IQ dimensions to ensure participants understood the metrics being measured, as outlined in section 3.2. To minimize the instrumentation threat, we captured measured variables automatically and accurately. The participants were not aware of the research goal; they simply aimed to achieve the highest score within the gamified environment.

External validity requirements were addressed properly; our experimental setup represents a real-world situation and our test population has all knowledge expected of the general population. To maximize external validity, we followed the requirements outlined by Carver et al. (2010). As far as a generalization is concerned, the findings in Mullinix et al. (2015) reveal a considerable similarity between many treatment effects obtained from the convenience and nationally representative population-based samples.

Concerning construct validity, participants were not subject to any pressure, and participation in the study was voluntary, which minimized mortality threat. In order to avoid unintentionally influencing the participants' behavior, there was no interaction between researchers and participants during the experiment or the study's goals. The problem domains in the gamified environment were selected to minimize any bias introduced by the familiarity of participants with given domains, which could have skewed the results in favor of some participants.

In terms of conclusion validity, we employed a robust measurement of inter-rater reliability agreement, ICC, to derive statistically correct conclusions based on the collected data. To compare the results with the findings in the literature, we included several ICC interpretations. We argue that the number of participants and the data collected were sufficient to draw reliable conclusions. We provide an explanation that a rater's motivation affects the measurability of IQ.

4.5 Limitations

Nevertheless, our study has some limitations that should be acknowledged. We wanted to include as large and as heterogeneous group of participants as possible, so we made the study available to the widest possible audience. Because our study did not have within a lab setup, we could not control all aspects of IQ assessment. For example, we were not able to assure that the participant completed the entire survey alone without assistance. However, by following the requirements outlined by Carver et al. (2010) and iteratively improving the study in the study design phase, we believe our findings are relevant. As discussed in section 3.2 we also addressed this issue by preprocessing and cleaning of obtained data.

One limitation of this study is that people of the same age groups are not fully equally represented with a median value of 20 years old as presented in section 3.2. Research was available to the widest possible audience, so we had limited influence on the age of the participants. In would thus prove insightful to replicate the study where all age groups are equally represented. Although we found no statistically significant differences, it has been found in the literature that some demographic factors may affect the perceived benefits of gamification (Koivisto and Hamari 2014).

We should also be aware of the limitations that come from the ability of the participant to assess the quality of the object according to information quality (Arazy and Kopak 2011). Information quality assessment proved to be difficult. In their study, (Arazy, Kopak, and Hadar 2017) showed that achieving agreement among assessors can be challenging.

For hints we used paragraph size documents instead of full size text documents. Fidler and Lavbič (2017) found that shortening the text from full-size text to paragraph-size text does not affect the agreemenet level of information quality evaluations. However, future studies should thus consider using full-size text hints, which might lead to better user experience despite retaining IQ perception.

Finally, only one problem domain has been used in our study, as presented in section 3.1. Creating gamified content for additional domains requires lots of effort, especially defining game levels for evaluating specific IQ dimensions. Hence, our study should also be applied to other problem domains in future study replications.

5 Conclusion

Reaching consensus on IQ assessment is challenging, and the factors that drive successful estimation of IQ have not been fully explored. This study extends related work and confirms the effect of motivation as a driving factor for improved IQ assessment. It concludes that the employment of innovative gamified IQ assessment was effective, particularly for IQ dimensions that proved to be more reliable to consistent judgment in the literature. It increased participant engagement through the assessment

content shortening and the inclusion of gamification features like points, levels, progress bars, and leader-board.

The level of agreement achieved with the most motivated group of participants (Q1) was superior in comparison with results from related work. Concerning the inter-rater agreement across the four IQ dimensions, we demonstrate that the relationship between individual IQ dimensions varies with motivation. With increasing motivation, the inter-rater agreement consistently improved for the dimensions of objectivity, completeness, and accuracy. For the representation dimension, inter-rater reliability improved in the initial three quartiles.

Overall, gamification proved to be very useful in the field of IQ assessment. Thus, we strongly recommend that further IQ assessment studies control for the influence of motivation, and consider including a gamification approach. With the investigation of a different IQ source, foreknowledge might also be a key factor. Further studies could investigate the association between the amount of foreknowledge and inter-rater reliability.

References

Alsawaier, Raed S. 2018. "The Effect of Gamification on Motivation and Engagement." *The International Journal of Information and Learning Technology*. https://doi.org/10.1108/IJILT-02-2017-0009.

Arazy, Ofer, and Rick Kopak. 2011. "On the Measurability of Information Quality." *Journal of the American Society for Information Science and Technology* 62 (1): 89–99. https://doi.org/10.1002/asi.21447.

Arazy, Ofer, Rick Kopak, and Irit Hadar. 2017. "Heuristic Principles and Differential Judgments in the Assessment of Information Quality." *Journal of the Association for Information Systems* 18 (5): 1. https://doi.org/10.17705/1jais.00458.

Ballou, D. P., and H. L. Pazer. 2003. "Modeling Completeness Versus Consistency Tradeoffs in Information Decision Contexts." *IEEE Transactions on Knowledge and Data Engineering* 15 (1): 240–43.

Batini, Carlo, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. 2009. "Methodologies for Data Quality Assessment and Improvement." *ACM Computing Surveys* 41 (3): 16:1–52. https://doi.org/10.1145/1541880.1541883.

Bovermann, Klaudia, and Theo J Bastiaens. 2020. "Towards a Motivational Design? Connecting Gamification User Types and Online Learning Activities." *Research and Practice in Technology Enhanced Learning* 15 (1): 1. https://doi.org/10.1186/s41039-019-0121-4.

Carver, Jeffrey C, Letizia Jaccheri, Sandro Morasca, and Forrest Shull. 2010. "A Checklist for Integrating Student Empirical Studies with Research and Teaching Goals." *Empirical Software Engineering* 15 (1): 35–59. https://doi.org/10.1007/s10664-009-9109-9.

Cicchetti, Domenic V. 1994. "Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology." *Psychological Assessment* 6 (4): 284. https://doi.org/10.1037/1040-3590.6.4.284.

Danniswara, Randy, Puspa Sandhyaduhita, and Qorib Munajat. 2020. "The Impact of EWOM Referral, Celebrity Endorsement, and Information Quality on Purchase Decision: A Case of Instagram." In *Global Branding: Breakthroughs in Research and Practice*, 882–905. IGI Global. https://doi.org/10.4018/978-1-5225-9282-2.ch042.

Deci, Edward L, and Richard M Ryan. 1985. *Intrinsic Motivation and Self-Determination in Human Behavior*. Springer Science & Business Media, NY. https://doi.org/10.1007/978-1-4899-2271-7.

———. 2000. "The" What" and" Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior." *Psychological Inquiry* 11 (4): 227–68. https://doi.org/10.1207/S15327965PLI1104_01.

Dedeoglu, Bekir Bora. 2019. "Are Information Quality and Source Credibility Really Important for Shared Content on Social Media?" *International Journal of Contemporary Hospitality Management*. https://doi.org/10.1108/IJCHM-10-2017-0691.

DePasque, Samantha, and Elizabeth Tricomi. 2015. "Effects of Intrinsic Motivation on Feedback Processing During Learning." *NeuroImage* 119: 175–86. https://doi.org/10.1016/j.neuroimage.2015.06.046.

Deterding, Sebastian, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. "From Game Design Elements to Gamefulness: Defining "Gamification"." In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, 9–15. MindTrek '11. New

York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2181037.2181040.
Fehrenbacher, D. D. 2016. "Perceptions of Information Quality Dimensions from the Perspective of Commodity Theory." *Behaviour & Information Technology* 35 (4): 254–67.
Fernandez-Rio, Javier, Esteban de las Heras, Tristan Gonzalez, Vanessa Trillo, and Jorge Palomares. 2020. "Gamification and Physical Education. Viability and Preliminary Views from Students and Teachers." *Physical Education and Sport Pedagogy* 25 (5): 509–24. https://doi.org/10.1080/17408989.2020.1743253.
Fidler, Miloš, and Dejan Lavbič. 2017. "Improving Information Quality of Wikipedia Articles with Cooperative Principle." *Online Information Review* 41 (6): 797–811. https://doi.org/10.1108/OIR-01-2016-0003.
Figueiredo, Flavio, Henrique Pinto, Fabiano BeléM, Jussara Almeida, Marcos GonçAlves, David Fernandes, and Edleno Moura. 2013. "Assessing the Quality of Textual Features in Social Media." *Information Processing & Management* 49 (1): 222–47. https://doi.org/10.1016/j.ipm.2012.03.003.
Floryan, Mark, Philip Chow I, Stephen M. Schueller, and Lee M. Ritterband. 2020. "The Model of Gamification Principles for Digital Health Interventions: Evaluation of Validity and Potential Utility." *Journal of Medical Internet Research* 22 (6). https://doi.org/10.2196/16506.
Francisco-Aparicio, Andrés, Francisco Luis Gutiérrez-Vela, José Luis Isla-Montes, and José Luis González Sanchez. 2013. "Gamification: Analysis and Application." In *New Trends in Interaction, Virtual Reality and Modeling*, 113–26. Springer. https://doi.org/10.1007/978-1-4471-5445-7_9.
Gilakjani, Abbas Pourhosein, Leong Lai-Mei, and Narjes Banou Sabouri. 2012. "A Study on the Role of Motivation in Foreign Language Learning and Teaching." *International Journal of Modern Education and Computer Science* 4 (7): 9. https://doi.org/10.5815/ijmecs.2012.07.02.
Gupta, Anchal, and S Gomathi. 2017. "A Review on Gamification and Its Potential to Motivate and Engage Employees and Customers: Employee Engagement Through Gamification." *International Journal of Sociotechnology and Knowledge Development (IJSKD)* 9 (1): 42–52. https://doi.org/10.4018/IJSKD.2017010103.
Hamari, Juho, David J Shernoff, Elizabeth Rowe, Brianno Coller, Jodi Asbell-Clarke, and Teon Edwards. 2016. "Challenging Games Help Students Learn: An Empirical Study on Engagement, Flow and Immersion in Game-Based Learning." *Computers in Human Behavior* 54: 170–79. https://doi.org/10.1016/j.chb.2015.07.045.
Helfert, Markus. 2001. *Managing and Measuring Data Quality in Data Warehousing.*
Hennessey, Beth, Seana Moran, Beth Altringer, and Teresa M Amabile. 2015. "Extrinsic and Intrinsic Motivation." *Wiley Encyclopedia of Management*, 1–4. https://doi.org/10.1002/9781118785317.weom110098.
Hwang, Jiyoung, and Laee Choi. 2020. "Having Fun While Receiving Rewards?: Exploration of Gamification in Loyalty Programs for Consumer Loyalty." *Journal of Business Research* 106: 365–76. https://doi.org/10.1016/j.jbusres.2019.01.031.
Ji-Chuan, Quan, and Wang Bing. 2016. "Problems and Measures of Information Service Quality Evaluation." *DEStech Transactions on Engineering and Technology Research.*
Kappen, Dennis L., and Lennart E. Nacke. 2013. "The Kaleidoscope of Effective Gamification: Deconstructing Gamification in Business Applications." In *Proceedings of the First International Conference on Gameful Design, Research, and Applications*, 119–22. Gamification '13. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2583008.2583029.
Kasurinen, Jussi, and Antti Knutas. 2018. "Publication Trends in Gamification: A Systematic Mapping Study." *Computer Science Review* 27: 33–44. https://doi.org/10.1016/j.cosrev.2017.10.003.
Koivisto, Jonna, and Juho Hamari. 2014. "Demographic Differences in Perceived Benefits from Gamification." *Computers in Human Behavior* 35: 179–88. https://doi.org/10.1016/j.chb.2014.03.007.
Koo, Terry K, and Mae Y Li. 2016. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine* 15 (2): 155–63. https://doi.org/10.1016/j.jcm.2016.02.012.
Landers, Richard N. 2019. "Gamification Misunderstood: How Badly Executed and Rhetorical Gamification Obscures Its Transformative Potential." *Journal of Management Inquiry* 28 (2): 137–40. https://doi.org/10.1177/1056492618790913.
Landis, J Richard, and Gary G Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics*, 159–74. https://doi.org/10.2307/2529310.
Larson, Kristi. 2020. "Serious Games and Gamification in the Corporate Training Environment: A Literature Review." *TechTrends* 64 (2): 319–28. https://doi.org/10.1007/s11528-019-00446-7.

Lee, Y. W., D. M. Strong, B. K. Kahn, and R. Y. Wang. 2002. "AIMQ: A Methodology for Information Quality Assessment." *Information & Management* 40 (2): 133–46. https://doi.org/10.1016/S0378-7206(02)00043-5.

Lian, Hao, Tieke He, Zemin Qin, Haoyu Li, and Jia Liu. 2018. "Research on the Information Quality Measurement of Judicial Documents." In *2018 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-c)*, 178–81. IEEE. https://doi.org/10.1109/QRS-C.2018.00043.

Lukyanenko, Roman, Andrea Wiggins, and Holly K Rosser. 2019. "Citizen Science: An Information Quality Research Frontier." *Information Systems Frontiers*, 1–23. https://doi.org/10.1007/s10796-019-09915-z.

Madhikermi, M., S. Kubler, J. Robert, A. Buda, and K. Framling. 2016. "Data Quality Assessment of Maintenance Reporting Procedures." *Expert Systems with Applications* 63: 145–64.

Mamaril, N. A., E. L. Usher, D. R. Economy, and M. S. Kennedy. 2013. "An Examination of Students' Motivation in Engineering Service Courses." In *2013 IEEE Frontiers in Education Conference (FIE)*, 1825–27. https://doi.org/10.1109/FIE.2013.6685152.

Mekler, Elisa D, Florian Brühlmann, Alexandre N Tuch, and Klaus Opwis. 2017. "Towards Understanding the Effects of Individual Gamification Elements on Intrinsic Motivation and Performance." *Computers in Human Behavior* 71: 525–34. https://doi.org/10.1016/j.chb.2015.08.048.

Metzger, Miriam J, and Andrew J Flanagin. 2013. "Credibility and Trust of Information in Online Environments: The Use of Cognitive Heuristics." *Journal of Pragmatics* 59: 210–20. https://doi.org/10.1016/j.pragma.2013.07.012.

Michnik, J., and M. C. Lo. 2009. "The Assessment of the Information Quality with the Aid of Multiple Criteria Analysis." *European Journal of Operational Research* 195 (3): 850–56.

Mitchell, Robert, Lisa Schuster, and Hyun Seung Jin. 2020. "Gamification and the Impact of Extrinsic Motivation on Needs Satisfaction: Making Work Fun?" *Journal of Business Research* 106: 323–30. https://doi.org/10.1016/j.jbusres.2018.11.022.

Mullinix, Kevin J, Thomas J Leeper, James N Druckman, and Jeremy Freese. 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2 (2): 109–38. https://doi.org/10.1017/XPS.2015.19.

Oppong-Tawiah, Divinus, Jane Webster, Sandy Staples, Ann-Frances Cameron, Ana Ortiz de Guinea, and Tam Y. Hung. 2020. "Developing a Gamified Mobile Application to Encourage Sustainable Energy Use in the Office." *Journal of Business Research* 106: 388–405. https://doi.org/10.1016/j.jbusres.2018.10.051.

Pelling, Nick. 2011. "The (Short) Prehistory of "Gamification." https://nanodome.wordpress.com/2011/08/09/the-short-prehistory-of-gamification/.

Peng, Wei, Jih-Hsuan Lin, Karin A Pfeiffer, and Brian Winn. 2012. "Need Satisfaction Supportive Game Features as Motivational Determinants: An Experimental Study of a Self-Determination Theory Guided Exergame." *Media Psychology* 15 (2): 175–96. https://doi.org/10.1080/15213269.2012.673850.

Pe-Than, Ei Pa Pa, Dion Hoe-Lian Goh, and Chei Sian Lee. 2014. "Making Work Fun: Investigating Antecedents of Perceived Enjoyment in Human Computation Games for Information Sharing." *Computers in Human Behavior* 39: 88–99. https://doi.org/10.1016/j.chb.2014.06.023.

Reiss, Steven. 2012. "Intrinsic and Extrinsic Motivation." *Teaching of Psychology* 39 (2): 152–56. https://doi.org/10.1177/0098628312437704.

Rosas, Ricardo, Miguel Nussbaum, Patricio Cumsille, Vladimir Marianov, Mónica Correa, Patricia Flores, Valeska Grau, et al. 2003. "Beyond Nintendo: Design and Assessment of Educational Video Games for First and Second Grade Students." *Computers & Education* 40 (1): 71–94. https://doi.org/10.1016/S0360-1315(02)00099-4.

Ryan, Richard M, and Edward L Deci. 2000a. "Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions." *Contemporary Educational Psychology* 25 (1): 54–67. https://doi.org/10.1006/ceps.1999.1020.

———. 2000b. "Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being." *American Psychologist* 55 (1): 68. https://doi.org/10.1037/0003-066X.55.1.68.

Sailer, Michael, Jan Hense, J Mandl, and Markus Klevers. 2014. "Psychological Perspectives on Motivation Through Gamification." *Interaction Design and Architecture Journal*, no. 19: 28–37.

Tilly, Roman, Oliver Posegga, Kai Fischbach, and Detlef Schoder. 2017. "Towards a Conceptualization

of Data and Information Quality in Social Information Systems." *Business & Information Systems Engineering* 59 (1): 3–21. https://doi.org/10.1007/s12599-016-0459-8.

Tokan, Moses, and Mbing Imakulata. 2019. "The Effect of Motivation and Learning Behaviour on Student Achievement." *South African Journal of Education* 39 (February): 1–8. https://doi.org/10.15700/saje.v39n1a1510.

Treiblmaier, Horst, and Lisa-Maria Putz. 2020. "Gamification as a Moderator for the Impact of Intrinsic Motivation: Findings from a Multigroup Field Experiment." *Learning and Motivation* 71: 101655. https://doi.org/10.1016/j.lmot.2020.101655.

Wang, Richard Y, Thomas J Allen, Wesley Harris, and Stuart Madnick. 2002. "An Information Product Approach for Total Information Awareness." https://doi.org/10.2139/ssrn.376820.

Wang, Richard Y., and Diane M. Strong. 1996. "Beyond Accuracy: What Data Quality Means to Data Consumers." *Journal of Management Information Systems* 12 (4): 5–33. https://doi.org/10.1080/07421222.1996.11518099.

Wang, Wei Tsong, and Ya Pei Hou. 2015. "Motivations of Employees' Knowledge Sharing Behaviors: A Self-Determination Perspective." *Information and Organization* 25 (1): 1–26. https://doi.org/10.1016/j.infoandorg.2014.11.001.

Werbach, K Hunter, and D Hunter. 2012. *For the Win: How Game Thinking Can Revolutionize Your Business*. Wharton Digital Press, Philadelphia.

West, Kathy, and Janet Williamson. 2009. "Wikipedia: Friend or Foe?" *Reference Services Review* 37 (3): 260–71. https://doi.org/10.1108/00907320910982758.

Yaari, E., S. Baruchson-Arbib, and J. Bar-Ilan. 2011. "Information Quality Assessment of Community-Generated Content - a User Study of Wikipedia." *Journal of Information Science* 37 (5): 487–98. https://doi.org/10.1177/0165551511416065.

Zainuddin, Zamzami, Samuel Kai Wah Chu, Muhammad Shujahat, and Corinne Jacqueline Perera. 2020. "The Impact of Gamification on Learning and Instruction: A Systematic Review of Empirical Evidence." *Educational Research Review* 30. https://doi.org/10.1016/j.edurev.2020.100326.

Zha, Xianjin, Haijuan Yang, Yalan Yan, Kunfeng Liu, and Chengsong Huang. 2018. "Exploring the Effect of Social Media Information Quality, Source Credibility and Reputation on Informational Fit-to-Task: Moderating Role of Focused Immersion." *Computers in Human Behavior* 79: 227–37. https://doi.org/10.1016/j.chb.2017.10.038.

Zhu, Linlin, He Li, Wu He, and Chuang Hong. 2020. "What Influences Online Reviews' Perceived Information Quality?" *The Electronic Library*. https://doi.org/10.1108/EL-09-2019-0208.