



HORIZON-CL4-2021-DIGITAL-EMERGING-01

AI, Data and Robotics for the Green Deal (IA)

AI-powered Robotic Material Recovery in a Box



D6.3: Assessment of the Recycling Data Game (intermediate)

<i>Contractual Date of Delivery:</i>	31/02/2024
<i>Actual Date of Delivery:</i>	19/04/2024
<i>Security Class:</i>	Public
<i>Editor:</i>	<i>Antonios Liapis (UM)</i>
<i>Contributors:</i>	Iro Voulgari (UM), Konstantinos Sfikas (UM), Nemanja Rasajski (UM)
<i>Quality Assurance:</i>	<i>Michail Maniadakis (FORTH)</i>
<i>Deliverable Status:</i>	Final

The *RECLAIM* Consortium

Part. No.	Short Name of Participant	Participant Organization name	Country
1	FORTH	Foundation for Research and Technology Hellas	EL
2	UoM	University of Malta	MT
3	KUL	Katholieke Universiteit Leuven	BE
4	HERRCO	Hellenic Recovery Recycling Corporation	EL
5	IRIS	Iris Technology Solutions, Sociedad Limitada	SP
6	RBNS	ROBENSO PC	EL
7	AIMPLAS	AIMPLAS - Technological Institute of Plastics	SP
8	AXIA	Axia Innovation UG	DE
9	ISWA	International Solid Waste Association	NL
10	ION	Periferiakos Foreas Diaxirisis Stereon Apovlition Ionion Nison Anonimi Eteria Ton Ota	EL

Document Revisions

Version	Date	Editor	Overview
0.1	18/02/2024	Antonios Liapis (UM)	Structure and Early Draft
0.2	15/03/2024	Iro Voulgari (UM), Konstantinos Sfikas (UM)	User Experience additions, experimental protocol details
0.3	15/04/2024	Iro Voulgari (UM)	Results from survey and focus groups discussions
1.0	17/04/2024	Antonios Liapis (UM), Konstantinos Sfikas (UM)	Final edits

Table of Contents

Executive Summary	6
1. Introduction	7
2.1. Intended readership	8
2.2 Relationship with other RECLAIM deliverables	8
3. Evaluation of Player and User Experience	9
3.1 User Experience	9
3.1.1 USE Questionnaire: Usefulness, Satisfaction, and Ease of use	9
3.1.2 Post-Study System Usability Questionnaire	10
3.1.3 User Experience Questionnaire	11
3.1.4 System Usability Scale	12
3.2 Player Experience in Digital Games	13
3.3 User Experience in “Serious” games	14
4. Evaluation Methodology	16
4.1 Data Collection	16
4.2 Participant recruitment	18
4.3 Data Analysis	18
4.4 Material	18
4.4.1 The game	19
4.4.2 Additions to the annotation challenges	20
4.5 Ethics	23
5. Analysis and Results	24
5.1 Survey	24
5.1.1 Demographics of the Sample	24
5.1.2 Game Preferences	24
5.1.3 Perceptions of the players on the Ease of Use, Satisfaction, and Ease of Learning	24
5.1.4 Game Experience, Age, and Environmental Experience in relation to Ease of Use, Ease of Learning, and Satisfaction	26
5.1.5 Positive Aspects	27
5.1.6 Negative Aspects	28
5.1.7 General Comments	28
5.2 Focus group	29

„D6.3: Deployment and evaluation of environmental games	RECLAIM – GA 101070524
5.2.1 Participants	29
5.2.2 Technical Aspects and Usability	29
5.2.3 Game Content and Mechanics	30
5.2.3 Positive Elements	30
5.2.4 Conclusions	31
6. Data Collected during Evaluation	32
7. Summary and Conclusions	37
8. Future Work	38
9. References	40

List of Abbreviations

Abbreviation	Definition
AI	Artificial Intelligence
DoA	Description of the Action
RDG	Recycling Data Game
prMRF	portable, robotic Material Recovery Facilities
RoReWo	Robotic Recycling Workers
GDPR	General Data Protection Regulation
IP	Intellectual Property

Executive Summary

RECLAIM is a Horizon Europe funded project with an objective to **develop** a portable, robotic Material Recovery Facilities (MRFs) (**prMRF**) tailored to small-scale material recovery. RECLAIM adopts a modular multi-robot/multi-gripper approach for material recovery, based on low-cost Robotic Recycling Workers (RoReWos). An **AI module** combines imaging in the visual and infrared domain to identify, localise and **categorize recyclables**. The output of this module is used by a multi-RoReWo team that implements efficient and accurate material sorting.

Further, RECLAIM englobes a citizen science approach to increase social sensitivity to the Green Deal. This is accomplished via a **novel Recycling Data-Game (RDG)** that enables and encourages citizens to participate in project RTD activities by providing annotations to be used in **deep learning** for the **re-training of the AI module**. Three different scenarios will attest its effectiveness and applicability in a broad range of locations that face material recovery challenges.

This deliverable is the first report of the activities undertaken under T6.3 of the RECLAIM Recycling Data-Game (WP6). As per the DoA for RECLAIM, T6.3 “will validate the RDG, through a series of dedicated pilots with focus groups as well as in-the-wild experiments carried out online”, allowing for “(a) targeted and specific feedback regarding the incremental advancement of AI-ILC, and rising recycling awareness (through focus groups) and (b) test RDG usability and data collection potential at scale (through in-the-wild experiments) and the exploitation of gathered data by other project developments”.

This deliverable, therefore, reports on the evaluation carried out at this early stage of RECLAIM, leveraging focus groups (as per DoA description) for targeted feedback. Following up on focus group studies in WP2 (see D2.1 for details), this deliverable uses the same focus groups consisting of interested stakeholders (mostly experts), thus ensuring that their earlier feedback regarding both usability and educational intent is validated in the current prototype of the RDG. The deliverable reports both on the evaluation methodology chosen, via **usability questionnaires and follow-up group discussions**, details on the experimental protocol, and results from this first round of evaluation.

It is worth noting that this first evaluation took advantage of focus groups consisting of experts in recycling and waste management, and focused on **usability issues** of the current version of the games (D6.2). Moreover, it acted as the first (fairly small-scale) **blind test** of the RDG infrastructure, in terms of software (the game itself), hardware (how it plays on different mobile devices) and database performance. We also assess the quality of data received and potential for improving AI algorithms. In the next iteration of D6.3, we intend to test usability and hardware similarly, but also focus on evaluating its potential for raising recycling awareness and long-term engagement.

1. Introduction

At its core, RECLAIM proposes the development of a low cost, portable, easy to install and increased productivity prMRF that can achieve full material recovery anywhere, even in the most remote areas. The developed prMRF is expected to have a key role in developing a global, leakage-free circular economy model benefiting businesses, the society, and the environment.

However, we do not consider that a circular economy is only limited to material waste. With the complementary RECLAIM pillar (PIL-4) for *Environmental gaming for social awareness and data collection*, we envision that **data can also form a positive feedback loop and be re-used in a circular fashion**. Therefore, recycling data games (RDG) are proposed as a novel approach introduced by RECLAIM to enrich collected waste data with users' own feedback and thus improve the AI algorithms. In turn, better algorithms can filter which collected data is most ambiguous and thus relevant for users' feedback, achieving a self-sustaining (assuming user engagement) cycle of data re-use.

The first version of the RDG has been reported in D6.2 (M9) and the updated version of the RDG, with a closed feedback loop via an online database for collecting and re-using player data, is reported in the updated D6.2 (M18) submitted concurrently. Since the RDG has a multitude of goals (see D6.2), including data collection, awareness, and fun, evaluating all of them would depend on the current development stage of the RDG and would leverage a different evaluation methodology. At the time of writing, a second version of the RDG is fully functional, with data collection integrated through an online database, but the awareness and fun aspects are still in different stages of design and development. Therefore, the evaluation methodology and experimental protocol described in this deliverable match the current stage of development and attempt to measure the usability of the RDG in terms of its data collection potential.

This deliverable is the first report of the activities undertaken under T6.3. As per the DoA, T6.3 *“will validate the RDG, through a series of dedicated pilots with focus groups as well as in-the-wild experiments carried out online”*, allowing for *“(a) targeted and specific feedback regarding the incremental advancement of AI-ILC, and rising recycling awareness (through focus groups) and (b) test RDG usability and data collection potential at scale (through in-the-wild experiments) and the exploitation of gathered data by other project developments”*. In this early phase of the RDG design and development, we conduct dedicated pilots with focus groups as indicated in DoA; however, the methodology however (especially via questionnaires) is scalable to in-the-wild experiments in the future.

This deliverable reports the evaluation methodology devised (along with related work on evaluation for games), and results from a survey and focus groups aiming to capture targeted and specific feedback regarding incremental advances to the RDG in the following months. The evaluation protocol followed is scalable, and thus can be re-used in future evaluation processes for in-the-wild data collection regarding the usability of the RDG. Additional work will be needed, however, to capture other aspects of the experience. We discuss these at the end of this report.

2.1. Intended readership

The present report is a public (PU) document. Its readership is considered to be the European Commission, the RECLAIM Project Officer, the partners involved in the RECLAIM Consortium, beneficiaries of other European funded projects, and the general public.

2.2 Relationship with other RECLAIM deliverables

The methodology and results allowed to capture the usability of the current version of the RDG. These results will certainly inform future design and development of the RDG, which will be reported in D6.2. Moreover, participants contacted in this evaluation step were largely based on focus groups of expert stakeholders used to define the intent and specifications of the RDG (and reported in D2.1). The “raw” data collected via the online database as part of the tests reported in this deliverable (see Section 6) will need to be assessed on how they can refine current AI algorithms, and is strongly linked to WP3 (Recyclable Waste Detection and Categorization). The image and AI data of the RDG prototype used in the reported test were collected as part of D6.1 (both M9 and M18 versions). Finally, since this is the first quantitative and qualitative analysis of the findings of the RDG so far, they can form the basis of a scientific publication on user experience and serious games (WP 7) Table 1 shows the main deliverables consulted (in case of past work), and impacted by (in case of future work) by this report.

Table 1: Other RECLAIM deliverables related.

Del. No	Deliverable Name	WP	Month
1.1	Data management plan and ethics/privacy manual	WP 1	M6/M36
2.1	prMRF and RDG requirements and systems specification	WP 2	M6
3.1	Material recognition based on RGB and Hyperspectral imaging	WP 3	M18
3.2	prMRF operation monitoring and repeating advancement	WP 3	M30
6.1	Waste Data for material recognition and Recycling Data Game	WP 6	M9/M18
6.2	Algorithms and pipelines for Recycling Data Games	WP 6	M9/M18/M30
7.1	Plan for the dissemination and communication activities	WP 7	M6/M18/M36
1.3	Final Project Report	WP 1	M36

In the following sections of this report, we discuss the approach, methodology, and results of the RDG evaluation. We review existing evaluation approaches of user and player experience of digital games, discuss the evaluation protocol applied, data collection tools and methods, data analysis and results, and further examine future work required for the next iterations of the game.

3. Evaluation of Player and User Experience

To identify the most relevant methodology for evaluating the RECLAIM project's RDG, we reviewed existing literature, specifically focusing on games of similar objectives and player experience (e.g. citizen science games, serious games).

This section presents various methodologies for systematically gathering user experience, using standardised questionnaires. Section 2.1 presents alternative methods for gathering user experience feedback in general, applying to any type of product. Section 2.2 focuses specifically on digital games, while Section 2.3 examines the evaluation of player experience regarding serious games.

3.1 User Experience

There is a wide range of standardised usability questionnaires, each with their own characteristics [Assila2016]. Some of them are more generic and therefore applicable to various types of systems, while others are targeting more specific use cases. The following sections [3.1.1 to 3.1.4] describe some alternatives that are generic enough to potentially apply to our use-case, although not specifically designed for serious games.

3.1.1 USE Questionnaire: Usefulness, Satisfaction, and Ease of use

The *"Usefulness, Usability, Satisfaction and Ease of Use"* (USE) questionnaire [Lund2001] is a standardised, non-proprietary tool designed to assess the subjective usability of products or services. Recent studies [Gao2018] have validated the reliability of the USE questionnaire, highlighting its strong correlation with other established metrics like the System Usability Scale (SUS) [Brooke1996]. As explained in [Assila2016], USE is a universal questionnaire, i.e. not bound to specific use-cases. Importantly, it evaluates four critical dimensions of user experience: usefulness, ease of use, ease of learning and satisfaction. Those dimensions have been deemed as particularly relevant during the current phase of the RECLAIM data collection game, offering valuable insights that can steer the subsequent stages of development.

The USE survey consists of 30-items in total, grouped into four categories, as shown in Table 2. Respondents are asked to rate each item on a 7-point Likert scale, ranging from 1 ("Strongly Disagree") to 7 ("Strongly Agree").

Table 2: Questions found in the USE questionnaire, as defined in [Lund2001].

Questions found in the USE questionnaire, as defined in [Lund2001].
<p>Usefulness</p> <ol style="list-style-type: none"> 1. It helps me be more effective. 2. It helps me be more productive. 3. It is useful. 4. It gives me more control over the activities of my life. 5. It makes the things I want to accomplish easier to get done. 6. It saves me time when I use it. 7. It meets my needs.

8. It does everything I would expect it to do.

Ease of Use

9. It is easy to use.

10. It is simple to use.

11. It is user friendly.

12. It requires the fewest steps possible to accomplish what I want to do with it.

13. It is flexible.

14. Using it is effortless.

15. I can use it without written instructions.

16. I don't notice any inconsistencies as I use it.

17. Both occasional and regular users would like it.

18. I can recover from mistakes quickly and easily.

19. I can use it successfully every time.

Ease of Learning

20. I learned to use it quickly.

21. I easily remember how to use it.

22. It is easy to learn to use it.

23. I quickly became skillful with it.

Satisfaction

24. I am satisfied with it.

25. I would recommend it to a friend.

26. It is fun to use.

27. It works the way I want it to work.

28. It is wonderful.

29. I feel I need to have it.

30. It is pleasant to use.

3.1.2 Post-Study System Usability Questionnaire

The Post-Study System Usability Questionnaire (PSSUQ) [Lewis 1992] is a 16-item standardised questionnaire used to measure users' perceived satisfaction of a software, system, website, or product at the end of a study. PSSUQ originated from an internal IBM project called SUMS (System Usability Metrics) in 1988. A few rounds of improvements have resulted in PSSUQ Version 3, which is the one used today. PSSUQ Version 3 (shown in Table 3) consists of 16 questions with 7 options (+ NA option) to choose from. Questions 1 to 6 refer to System Usefulness, 7 to 12 refer to Information Quality, while 13 to 15 refer to Interface Quality.

Table 3: Questions found in PSSUQ V.3.

Questions found in PSSUQ V.3
<p>On a scale between Strongly Agree to Strongly Disagree, please rate the following statements regarding Amazon:</p> <ol style="list-style-type: none"> 1. Overall, I am satisfied with how easy it is to use this system. 2. It was simple to use this system. 3. I was able to complete the tasks and scenarios quickly using this system. 4. I felt comfortable using this system. 5. It was easy to learn to use this system. 6. I believe I could become productive quickly using this system. 7. The system gave error messages that clearly told me how to fix problems. 8. Whenever I made a mistake using the system, I could recover easily and quickly. 9. The information (such as online help, on-screen messages, and other documentation) provided with this system was clear. 10. It was easy to find the information I needed. 11. The information was effective in helping me complete the tasks and scenarios. 12. The organization of information on the system screens was clear. 13. The interface of this system was pleasant. 14. I liked using the interface of this system. 15. This system has all the functions and capabilities I expect it to have. 16. Overall, I am satisfied with this system.

3.1.3 User Experience Questionnaire

The User Experience Questionnaire (UEQ) [Schrepp2015] is a versatile tool designed to evaluate the user experience of interactive products. Developed by a team of researchers led by Martin Schrepp, it allows for quick assessment of a product's user experience across six key dimensions: attractiveness, perspicuity (ease of understanding), efficiency, dependability, stimulation (the degree to which it is exciting or motivating), and novelty (innovation). The questionnaire consists of 26 items that users rate on a scale, providing a comprehensive view of how users perceive the usability and appeal of a product.

The UEQ stands out for its ability to cover a broad range of user experience aspects with relatively few items. It's particularly useful in comparative studies where different versions of a product or different products are being evaluated against each other. The results can give direct insights into the strengths and weaknesses of a product from the user's perspective,

guiding designers and developers in making informed improvements. Table 4 showcases all questions included in the UEQ, grouped by category.

Table 4: Questions found in the User Experience Questionnaire, grouped by category.

Questions found in the User Experience Questionnaire, grouped by category.	
Attractiveness	<ul style="list-style-type: none"> 1. Annoying / Enjoyable 2. Bad / Good 3. Unlikeable / Pleasing 4. Unpleasant / Pleasant 5. Unattractive / Attractive 6. Unfriendly / Friendly
Pragmatic Quality	<ul style="list-style-type: none"> Efficiency <ul style="list-style-type: none"> 7. Slow / Fast 8. Inefficient / Efficient 9. Impractical / Practical 10. Cluttered / Organized Perspiciuity <ul style="list-style-type: none"> 11. Not Understandable / Understandable 12. Difficult to learn / Easy to learn 13. Complicated / Easy 14. Confusing / Clear Dependability <ul style="list-style-type: none"> 15. Unpredictable / Predictable 16. Obstructive / Supportive 17. Not secure / Secure 18. Does not meet expectations / Meets expectations
Hedonic Quality	<ul style="list-style-type: none"> Stimulation <ul style="list-style-type: none"> 19. Inferior / Valuable 20. Boring / Exciting 21. Not interesting / Interesting 22. Demotivating / Motivating Novelty <ul style="list-style-type: none"> 23. Dull / Creative 24. Conventional / Inventive 25. Usual / Leading edge 26. Conservative / innovative

3.1.4 System Usability Scale

The System Usability Scale (SUS) [Brooke1996] is a concise, ten-item Likert scale that provides a comprehensive overview of subjective usability assessments. It was created as a component of the usability engineering program during the development of integrated office systems at Digital Equipment Co Ltd., Reading, United Kingdom. The design of SUS is grounded in the

definition of usability as outlined in [ISO 9241-11:2018]. Accordingly, usability can only be accurately measured by considering the context in which the system is used—specifically, the users, their purposes for using the system, and the environment of its use. Usability measurement encompasses several distinct dimensions: (1) Effectiveness: Can users successfully achieve their objectives?, (2) Efficiency: How much effort and resources are required to achieve these objectives?, (3) Satisfaction: Was the overall user experience satisfactory? Table 5 presents the specific ten questions included in the SUS, each responded to on a five-point Likert scale. The cumulative scores range from 0 to 100, reflecting an aggregate measure of system usability.

Table 5: Questionnaire of SUS.

Questionnaire of SUS
1. I think that I would like to use this system frequently
2. I found the system unnecessarily complex
3. I thought the system was easy to use
4. I think that I would need the support of a technical person to be able to use this system
5. I found the various functions in this system were well integrated
6. I thought there was too much inconsistency in this system
7. I would imagine that most people would learn to use this system very quickly
8. I found the system very cumbersome to use
9. I felt very confident using the system
10. I needed to learn a lot of things before I could get going with this system

3.2 Player Experience in Digital Games

Borrowing from the broader concept of user experience, **player experience** (PX) aims to describe "the individual, personal experience held by the player during and immediately after the playing of the game" [Wiemeyer2016]. Due to the differing goals of a game (to entertain, engage, etc.) compared to a productivity application or a website, and the different emotions that games elicit, "conceptualization of player experience requires differentiating specific dimensions like (game-) flow, immersion, challenge, tension, competence, and emotions" [Wiemeyer2016]. Since both user experience and PX originate from the Human-Computer Interaction (HCI) discipline, PX conceptualization and PX evaluation has focused on digital games as its application domain.

While early research aimed to capture "fun" in games, e.g. via constructs of the FUN construct of Newman [Newman2005], modern views move beyond a singular dimension of "fun" and question the usefulness of the term. Recent research on eudaimonic experiences [Cole2021, Cole2022, Daneels2021] considers more nuanced and subtle player experiences "beyond fun",

such as emotionally moving [Bopp2016], emotionally challenging [Bopp2018, Denisova2020, Denisova2021] and discomforting experiences [Gowler2019].

A plethora of research methods is used in the digital games industry for evaluating PX, as surveyed by Medlock [Medlock2018]. Importantly, different methods are applicable in different stages of game development. One way to assess one's PX is by looking at the 'objective' data in the form of physiological responses from players, such as heart rate or skin conductance. The downside of this approach is the lack of subjective context, i.e. why participants are feeling something and what it is that they are experiencing or thinking exactly. To address this shortcoming, qualitative evaluation methods can be used to complement these 'objective' responses, including interviews, focus groups, and ad-hoc surveys; however, results from these methods can lack standardisation and comparability. Validated questionnaires, on the other hand, exist to address this issue.

Questionnaires are perhaps the most common method for assessing subjective experiences of playing digital games. These instruments can quantify experiences and compare these experiences between groups of players or between sessions. Questionnaires are usually created based on a specific theory with a view to test and refine this theory and to be able to compare experiences across different games, features, and player types. The most common experiences that are measured through questionnaires are motivation [Azadvar2018, Ryan2006, Yee2012], immersion [Jennett2008], engagement [Brockmyer2009], flow [Jackson1996], spatial presence [Vorderer2004, Witmer1998], social presence [DeKort2007, Hudson2014], and overall gaming experience [Abeele2020, IJsselsteijn2013]. Specialised questionnaires assessing more nuanced PX include challenge [Denisova2020], demand [Bowman2018], attribution of failure [Depping2017], character attachment [Lewis2008], character morality [Graham2011, Joeckel2012], player-avatar interaction [Banks2016, Banks2019], creativity [Hall2022], embodiment [Peck2021], uncertainty [Power2017], fantasy [Choi2013, Plante2017], and more.

Despite extensive work and validated questionnaires for these play-specific notions, we acknowledge that the current state of the RDG is not targeting such specialised experience constructs. Moreover, using intrusive “objective” data collection methods would likely not answer our core evaluation questions and moreover would not scale for in-the-wild data collection experiments planned for the RECLAIM DoA. Questionnaires, however, are designed to be scalable and it is worth considering whether such of the specialised PX questionnaires surveyed could be used for capturing player engagement and flow, as well as awareness in future versions of RDG that are explicitly designed to target these experiences.

3.3 User Experience in “Serious” games

As surveyed by [Deterding2011], games used for serious purposes or “serious games” date back several millennia [Halter2006], migrating from mainly military uses into education and business in the second half of the 20th century. In the early 2000s, the rise of digital games has reinvigorated this into a substantial industry and research field of its own. Such digital, serious games can be defined as “any form of interactive computer-based game software for one or multiple players to be used on any platform and that has been developed with the

intention to be more than entertainment” [Ritterfeld2009]. Since all of the goals (see D6.2) of the Recycling Data Game transcend entertainment, observing how user experience is assessed in serious games is useful for designing a methodology for evaluating the RDG in the coming years.

The findings of different systematic literature reviews in surveys on the evaluation of serious games—which included papers focused on the assessment of their educational effectiveness—show that questionnaires are the main assessment method (in 90% of cases surveyed), followed by interviews [Calderon2015], or quantitative approaches such as the quasi-experimental design [Boyle2016]. The most commonly assessed quality characteristics include usability, learning outcomes, and user engagement [Connolly2012]. When questionnaires are used as an evaluation instrument to assess educational effectiveness, they need to be standardised [Paliokas2016] or mapped to an accepted education framework [Coenen2013].

Quantitative quasi-experimental methods often adopted to assess games’ learning effectiveness [Backlund2013] may not have the rigour of controlled experiments, but they maintain the argument and logic of experimental research. Whether the research design is experimental or quasi-experimental, the most common strategy in this type of investigation is a comparison between groups: one group provides baseline information (acting as the control group), whilst the other group is given the experimental treatment [Froschauer2013, Volkmar2018].

Another key method in quantitative research is the (large- or small-scale) survey [Andreoli2017] as it provides information on the distribution of a wide range of respondent characteristics. However, it is argued [Eliean1999] that small-scale surveys often under-report the methodology and findings; most importantly, such surveys remain within the institution, inaccessible and unavailable.

One of the most important sources of information in qualitative research is the interview, which can take a number of forms, including open-ended, focused, or survey [Kiili2007]. In-depth interviews are typically a means of understanding users’ experience and gaining insights from individuals. Sometimes, interviews use focus groups when the aim is to detect users’ behavioural patterns or insights into their attitudes and perceptions. Significant problems common to all semi-structured and unstructured interviews are issues of memory; especially in museum studies where participants must remember their entire experience throughout the museum visit [Macdonald2007], memories of the visit are partial and visitors cannot always remember what they have just seen. In such cases, users are eager to please and will offer answers they think the interviewer expects to hear. In our case, since the game is a much shorter playthrough compared to a museum visit, we expect that memory plays a lesser role. However, during focus group discussions (see Section 5.2) we also remind focus group participants of the games through a live demo of the game by one of our team members.

4. Evaluation Methodology

As previously described, quantitative and qualitative data were collected for the evaluation of the RDG by a sample of users. Two data collection approaches were used: an online survey and a focus group. This approach would allow us to systematically evaluate specific aspects of the game as well as gain a more in-depth understanding of the player experience. This evaluation approach and methodology, as well as the analysis of the data, is scalable and appropriate for future evaluations of the RDG with larger samples and populations.

4.1 Data Collection

The online survey was designed based on the survey by [Lund2001]. Following up on the review of existing surveys discussed in the previous section, we selected Lund's questionnaire as it included all the relevant to our study constructs and adapted for the needs of this project. The first section of the survey contained detailed information about the project, the confidentiality and anonymity of the respondents, and contact information. At this point, the informed consent of the participants was required, before moving to the next section of the survey. The next section contained information about the testing protocol and namely, the links for downloading the game, guidelines on how to play the game for the purposes of the testing (i.e. complete all the 10 challenges, guide on the materials presented in the game, guidelines on how to play the game), and guidelines on the process of testing (i.e. complete the game and then complete the survey). The final section of the survey included the survey items. Further to the items regarding the evaluation of the Ease of Use, Ease of Learning, and Satisfaction as the main constructs evaluated, demographic details (gender, age) and information about their previous experience with digital games, citizen science games, and environmental activism were collected. Three open-ended questions were also added to the survey where participants could express their insights about the positive and negative aspects more freely (Q1 List the positive aspects of the game, Q2 List the negative aspects of the game, Q3 Do you have any other comments?). This would allow us a more in-depth understanding of their perceptions towards the game. The survey was accessed online at <https://forms.gle/S3gcjg3NEVpPaYNK9> and the questions included were the following:

Survey Items

- 1) Do you have any experience with digital games (e.g., on mobile, PC, game consoles)? (1 Not at all-4 Yes, I play a lot of games)
- 2) If you have played any digital games, can you write some of their titles? (open-ended question)
- 3) If you have played digital games, which one would you say was your favourite game? (open-ended question)
- 4) Do you have any experience with citizen science games (e.g., Foldit, EyeWire, Zooniverse)? (1 No I have not participated in any environment protection activities – 7 Yes, I have played citizen science games)
- 5) If you have played any citizen science games, can you write some of their titles? (open-ended question)

- 6) Are you involved in any environmental activities or environment protection groups? (1 Not at all - 4 Yes, I am an activist in environment and sustainability activities.)
- 7) What is your gender? (Female, Male, Non-binary, Prefer not to say)

EASE OF USE (1 strongly disagree – 7 strongly agree)

- 8) It is easy to use.
- 9) It is simple to use.
- 10) It is user friendly.
- 11) It requires the fewest steps possible to accomplish what I want to do with it.
- 12) It is flexible.
- 13) Using it is effortless.
- 14) I can use it without written instructions.
- 15) I don't notice any inconsistencies as I use it.
- 16) Both occasional and regular users would like it.
- 17) I can recover from mistakes quickly and easily.
- 18) I can use it successfully every time.

EASE OF LEARNING (1 strongly disagree – 7 strongly agree)

- 19) I learned to use it quickly.
- 20) I easily remember how to use it.
- 21) It is easy to learn to use it.
- 22) I quickly became skillful with it.

SATISFACTION (1 strongly disagree – 7 strongly agree)

- 23) I am satisfied with it.
- 24) I would recommend it to a friend.
- 25) It is fun to use.
- 26) It works the way I want it to work.
- 27) It is wonderful.
- 28) I feel I need to have it.
- 29) It is pleasant to use.

Comments (open-ended questions)

- 30) List the positive aspects of the game
- 31) List the negative aspects of the game
- 32) Do you have any other comments?

A focus group was further organised where the participants could elaborate further on the positive and negative aspects of the game. The duration of the focus group was 1 hour. 7 people participated. They were experts in AI and environmental sustainability. After the welcome and introductions and as a reminder, the gameplay of the RDG game was presented for approximately 10 minutes by one of the researchers. The focus group followed a semi-structured interview approach. The main 2 axes discussed were a) positive and b) negative aspects of the game. Further questions were addressed to the participants for clarifications

e.g. their experience with the interface and the interaction with the game, the aesthetics, the learnability of the game, the format of the tasks, and the motivating or demotivating elements of the game. The focus group took place online via Zoom (for the ethics see section 4.5).

4.2 Participant recruitment

We addressed the invitations for the survey and the focus group to the same sample of people that participated in the game design requirements survey and focus group (see D2.2). We made this decision in order to ensure a consistency and continuity of the feedback and engage participants already familiar with the goal and requirements of the game.

4.3 Data Analysis

Quantitative data from the survey were analysed using SPSS29. Mainly descriptive analysis and correlation analyses were conducted.

The qualitative data of the survey (i.e. open-ended questions) and the feedback from the focus group (transcript) were analysed through a thematic analysis [Braun2006] to identify trends and patterns emerging from the experience of the participants with the game. In this case, indicative excerpts (quotes) of the participants' comments are included, to establish the validity of the themes.

4.4 Material

The participants of the survey and focus group were first invited to download the game and the accompanying guide, before playing the game (following the guidelines) and participating in the survey and/or focus group.

The guide aimed to familiarize the participants with the game interface and the process of the trials. The goal of the trial was that the participants complete all 10 in-game challenges (described in Section 4.4.2). The steps and relevant resources for the participants were:




1. Download the user guide and read carefully
<https://drive.google.com/file/d/1lqZViCBso4A2VzN3TLxvGEYdAChJbaC6/view?usp=sharing>.
2. Download and install the application
<https://drive.google.com/file/d/1lqZViCBso4A2VzN3TLxvGEYdAChJbaC6/view?usp=sharing>
3. Complete the testing session, as prescribed in the user-guide.
4. Return to this form, to complete the survey.

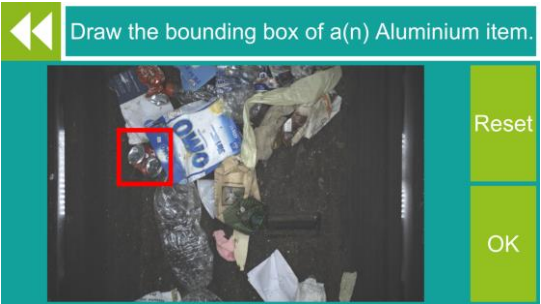

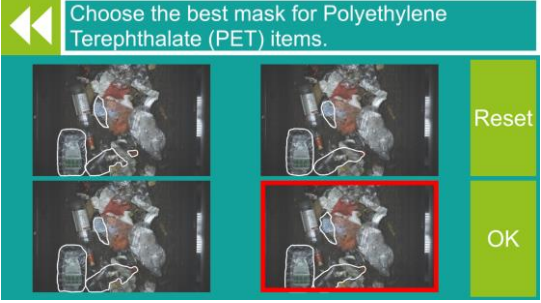

In addition, the focus group participants, before participating in the focus group, they had to submit their informed consent which detailed information about the process, data collection and management, confidentiality, and anonymity, following the University of Malta Research Code of Practice (<https://www.um.edu.mt/media/um/docs/research/urec/ResearchCodeofPractice.pdf>) and the University of Malta Research Ethics Review Procedures (<https://www.um.edu.mt/media/um/docs/research/urec/ResearchEthicsReviewProcedures>).

[pdf](#)). The informed consent form was uploaded online: <https://forms.gle/X4csbBwtqsr4LSCM9>.

4.4.1 The game

The Recycling Data Game will be a series of challenges for the player, thus offering short interactions that can be paused in-between challenges. Below, we describe the developed annotation challenges, each of which is described in D6.2 and D6.5. We include screenshots for each game based on the updated build of the game, which features functionality improvements and technical fixes found during internal testing.

<p>Paint: the user must highlight all items of a specified material in each image using their finger (via a paintbrush and an eraser tool).</p>	
<p>Detect: the user must answer whether they can detect any item of a specific material in each image, using a Yes or No button.</p>	
<p>Count: the user must answer how many items of a specific material they can see in each picture, using a + and - button to increase/decrease the number.</p>	

<p>Outline: the user must choose one item of a specific material and draw a bounding rectangle around it, using their finger.</p>	
<p>Locate: the user must identify the center of a single item of a specific material using their finger (and a helper target graphic).</p>	
<p>Choose: the user is shown four different AI-generated masks around objects of a specific material, and must choose the best one among them.</p>	
<p>Categorize: the user is shown one AI-generated mask for all identified materials, and must choose which material each mask is via a "material" colour palette.</p>	

4.4.2 Additions to the annotation challenges

The main additions to the challenges have been on the graphics side before and after the actual annotation challenge. We present all challenges' graphics changes below as they are very similar across challenges.

The introduction screen now offers more information about the challenge, including the material and the basic operation of the game.

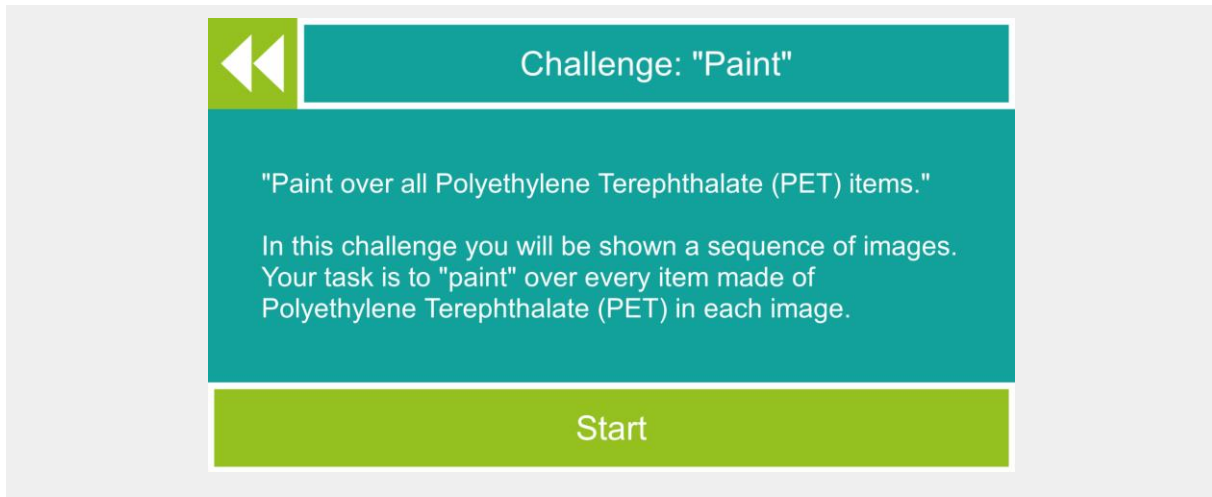


Figure 1: Intro Screen of the “Paint” challenge, explaining the basic operation of the game.

When the player has finished annotating an image, they press the Submit button, and submit their response to the server. The uploading process takes a few milliseconds, displaying a loading screen while in progress. After the user’s response is submitted, it is compared against other players’ responses to the same challenge, and a score is granted to the player accordingly, as shown in Figure 2. Detailed information regarding how the player’s response is evaluated can be found in D6.5. This process is repeated for a sequence of 5 images. As soon as the session is complete, the player is shown a thank you message, as shown in Figure 3, explaining that the session is over and that their valuable input will be used to train better AI models.

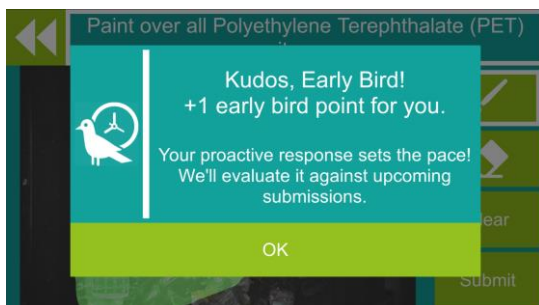


Figure 2: The user’s response is evaluated and the user is assigned a score.

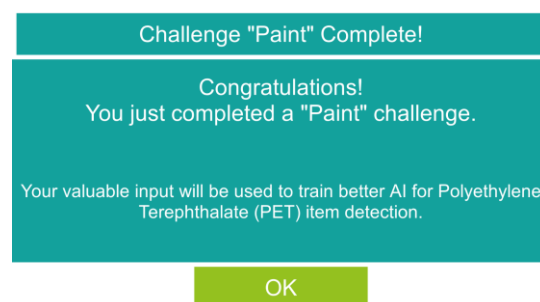


Figure 3: Outro Screen of the “Paint” mini-game.

In addition, a player profile is added to the game. While this player profile will be enhanced in future work to increase the engagement (via gamification elements such as achievements etc., as discussed in D6.5), it serves an important purpose currently to inform the player about their rewards (points) in the different challenges. We revisit how points are calculated per challenge in D.6.5. The player profile is shown in Figure 4.

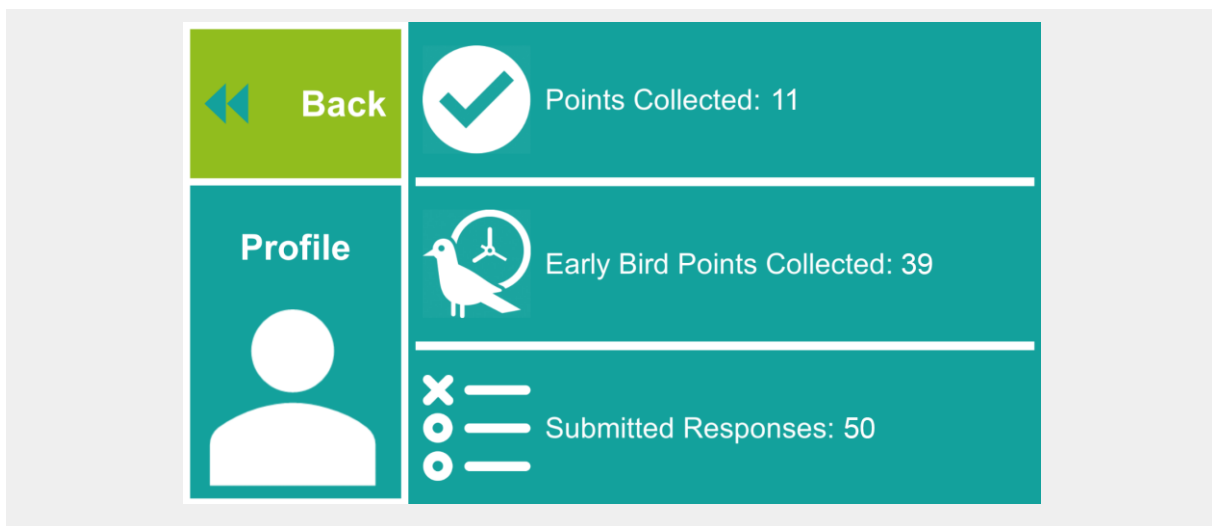


Figure 4: Player profile after completing all challenges.

Since the current game is intended to be evaluated for usability, some changes to the main interface were made. While in earlier versions of the RDG all mini-games were available on the start page, now the user must perform each challenge sequentially. The sequence of challenges is hand-crafted to ensure the easier challenges are first (e.g. detect, count) and on simpler images (e.g. images of isolated streams for PET bottles, as described in D6.4), increasing in difficulty and also moving to other materials (indicatively, glass, LDPE) and on mixed waste images where any challenge is more difficult. When the user exits the game, the system stores which challenge they were in. When the player re-opens the RECLAIM RDG, the system restores the player's local load file (their last completed challenge) and checks their player profile from the server, thus letting the player continue where they left off. In total, 10 challenges are implemented in the following order:

1. **“Count”** PET items (data: isolated stream with PET Only)
2. **“Locate”** PET items (data: isolated stream with PET Only)
3. **“Outline”** PET items (data: isolated stream with PET Only)
4. **“Detect”** HDPE items (data: Mixed Materials stream)
5. **“Count”** GLASS items (data: Mixed Materials stream)

6. **“Locate”** HDPE items (data: Mixed Materials stream)
7. **“Outline”** GLASS items (data: Mixed Materials stream)
8. **“Choose”** PET items (data: Mixed Materials stream)
9. **“Paint”** PET items (data: Mixed Materials stream)
10. **“Categorize”** all materials (data: Mixed Materials stream)

4.5 Ethics

This study was reviewed by the Faculty Research Ethics Committee, was deemed to be in conformity with the University of Malta’s Research Code of Practice and Research Ethics Review Procedures, and consequently approved.

5. Analysis and Results

5.1 Survey

Although our survey sample is extremely small for any statistically significant findings and generalisations, we used SPSS for descriptive analyses and correlations, so as to draw more objective conclusions about any trends and themes emerging.

5.1.1 Demographics of the Sample

The sample size was 11 respondents. Their average age was 42 years old, ranging from 27 to 57 years old. 9 were male, 1 was female, and one preferred not to say. The participants had varying levels of expertise with digital games on any console ($M=2.55$, $SD=.93$), as measured in a 4-point scale). They had very little experience with citizen science games (e.g., Foldit, EyeWire, Zooniverse), ($M=1.18$, $SD=.40$), and varied levels of involvement with environmental activities or participation in environmental protection groups ($M=2.45$, $SD=.934$).

The different levels of expertise and environmental engagement of the participants ensured valuable insights that may help refine the game to target a mixed and wider audience.

5.1.2 Game Preferences

The game preferences of the participants were examined through an open question, where they could list games they have played (“If you have played any digital games, can you write some of their titles?”).

The games reported by the participants ranged from more casual, simple, and easy to learn to more complex and sophisticated games. Indicatively, the participants reported casual games such as Mahjong, Candy Crush, Tetris, Solitaire, and Chess, strategy games such as Age of Empires, Stellaris, Europa Universalis, simulation games such as The Sims, Flight Simulator, action/adventure games such as Super Mario, Warcraft, Grand Theft Auto, sports games such as FIFA, and music games such as Guitar Hero. This range indicates that the respondents had diverse and a broad range of interests, from casual, relaxing gaming and temporary entertainment to more specialised, complex, immersive, and intellectually challenging games. This broad range of gaming interests among the participants is again valuable for the evaluation of the RDG. Understanding the perceptions and attitudes of diverse players and their preferences can be crucial in tailoring the game development to meet diverse user needs and expectations.

5.1.3 Perceptions of the players on the Ease of Use, Satisfaction, and Ease of Learning

For examining the perceptions of the participants on the three main constructs of the evaluation (i.e. Ease of Use, Satisfaction, Ease of Learning) the mean scores of the participants for all the construct items were calculated. The descriptive analysis shows average to positive attitudes for each construct (Table 6). Most users found the app easy to use and easy to learn, and while on average they were less satisfied it is still above the mean of 4 (given the 7-point Likert scale used).

Table 6: Descriptive Statistics of the main constructs.

Descriptive Statistics of the main constructs					
	N	Min.	Max.	M	SD
Ease of Use	9	1	7	5.45	1.86
Satisfaction	10	3	6	5.05	1.24
Ease of Learning	9	1	7	5.94	1.89
Valid N (listwise)	9				

Most of the participants were generally positive (fairly agree, agree, strongly agree) towards all the statements of the 3 constructs. We list frequency of each response for each statement in figure 5 below.

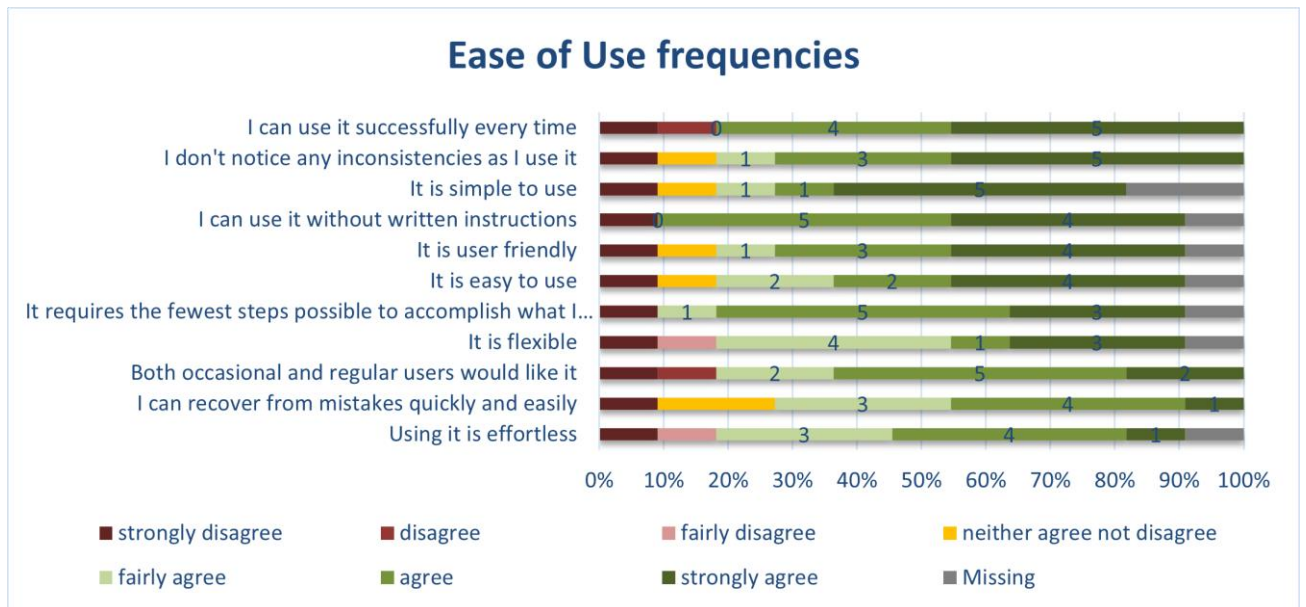


Figure 5a: Frequency of responses for Ease of Use

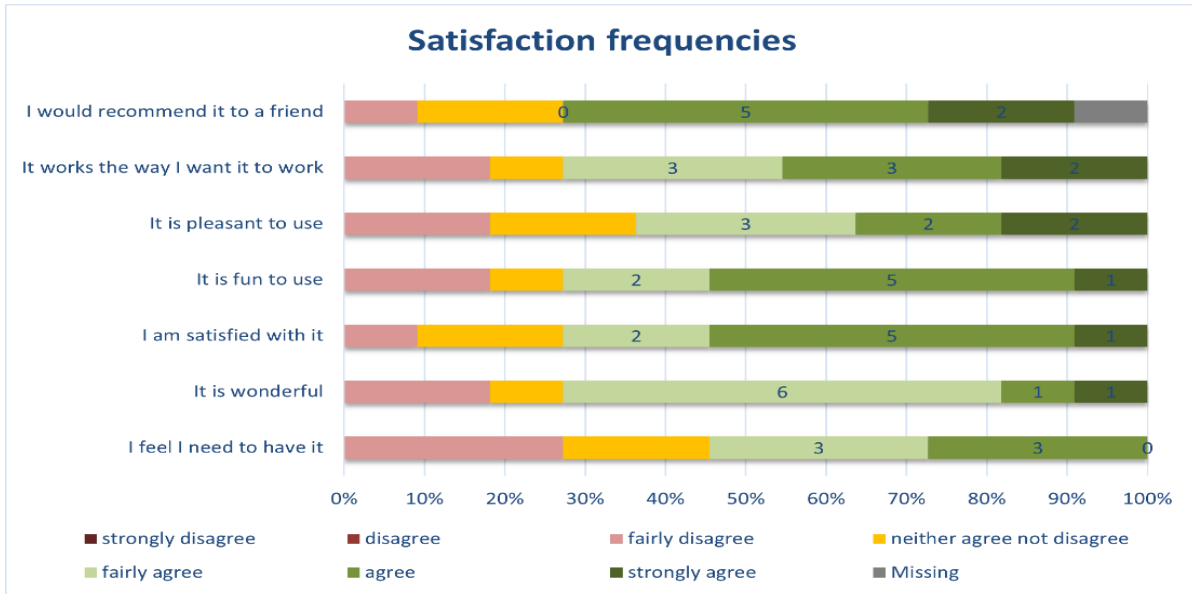


Figure 5b: Frequency of responses for Satisfaction

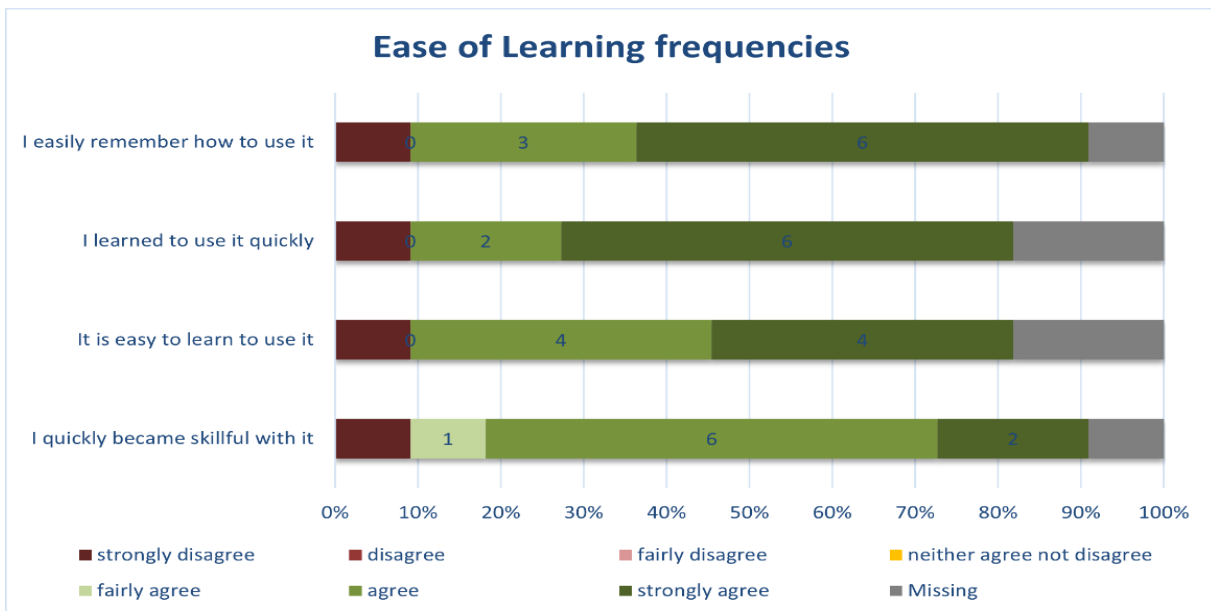


Figure 5c: Frequency of responses for Ease of Learning

5.1.4 Game Experience, Age, and Environmental Experience in relation to Ease of Use, Ease of Learning, and Satisfaction

Although our sample is quite small for generalisations, it seemed that *the participants who reported higher gaming experience also reported lower perceived ease of use*. A Pearson correlation coefficient was computed to assess the linear relationship between experience

with digital games and Ease of Use. There was a negative correlation between the two variables. ($r(7) = -.567, p = .111$). This could imply that more experienced gamers have higher standards or expectations for ease of use, or it could mean they are more critical in their evaluations.

Similarly, although again with no statistical significance, *more experienced gamers reported lower levels of satisfaction* by the game. A Pearson correlation coefficient was computed to assess the relationship between experience with digital games and player Satisfaction. The analysis revealed a negative correlation. ($r(8) = -.375, p = .286$). This suggests a weak to moderate negative correlation, indicating that as gaming experience increases, satisfaction ratings tend to decrease slightly. This may also be an indication that experienced players have higher expectations and a more critical eye when assessing the game, leading to lower satisfaction scores.

Games experience was further moderately and negatively correlated with the ease of learning the game, with no statistical significance ($r(7) = -.507, p = .164$). This finding suggests that participants with less gaming experience tended to rate the Ease of Learning as higher. It is encouraging to see that participants who were more novice players found the game less challenging and more straightforward to learn.

The age of the participants did not seem to have any significant impact on the 3 main constructs in our sample (very weak correlations). The age of the participants did not influence their perceptions about the game.

Participants with more experience with environmental activities tended to find the game easier to use ($r(7) = .470, p = .201$) and learn ($r(7) = .564, p = .113$), and were slightly more satisfied by the game than less experienced players ($r(8) = .341, p = .334$). This finding is mirrored by the comments received during the focus group discussion (described in the next section); focus group participants involved in environmental organisations expressed their satisfaction that the game focuses specifically on facilitating research on recycling and on raising public awareness for the process and implications of recycling.

For assessing the positive and negative aspects of the game in more depth, two open-ended questions were included, where the participants were free to discuss such aspects. Through the thematic analysis of the positive aspects of the game the following elements emerged:

5.1.5 Positive Aspects

Educational Value: in 4 cases, the participants appreciated the educational aspect of the game and particularly the information about recycled materials (*“learn the categories of recyclables, learn the abbreviations of materials (PET.PE.LDPE etc)”*).

Ease of Use: In 6 cases, the game was described as easy to understand and user-friendly (*“It is really easy to understand. I really had a nice time while playing”*)

Engagement: Three of the participants found the game fun and engaging due to its variety and challenge level (*“It provides a fun and easy experience”*).

Contribution to Science: In one case, the participant acknowledged and appreciated the value of the game for the data collection and analysis (*“I really had a nice time while playing and I would be willing to do it again if it would help collecting data.”*).

5.1.6 Negative Aspects

The main areas discussed in this section were technical and interface aspects, and the content engagement. Specifically:

Technical Aspects: multiple participants mentioned that the game was very slow in providing feedback and moving to the next screen. This problem has more to do with the network and less with the game performance itself and is expected to be optimized with the appropriate image compression in the next versions of the game. Other drawbacks mentioned were the visual design and specifically the images size which was too small for the mobile screen and hard for the players to see clearly and interact with effectively (*“the object seem quite small so most of the times the answer was not clear to me because of the size”*). The Clear button was also commented on in one case (*“at mask drawing the clear button could have a second verification so it wont erase all mask at once by mistake.”*). These could be addressed with an adjustment of the interface and the selection of the appropriate images.

The second axis of comments was relevant to the **engagement with the content:** in one case, the participant suggested a *“better game environment”* which could refer to the aesthetic, thematic, or narrative settings of the game. In another case, one participant referred to the Mask challenge (*“I think that the MASK CHALLENGE was the least interesting”*) indicating that this segment of the game might not be as engaging or relevant to the player's interests (note that the “Mask challenge” refers to the *Choose* challenge, described in section 4.4.1).

The feedback indicates that while there are specific aspects of content that could be improved, most of the criticisms are related to technical aspects of the game. This suggests that before adding new content or dramatically altering the game's mechanics, it would be beneficial to focus on improving the overall quality and user experience through technical enhancements to the visuals and interface. Addressing these concerns might not only resolve specific complaints but could also indirectly improve content engagement by making the game more enjoyable and easier to interact with.

5.1.7 General Comments

The overall impression of the comments added to the 3rd open-ended question (Do you have any other comments?) was positive suggesting a general approval and a replay value of the game. Comments such as *“Very nice job :)”*, *“generally i liked it and i would play it more!”*, and *“very nice game”* **express a satisfaction with the game** and a willingness to engage with it further. Most of the comments indicated **areas of improvement and suggestions** on the technical and aesthetic aspects of the game such as *“It would be great to see some strange global statistics on waste production or management”* and *“image should be bigger (full screened)”*. In summary, the comments provide a mixed view with some participants indicating overall satisfaction and others suggesting specific areas for improvement. Technical

and aesthetic quality concerns remain prominent and there's a desire for more content or features to enhance the game's educational and engagement potential.

5.2 Focus group

5.2.1 Participants

The participants of this focus group were recruited by the pool of participants of the previous focus groups on the specifications and requirements of the game (D2.1). This ensured the coherence, consistency, and continuity of the feedback and engaged participants already familiar with the goal and requirements of the game. 7 people participated. They were experts in AI and environmental sustainability, as shown in Table 7.

Table 7: Details on the process and participants of the focus group

Label	Date	Number of Participants	Main expertise
FocusGroup1	12/Apr/2024	7	AI, Recycling, Games

The feedback gathered from the focus group appears to align with the themes identified in the survey's positive and negative responses. Participants appreciated the educational value and the potential for fun, but desired a more refined user experience, particularly regarding the technical performance and visual clarity. The suggestions for additional features and mechanics highlighted a desire for a more engaging and interactive experience. The comments also reflect an interest in competitive elements, such as rankings and time-based challenges, which could increase the game's appeal and replay value. More specifically, through a thematic analysis of the focus group transcript, the following themes were identified:

5.2.2 Technical Aspects and Usability

Issues with the game speed and responsiveness were noted such as delays in receiving feedback. Some of the participants noted significant loading times, while for other participants, there was not such issue (*"I encountered this in the 1st challenge. After that they run quickly."*, *"It didn't stick anywhere."*). As also discussed in the previous section, this relates to the network and will be resolved with the appropriate image compression. Technical difficulties were further noted when multitasking on devices; the game does not appear in the list of open apps. Image size and the need for a zoom feature for better visibility was also commented here (*"I struggled to see the image clearly. I tried to zoom in. I saw that it wasn't possible."*), as by the survey participants. In one case, the Clear button functionality was again mentioned indicating a need for a confirmation step to prevent progress loss (*"the clear button should have an exit that says 'are you sure?'. Because if you've made progress and you accidentally press clear, you've lost your progress and there you'll get quite frustrated."*), and the Profile button on the starting screen (*"The profile confused me at the beginning of the game. As soon as you enter, you see the profile and I thought it was something important that I needed to set up the profile"*). Again, these elements suggest slight modifications to the interface design. There was also as a strategic improvement recommended: the availability of

the game for iOS platforms to reach a wider audience, although this raised concerns about licensing issues.

5.2.3 Game Content and Mechanics

The participants of the focus group requested more sophisticated game elements such as statistics, ranking, fun facts, educational snippets and feedback to motivate continued play and competition and to enhance learning (“Yes, we want fun facts something to learn.”, “a fun fact regarding the material part e.g., recycling glass saves so many lamps something like that”). It has to be noted that fun facts, statistics, and educational snippets have already been designed and developed for the game (as described in D6.5) and will be implemented in the next version of the game. Certain challenges of the game, such as the Choose Challenge, received mixed reviews by the participants.

5.2.3 Positive Elements

Summarising the positive feedback and the game elements appreciated by the participants, the following themes emerged:

Simplicity, Ease of Use Simplicity, and User-Friendliness: Participants found the game generally easy to understand and play. This ease of use was noted as a strong point, making the game accessible to new users (“The game was easy. You understood what you had to do.”, “For the first question it seemed quite simple to me. In my case it went very fast.”). The game was further described as simple, user-friendly, and fast, which are crucial factors for keeping players engaged without causing frustration (“It’s simple user-friendly fast and pleasant.”).

Engagement and Educational Value: The game was seen as potentially engaging or “addictive”, especially with features that could rank players and show their performance over time. This aspect was valued as it could encourage players to improve their skills and invest more time in the game. Educational content, such as fun facts about materials, was suggested and appreciated as a way to enhance the learning experience while playing (“If it throws an index that says you have so many correct in so much time and you’ve entered the top charts it encourages you to improve the time.”, “Yes we want fun facts something to learn.”)

Visual Design and Aesthetics: The visual design, including vibrant colours and the overall aesthetic, was highlighted as appealing. This aspect of the game contributed to a pleasant user experience (“Another positive for me was the colours. Because it’s so simple and minimal the image affects me a lot and the fact that the colours were so vibrant helped me.”).

Game Variety and Fun: The variety in game tasks and the different challenges offered were seen as positives, making the game more interesting and less monotonous. Participants liked that the game wasn’t repetitive and offered different types of interactions. The game’s fun factor was frequently mentioned, with some games within the RDG game particularly highlighted as enjoyable (“I find it pleasant that there was variety. It wasn’t one thing you were forced to play the same thing over and over.”, “The last challenge is the most fun to play.”).

Game Concept and Theme: The theme of recycling and sustainable development, and the related tasks were particularly appreciated by users who were interested in environmental topics. This relevance to personal interests and everyday life was seen as a positive aspect (*“A positive is that it deals with a subject that I really like. I don't know if there is another game [like this] in the market. Just the fact that it deals with something I am involved with in everyday life is positive for me.”*, *“About recycling which I think should interest more people.”*).

Overall, most of the participants found the game enjoyable and would be open to playing more. Criticisms focused on aspects that could be improved, like the profile setup and navigation between games. The concept of incorporating a competitive element and social features such as sharing achievements with friends was mentioned as potentially motivating. Innovative ideas for game mechanics were further suggested, such as new challenges and mini-games (e.g., a conveyor belt simulation for a more immersive experience).

5.2.4 Conclusions

Throughout the discussion, the participants felt comfortable expressing their insights. Having already participated in the previous phase focus group, they built upon each other's comments (*“I liked what N said about time. Indeed, it would make some games more interesting. That I need to do something within the time and not just do it correctly.”*), agreed or disagreed in certain aspects and moved the discussion forward. The consensus seemed to be a positive overview of the game and the desire to become even better and more engaging so as to reach a wider audience (*“A positive is that it deals with a subject that I really like. I don't know if there was another one in the market. Just the fact that it deals with something I am involved with in everyday life is positive for me. About recycling which I think should interest more people.”*).

This thematic analysis highlights key areas where participants felt improvements could enhance their experience. These insights can guide the development team in prioritizing which features to refine or introduce in future updates. The positive feedback reflects well on the game's design and execution, indicating strong points that can be leveraged in further development and game dissemination to enhance user satisfaction and engagement.

6. Data Collected during Evaluation

We analyze below the actual annotations, stored in the RECLAIM database (see D6.5 for details), that were collected during the period of both the survey and the focus group. As noted above, the survey included 11 participants playing with the RDG, but additional data during demonstrations for the sake of the focus group (and internal testing) were also collected during this period.

Table 8 summarizes the data collected through the use of the RECLAIM data game so far. 17 users have interacted with the game, considering 13 different images from the conveyor belt, in various Challenges (and their corresponding modes of annotations). We note that multiple challenges use the same images, as a way to assess whether the challenges make annotation more or less difficult on the same data. In total, 1031 annotations were stored in our database. The majority of annotations (96.02%) was considered to be an Early Bird annotation, while 15.52% was considered correct, and 4.07% was considered to be incorrect. The large amount of Early Bird annotations is due to the relatively small number of users so far. As more users are involved in the game, this number will reduce. We analyse this further below. Other than that, we can observe that the number of correct annotations is significantly higher than that of the incorrect ones. This shows a high degree of inter-user agreement, which is a positive observation, supporting the overall strategy of deriving a ground truth based on a group consensus among players. It also validates, in part, the strategy used for deriving the ground truth in each challenge, presented in D6.5. However, there are more insights when we analyze this data on a per-challenge basis below.

Table 8: Overview of data collected so far through the RECLAIM data collection game, across all types of challenges.

Images: 13		Annotators: 17	
All Annotations	Early Bird Annotations	Correct Annotations	Incorrect Annotations
1031	990 96.02%	160 15.52%	42 4.07%

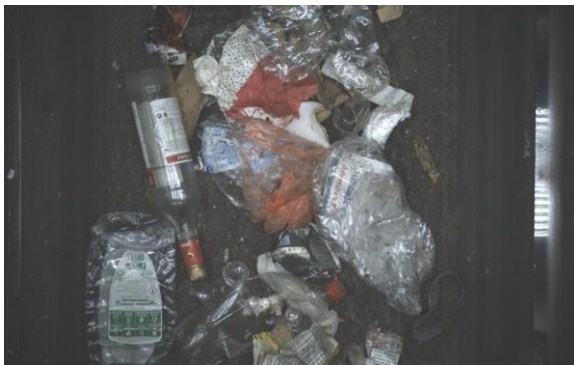
Table 9 presents the same statistics per Challenge, which allows us to extract a more precise view of the results so far. All challenges have been taken by 17 annotators, although since some challenges were presented for both isolated and mixed streams (see Section 4.4.2), the number of images annotated per challenge is not the same. We note that most annotators have gained early bird points, which are assigned when a “definitive” ground truth (based on a threshold of agreement among annotators so far) has not been achieved. It is expected that the first few annotators will all invariably gain early bird points, which is verified as over 50% of annotations are awarded early bird points (and as high as 100% for Outline and Locate challenge). The high early number of early bird points for the Outline and Locate challenge indicates that the way of calculating group consensus (presented in D6.5) is especially strict

and should be more lenient in these two challenges. These two challenges are comparing coordinates (either the center of an object or the center of a bounding box) of an image that likely contains more than one of these types of objects. It is very likely, therefore, that each user chooses a different object to locate its center (Locate challenge) or identify its bounding box (Outline challenge). This points to revisions that should be done on these types of challenges on the development side. A similar high number of early bird points is found for Categorize challenge, which is less surprising due to the fact that for this type of challenge agreement is calculated per item rather than on the entire image (and there are many items in each image) which means that reaching sufficient group agreement for all items is unlikely. By comparison, we note that the “easiest” challenges to reach group consensus are the Detect challenge and the Count challenge. For the Detect challenge, this is not surprising since the options for the user annotations are True (there is an item of this type in this image) and False (there is no item of this type in this image); reaching consensus is trivial on a 50/50 split. For the Count challenge, however, it is promising that not only is a group consensus easily reached (even though the user can put any number from 0 to infinity as a response to this challenge), but also that high rates of correct responses (i.e. users annotating the same number of objects of this type in the image) are attained. Surprisingly, players were adept at identifying the number of items correctly; we should explore how this holds in more diverse sets of images (e.g. including more noise or more diverse/mixed streams). A positive surprise comes from the Paint challenge, which is somewhat involved (the user paints all objects of a specific type on their mobile phone with their finger); therefore, the fact that group consensus is reached (only 69% of annotations receive early bird points) and more importantly that users tend to agree with the ground truth (25% of annotations are correct according to the group consensus) is very promising. However, this may also indicate that calculations of group consensus are too lenient (in the same way that for Outline challenge it is too strict) and we will be reviewing this in future iterations.

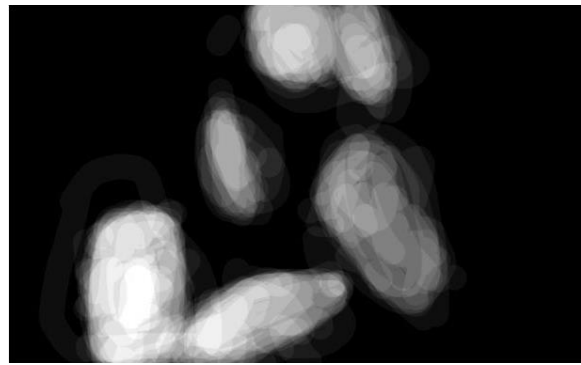
Table 9: Overview of data collected so far from each challenge of the RECLAIM data collection game.

Detect Challenge			
Images: 8		Annotators: 17	
All Annotations	Early Bird Annotations	Correct Annotations	Incorrect Annotations
110	70 (63.6%)	30 (27.3%)	10 (9.1%)
Count Challenge			
Images: 13		Annotators: 17	
All Annotations	Early Bird Annotations	Correct Annotations	Incorrect Annotations
228	132 (57.9%)	85 (37.3%)	11 (4.8%)
Outline Challenge			
Images: 13		Annotators: 17	
All Annotations	Early Bird Annotations	Correct Annotations	Incorrect Annotations
215	215 (100%)	0 (0%)	0 (0%)
Locate Challenge			
Images: 13		Annotators: 17	
All Annotations	Early Bird Annotations	Correct Annotations	Incorrect Annotations
231	231 (100%)	0 (0%)	0 (0%)
Paint Challenge			
Images: 13		Annotators: 17	
All Annotations	Early Bird Annotations	Correct Annotations	Incorrect Annotations
135	93 (68.9%)	34 (25.2%)	8 (5.9%)
Choose Challenge			
Images: 5		Annotators: 17	
All Annotations	Early Bird Annotations	Correct Annotations	Incorrect Annotations
150	127 (84.7%)	10 (6.7%)	13 (8.7%)
Categorize Challenge			
Images: 5		Annotators: 17	
All Annotations	Early Bird Annotations	Correct Annotations	Incorrect Annotations
123	122 (99.2%)	1 (0.8%)	0 (0%)

We would also like to show some indicative data collected from all users during this evaluation period. Figure 6 illustrates the outcomes of a single Paint challenge instance. In this case, participants were shown the conveyor belt image in Figure 6.a and instructed to mark all PET items. To assess a participant's response and assign a score, we compare their annotation with those from other users. Our comparison method proceeds as follows: Initially, we superimpose all user responses, creating a composite heatmap, as depicted in Figure 6.b. We then determine the ground truth by applying a threshold to the heatmap to produce a binary map. Figure 6.c displays the ground truth with a threshold of 0.5 (our chosen threshold for early bird points), and Figure 6.d illustrates it with a threshold of 0.7. Subsequently, each new user annotation is compared against this ground truth on a pixel-by-pixel basis.



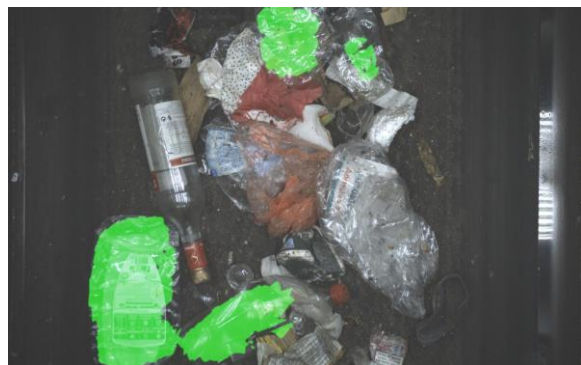
6.a. Reference conveyor belt image.



6.b. Aggregated heatmap, generated by overlaying all user's responses.



6.c. Ground truth with a threshold of 0.5 (in green)



6.d. Ground truth with a threshold of 0.7 (in green)

Figure 6: Indicative data collected from a single instance of the Paint challenge, across all users, when asked to annotate PET items. We note that currently an agreement threshold of 0.5 is used in Paint challenges for assigning correct points.

Referring to the specific example shown in Figure 6, it is apparent that the users' annotations generally capture the areas containing PET items effectively. Although individual responses vary in precision, as shown in Figure 6.b, the application of an agreement threshold results in a much more accurate mask, as demonstrated in Figure 6.c. These initial results indicate that crowd-sourced annotations can serve as a reliable means of generating high-quality data for training machine learning models. However, too high a threshold (0.7) may miss some PET items which could be useful for training. This threshold can be adjusted in future iterations of the RDG based on insights from evaluation runs such as this one.

7. Summary and Conclusions

This evaluation of the Recycling Data Game (RDG) combined both quantitative and qualitative methodologies to assess user interaction, satisfaction, and player experience. The assessment approach included a survey and a focus group, providing a comprehensive understanding of user experiences. Our key findings are summarised in the following themes:

Ease of Use and Learning the game: The game was noted for its user-friendliness and simple interface, which facilitated quick learning and ease of use. Users found the game straightforward to navigate and interact with, which is critical for engaging a broader audience.

Satisfaction: Overall, satisfaction levels among participants were rather positive. The game's design, which incorporates short, manageable challenges, aligns well with the needs of users looking for casual gaming experiences. Satisfaction correlated inversely with the gaming experience, indicating that more experienced gamers had higher expectations.

Educational Value: The game was appreciated for its educational content, particularly its ability to inform players about recycling practices and materials. This aspect was crucial for users interested in environmental sustainability, highlighting the game's role in raising awareness and educating the public.

Technical Performance: Technical issues related to game speed and responsiveness were noted, particularly in transitioning between tasks. These issues are attributed to network speeds and are expected to be addressed in future updates with improved image compression techniques.

Engagement and Motivation: Game features, such as ranking and feedback, were suggested as motivating factors that could enhance user engagement. The variety in game tasks and the integration of educational snippets, which are scheduled for the next iterations of the game, were also suggested for supporting a more dynamic user experience and increased educational and awareness impact to the public.

Visual and Interface Design: While the visual design was generally well-received, highlighted by vibrant colours and clear graphics, there were calls for improvements in image sizing and the interactive elements of the interface to better accommodate mobile users.

The RDG has demonstrated significant potential as an educational and research tool in the field of environmental sustainability. The game's structure and content effectively engaged users, giving them a number of opportunities to annotate waste data on different levels of cognitive challenge. Future iterations of the game should, additionally, focus on imparting valuable knowledge about recycling while providing an enjoyable gaming experience. The feedback also underscored the necessity for technical enhancements to optimise interaction and performance. The data collected and processed also raise some issues regarding e.g. how ground truth is calculated per challenge. Future iterations will focus on refining these elements to improve overall user satisfaction and extend its educational impact.

8. Future Work

The evaluation and feedback received for the Recycling Data Game (RDG) highlights several areas for improvement. As we progress, our focus will be on refining the game's design, expanding its content, and improving user interaction to elevate the gaming experience and educational value. The following points outline the main directions for the upcoming developments, driven from the user requirements highlighted during the focus group discussion and free-text responses of the online survey:

Interface and Accessibility Improvements: Enhancements to the user interface will focus on improving visibility in multitasking environments and adding functionality such as image zoom, which is crucial for smaller screens. These improvements aim to make the game more accessible and user-friendly, particularly for mobile users.

Technical Optimizations: Addressing the technical issues related to loading times and responsiveness is a priority. Future versions will incorporate optimised image compression techniques to enhance performance across various network conditions and devices.

Educational Content Expansion: We will continue to integrate educational content dynamically throughout the game. Planned additions include fun facts, quizzes, and informational tidbits between challenges to enrich the player's learning experience and maintain engagement.

Gameplay Diversification: To cater to a broader audience and keep the gameplay engaging, we will refine the challenges and mini-games. These will include more complex tasks that combine different types of content, providing a structured and varied gaming experience.

Enhanced Feedback and Motivation Systems: Implementing a more sophisticated feedback mechanism, including a ranking system and possibly a global leaderboard, will motivate players by making the gameplay more competitive and rewarding.

Player Profile and Social Features: The development of a player profile that tracks achievements and game progress is already planned. While initially not planned, feedback during the focus group indicated that the player profile could also facilitate social interactions, such as sharing achievements and competing with friends, enhancing the communal and competitive aspects of the game.

Continuous User Feedback Integration: Ongoing collection and integration of user feedback will remain a cornerstone of our development process. This iterative feedback loop will ensure that the game continually evolves to meet user needs and preferences.

Adjustments to ground truth calculation: Using the data from this evaluation round (see Section 6) and upcoming tests, we identified that there may be a discrepancy with the threshold for assessing a group agreement between challenges. Internal tests and new rounds of feedback will address this so that the data collected is useful for AI training.

The future work outlined aims to build upon the current strengths of the RDG while addressing the areas for improvement identified through user feedback. These efforts are expected to

enhance the overall quality of the game, making it more engaging, educational, and accessible to a diverse audience.

9. References

- [Braun2006] Braun, V. and Clarke, V. (2006) Using Thematic Analysis in Psychology. *Qualitative Research in Psychology*, 3, 77-101.
- [Gao2018] Gao, M., Kortum, P. and Oswald, F., 2018, September. Psychometric evaluation of the USE (usefulness, satisfaction, and ease of use) questionnaire for reliability and validity. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 62, No. 1, pp. 1414-1418). Sage CA: Los Angeles, CA: SAGE Publications.
- [Lund2001] Lund, Arnold M. "Measuring usability with the use questionnaire12." *Usability interface 8.2* (2001): 3-6.
- [Assila2016] Assila. A. and Ezzedine. H.. 2016. Standardized usability questionnaires: Features and quality focus. *Electronic Journal of Computer Science and Information Technology*. 6(1).
- [Lewis 1992] James Lewis. 1992. Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ. In *Proceedings of the human factors society annual meeting* (Vol. 36. No. 16. pp. 1259-1260). Sage CA: Los Angeles. CA: Sage Publications.
- [Lund2001] Lund, A. M. (2001). Measuring usability with the use questionnaire12. *Usability interface*, 8(2), 3-6.
- [PSSUQV3] <https://uiuxtrend.com/pssuq-post-study-system-usability-questionnaire/>
- [Brooke1996] John Brooke. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry*. 189(194). pp.4-7.
- [Schrepp2015] Martin Schrepp. *User experience questionnaire handbook - All you need to know to apply the UEQ successfully in your project.* (2015).
- [ISO9241-11:2018] International Organization for Standardization. 2018. *Ergonomics of Human-System Interaction — Part 11: Usability: Definitions and Concepts*. ISO 9241-11:2018. Geneva. Switzerland: ISO.
- [Wiemeyer2016] Josef Wiemeyer. Lennart Nacke. Christiane Moser. and Florian 'Floyd' Mueller. 2016. Player Experience. In *Serious Games: Foundations. Concepts and Practice*. Ralf Dörner. Stefan Göbel. Wolfgang Effelsberg. and Josef Wiemeyer (Eds.). Springer International Publishing. 243–271.
- [Newman2005] Ken Newman. 2005. Albert in Africa: Online role-playing and lessons from improvisational theatre. *Computers in Entertainment* 3. 3 (2005).
- [Cole2021] Tom Cole and Marco Gillies. 2021. Thinking and doing: Challenge, agency, and the eudaimonic experience in video games. *Games and Culture* 16. 2 (2021). 187–207.
- [Cole2022] Tom Cole and Marco Gillies. 2022. Emotional exploration and the eudaimonic gameplay experience: A grounded theory. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

- [Daneels2021] Rowan Daneels, Nicholas D Bowman, Daniel Possler, and Elisa D Mekler. 2021. The 'eudaimonic experience': A scoping review of the concept in digital games research. *Media and Communication* 9. 2 (2021). 178–190.
- [Bopp2016] Julia Ayumi Bopp, Elisa D. Mekler, and Klaus Opwis. 2016. Negative emotion, positive experience? Emotionally moving moments in digital games. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 2996–3006.
- [Bopp2018] Julia Ayumi Bopp, Klaus Opwis, and Elisa D. Mekler. 2018. "An odd kind of pleasure": Differentiating emotional challenge in digital games. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.
- [Denisova2020] Alena Denisova, Paul Cairns, Christian Guckelsberger, and David Zendle. 2020. Measuring perceived challenge in digital games: Development & validation of the challenge originating from recent gameplay interaction scale (CORGIS). *International Journal of Human-Computer Studies* 137 (2020).
- [Denisova2021] Alena Denisova, Julia Ayumi Bopp, Thuy Duong Nguyen, and Elisa D Mekler. 2021. "Whatever the emotional experience, it's up to them": Insights from designers of emotionally impactful games. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.
- [Gowler2019] Chad Phoenix Rose Gowler and Ioanna Iacovides. 2019. "Horror, Guilt and Shame" – Uncomfortable Experiences in Digital Games. In *Proceedings of the Symposium on Computer-Human Interaction in Play*. 325–337.
- [Medlock2018] Michael C. Medlock. 2018. User experience maturity levels: Evaluating and improving Game User Research practices. In *Serious Games: Foundations, Concepts and Practice*. Anders Drachen, Pejman Mirza-Babaei, and Lennart Nacke (Eds.), Oxford University Press.
- [Azadvar2018] Ahmad Azadvar and Alessandro Canossa. 2018. UPEQ: Ubisoft perceived experience questionnaire: A self-determination evaluation tool for video games. In *Proceedings of the International Conference on the Foundations of Digital Games*.
- [Ryan2006] Richard M. Ryan, Scott Rigby, and Andrew Przybylski. 2006. The motivational pull of video games: A self-determination theory approach. *Motivation and Emotion* volume 30 (2006). 344–360.
- [Yee2012] Nick Yee, Nicolas Ducheneaut, and Les Nelson. 2012. Online gaming motivations scale: Development and validation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.
- [Jennett2008] Charlene Jennett, Anna L. Cox, Paul Cairns, Samira Dhoparee, Andrew Epps, Tim Tijs, and Alison Walton. 2008. Measuring and defining the experience of immersion in games. *International Journal of Human-Computer Studies* 66. 9 (2008). 641–661.
- [Brockmyer2009] Jeanne H. Brockmyer, Christine M. Fox, Kathleen A. Curtiss, Evan McBroom, Kimberly M. Burkhart, and Jacquelyn N. Pidruzny. 2009. The development of the Game

Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology* 45 (2009). 624–634.

[Jackson1996] Susan A. Jackson and Herbert Marsh. 1996. Development and validation of a scale to measure optimal experience: The Flow State Scale. *Journal of sport and exercise psychology* 18. 1 (1996). 17–35.

[Vorderer2004] Peter Vorderer. Werner Wirth. Feliz Gouveia. Frank Biocca. Timo Saari. Lutz Jäncke. Saskia Böcking. Holger Schramm. Andre Gysbers. Tilo Hartmann. Christoph Klimmt. Jari Laarni. Niklas Ravaja. Ana Sacau. Thomas Baumgartner. and Petra Jäncke. 2004. MEC spatial presence questionnaire (MEC-SPQ): Short documentation and instructions for application. Report to the European Community. Project Presence: MEC (IST-2001-37661).

[Witmer1998] Bob G Witmer and Michael J Singer. 1998. Measuring presence in virtual environments: A presence questionnaire. *Presence* 7. 3 (1998). 225–240.

[DeKort2007] Yvonne De Kort. Wijnand IJsselsteijn. and Karolien Poels. 2007. Digital games as social presence technology: Development of the social presence in gaming questionnaire (SPGQ). In *Proceedings of the International Workshop on Presence*.

[Hudson2014] Matthew Hudson and Paul Cairns. 2014. Measuring social presence in team-based digital games. In *Interacting with Presence: HCI and the sense of presence in computer-mediated environments*. Giuseppe Riva. John Waterworth. and Dianne Murray (Eds.). De Gruyter Open Ltd.. 83–101.

[Abeele2020] Vero Vanden Abeele. Katta Spiel. Lennart Nacke. Daniel Johnson. and Kathrin Gerling. 2020. Development and validation of the player experience inventory: A scale to measure player experiences at the level of functional and psychosocial consequences. *International Journal of Human-Computer Studies* 135 (2020).

[IJsselsteijn2013] Wijnand IJsselsteijn. Yvonne de Kort. and Karolien Poels. 2013. The Game Experience Questionnaire. Technical Report. Technische Universiteit Eindhoven.

[Denisova2020] Alena Denisova. Paul Cairns. Christian Guckelsberger. and David Zendle. 2020. Measuring perceived challenge in digital games: Development & validation of the challenge originating from recent gameplay interaction scale (CORGIS). *International Journal of Human-Computer Studies* 137 (2020).

[Bowman2018] Nicholas David Bowman. Joseph Wasserman. and Jaime Banks. 2018. Development of the Video Game Demand Scale. In *Video games: A medium that demands our attention*. Nicholas David Bowman (Ed.). Routledge. 208–233.

[Depping2017] Ansgar E. Depping and Regan L. Mandryk. 2017. Why is this happening to me? How player attribution can broaden our understanding of player experience. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

[Lewis2008] Melissa Lewis. Rene Weber. and Nicholas Bowman. 2008. They may be pixels. but they're MY pixels: Developing a metric of character attachment in roleplaying video games. *Cyberpsychology & behavior: The impact of the Internet. multimedia and virtual reality on behavior and society* 11 (2008).

- [Graham2011] Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. 2011. Mapping the moral domain. *Journal of Personality and Social Psychology* 101. 2 (2011). 366–85.
- [Joeckel2012] Sven Joeckel, Nicholas David Bowman, and Leyla Dogruel. 2012. Gut or game? The influence of moral intuitions on decisions in video games. *Media Psychology* 15. 4 (2012). 460–485.
- [Banks2016] Jaime Banks and Nicholas David Bowman. 2016. Emotion, anthropomorphism, realism, control: Validation of a merged metric for player–avatar interaction (PAX). *Computers in Human Behavior* 54 (2016). 215–223.
- [Banks2019] Jaime Banks, Nicholas David Bowman, Jih-Hsuan Tammy Lin, Daniel Pietschmann, and Joe A. Wasserman. 2019. The common player-avatar interaction scale (cPAX): Expansion and cross-language validation. *International Journal of Human-Computer Studies* 129 (2019). 64–73.
- [Hall2022] Johanna Hall, Christothea Herodotou, and Ioanna Iacovides. 2022. Measuring player creativity in digital entertainment games using the Creativity in Gaming Scale. In *Open world learning: Research, innovation and the challenges of highquality education*. Bart Rienties, Regine Hampel, Eileen Scanlon, and Denise Whitelock (Eds.). Routledge.
- [Peck2021] Tabitha C. Peck and Mar Gonzalez-Franco. 2021. Avatar embodiment. A standardized questionnaire. *Frontiers in Virtual Reality* 1 (2021).
- [Power2017] Christopher Power, Alena Denisova, Themis Papaioannou, and Paul Cairns. 2017. Measuring uncertainty in games: Design and preliminary validation. In *Proceedings of the CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 2839–2845.
- [Choi2013] Beomkyu Choi, Jie Huang, Annie Jeffrey, and Youngkyun Baek. 2013. Development of a scale for fantasy state in digital games. *Computers in Human Behavior* 29 (2013).
- [Plante2017] Courtney N. Plante, Stephen Reysen, Christopher L. Groves, Sharon E. Roberts, and Kathleen Gerbasi. 2017. The Fantasy Engagement Scale: A flexible measure of positive and negative fantasy engagement. *Basic and Applied Social Psychology* 39. 3 (2017). 127–152.
- [Abt1970] Abt, C.C. *Serious Games*. Viking, New York. 1970.
- [Halter2006] Halter, E. *From Sun Tzu to Xbox: War and Videogames*. Thunder’s Mouth Press, New York. 2006.
- [Ritterfeld2009] Ritterfeld, U., Cody, M., and Vorderer, P. *Serious Games: Mechanisms and Effects*. Routledge, London. 2009.
- [Calderon2015] Alejandro Calderón and Mercedes Ruiz. 2015. A systematic literature review on serious games evaluation: An application to software project management. *Computers & Education* 87 (2015). 396–422.
- [Boyle2016] Elizabeth A Boyle, Thomas Hainey, Thomas M Connolly, Grant Gray, Jeffrey Earp,

Michela Ott. Theodore Lim. Manuel Ninaus. Claudia Ribeiro. and João Pereira. 2016. An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. *Computers & Education* 94. C (2016). 178–192.

[Connolly2012] Thomas M Connolly. Elizabeth A Boyle. Ewan Macarthur. Thomas Hainey. and James M Boyle. 2012. A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education* 59. 2 (2012). 661–686.

[Paliokas2016] Ioannis Paliokas and Stella Sylaiou. 2016. The Use of Serious Games in Museum Visits and Exhibitions: A Systematic Mapping Study. In *Proceedings of the International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES)*.

[Coenen2013] Tanguy Coenen. Lien Mostmans. and Kris Naessens. 2013. MuseUs: Case study of a pervasive cultural heritage serious games. *Journal on Computing and Cultural Heritage* 6. 2 (2013).

[Backlund2013] Per Backlund and Maurice Hendrix. 2013. Educational games - Are they worth the effort? A literature survey of the effectiveness of serious games. In *Proceedings of the International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES)*.

[Froschauer2013] Josef Froschauer. Dieter Merkl. Max Arends. and Doron Goldfarb. 2013. Art History Concepts at Play with ThIATRO. *Journal on Computing and Cultural Heritage* 6. 2 (2013)

[Volkmar2018] Georg Volkmar. Nina Wenig. and Rainer Malaka. 2018. Memorial Quest - A Location-Based Serious Game for Cultural Heritage Preservation. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*. 661–668.

[Andreoli2017] Roberto Andreoli. Angela Corolla. Armando Faggiano. Delfina Malandrino. Donato Pirozzi. Mirta Ranaldi. Gianluca Santangelo. and Vittorio Scarano. 2017. A Framework to Design, Develop, and Evaluate Immersive and Collaborative Serious Games in Cultural Heritage. *ACM Journal on Computing and Cultural Heritage* 11. 1 (2017).

[Schuster1991] J. Mark David Schuster. 1991. *The audience for American art museums*. Seven Locks Press. Washington.

[Eliean1999] Hooper-Greenhill Eilean (Ed.). 1999. *The educational role of the museum* (2nd ed.). Routledge. London ; New York.

[Kiili2007] Kristian Kiili. Harri Ketamo. and Timo Lainema. 2007. Reflective thinking in games: triggers and constraints. In *Proceedings of the European Conference on Games Based Learning*. 169–176.

[Macdonald2007] Sharon Macdonald. 2007. Studying Visitors. In *A Companion to Museum Studies*. Blackwell Publishing Ltd. 362–376