# Little Big Data – a secondary school game raising awareness on data collection and analysis

Robin Gerster, Bart Swinkels, Thomas Streefkerk, Thomas Sjerps, Kris van Melis, Rafael Bidarra
Delft University of Technology, The Netherlands
Email: r.bidarra@tudelft.nl

***Abstract*—The science of collecting and analyzing vast amounts of data to make predictions or derive other types of conclusions is often called Big Data. It is expected that a new generation of college students will be studying and working with Big Data. However, at the secondary school level, a serious lack of knowledge about it has often been detected. We designed a serious game to introduce students to this topic. Because Big Data is far too complex to teach in a single game, we focused our game design on two main elements: *data collection* and *data analysis*. The game raises awareness on these topics by immersing players in realistic situations, in which they strategically place data collection centers and use the collected data to answer various appealing queries and scenarios. These scenarios exemplify several opportunities and risks of real-world uses of Big Data. From a preliminary evaluation of the game, we conclude that the current game design leads to a statistically significant increase in awareness.**

***Index Terms*—Serious Game, Big Data, Data collection, Data analysis, Secondary education.**

## I. Introduction

Big Data is of increasing importance and impact on everyone's life [16]. However, it is a very complex topic, with many interconnected facets. Big data has the potential to become more accessible in the future like traditional data processing pipelines already are [21]. Laney et al. [15] compare Big Data with more conventional data processing as having larger:

- **volume**: there is more data, and it may have unknown value
- **velocity**: the data is received and processed faster
- **variety**: there are more types of data, and it is often unstructured

A slew of data integration, management, and analysis techniques are often deployed within Big Data. Big Data fuels many advances [16], as large amounts of data allow finding unexpected patterns and deriving interesting and useful conclusions. Large companies use Big Data for security, crime prevention [10], better user recommendations [13], and solving other complex problems [19]. However, Big Data can also have its downsides, as data may be biased and its analysis may lead to poor or prejudiced performance. For instance, Big Data systems have been developed for US law enforcement using unlawful and racially discriminatory data [20].

Progress has been made in lowering the technical skills needed to work with Big Data. Just as the bar has been lowered for website creation, due to developments in computational power and accessible website builders, similar progress is expected to also occur with Big Data [21]. This increase in accessibility will allow less technically savvy end-users to work with its technologies.

In more practice-oriented secondary schools, like VMBO-level schools in the Netherlands [18], students are usually not taught about Big Data concepts. However, due to the potential that Big Data has in the future and the predicted increase in accessibility and ease of use, this less technically savvy group could be a potential target audience for professional work with Big Data. For that, it would be important to raise awareness among students about Big Data notions. It is known that these students generally enjoy computer/mobile games, which have proven to be good media for educational purposes [3]. We, therefore, believe that a serious game has the potential to achieve this Big Data awareness goal.

The kind of Big Data awareness that is needed for these students includes knowledge (as defined by Bloom's taxonomy [11]) and insight into Big Data possibilities: you do not need to apply many concepts to be aware of Big Data, but you do need more than simply memorizing information. With this in mind, we have chosen the following awareness goals:

1) **Understand that Big Data uses vast amounts of varied data**
Students have to understand that Big Data uses a vast amount of varied, possibly non-informative, data to give an informative answer to a complex question. This quantity of data makes Big Data different from traditional data analysis, which thrives on informative data from a single source.
2) **Give examples of data collected and used in Big Data**
Students should understand which data might get collected to be used in Big Data. This goal also shows that data might sometimes get collected unconsciously, which might turn out to be undesirable.
3) **Realize Big Data analysis helps to answer otherwise complex questions**
Students should understand that analyzing a vast amount of varied data can help answer a complex question. This, in turn, answers why Big Data is used and why all this data is collected.
4) **Understand that Big Data can have a positive and negative impact on society**
Students should identify which positive and negative impacts Big Data can have on society and themselves.

979-8-3503-2277-4/23/$31.00 ©2023 IEEE

This, in turn, makes them realize why it is essential to understand Big Data, since it can be helpful in some cases but harmful in others.

Our main contributions are:

1) a simple but effective game design to make secondary school students aware of Big Data
2) the following insights for creating serious games that raise awareness of Big Data among practical learners:
    - A serious game is a good means for raising awareness.
    - More emphasis on the harmful and ethically questionable use cases is needed to raise students' awareness regarding the risks of Big Data.
    - A standardized evaluation model for educational serious games could prove extremely useful for this and similar research.

## II. Related Work

Laamarti et al. [14] survey several serious games, many of which are education-focused, and provide an overview of what made these games successful. For example, the gameplay experience should bring players into a flow state. The game should not be too informative (as this would be hard for the students to engage in) or too easy. Visuals and sound are also considered essential features in maintaining and guiding attention. Additionally, collaborative and competitive games are more engaging than single-player experiences. Finally, negative feedback should be avoided during learning because this tends to discourage the players from performing better.

Gloria et al. [8] also provides an overview of successful educational serious games. An emphasis is placed on how specific game architectures teach players. In general, the type of serious game one builds should be clearly related to the domain of the topic it teaches. They also recommend methods to measure the impact of serious games.

Catalano et al. [7] made an extensive survey of serious games currently proposed in different application domains. As a result, they provide various useful recommendations for making educational serious games more effective and thus increase their learning impact.

Unfortunately, not many serious games have been adapted to teaching Big Data to secondary school students. However, some games for teaching data analysis have been implemented and reported on. An example of such a game is Proximity, proposed by Ericksson [9]. In this game, the player shoots a ball at the center of a target. Data analysis techniques are available to the player to improve the shot's accuracy. Ericksson asserts that data should serve as a fundamental element in these types of games, integral to the gameplay itself, rather than being a peripheral or secondary feature. Furthermore, it is also pointed out that a player's skills should dominate over luck, since too much reliance on luck would discourage the players from performing better. This same design principle has been applied in other real-world data-related serious games, to improve player's insight into the phenomena revealed by that same data [4], [5].

Adisusilo et al. [1] discuss how immersion is vital when teaching players via a serious game: 'distracting' the players from the fact that they are being taught about a specific topic improves the overall experience. Additionally, an inverse correlation between "fun" and "reality" is noted: entertaining games often differ more from reality than those that are less fun. This design tension, also identified by Harteveld et al. [12], is important since a more fun game is obviously more enjoyable for the user, although a very realistic game might seem better at teaching a topic such as Big Data.

## III. Game Design

It is quite challenging to engage and tangibly incorporate into a single game all the awareness goals mentioned in Section I. One of the main challenges is the balance between education and enjoyment within a game. We developed the serious game *Little Big Data* with the special concern of aiming at our awareness goals without taking players away from their enjoyment. Considering the complexity of the topic, and our target audience, we realized the importance of keeping players interested and focused. As a result, we decided to highlight a choice of aspects of Big Data in a simplified manner whilst leaving others out. The game focuses on fostering a high-level understanding of Big Data, as well as pointing out the main processes that drive it in the real world. As such, *Little Big Data* aims to emphasize the data *collection* and *analysis* phases.

To achieve the effect mentioned above, the game loop has been divided into several phases; see Figure 1. Firstly, students are prompted with a scenario that exemplifies the application of Big Data during the building phase. For example, they could be asked to help a company collect and analyze data to find the source of a disease. To fulfill this scenario, students are first instructed to place data collection centers on a map spanning several cities. Next, students enter a collection stage in which they are shown real-life examples of collected data. Finally, they enter a data analysis phase where they utilize the collected data to find an answer to the prompted scenario, after which they are rewarded with points. The score is calculated as a function of the quality and speed of their analysis. Rewarding the quality of analysis ensures that students exercise their critical thinking and reflect on the scenario. Speed is also rewarded for discouraging *brute-forcing* the solution.

This game design was chosen to comply with several constraints. For one, the game must be intuitive and playable without supervision, to streamline the learning experience. As

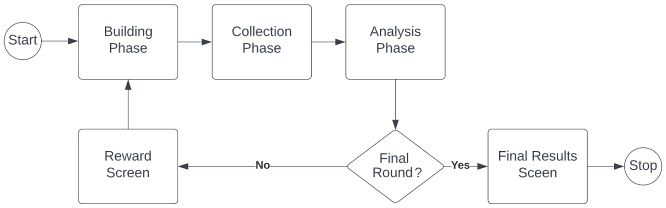
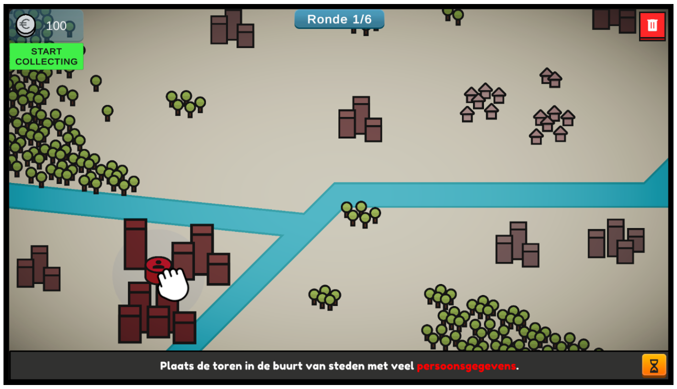


Fig. 1. Core game loop of *Little Big Data*



Plaats de toren in de buurt van steden met veel persoonsgegevens.

Fig. 2. A data center is about to be placed by the player.

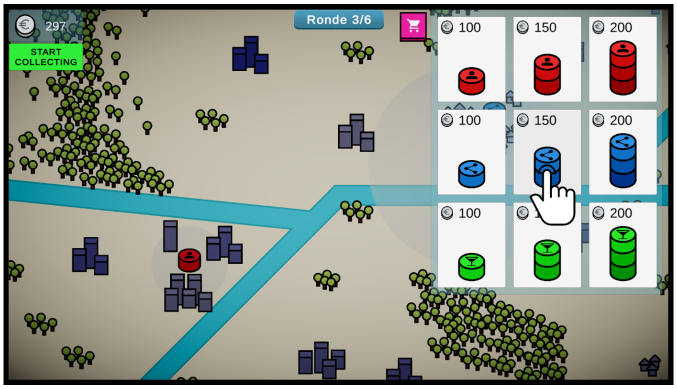
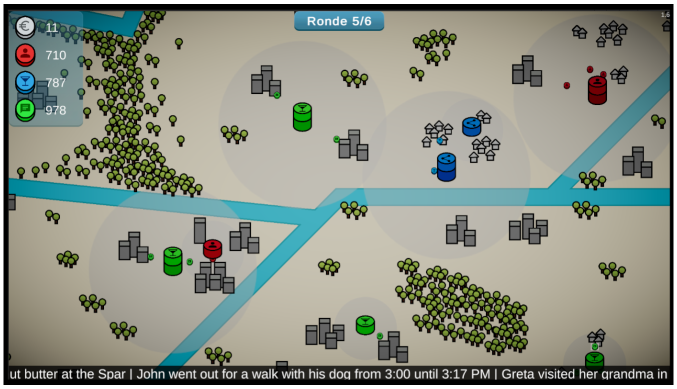


Fig. 3. The data center shopping interface.



ut butter at the Spar | John went out for a walk with his dog from 3:00 until 3:17 PM | Greta visited her grandma in

Fig. 4. The data collection phase in the 5th round.

such, the game features a set of easy-to-execute tasks, and the necessary instructions to perform them. The small set of game stages makes the game functional on low-end hardware and deployable to the web so that it can be easily accessed across secondary schools. Typically, students only require 10 minutes to complete the game which easily fits into the classroom schedule. In order to further align the game genre with the domain, we integrated strategic, spatial and otherwise cognitive tasks into the game loop. These are common concerns within similar educational games [17]. More specific details on the main game stages are now discussed.

### A. Big Data collection

Each round of the game starts with a scenario to give the student a motive for why they would want to collect big data. These scenarios range from finding the origin of a virus outbreak to predicting the most popular upcoming movies. Then, the student is presented with a map of cities and villages. This map clearly shows the large-scale Big Data collection covers in real life. The student can place data centers that collect data from nearby cities. Some cities have more potential data to collect, as visually indicated by the city density; see Fig. 2. Larger cities have more citizens and, therefore, more data. Exploring game mechanics of a new game is a difficult task, especially for students with a limited attention span [6]. We, therefore, included a tutorial to guide them in interactively learning the game mechanics, as shown in Fig. 2.

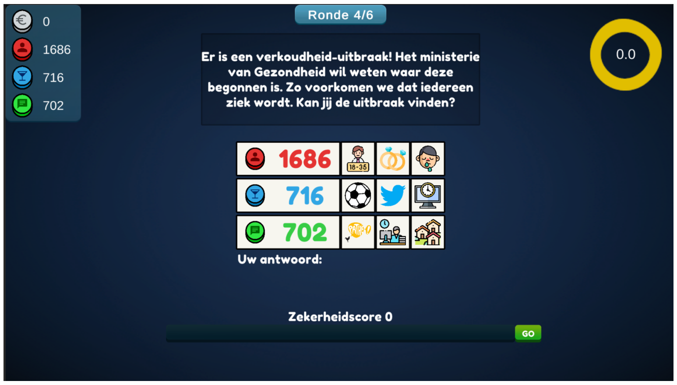
The player can collect three different types of data: *personal*, *online*, and *location*. This distinction shows that many different types of data can be used in Big Data analysis, thus addressing the first awareness goal: *Understand that Big Data uses vast amounts and varied data*. To ensure the game stays comprehensible, we have limited the number of categories to three. Later in the game, the student will combine these data types to achieve this objective. We created three sizes for each data type to engage the students with decision-making and, thus, give a more personalized feeling to the game; see Fig. 3.

When the player is satisfied with their data center placement, the game enters an automatic collection phase. Here the second awareness goal is realized: as data is being collected, a news ticker, with real-life examples of how data can be collected, is scrolled at the bottom, as shown in Fig. 4.

### B. Big Data analysis

At this stage, we address our third awareness goal: *Big Data analysis helps to answer otherwise complex questions*. In this phase, the player can use the data collected to solve the prompted scenario. These prompts can show the negative and positive impact of Big Data, which is the fourth awareness goal. Students have to use their reasoning skills to select a set of data sources relevant for the scenario at hand. A good selection will lead to a stronger model with more accurate predictions. This notion is communicated to the student through a dynamic display that shows their model's certainty score; see Fig. 5.

This selection task was chosen because data selection goes together with Big Data analysis in the real world [2]. Additionally, the student can draw parallels between the sources and media with which they interact. This cognitive task aligns with the educational domain as students must use their reasoning skills, and it is similar to that of other games around data analysis, such as Proximity [9]. Here we opted to keep it a simple and approachable task, not to compromise the learning purpose of this educational game [23]. After a satisfactory selection is made, a reward window will summarize the stu-



Er is een verkoudheid-uitbraak! Het ministerie van Gezondheid wil weten waar deze begonnen is. Zo voorkomen we dat iedereen ziek wordt. Kan jij de uitbraak vinden?

Fig. 5. The scenario prompts a medical question, asking the player to find the source of a disease. Choosing pertinent information, for instance, personal data including ”age” or determining who is ”sick”, will substantially boost the model’s confidence score. Making irrelevant selections will have an adverse effect, making the model less confident and consequently leading to a poorer reward.

dent’s performance. The algorithm that calculates the reward is fully deterministic to make the learning process performance-based, a feature of most successful serious games [14]. This reward is further underlined by an online leaderboard aimed at encouraging competition and engagement.

## IV. Evaluation

We have evaluated how *Little Big Data* raises awareness regarding Big Data by assessing the extent to which the game reaches the awareness goals set in Section I. For a serious game to be effective in raising awareness about a topic, it must be intuitive and engaging to play [22], or this might unwittingly hinder achieving those awareness goals. We, therefore, have also evaluated the intuitiveness and engagement of the game.

To test whether our awareness goals were achieved, we conducted an experiment. Convenience sampling was used to sample 12 students from Dutch secondary education, in cooperation with a teacher, who confirmed this sample generalizes well to the target population. We tested the students’ awareness before and after playing the game with a questionnaire, see Appendix A. There was no delay between the pre-test, game interaction and post-test. Due to the limited sample size, a within-group design was chosen. We have graded the results according to the indicated metrics (for convenience, now on the questionnaire). To minimize bias, we employed a blind approach in grading the responses. This approach ensured that when grading, we were unaware of whether a student’s answer was given before or after interacting with the game as well as which student had given the response. This ensured that our grading process was not influenced by our knowledge of the intervention (the game) or any preconceived notions about individual students, thereby enhancing the objectivity of our evaluation.

Participants were also asked to score how intuitive and engaging the game was with an evaluation form A. The form was answered on a 7-point scale from ’strongly disagree’ to ’strongly agree’.

TABLE I
Significant Improvement for Test Questions

| Question | Significant Improvement |
|---|---|
| What’s important when collecting data for Big Data analysis? | True |
| How can Big Data have a positive impact on our lives? | True |
| How can Big Data have a negative impact on our lives? | False |
| How could you collect data in a supermarket? | True |
| Give an example of what a data point could look like. | False |
| Give an example question difficult to answer without Big Data. | False |
| Why is this question difficult to answer without Big Data? | False |

### A. Results

The one-tailed, paired sampled t-test is used to check for a statistically significant increase in the means between the before and after test scores. A t-test was possible given that the data is continuous, the observations are independent, the variance is homogeneous, and the score distributions are approximately normal.

Students show a significantly higher awareness of Big Data after playing the game (M=3.42, SD=3.18) than before interacting with the game (M=0.83, SD=1.40). Table I does not show improvement for all test questions: students struggled to give negative examples of Big Data uses, and could not describe what collected data might look like.

Table II shows the result of the subjective evaluation. Generally, students felt neutral about all questions, given that all scores were less than one standard deviation from 4.0.

TABLE II
Results of the subjective evaluation form
(questions translated from Dutch)

| Question | Mean | Standard deviation |
|---|---|---|
| The game is fun to play | 4.5 | 1.4 |
| I felt in control of the game | 3.6 | 1.4 |
| The game was easy to play | 4.5 | 1.3 |
| I understood how the game worked | 3.6 | 1.7 |
| I learned something new | 4.25 | 2.0 |

### B. Discussion

Table III summarizes our findings with regard to achieving the awareness goals.

The results show a significant increase in the student’s awareness of Big Data before and after playing the game. We can, therefore, state that the students’ awareness has been raised. This increase can be seen by the 95% significant rise in test results for three out of six questions (see Table I). However, because the significance of this increase is not present for all questions, we cannot say that all awareness goals have been reached.

Table III shows that the first two awareness goals have been reached: *Big Data uses vast amounts of varied data* and *What*

TABLE III
AWARENESS GOALS AND ACHIEVEMENT LEVELS

| Awareness Goal | Achievement Level |
|---|---|
| Understand that Big Data uses vast amounts and varied data | Yes |
| Give examples of data collected and used in Big Data processing | Yes |
| Realize Big Data analysis helps to answer otherwise complex questions | No |
| Understand that Big Data can have positive and negative impact on society | Partially |

*examples of data collected and used in Big Data processing are*. The awareness goal *Big Data can have a positive and negative impact on society* has been partially achieved since the students were able to name examples of positive impacts but not quite of negative impacts. Finally, goal *Big Data analysis can answer complex questions* has not been reached.

However, several objections could be made as to why this result is less than conclusive. The first is a general lack of standardization in evaluating serious games. We designed our own evaluation questions, and it is difficult to judge how well they make the learning goals measurable. For example, the phrase "Give an example of what a data point could look like" seems not to have been well understood by most students, making it a poor measure of the game's efficacy. Moreover, even though we took steps to reduce the bias in the results, there may still be some bias due to our designing and grading the questions ourselves. This is, again, difficult to account for.

Secondly, our sample size was a mere 12 students. In combination with possible distracting factors during the evaluation, this reduces the accuracy and reliability of the test results. In addition, the students had little motivation to perform the evaluation to the best of their ability. Furthermore, they might have been more motivated to get the questions right the second time they were answering, once they most likely figured out why we asked them twice (before and after playing the game). Again, this reduces the certainty of the test results.

Thirdly, the way the questions were structured in some cases (partially) gave away the answer we were looking for. For example, by asking the students, "What is important when collecting data for Big Data analysis?" before they played the game, they might already be on the alert to look for answers in the game. As a result, their score on this question when answering it after playing the game will most likely be (much) higher.

Besides testing the increase in awareness, we also evaluated the students' subjective experience of playing *Little Big Data* (see Table II). Considering the mean and standard deviation of these results, the game scored somewhat neutral on all elements. This neutral score can be explained by the challenge of balancing education and enjoyment throughout the game. The educational component of the game may have somehow negatively affected the enjoyment and vice versa. Although a higher average score would have been preferred, this neutral score is entirely satisfactory given the challenges of including all awareness goals within a single experience and keeping the game simple and enjoyable. Another potential contributing factor to these scores not being higher is the game's balancing. It proved rather tricky to strike a balance between game elements' effectiveness that both supported the chosen awareness goals and made the game enjoyable.

From the results, it has become clear that a more effective and standardized method of evaluating the game is desirable. Some questions from Table I proved helpful in answering whether a teaching goal was achieved, whilst others were either not understood by the students or produced unhelpful answers. Furthermore, future evaluation should be done with a larger sample size and potentially with some additional reward, to encourage students to do their best.

## V. CONCLUSION

We have presented a serious game, *Little Big Data*, with the purpose of raising awareness about Big Data among secondary school students which can be accessed online [1]. The game focuses on two important aspects of Big Data, *data collection* and *data analysis*. For this, it exposes players to a scenario and lets them decide on the best data collection strategy to address a given prompt, as well as query the collected data to find answers for it.

The results of the evaluation of *Little Big Data* indicate that the game was effective in raising awareness about Big Data among the students. Based on this finding, we consider that the use of serious games can be explored and recommended as a viable option for introducing Big Data to secondary school students.

The current version of *Little Big Data* could likely be improved to also emphasize some negative or ethically questionable uses of Big Data. Similarly, additional design effort should be made to clarify that Big Data can answer complex questions. Finally, it would be also interesting to investigate which components of the game were the most contributing factors for raising Big Data awareness.

## REFERENCES

[1] Anang Kukuh Adisusilo and Santirianingrum Soebandhi. A review of immersivity in serious game with the purpose of learning media. *International Journal of Applied Science and Engineering*, 18(5):1–11, 2021. Publisher: Chaoyang University of Technology.

[2] Waleed Albattah, Rehan Ullah Khan, Mohammed F. Alsharekh, and Samer F. Khasawneh. Feature selection techniques for big data analytics. *Electronics*, 11(19), 2022.

[3] Fran C Blumberg, Mark Blades, and Caroline Oates. Youth and new media. *Zeitschrift für Psychologie*, 2015.

[4] Dennis Bohm, Bob Dorland, Rico Herzog, Ryan B. Kap, Thijmen S. L. Langendam, Andra Popa, Mijael Bueno, and Rafael Bidarra. How can you save the world? Empowering sustainable diet change with a serious game. In *Proceedings of CoG 2021 - IEEE Conference on Games*. IEEE, 2021.

[5] Hidde Bolijn, Martin Li, Andries Reurink, Cas van Rijn, and Rafael Bidarra. Benni's forest – A serious game on the challenges of reforestation. In *Proceedings of CoG 2022 - IEEE Conference on Games*. IEEE, 2022.

[6] Diane M Bunce, Elizabeth A Flens, and Kelly Y Neiles. How long can students pay attention in class? A study of student attention decline using clickers. *Journal of Chemical Education*, 87, 2009.

[1] The game is available at: https://simmer.io/@RobinGerster/little-big-data

[7] Chiara Eva Catalano, Angelo Marco Luccini, and Michela Mortara. Guidelines for an effective design of serious games. *International Journal of Serious Games*, 1(1), Feb. 2014.

[8] Alessandro De Gloria, Francesco Bellotti, and Riccardo Berta. Serious games for education and training. *International Journal of Serious Games*, 1(1), 2014.

[9] Tim Erickson. Designing games for understanding in a data analysis environment. *Material Research Innovations*, 7, 01 2013.

[10] Mingchen Feng, Jiangbin Zheng, Yukang Han, Jinchang Ren, and Qiaoyuan Liu. Big data analytics and mining for crime data analysis, visualization and prediction. In *International conference on brain inspired cognitive systems*, pages 605–614. Springer, 2018.

[11] Mary Forehand et al. Bloom's taxonomy: Original and revised. *Emerging perspectives on learning, teaching, and technology*, 8:41–44, 2005.

[12] Casper Harteveld, Rui Guimaraes, Igor Mayer, and Rafael Bidarra. Balancing play, meaning and reality: the design philosophy of levee patroller. *Simulation and Gaming*, 41(3):316–340, 2010.

[13] Leonard Heilig and Stefan Voß. Managing cloud-based big data platforms: a reference architecture and cost perspective. In *Big data management*, pages 29–45. Springer, 2017.

[14] Fedwa Laamarti, Mohamad Eid, and Abdulmotaleb El Saddik. An overview of serious games. *International Journal of Computer Games Technology*, 2014, 10 2014.

[15] Doug Laney et al. 3d data management: Controlling data volume, velocity and variety. *META group research note*, 6(70):1, 2001.

[16] David Lazer and Jason Radford. Data ex machina: Introduction to big data. *Annual Review of Sociology*, 43(1):19–39, 2017.

[17] Richard E. Mayer. Computer games in education. *Annual Review of Psychology*, 70(1):531–549, 2019. PMID: 30231003.

[18] Cultuur en Wetenschap Ministerie van Onderwijs. Hoe zit het vmbo in elkaar?, 1 2023.

[19] Muhammad Imran Razzak, Muhammad Imran, and Guandong Xu. Big data analytics for preventive medicine. *Neural Computing and Applications*, 32(9):4417–4451, 2020.

[20] Rashida Richardson, Jason Schultz, and Kate Crawford. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice.

[21] Daniel L Rubinfeld and Michal S Gal. Access barriers to big data. *Ariz. L. Rev.*, 59:339, 2017.

[22] Amri Yusoff. *A Conceptual Framework for Serious Games and its Validation*. PhD thesis, University of Southampton, Faculty of Engineering, Sciences and Mathematics, 2010.

[23] Yu Zhonggen. A meta-analysis of use of serious games in education over a decade. *International Journal of Computer Games Technology*, 2019, 2019.

## Appendix

### A. *Awareness Goals Evaluation Form*

1) **What is important while collecting data for Big Data analysis?**

   The student names the following:

   - Vast amounts of data (1p)
   - Varied data (1p)

2) **In what positive way can Big Data impact our lives?**

   Any realistic positive example (2p) such as:

   - Help find cures
   - Improve transit
   - etc.

3) **In what negative way can Big Data impact our lives?**

   Any realistic negative example (2p) such as:

   - Manipulate people
   - Less privacy
   - etc.

4) **Give an example of how you can collect data for Big Data processing in a supermarket. Also give a data point example.**

   Example answer:

   - We can keep track of what everyone buys to recommend them discounts which will make them buy more (1p)
   - An example data point could be: Mark buys 5 apples every week (1p)

5) **Give an example of a question you can answer using Big Data, which otherwise would be very difficult. Explain why.**

   An example could be as follows: What cure for the flu works the best? (1p) Normally this is very difficult because there are many cures and many different types of flu. (1p)

### B. *Intuition and Engagement Evaluation Form*

Rate the following statements from one to seven. One is fully disagree and seven is fully agree. Four is an undecided answer.

1) The game is fun to play
2) I felt engaged while playing the game
3) I feel in control of my actions
4) The game felt intuitive to play
5) I know what the goals of the game are
6) I felt like I learned something new