DATA MANAGEMENT PLAN

ACADEMY OF FINLAND: ACADEMY OF FINLAND DATA MANAGEMENT PLAN GUIDELINES

RESEARCH TITLE: EMOTIONAL MODELLING TO ENHANCE LEARNING WITH GAMES

RESEARCH ABSTRACT

A UN Sustainable Development Goal is to ensure inclusive and equitable quality education and promote lifelong learning opportunities for all. The project results contribute to improving the future of education for all and have direct scientific, societal, and technological impacts. Current education solutions are limited to a one-size-fits-all paradigm, unable to support diverse learning needs, and contribute to learning exclusion and gender inequality. State-of-the-art (SOA): Game-based learning environments (GBLEs) bring to bear capabilities that not only shatter physical barriers to education via digital learning, but game mechanics built into GBLEs can be customized and adapted to support diverse learning needs. Game mechanics are important because they sustain learners' engagement and interest in learning activities, which elicit cognitive processes that water the seeds of learning. Still, more research is needed to develop adaptive game mechanics to support diverse learning needs. The mission of this research is to study interactions between game mechanics and emotions as a non-linear dynamical system and observe the impact of such interactions between diverse learners' cognition and learning outcomes with GBLEs.

In this research project, we will conduct a within-subjects mixed-methods study design to observe relations between game mechanic interactions and in-game emotions, cognition, and learning outcomes by collecting multiple data channels. Specifically, we will recruit 120 university students (50% women) from Tampere University and they will be asked to learn with a game-based learning environment called Antidote COVID-19. Prior to gameplay, a series of survey measures and prior knowledge assessment will be collected, and during gameplay, game interaction data and multimodal time series data will be captured (i.e., video, audio recordings of think-alouds, computer screen recordings, facial expressions, electrodermal activity, and heart rate). After gameplay, all participants will be asked to complete a knowledge assessment of material they learned in the game, and survey measures. Next, the project will merge the component process model of emotion and complex systems theory to examine the temporal emergence of emotions over the course of interacting with game mechanics using nonlinear dynamical analyses.

Theoretically, the project will advance game-based learning research by merging an interdisciplinary approach to gain fresh insights on how emotions emerge in relation to learners' interactions with game mechanics that foster engagement, cognition, and learning. Methodologically, this project will contribute to advancing the measurement and modelling of emotions and their dynamics. Practical guidelines for designers, educational game companies, instructors, and researchers on how to leverage adaptive game mechanics that support inclusive, equitable, and high-quality, GBLEs designed to support diverse learning needs. The results of this project will allow for the further development of game mechanics that foster specific emotions (e.g., joy, frustration) that benefit cognition and learning based on diverse learning needs.

1. GENERAL DESCRIPTION OF DATA

1.1 What kinds of data is your research based on? What data will be collected, produced or reused? What file formats will the data be in? Additionally, give a rough estimate of the size of the data produced/collected.

The research data in the current project will consist of self-reports, content knowledge assessment scores (pre/post), video, audio, computer screen recordings, physiological data, digital text data, and open-source code data. Data will be collected before, during, and after game-based learning, and to clarify the nature of the data, they are divided as such below:

Before game-based learning: anonymous data will be collected before game-based learning. These data will be collected using a demographic questionnaire (i.e., age, sex, and grade), validated survey instruments (Intrinsic Motivation, Achievement Goal orientation, Emotions and Values, and Multidimensional Gender Expression) – to gauge the students' gender identity, emotions and values, and motivation – and a content knowledge assessment to capture the participants' level of knowledge about the domain topic of the game environment. These data will be collected directly from the participants and an ID key (randomly generated) will be used once they enter the laboratory after the consent form has been signed.

During game-based learning: sensitive data will be collected during game-based learning, involving a video of the participants' face, an audio recording of their concurrent verbalizations using a think-emote-aloud protocol, a computer-screen recording of their game-based learning interactions, as well as electrodermal activity (heart rate).

Once the participant has finished game-based learning, the video recording will be immediately anonymised on the local data workstation by post-processing the video data into digital numeric and textual data. The new labels will be produced by a machine learning algorithm (AffDex) that will classify momentary facial expressions of participants' emotions collected during gameplay. The classification algorithm will produce textual data indicating the presence of facial expressions of seven basic emotions: happy, sad, angry, surprised, contempt, fear, and disgust, and numerical data, involving scores (0 to 100) that indicate that a facial action unit moved from one moment to another over time. Once the post-processing of the video recording data is completed, the video recording will be permanently deleted from the local data workstation. The anonymized digital labels will be then linked to the participant's ID key used during the pre-test phase, and these data will be stored in two locations to reduce data loss: password-protected external hard drive (stored in a locked desk in Dr. Cloude's locked office) as well as the university data server.

The audio files will be pseudonymized following data collection, where they will be stored in two locations to reduce data loss risk: password-protected external hard drive (stored in a locked desk in Dr. Cloude's office) as well as the university data server. The audio files will be permanently deleted from the local workstation in which they were collected. These data will not be anonymized immediately since they will require additional post-processing of the audio files into transcriptions of the verbalizations made during game-based learning. Once the audio files are fully transcribed and verified to contain little to no errors, the audio files will be permanently deleted from the external hard drive and university data server. The transcription files will be linked to the participant's ID key and these data will then be anonymized.

The computer-screen recordings will be fully anonymized since they do not contain sensitive data (e.g., participant's face or voice). These files will be linked to the participant's ID key and will be stored on the external hard drive and university data server immediately following data collection to reduce any risk of

data loss. The computer screen recording will be permanently deleted from the local workstation in which it was collected.

In addition, the electrodermal data will be fully anonymized since it does not contain sensitive data. These files will be linked to the participant's ID key and will be immediately post-processed following data collection using the AcqKnowledge software (GDPR compliant). The software will produce digital numerical data to define heart rate variability across the learning session. Once data has been post-processed, they will be stored on the external hard drive and the university data server. Following the successful transfer, these files will be permanently deleted from the local workstation.

After game-based learning: anonymous data will be collected after game-based learning. These data will be collected using validated survey instruments – to gauge the students' gender expression, concurrent emotions, and intrinsic motivation – and a content knowledge assessment to capture the participants' level of content knowledge about the domain. These data will be collected directly from the participant and they will use an ID key (randomly generated) after they complete the game.

All data will be anonymous during the analysis stage.

Open source-code data will be made available by providing scripts of data processing and analysis procedures with various software, including RStudio, Matlab, and Python.

The project team aims to collect data from 120 undergraduate students. It is estimated that the total duration of video data will be nearly 120 hours. The datasets could have several terabytes (TB) original data, plus several TB intermediate data and study results data, in total there could be around 25 TB data.

Table 1 presents the data types and software necessary to view the data in the current project.

Table 1. Data types and software necessary to view the data

Data type	File format	Software
Validated Surveys	.csv; .xlsx, .sav, .txt	Excel; Matlab; R; Python
Content Knowledge Scores	.csv; .xlsx; .sav, .txt	Excel; Matlab; R; Python
Audio	.Mpeg-4	Noldus Observer XT, Windows media player
Video	.Mpeg-4	Noldus Observer XT, Windows media player
Computer screen recording	.Mpeg-4	Noldus Observer XT, Windows media player
Electrodermal activity	.xlsx; .sav,.csv	Excel, R, Matlab, OriginLab, Python
Heart rate	.xlsx; .sav,. csv	Excel, R, Matlab, OriginLab, Python
Digital text data	.txt, .csv	Excel, Notepad, Word, R; Python
Open-source code	.txt, .csv	Notepad, Word, R, Python, GitHub

In addition to the research data summarized above, we will also collect manual consent forms from the participants. The consent forms will include participants' name and surname, e-mail address, and

signature. The consent forms will be collected and stored in a locked cabinet in Dr. Cloude's locked office. The consent forms will be kept on a locked shelf in Dr. Cloude's locked office for five years after the data collection is completed. Following this timeline, they will be destroyed.

1.2 How will the consistency and quality of data be controlled?

The state-of-the-art infrastructure in Ludus laboratory guarantees collection of high quality video, audio, and physiological data. Researchers will carefully install the Biopac wristbands on the participants' wrists and check the electrodermal activity and heart rate signal quality before the data collection starts. Similarly, researchers will place the video cameras to the best positions in Ludus data collection cabins to ensure that facial expressions and concurrent verbalizations of the participants will be recorded from the best possible point of view.

Researchers will collect data by following a written protocol that explains data collection procedures step by step in a detailed manner, including scripts to read to the participant to minimize inconsistencies between data collections. This will ensure that all the participants will go through the exact same data collection process.

After each data collection session, the protocol will also require researchers to download, store, and assess the quality of data. In case of missing or bad quality data, new participants will be recruited for the following sessions.

2. ETHICAL AND LEGAL COMPLIANCE

2.1 What legal issues are related to your data management? (For example, GDPR and other legislation affecting data processing.)

Participation in this study will be completely voluntary. Prior to data collection, an ethics statement for the research needs to be obtained from The Ethics Committee of the Tampere Region. The consent forms asking for participants' permission will be structured according to the recommendations of the ethical committee. Participants of the study must be at least 15 years of age. Thus, it is not necessary to ask their parents' permission for them to participate in the study.

We will abide by the GDPR of the European Union during the research. To ensure anonymity, the participants' names will be replaced with randomly-generated ID numbers from the beginning to the end of data collection, and access to the original video, audio, and computer screen recording data will only be available to authorized researchers. Furthermore, once these data channels have been post-processed and transformed into digital numerical and textual data, they will be permanently destroyed to maintain the privacy of participants.

Other research data (i.e., self-reports, learning test scores, physiological data, digital text data and open-source code data) can be released without privacy restrictions for full public release after the research data has been anonymized upon participant consent since it is not personally identifying. Anonymized data will be shared in legitimate and international open-science archives (e.g., Australian Social Science Data Archive, Research Data Storage Service of the Finnish Ministry of Education and Culture IDA, the Finish Social Science Data Archive, TAU repository Trepo, and the Pittsburgh Science of Learning Center DataShop) for future use upon participant consent.

In addition, data will not be shared unless it is fully anonymized with anyone outside of the research team at Tampere University, or through any form of electronic transfer outside of the Tampere University IT network to avoid any kind of data breach.

Since some pseudonymized data (i.e., audio files) will be used, these data not be stored outside of the university data infrastructure, and all data (anonymized and pseudonymized) will be stored on an encrypted, password-protected and encrypted external hard drive (stored in a locked file cabinet in Dr. Cloude's locked office) and university data server with participant ID numbers.

The research project will not use any data which is covered by the copyright, patents, or any other similar legislation. Research team will avoid any kind of unprotected electronic data file transfer in order to ensure that confidentiality of participants is not jeopardized.

2.2 How will you manage the rights of the data you use, produce, and share?

The research project will not use any data which is covered by the copyright, licenses and patents or any other similar legislation. Tampere University will own the collected data. The intellectual property of the data generated will remain with researchers. Every authorized research partner will sign a contract agreeing that data arising from research projects will be fully anonymized and made openly available. There are no copyright or license issues that restrict the research team from sharing anonymized data publically. However, data sharing will be limited to the data of participants that grant permission for sharing and reuse of their data in future research.

3. DOCUMENTATION AND METADATA

3. How will you document your data in order to make the data findable, accessible, interoperable and re-usable for you and others? What kind of metadata standards, README files or other documentation will you use to help others to understand and use your data?

In the current project, variables and value names will be constructed logically following the data processing guidelines of the Finnish Social Science Data Archive. To ensure the continued use of the data, data collection methods and the contents of the datasheets will be documented carefully and these protocols will be made publicly available on a public repository on the project website on Github (https://ecloude.github.io/portfolio/portfolio-1/).

It will be organized by both data type and collection times. Data Documentation Initiative standards will be followed for the documentation of the data. The metadata will be created according to the METS (Metadata encoding and transfer standard) as suggested by Finnish National Digital Preservation Services. The same metadata description will be used for all sorts of data collected during the project.

Video, audio, and computer screen recordings of participants will be stored within the data corpus in MPEG-4 (.mp4), which is an International Standards Organization (ISO) specification and this format is readable for most media players. MPEG-4 files will include metadata including participant ID number, dates and time of data collection.

The open-source scripts produced during the project will be stored in M-file (.m), python format (.py) or r format (.r), depending on the programming language used in data processing and analysis stages. The version control system of the script will be achieved by sharing the open-source code in a Github repository upon submission for publication review. We will carefully document and explain our processing and analysis procedures with README files to illustrate the data for the convenience of reuse.

Features generated in experiments will be in MATLAB (.mat) and TEXT (.txt) formats. All variables will be described, and aforementioned metadata standards will be used, if available.

The research methods and data results will be published in open-access and international conference proceedings, journals, and technical reports. All of them will also be included in the open institutional repository of Tampere University (https://trepo.tuni.fi/), and made publicly available on the project website on Github.

4. STORAGE AND BACKUP DURING THE RESEARCH PROJECT

4.1 Where will your data be stored, and how will the data be backed up?

Manual materials (e.g., signed consent forms) will be stored in a locked file cabinet in Dr. Cloude's locked office.

Digital data will be primarily stored in two locations: a password-protected external hard drive (stored in a locked cabinet in Dr. Cloude's locked office) and Tampere University data server, linux-ssh.tuni.fi. To provide further protection, the external hard drive will be encrypted with an encryption software (Boxcryptor or Cryptomator) suggested by Tampere University.

Survey and knowledge assessment data will be collected using Lime Survey, an online server environment that is GDPR compliant. Survey and assessment data will be anonymous and stored on the external hard drive and Tampere University data server, linux-ssh.tuni.fi, once the study is completed. All data stored in LimeSurvey will be permanently deleted once data are transferred to the external-hard drive and university data server.

The local workstation in Ludus laboratory will collect video, audio, computer-screen recordings, and electrodermal activity. The video recordings will be immediately anonymized following data collection and the digital text and numerical data generated from post-processing will be immediately stored on the external hard drive and university server using the participant's ID key. Afterwards, the video recording will be permanently destroyed from the local workstation.

The audio recording will be pseudonymised (using participant's ID key) and stored on the external hard drive and university data server immediately following data collection. The audio recording will be permanently deleted from the local workstation after the study. The audio files will be anonymized once they are transcribed into textual files. Once the transcription processing is completed, the audio files will be permanently destroyed on the external hard drive and university data server.

The computer-screen recordings and electrodermal data will be anonymous and will be stored on the external hard drive and university data server immediately after the study. The data collected on the local workstation will be permanently deleted after the data are securely stored on the external hard drive and university data server.

There is minimal security risk associated with storing data. Since some pseudonymized data (i.e., audio files) will be used, these data not be stored outside of the university data infrastructure, and all data (anonymized and pseudonymized) will be stored on an encrypted, password-protected and encrypted external hard drive (stored in a locked file cabinet in Dr. Cloude's locked office) and university data server with participant ID numbers. In addition, data will not be shared unless it is fully anonymized with anyone outside of the research team at Tampere University, or through any form of electronic transfer outside of the Tampere University IT network to avoid any kind of data breach.

4.2 Who will be responsible for controlling access to your data, and how will secured access be controlled?

The project's principal investigator Dr. Elizabeth Cloude will have the ultimate control on access to the project data. During data analysis, the data related to analysis and study output will be accessible only by certified members of the project team. The project team will adhere to Tampere University's information security policy when sharing data among themselves.

All the anonymized data will be stored on secure, password-protected servers. There will be appropriate backups and firewall protection. Third-party access to data will only be granted to *Tutkimustie Oy*, a transcription service provider to transcript the audio files into text files to facilitate coding procedures. Tampere University has established a data protection agreement with Tutkimustie Oy. Additional third parties will not be granted access to the data.

5. OPENING, PUBLISHING AND ARCHIVING THE DATA AFTER THE RESEARCH PROJECT

5.1 What part of the data can be made openly available or published? Where and when will the data, or its metadata, be made available?

The data and metadata will be available to the scientific community after the data are secured, organized and anonymity of the participants is ensured. All the data except for video of participants and their computer screen recordings will be available in an anonymous format and cited in publication. Anonymized data (i.e., , self-reports, learning achievement test scores, performance scores, digital text output for the machine learning methods, digital text data produced by coding of group interaction from video recordings, electrodermal activity and heart rate) will be only shared in open science platforms acknowledge by Tampere University or Academy of Finland (e.g., Finnish Social Sciences Data Archive) with the permission of research participants. We will provide metadata for the archival in such platforms (e.g. DDI- standard). We will publish metadata publicly (i.e., in https://www.fairdata.fi/en/etsin/) to make it discoverable. Open source-code produced during the project can be shared in open science platforms including GitHub and osf.io.

5.2 Where will data with long-term value be archived, and for how long?

The pseudonymised data will be stored in Tampere University's home storage server (linux-ssh.tuni.fi) until post-processing procedures have been completed. Afterwards, they will be permanently destroyed.

Data will be stored only in formats that are recommended by Ministry of Education Guidelines on open science and cultural heritage entities (http://digitalpreservation.fi/files/File-Formats-1.6.1-en.pdf) to ensure long-term use. The original video, audio, and computer screen recording data will be permanently destroyed following post-processing procedures since these data may include personally-identifying information.

6. DATA MANAGEMENT RESPONSIBILITIES AND RESOURCES

6.1 Who (for example role, position, and institution) will be responsible for data management?

Dr. Elizabeth Cloude is a Marie Curie research fellow in the Faculty of Education and Culture at Tampere University will have the utmost responsibility regarding the management and preservation of data. Dr.

Cloude will allocate her time to deal with data management, documentation and storage during and after the project. She has sufficient expertise in managing, preserving and sharing data in digital platforms. Thus, the current research project does not require a special expert to manage the data collected in the current study.

6.2 What resources will be required for your data management procedures to ensure that the data can be opened and preserved according to FAIR principles (Findable, Accessible, Interoperable, Re-usable)?

Ludus laboratory has all the technical infrastructure to record high quality data for this study. Pseudonymized data will be stored at the password protected, secure Tampere University servers for free. External hard disks for storing video, audio, and computer screen recording data will be budgeted from the Faculty of Education and Culture. No other financial resources are necessary to ensure that the data can be opened and preserved according to FAIR principles.