Co-funded by the Horizon 2020
Framework Programme of the European Union

HORIZON2020 FRAMEWORK PROGRAMME

ICT – 21 -2014

Advanced digital gaming/gamification technologies

ProsocialLearn

Gamification of Prosocial Learning

for Increased Youth Inclusion and Academic Achievement

D2.5

Evaluation Strategy and Protocols

Document Control Page

WP/Task	WP2 / T2.4
Title	D2.5-Evaluation Strategy and Protocols
Due date	30/09/2015
Submission date	30/09/2015
Abstract	This document describes the evaluation strategy for the assessment of game effectiveness, market value impact and ethics procedure to drive detailed planning of technical validation, short and longitudinal studies and market viability tests.
Author(s)	Kosmas Dimitropoulos (CERTH)
Contributor(s)	Laura Vuillier (UCAM), Stefano Modeferri (ITINNOV), Lee Middleton (ITINNOV), Christopher Peters (KTH), Francesco D'Andria (ATOS), Anna Zoakou (EA), Kam Star (PG) Spyridon Thermos (CERTH), Kyriaki Kaza (CERTH), Athanasios Psaltis (CERTH), Kyriakos Stefanidis (CERTH), Kostas Apostolakis (CERTH), Petros Daras (CERTH)
Reviewer(s)	Evangelia Dimaraki (EA), Stefano Cobello (EUR)
Dissemination level	☐ internal ☒ public ☐ confidential

Document Control Page

Version	Date	Modified by	Comments
0.1	31/08/2015	Kosmas Dimitropoulos (CERTH)	TOC
0.2	25/09/2015	Kosmas Dimitropoulos (CERTH)	First full version
0.3	22/10/2015	Kosmas Dimitropoulos(CERTH), Laura Vuillier (UCAM)	Contributions to the evaluation of the scientific effectiviness
0.4	29/10/2015	Kosmas Dimitropoulos (CERTH)	Update of the evaluation protocol
0.5	3/11/2015	Laura Vuillier (UCAM) and Anna Zoakou (EA)	Update of the methodology for the evaluation of the scientific effectiveness of prosocial games.
0.6	29/04/2015	Pilar Pérez (ATOS)	Format review final version

List of Abbreviations

Abbreviation	Description
QoS	Quality of Service
QoE	Quality of Experience
PLO	Prosocial Learning Objective
LT	Laboratory Tests
SSE,	Small Scale Experiments
LS	Longitudinal Studies
SaaS	Software as a Service

Executive summary

This document aims to present the assessment plan, developed within Task 2.4, for the assessment and evaluation of the system's independent modules, the integrated platform and the prosocial games. In particular, it defines the evaluation strategy for the game effectiveness, market value impact and ethics procedures to drive detailed planning of technical validation, short and longitudinal studies and market viability tests.

Index

1 Introduction

The aim of the ProsocialLearn project is to create a ground-breaking digital gaming genre in order to help children (7-10 years old) to acquire prosocial skills necessary for positive relationships, team working, trustworthiness and emotional intelligence. The project will deliver a series of disruptive innovations building on a game development and distribution platform for the production of prosocial games that engages children and stimulates technology transfer from traditional game industry to the education sector. ProsocialLearn will also offer game developers scientifically proven prosocial game elements for the development of digital games. An application programming interface (API), named ProsocialAPI, will allow developers to integrate functions into games including visual sensing, identification of prosocial signals from in-game actions, personalized adaptation of game elements, player profiles, game mechanics, expressive virtual characters, and support for data collection with protection of personal data.

The role of WP2 "Gamification of Prosocial Learning" is crucial in the project since it aims to elicit user and system requirements for the gamification of prosocial learning and skill development based on the theoretical understanding of prosociality and its application to the goal of increased youth inclusion and academic achievement. These requirements will provide the foundation for the system architecture, gamification methodology, and validation metrics within the evaluation strategy. The hitherto progress of WP2 includes the successful submission of D2.1 "User requirements" (M3), D2.2 "Prosocial Game Scenarios" (M6) and D2.3 "1st System Requirements and Architecture" (M6). These first user and system requirements and architecture developed within WP2 form the stepping stones for the design of an efficient and realizable technical assessment and evaluation strategy.

1.1 Purpose of the document

This document, D2.5 Evaluation Strategy, is the fourth deliverable of WP2. The scope of this deliverable is to provide the assessment framework, developed within Task 2.4 "Evaluation Strategy and Protocols", for the assessment - evaluation of the ProsocialLearn platform, its modules, the proposed sensor technologies, as well as the effectiveness of the prosocial games that will be developed in WP6 for improving youth inclusion and increasing education achievement of children. The user and system requirements and architecture determined during Tasks 2.1 and 2.3 and described in deliverables D2.1 and D2.3, respectively, are considered as a starting point in order to define appropriate assessment categories, objectives and measurable indices towards the construction of a detailed evaluation strategy.

More specifically, the main objective of this deliverable is to define an evaluation strategy for the assessment of game effectiveness, market value impact and ethics procedures to drive detailed planning of technical validation (WP5), short and longitudinal studies (WP7) and market viability tests (WP1).Moreover, a set of formalized Quality of Experience metrics, derived from the user requirements (T2.1) and a set of formalized Quality of Service metrics, derived from the system's architecture, are defined to play a key role in the design of experimental studies to be carried out in WP7.

1.2 Scope and Audience of the document

The dissemination level of this document is public. The final outcome of this deliverable will be an evaluation strategy to assess the socio-economic impact of the ProsocialLearn platform in trials conducted within education markets in schools throughout Europe.

1.3 Structure of the document

The structure of this document is the following:

Section 2: Overall ProsocialLearn Evaluation Strategy - provides an overview of the assessment-evaluation strategy engaged as well as the organisation and scheduling of the assessment-evaluation process.

Section 3: Assessment of ProsocialLearn Technology and Game Effectiveness - describes the general evaluation plan for the assessment of the ProsocialLearn platform and its components (e.g. player input modalities, data fusion, adaptation mechanism etc.) as well as of the games effectiveness. The evaluation plan includes laboratory tests, small experimental studies and longitudinal studies. The ultimate goal of the proposed plan is to evaluate the effectiveness of prosocial skill development using digital games.

Section 4: Assessment of Market Value Impact –presents the market viability tests and the strategies aiming to explore KPIs, defined by Task 1.3, for service operational performance, cost and pricing characteristics.

Section 5: Assessment of Ethics and Experiments Procedures –presents the evaluation protocols for a series of short and longitudinal experimental studies (pilots) that will be conducted in the different evaluation phases of the ProsocialLearn project. Moreover, this section describes the methodology that will be adopted for the assessment of the ethical procedure during the experiments.

Section 6 and 7: Conclusions – References – contain the conclusions and the references of this report.

At last**, Section 8** is the **Appendix** that presents a questionnaire for social inclusion, a technical assessment report template and an experimental study evaluation report template.

2 Overall ProsocialLearn Evaluation Strategy

2.1 Overview of interdependencies with other WPs

This document is the final outcome of Task 2.4 "Evaluation Strategy and Protocols" and aims to define an evaluation strategy for the assessment of game effectiveness, market value impact and ethics procedures to drive the detailed planning of technical validation, short and longitudinal studies and market viability tests. As shown in Figure 1,the proposed evaluation strategy uses Task 2.1 and Task 2.3 as starting points and has direct interconnection with WP7 "Experimentation and Validation" (design of experimental studies to be carried out in WP7, evaluation of ethics procedure and evaluation of scientific effectiveness), WP5 "Prosocial Platform Development and Operations" (platform testing and operations), WP6 "Prosocial Game Development" (technical validation of prototype prosocial games) and WP1 "Prosocial Game Market Analysis, Exploitation and Business Modeling (market viability tests and strategies to explore KPIs).

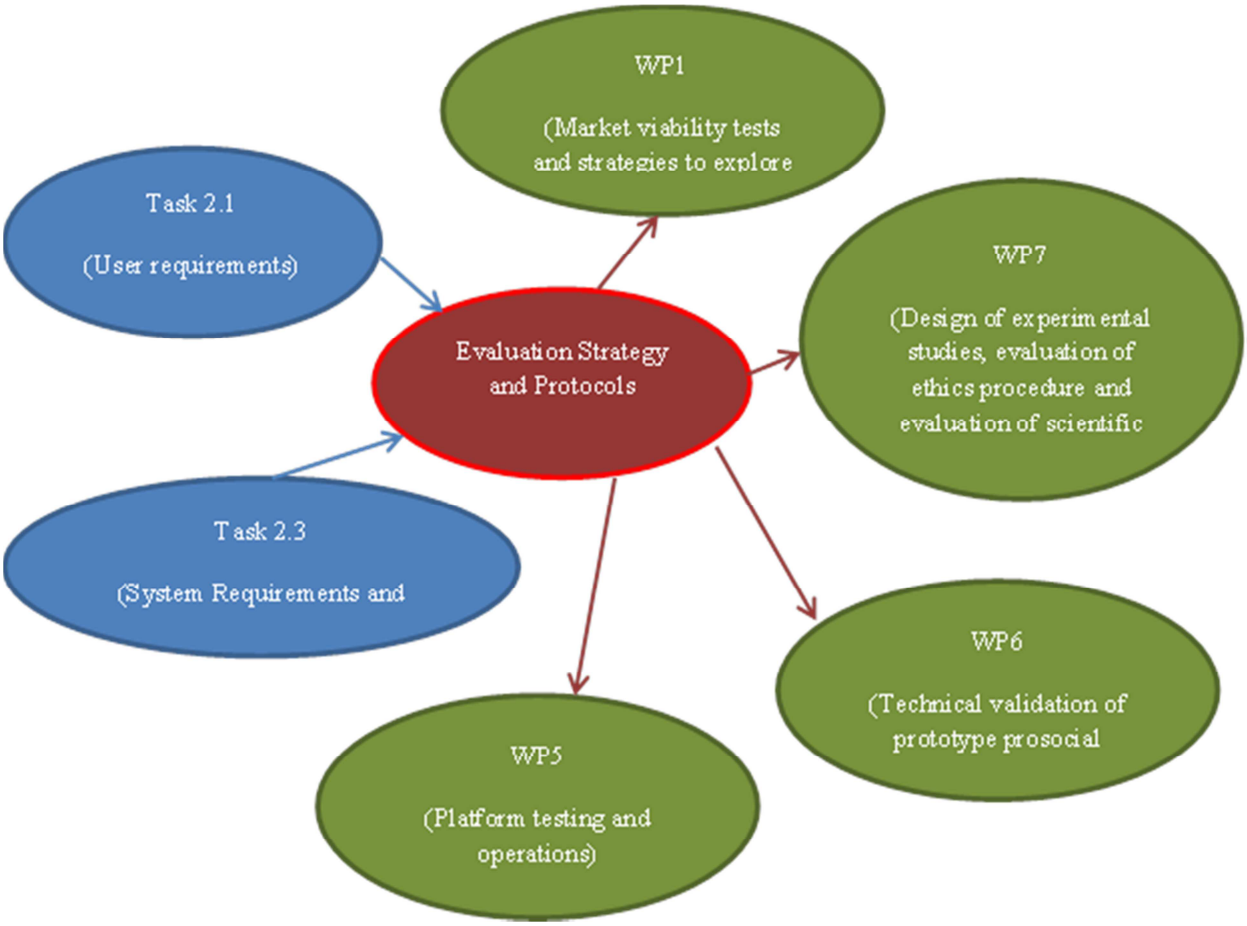


Figure 1 - Interdependencies with other WPs.

2.2 Overall assessment – Evaluation Framework

In this Section an overview of the assessment – evaluation framework is presented by introducing the general categories that are taken under consideration, the rational, as well as the chronological organization of the assessment – evaluation process.

2.2.1 General assessment – Evaluation Methodology

The ProsocialLearn project aims to increase social inclusion and individual empowerment by helping children learn prosocial skills through digital games. The gamification of prosocial learning will be driven by a set of well-defined prosocial learning objectives that are designed for the development of specific prosocial skills, in terms of prosocial theory, gameplay and game mechanics. To this end, the ProsocialLearn platform incorporates different technology modalities that will be recruited to accomplish an efficient, effective and satisfactory conveyance of the intended information to children. To foster the optimization of the aforementioned characteristics of the platform, i.e., efficiency, scientific effectiveness, satisfaction etc, an assessment - evaluation process has to be implemented during the development and testing phases of the system. Figure 2 offers an overview of the expected evolution of the development and assessment – evaluation processes within the ProsocialLearn project.

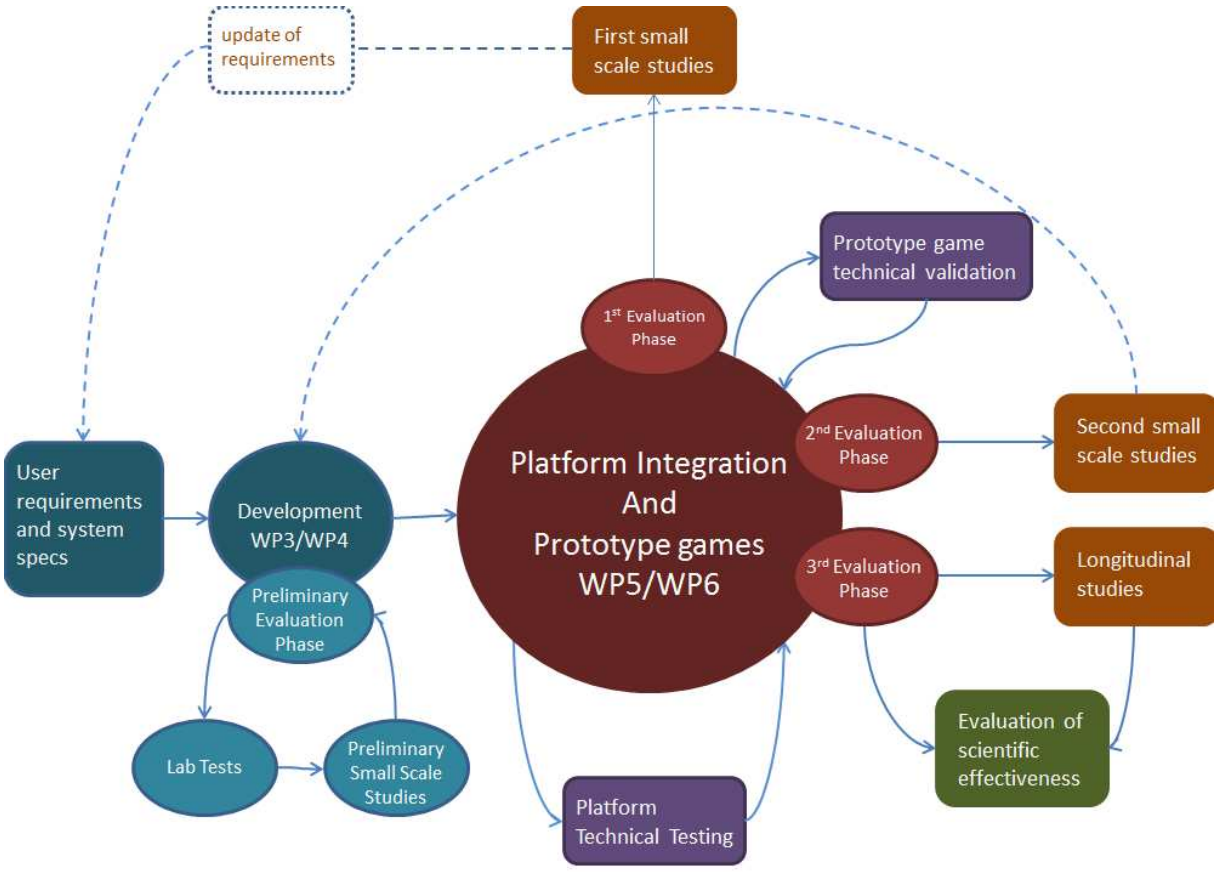


Figure 2 – ProsocialLearn Evaluation Strategy

The methodology adopted here includes a preliminary evaluation phase during the development phase of WP3 and WP4 as well as three successive evaluation phases aiming to provide a multilateral assessment process covering the technical validation of the platform and the proposed technology,

as well as the overall scientific effectiveness of the games. More specifically, the proposed evaluation methodology consists of the following phases:

- **Preliminary evaluation phase (M1-M8):** This evaluation phase deals with the assessment of WP3 modules and will be performed during the development stage of the project. Initially, tests will be conducted in laboratory conditions, while subsequently small scale trials will be performed to evaluate the modules' performance and functionalities. The measurements collected through this process will be analyzed in order to optimize the performance of the technological modalities, such as sensors, processing algorithms and interfaces.
- **First evaluation phase (M9-M15):** The first evaluation phase aims to assess the performance of the first version of the platform and its modules, as well as the effectiveness of the initial prosocial games (e.g. path of trust, Kitty King's Candy Quest, Cooperative game). In this evaluation phase, a series of small scale experiments will be launched in operational or near operational school conditions. The studies will run to collect data for WP3 and WP4 modules, e.g., data fusion, adaptation etc., and validate the functionalities of the initial prosocial games. The analysis of the data collected through the first phase of small scale studies will be used as feedback for the update of the system requirements. This will allow the consideration of any upcoming problems and limitations as well as additional requirements experienced during the first evaluation phase. In addition, valuable feedback is expected after the completion of this phase for the optimization of the first version of the platform and its modules. Anonymized data sets for the development of user modelling, fusion and adaptation algorithms will be recorded.
- **Second evaluation phase (M16-M24):** The second evaluation phase will assess the final version of the platform and its components as well as the prosocial games developed in Task 6.2 in operational or near operational school conditions. The main objective of this evaluation phase is to validate the functionality and user acceptance of the prototype games developed in Task 6.2. Moreover, user feedback will be of vital importance for improving/adjusting platform aspects related to graphics, virtual characters, adaptation and natural interaction. The collected data will be reported by the students and the system in the form of logs regarding affective and game-related cues, quality of experience/service reports, as well as functional validation.
- **Third evaluation phase (M25-M36):** In this final evaluation phase a series of longitudinal studies will be conducted using the prosocial games developed in WP6. This evaluation process is planned in two distinct stages:
    - In the first stage a set of studies will commence using games developed in Task 6.2,
    - In the second stage a set of studies will commence based on games developed in Task 6.3 by partners involved in the third year.
- In both stages, the studies will be conducted using a mature platform and tested in real school conditions. The main objective of these studies will be the collection of data indicating prosocial learning outcomes of students resulting from prosocial game playing in real-world conditions. The collected data will be used as input for the evaluation of the scientific effectiveness of the games and will be the final outcome of the ProsocialLearn evaluation process. The evaluation of the scientific effectiveness of the games will be based on the analysis of the collected data and will assess the ProsocialLearn's potential to have a societal impact (i.e., increase social inclusion and academic achievement in young children), ii) derive correlation among Quality of Experience (QoE), Quality of Service (QoS) and tutors' feedback

and iii) make recommendations for game certification procedures to be applied by the platform.

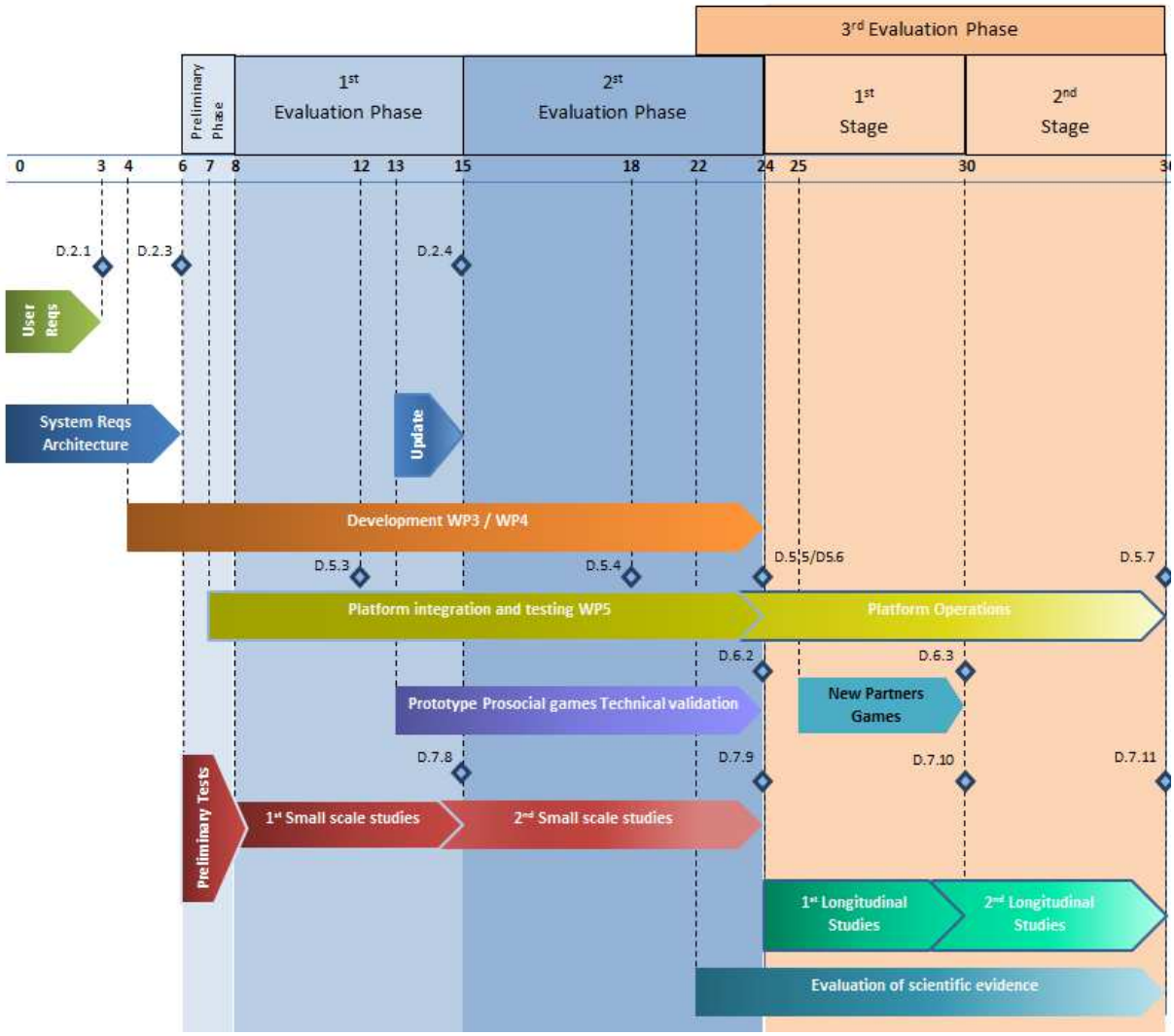


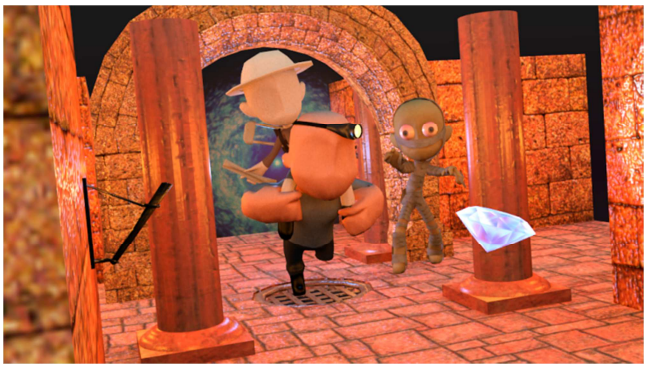
Figure 3 - Organization of the ProsocialLearn evaluation plan.

2.2.2 Organization and Scheduling of Assessment – Evaluation Methodology

The assessment - evaluation process of the ProsocialLearn project is planned in three distinct phases along with a preliminary phase with respect to time. Although the preliminary phase was not initially foreseen in the DoW (Description of Work) of the project, the consortium decided to conduct a series of small scale experiments in order to collect data to be used for the development and optimization of various technological modalities, such as sensors configuration, processing algorithms and interfaces.

More specifically, between **month 6 and 8,** WP3 modules were tested in laboratory conditions, while subsequently small scale trials were performed in different schools in Greece to evaluate the modules' performance and functionalities. For the collection of data, a first prototype game, the "Path of Trust", was developed by CERTH aiming to build up trustworthiness and teamwork among children aged 7-10.

The storyline of the game focuses on two adventurers who agree on working together in order to explore an ancient Egyptian tomb and collect the treasures hidden within. It just so happens that one of the two (an Indiana Jones wannabe old-timer) has suffered a serious injury during a past attempt at exploring the tricky corridors and has to be carried around by the other character, portraying a traditional, muscle-bound mercenary with practically zero experience in dungeon crawling. Together, these two agree on embarking on a treasure hunting quest, where one player has to properly provide directions as to where to go to next in order to avoid roaming mummies and traps, while the other has to navigate the environment and try to collect as much treasure pieces as possible. The game features colorful, immersive 3D graphics, cheerful cartoon characters as the main protagonists and up to five different endings in response to players' cooperation efforts and mutual expression of trust. The game supports both traditional and gesture-driven gameplay through three game input configurations including Keyboard, LEAP Motion and Microsoft Kinect sensors.



(a)

(b)

Figure 4: (a) The "Path of Trust" game and (b) Experiments with children in Portaria Elementary School, in Greece.

For the collection of data, four experiments with children (7-10 years old) were organized by CERTH and EA in Greece:

- The first small scale experiment was organized in Portaria Elementary school on 9th June 2015 where18 children were tested.
- For the optimization of face recognition and body motion analysis algorithm, additional tests were conducted at the premises of CERTH, initially with two (17 June 2015) and then with six children (21 June 2015).
- Finally, a small scale experiment with 16 children was also organized at Ellinogermaniki School on 24th June 2015, in Athens.

The first evaluation phase will focus on the technical performance assessment of WP3 and WP4 modules and validate the functionalities of the first prosocial games (e.g. Path of Trust, Kitty King's Candy Quest, Cooperative game) and the first version of the platform (D5.3 "1st Prosocial platform release, M12). It will start in **month 9** and finish in **month 15** in order to give feedback to WP2 and, particularly, to use in the deliverable D2.4 "2nd System Requirements and Architecture" (due M15). The results of this evaluation phase will be described in deliverable D7.8 "1st Results of small experimental studies", which will be submitted in month 15.

In the second evaluation phase, which will start in **month 16 and finish in month 24,** the technical assessment of the final version of the platform and its components will be performed. Experimental

results in operational or near operational school conditions will be conducted to improve/adjust the ProsocialLearn platform (D5.4 "2nd Prosocial platform release", month 18, D5.5 "3rd Prosocial platform release, month 24, and D5.6 "1st Platform operations report") and validate the functionality and user acceptance of the prototype games (D6.2 Prototype Prosocial Games, month 24). The results of this evaluation phase along with the procedures for acquiring, using and evaluating components and technologies as platform and game prototypes will be described in detail in deliverable D7.9 "2nd Results of small experimental studies".

After the completion of the small scale studies, the third evaluation phase will start in **month 22 and finish month 36** consisting of the two stages of the longitudinal studies as well as the evaluation of the scientific effectiveness. More specifically, the first stage of the longitudinal studies is expected to last from **month 25 to month 30** in order to evaluate the effectiveness of the games developed in Task 6.2, while a second set of studies will be conducted from **month 31 to month 36** using the games developed in Task 6.3 by partners involved in the third year of the project. In parallel, the evaluation of the scientific effectiveness of ProsocialLearn will be performed from **month 22 to month 36** and the results of this analysis will be described in D7.10 "1st Validation activities in operating school conditions", month 30, and in D7.11 "Validation activities in operating school conditions", month 36.

3 Assessment of ProsocialLearn Technology and Game Effectiveness

The methodology adopted here focuses on three perspectives: i) the technical performance assessment of the platform and its components (technical performance assessment), ii) the usability/acceptability evaluation of the proposed technology and iii) the scientific effectiveness of prosocial games. As far as the technical performance is concerned, scientific expertise is required in order for the proposed technology to be properly assessed, while the usability evaluation mandates valuable feedback from the users' perspective. Finally, the evaluation of the scientific effectiveness of games requires the analysis from expert psychologists of the data collected during the technical performance assessment as well as the feedback received from teaching professionals. These three perspectives -which will be performed in the evaluation phases described in the previous section - are described in more details below.

- **Technical performance assessment:** Technical performance assessment is critical for an optimized implementation of the technological modalities, such as sensors, processing algorithms, mechanisms and interface.. As these modalities require scientific knowledge and expertise, it relies mainly on the researchers involved in the project to perform the assessment. To this end, assessment categories and corresponding indices are introduced which are based on the system requirements and architecture defined in deliverable D2.3 "First System Requirements and Architecture". Due to the different characteristics of each module/platform's component, specific technical performance categories and indices are introduced for each one of them e.g., facial expression analysis, data fusion, adaptation algorithm etc., while specific QoS (Quality of Service) metrics are defined for the assessment of the platform's performance.
- **Usability/Acceptability evaluation:** Usability/acceptability is a crucial characteristic of ProsocialLearn platform and games, which aim to increase social inclusion and individual empowerment by helping children learn prosocial skills. In order to evaluate the usability, a series of small scale and longitudinal experiments will be organized in operational or near operational school conditions. Usability/acceptability data will be acquired using both traditional techniques (e.g. questionnaires) and automated tools (software tools for usability data collection), while a set of QoE (Quality of Experience) metrics derived from the user requirements, and specifically deliverable D2.1 "User Requirements", will be defined.
- **Scientific effectiveness of ProsocialLearn games:** This assessment category aims to evaluate the effectiveness of prosocial skill development using digital games for increasing youth inclusion and academic achievement. This evaluation requires the collaboration of expert psychologists and teaching professionals to assess the impact of prosocial games. The analysis of data collected by the ProsocialLearn platform, the feedback received from teaching professionals and the correlation among QoE, QoS and tutor's feedback will play a crucial role in the evaluation of scientific effectiveness.

3.1 Technical performance assessment of modules

This section presents the assessment criteria and indices that will be used for the technical assessment of the platform and its individual modules. Every assessment index is accompanied by:

- A short description explaining which quality/feature is measured/assessed and the type of data (Numerical, Qualitative, Continuous, Binary, Discrete, Ordinal etc.),
- The values that the index may acquire,

- The codes of the user or system requirements that might be examined by the specific index (if applicable), as defined in D2.1 "User Requirements" and D2.3 "First System Requirements and Architecture".
- The suggested type of experiments (Laboratory Tests - LT, Small Scale Experiments – SSE, Longitudinal Studies – LS) that the index will be used. However, this declaration is not binding. Every assessment index can and may be used in every assessment phase if the circumstances require so.

3.1.1 Player input modalities

In the context of ProsocialLearn project, the main input modalities that will be assessed can be divided in two broad categories: i) visual input and ii) audio input.

3.1.1.1 Visual input

Facial Expression Analysis

The following table presents critical performance indices that will be used for the technical assessment of facial expression analysis module.

Module	**Facial Expression Analysis**					
Assessment Category	**Assessment Indices**				**Requirements Examined/ KPIs**	**Type of experiments**
	ID	**Description**	**Type**	**Values**		
Facial featuretracking accuracy	LocalError	Mean facial feature localization error in mm and/or pixels. To measure the performance of the face tracking algorithm, we compare the estimated feature positions against their real (ground-truth) positions.	Numerical	0-Inf	p.REQ4 KPI 3.2	LT / SSE
Facial Action Unit (AU) recognition accuracy	ConfMatAU	Confusion matrix: The element *(i,j)* of the confusion matrix represents the percentage of instances that AU *i*(ground truth) was classified as AU *j* (result of classifier).	Matrix with numerical values	0-1	p.REQ4 KPI 3.2	LT / SSE
	AccAU	Classification accuracy: ratio of correctly predicted AUs over the total number of predictions.	Numerical	0-1	p.REQ4 KPI 3.2	LT / SSE
	Conf	Confusion matrix: The	Matrix	0-1	p.REQ4	LT

Basic emotion recognition accuracy	MatEmo	element *(i,j)* of the confusion matrix represents the percentage of instances that emotion *i*(ground truth) was classified as emotion *j* (result of classifier).	with numerical values		KPI 3.2	/ SSE
	AccEmo	Classification accuracy: ratio of correctly predicted emotions over the total number of predictions.	Numerical	0-1	p.REQ4 KPI 3.2	LT / SSE

Confusion matrix: a specific matrix layout that allows visualization of the performance of a machine learning algorithm. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class (ground truth). The element (*i,j*) of the confusion matrix represents the ratio of instances that a sample from class *i* was classified as class j over the total number of instances of class *i*. The matrix makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another).

For the assessment of the module, the following test data sets will be used:

- Datasets of image sequences recorded by CERTH-ITI. The database will be comprised of sequences of 2D images showing children mimicking basic facial expressions and performing a subset of the action units of the FACS system. This data set will be used for the laboratory testing of this module (Preliminary evaluation phase).
- Existing datasets (e.g. Facial Expression Recognition and Analysis challenge, IEEE Int'l. Conf. Face and Gesture Recognition, FG'11,[1])for AU detection and expressions of discrete emotion recognition will also be used during the laboratory testing of the facial expression analysis module.
- ProsocialLearn data recordings. In the context of the preliminary evaluation tests as well as the first and second evaluation phase (small scale experiments), a set of video sequences will be recorded and will be used for algorithm assessment.

Gaze Analysis

For the evaluation of the gaze analysis module the following indices will be used:

Module	**Gaze Analysis**					
Assessment Category	**Assessment Indices**				**Requirements Examined/ KPIs**	**Type of experiments**
	ID	**Description**	**Type**	**Values**		
Remote Gaze Tracker Accuracy	MeanAngularError	Using the gaze tracker experiment defined in [2] we estimate the accuracy of our gaze tracker by extracting the mean gaze angle deviation that corresponds to the distance of the estimated	numerical	0-Inf (degrees)	p.REQ4 KPI 3.2	LT / SSE

		gaze location on the screen from the center of the depicted circle displayed following a circular trajectory.				
Blink Detection accuracy	Conf MatAU	Confusion matrix: The element *(i,j)* of the confusion matrix represents the percentage of instances that AU *i*(ground truth) was classified as AU *j* (result of classifier).	Matrix with numerical values	0-1	p.REQ4 KPI 3.2	LT / SSE
	AccAU	Classification accuracy: ratio of correctly predicted AUs over the total number of predictions.	Numerical	0-1	p.REQ4 KPI 3.2	LT / SSE
Visual Attention	Conf MatEmo	Confusion matrix: The element *(i,j)* of the confusion matrix represents the percentage of instances that attention *I* (ground truth) was classified as attentive/non-attentive *j* (result of classifier).	Matrix with numerical values	0-1	p.REQ4 KPI 3.2	LT / SSE
	AccEmo	Classification accuracy: ratio of correctly predicted emotions over the total number of predictions.	Numerical	0-1	p.REQ4 KPI 3.2	LT / SSE

For the assessment of the gaze analysis module, we will use the same datasets described in the previous section for the evaluation of the facial expression analysis.

Body Motion Analysis

The main criteria for the evaluation of the adaptation algorithm are presented in the following table:

Module	**Body Motion Analysis**					
Assessment Category	**Assessment Indices**				**Requirements Examined/ KPIs**	**Type of experiments**
	ID	**Description**	**Type**	**Values**		
Tracking accuracy	Tr_Acc	The tracking accuracy index in each frame measures the sum of tracked joints confidence values divided by the total	Numerical	0-1	p.REQ4 KPI 3.2	LT / SSE

		number of skeletal joints.				
Validation of body motion analysis feature accuracy	Feat_ Acc	To evaluate the importance and the accuracy of information that body motion analysis features can offer, we compare the estimated features waveforms against annotated video recordings.	p.REQ4	LT / SSE	p.REQ4 KPI 3.2	LT / SSE
Basic emotion recognition accuracy	Conf MatEmo	Confusion matrix: The element *(i,j)* of the confusion matrix represents the percentage of instances that emotion *i*(ground truth) was classified as emotion *j* (result of classifier).	Matrix with numerical values	0-1	p.REQ4 KPI 3.2	LT / SSE
	AccEmo	Classification accuracy: ratio of correctly predicted emotions over the total number of predictions.	Numerical	0-1	p.REQ4 KPI 3.2	LT / SSE

For the assessment of the body motion analysis module, the following test data sets will be used:

- Datasets recorded by CERTH-ITI. The database will be comprised of Kinect data (skeletal data, depth and RGB video sequences) from subjects mimicking basic body motion expressions. This dataset will be used for the laboratory testing of this module (Preliminary evaluation phase).
- ProsocialLearn data recordings. In the context of the preliminary evaluation tests as well as the first and second evaluation phase (small scale experiments), a set of Kinect data recordings will be used for the assessment of the algorithm.

3.1.1.2 Audio input

In the following table there is a description of the main methodologies for validating the detection of emotion from voice.

Module	**Audio input**					
Assessment Category	**Assessment Indices**				**Requirements Examined/ KPIs**	**Type of experiments**
	ID	**Description**	**Type**	**Values**		
Voice emotion evaluation	Conf MatVoice	Confusion matrix: The element *(i,j)* of the confusion matrix represents the percentage of instances that emotion *i*(ground truth) was	Numeric	[0,1]	p.REQ4 KPI 3.2	SSE/LT

		classified as emotion *j* (result of classifier).				
	AccVoice	Classification accuracy: ratio of correctly predicted emotions over the total number of predictions.	Numeric	[0,1]	p.REQ4 KPI 3.2	SSE/LT

The following datasets will be employed in the validation:

- FAU AIBO: this dataset contains recordings of children's interactions with an AIBO rrobot. Ground truth labels are provided for a variety of different emotional classes.
- Prosocial Learn data recordings: this dataset is captured as part of the experiments undertaken. These will also be labeled and used for evaluation and improving classifier performance.

3.1.2 Dynamic data fusion

In order to evaluate the effectiveness of the fusion algorithms the methods described in the following table will be applied.

Module	**Dynamic data fusion**					
Assessment Category	**Assessment Indices**				**Requirements Examined/ KPIs**	**Type of experiments**
	ID	**Description**	**Type**	**Values**		
Basic emotion recognition accuracy	ConfMatFusionEmo	Confusion matrix: The element *(i,j)* of the confusion matrix represents the percentage of instances that emotion *i*(ground truth) was classified as emotion *j* (result of classifier).	Matrix with numerical values	0-1	p.REQ4 KPI 3.3	LT / SSE
	AccFusionEmo	Classification accuracy: ratio of correctly predicted emotions over the total number of predictions.	Numerical	0-1	p.REQ4 KPI 3.3	LT / SSE
Valence Arousal emotional space recognition accuracy	AccVA	Classification accuracy: ratio of correctly predicted valance arousal values over the total number of predictions.	Numerical	0-1	p.REQ4 KPI 3.3	LT / SSE
Engagement recognition accuracy	AccEng	Classification accuracy: ratio of correctly predicted level of engagement over the total number of	Numerical	0-1	p.REQ4 KPI 3.3	LT / SSE

		predictions.				

Confusion matrix: a specific matrix layout that allows visualization of the performance of a machine learning algorithm. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class (ground truth). The element (i,j) of the confusion matrix represents the ratio of instances that a sample from class i was classified as class j over the total number of instances of class i. The matrix makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another).

For the assessment of the module, the following test data sets will be used:

- Multimodal datasets recorded by CERTH-ITI. The overall recording procedure will be based on the GEMEP corpus, a multimodal collection of portrayed emotional expressions: we will record data on facial expressions, body movement and gestures and speech. The database will be comprised of sequences of 2D images, Kinect data streams and audio signals showing children and adults mimicking basic affective states and performing specific gestures that exemplify each emotion. This data set will be used for algorithm assessment.
- ProsocialLearn data recordings. In the context of the preliminary evaluation tests as well as the first and second evaluation phase (small scale experiments), a set of video sequences will be recorded, annotated and will be used for algorithm assessment.

3.1.3 Adaptation mechanism

The table below presents the major criteria that will be used for the technical assessment of the adaptation algorithm.

Module	**Adaptation mechanism**					
Assessment Category	**Assessment Indices**				**Requirements Examined/ KPIs**	**Type of experiments**
	ID	**Description**	**Type**	**Values**		
User Satisfaction Index	USI_Adapt	Prediction Accuracy: Sum of squared differences between user affective states before and after the adaptations performed in a game	Numerical	[-1,-1]	u.REQ27 KPI 4.1 KPI 4.2	LT / SSE
Confidence	ConfAdapt	Sign Test: The significance level that algorithm A is not truly better than B, i.e. the probability of at least n_A out of n 0.5-probability Binomial tests succeeding. (where $n_A$is the number of users that preferred algorithm A over B, and n = n_A+ n_B)	Numerical	0-100	u.REQ27 KPI 4.1 KPI 4.2	LT / SSE

Trust	T_Adapt	How much do the users trust the mechanism's suggestions	Numerical	0-10	u.REQ27 KPI 4.1 KPI 4.2	LT / SSE
Scalability	S_Adapt	Time complexity	Numerical		u.REQ27 KPI 4.1 KPI 4.2	LT / SSE

For the assessment of the module, the following approaches will be used to determine user state (i.e. engagement, prosocial state, etc.):

- Implicitly during the game from the fusion module estimation.
- Explicitly using questionnaires at the end of the game.

User Satisfaction Index: a custom measure of accuracy that will be used to compare adaptation mechanisms. The measure gives a view of algorithmic performance for a user in a single game, emphasizing in the magnitude of change in user state that each game adjustment introduced. User state is determined implicitly by the fusion mechanism.

Confidence: a comparative study between two algorithms. Each user plays a game twice, in each game of which a different adaptation algorithm is used. At the end of the game the preference of the user is determined via questionnaires querying the user on choosing which of the two games she preferred towards the personalization to her needs.

Trust: at the end of each game the users are asked to rate via questionnaires the level of personalization that the game managed to achieve.

For the evaluation of the adaptation algorithm, data recordings from prosocial games during the first and second evaluation phase (small scale experiments) will be used.

3.1.4 Prosociality Mechanism

Ground truth evidence for the evaluation of prosocial models in kids is not available and therefore experts will define the correct procedures according to the conducted experiment. The following table summarizes the main criterion.

Module	**Prosociality Mechanism**					
Assessment Category	**Assessment Indices**				**Requirements Examined/ KPIs**	**Type of experiments**
	ID	**Description**	**Type**	**Values**		
Confidence	AccProso	Accuracy: ratio of correctly predicted level of prosociality over the total number of predictions.	Numeric	[0,1]	p.REQ2 p.REQ4 KPI 3.1	SSE

For the assessment of the mechanism, the approach will consider:

- Game events.
- Possibly questionnaires at the end of the game.

3.1.5 Natural game interaction

The following table presents the performance indices that will be used for the technical assessment of natural game interaction module.

Module	**Natural game interaction**					
Assessment Category	**Assessment Indices**				**Requirements Examined/ KPIs**	**Type of experiments**
	ID	**Description**	**Type**	**Values**		
Tracking accuracy	Tr_Acc	The tracking accuracy index in each frame measures the sum of tracked joints confidence values divided by the total number of skeletal joints.	Numerical	0-1	p.REQ4 KPI 2.3	LT / SSE
Recognition of human action	ConfMatEmo	Confusion matrix: The element *(i,j)* of the confusion matrix represents the percentage of instances that action *i*(ground truth) was classified as action *j* (result of classifier).	Matrix with numerical values	0-1	p.REQ4 KPI 2.3	LT / SSE
	AccEmo	Classification accuracy: ratio of correctly predicted actions over the total number of predictions.	Numerical	0-1	p.REQ4 KPI 2.3	LT / SSE

For the assessment of the natural game interaction module, the following test data sets will be used:

- Existing Kinect data sets, e.g., MSRC-12 [9], G3D [10] and MSR Action3D [11] Datasets for gesture and action recognition will also be used during the laboratory testing of the natural game interaction module.
- ProsocialLearn data recordings. In the context of the preliminary evaluation tests as well as the first and second evaluation phase (small scale experiments), a set of Kinect data recordings will be used for the assessment of the algorithm.

3.1.6 Expressive virtual characters

The main output modalities of virtual character expression that will be assessed can be divided into the facial area; the body (excluding facial expressions); full face and bodily expressions; and higher-level expressions potentially associated with impressions of prosocial character [12][14] and traits (e.g. trustworthiness, cooperation). Such expressions may be attentive, as well as emotional, in nature – for example, expressing social engagement through appropriately maintained eye contact (a subcategory of 'Facial expression recognition' below). The appearance and embodiment of virtual characters are also of importance: see, for example [13]. The definition of a small set of test characters with varying characteristics (see the characters in [14] for example) is one option that will be explored for this purpose.

It should be noted that in contrast to 3.1.1 'Player input modalities', the expressive behaviors of virtual characters are classified according to ratings made by human participants, who view them during controlled user studies. The overall purpose is to ensure appropriate control in the final system i.e. so that the integrated system can select the appropriate expressions in order to ensure characters and behaviors that provide the desired prosocial impressions to viewers.

Module	**Expressive virtual characters**					
Assessment Category	**Assessment Indices**				**Requirements Examined**	**Type of experiments**
	ID	**Description**	**Type**	**Values**		
Facial expression recognition accuracy	Conf MatFACE	Confusion matrix: The element *(i,j)* of the confusion matrix represents the percentage of instances that facial expression *i*(ground truth) was classified as expression *j* (result of participant ratings).	Matrix with numerical percentage values	0-100	p.REQ4 KPI 4.1	LT / SSE
	AccFACE	Classification accuracy: ratio of correctly predicted facial expressions over the total number of predictions.	Numerical	0-1	p.REQ4 KPI 4.1	LT / SSE
Bodily expression recognition accuracy	Conf MatBODY	Confusion matrix: The element *(i,j)* of the confusion matrix represents the percentage of instances that bodily expression *i*(ground truth) was classified as expression *j* (result of participant ratings).	Matrix with numerical percentage values	0-100	p.REQ4 KPI 4.1	LT / SSE
	AccBODY	Classification accuracy: ratio of correctly predicted bodily expressions over the total number of predictions.	Numerical	0-1	p.REQ4 KPI 4.1	LT / SSE
Basic expression recognition	Conf MatEBASIC	Confusion matrix: The element *(i,j)* of the confusion matrix represents the percentage of instances that full-body, basic expression *i*(ground truth) was classified as expression *j* (result of participant ratings).	Matrix with numerical percentage values	0-100	p.REQ4 KPI 4.1	LT / SSE

	AccEBASIC	Classification accuracy: ratio of correctly predicted basic expressions over the total number of predictions.	Numerical	0-1	p.REQ4 KPI 4.1	LT / SSE
Recognition of nonverbal signs suggestive of prosocial character	ConfMatEPROSOCIAL	Confusion matrix: The element *(i,j)* of the confusion matrix represents the percentage of expressions suggestive of prosocial character [8] (e.g. trustworthiness) *i*(ground truth) was classified as expression *j* (result of participant ratings).	Matrix with numerical percentage values	0-100	p.REQ4 KPI 4.1	LT / SSE / LS
	AccEPROSOCIAL	Classification accuracy: ratio of correctly predicted expressions over the total number of predictions.	Numerical	0-1	p.REQ4 KPI 4.1	LT / SSE / LS

Confusion matrix: a specific matrix layout that allows visualization of the performance of participants in recognizing expressions. The matrix makes it easy to see if participants confuse different cases of expressive stimuli (i.e. for example, by mislabelling a sad facial expression as one of disgust [15]).

Annotated datasets will be used for the development of ground-truth virtual expressions, where possible. In the case of bodily expressions, options include the Carnegie-Mellon Graphics Lab Motion Capture Database[1] and the UCLIC Affective Posture and Body Motion Database [16], an annotated database of acted expressions (e.g. angry, fearful, happy and sad expressions) recorded using a motion capture system. Archetypal facial expressions will be constructed offline from reference facial expression datasets and/or face capture technologies (from WP3 modules and external programs as required). Further stages of development, involving the investigation of more direct, possibly real-time, user behavior mappings onto virtual characters via WP3 modules, will involve similar evaluation criteria as above.

3.2 Prosocial games evaluation

3.2.1 Usability Evaluation through prosocial games

Through time many definitions for 'usability' have been proposed. Two of the most established definitions can be found in international standard for the evaluation of software ISO 9241-11[3] and ISO 9126[4]. ISO 9241-11defines usability as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use". In ISO 9126, usability is defined as "the capability of the software product to be understood, learned, used and attractive to the user, when used under specified conditions". In other words, usability studies relate to evaluating a product by testing it on representative users while they focus not only on how well users can learn and use a product to achieve their goals but also on how

[1] http://mocap.cs.cmu.edu/

satisfied users are with that process. This can be seen as an irreplaceable usability practice since it gives direct input on how real users use the system. Usability studies examine three principles: effectiveness, efficiency and overall satisfaction of the user.

- Effectiveness is the capability of the product to enable users to achieve specified goals with accuracy and completeness in a specified context of use.
- Efficiency is the capability of the product to enable users to expend appropriate amounts of resources in relation to the effectiveness achieved in a specified context of use.
- Satisfaction is the capability of the product to adequately satisfy users in a specified context of use.

In this context, usability evaluation will be performed through a series of short and longitudinal studies. These studies/experiments will engage an adequate number of real users/children, so as to extract valid conclusions. The objectives of the evaluation procedure adopted in these studies are mainly based on the user and system requirements that are identified within deliverables D2.1 and D2.3.

Below we describe a series of methods for gathering usability data that will be considered within the framework of the ProsocialLearn project. The final choice of the method depends on many factors e.g. the type of the experiment, the place where the experiment will be conducted, the available tools, the number and expertise of evaluators, the number of children/participants, the available time etc.

3.2.2 Usability Data Collection Methods

It is common during a usability study to ask participants to complete the tasks to be evaluated while observers watch, listen and take notes. The goal is to identify any usability problems, collect qualitative (that approximate or characterize but does not measure the attributes, properties, and characteristics of a thing or a phenomenon) and quantitative (that quantify and verify the attributes, properties, and characteristics of a thing or a phenomenon) data and better understand the users' satisfaction with the product and their motivations/perceptions in addition to their actions. The methods popularly used to gather usability data can be divided into two categories, namely testing and inquiry, and are described below.

3.2.2.1 Testing

In usability testing approach, representative users work on typical tasks using the system (or the prototype) and the evaluators use the results to understand how the user interface as well as the system in general supports the users to perform their tasks. The most popular techniques used to gather data during a usability test are the following.

Think Aloud Protocol

Think Aloud Protocol was introduced in the usability field by Clayton Lewis [5] and was based on the techniques of protocol analysis by Ericsson and Simon [6]. Think Aloud Protocol involves participants thinking aloud as they are performing a set of specified tasks. During the course of a usability test, the test users are asked to verbalize their movements, thoughts, feelings, and opinions while interacting with the system. That is the reason why it is also referred as Concurrent Think Aloud Protocol so as to differentiate it from Retrospective Think Aloud described in Section 5.2.1.2. More specifically, the test users are provided with the product to be tested and a set of tasks to perform. Then, they are asked to perform the tasks using the product and explain what they are thinking

about while working with the product's interface. Observers of such a test are asked to objectively take notes of everything that users say, without attempting to interpret their actions and words. Test sessions are often audio- and video-recorded so that developers can go back and refer to what participants did and how they reacted. The purpose of this method is to make explicit what is implicitly present in subjects who are able to perform a specific task.

Thinking aloud is very useful in capturing a wide range of cognitive activities and enables observers to see first-hand the process of task completion (rather than only its final product). Furthermore, it allows observers to understand how the user approaches the interface and what considerations the user keeps in mind when using the product. If the user expresses that the sequence of steps dictated by the product to accomplish their task goal is different from what they expected, perhaps the interface is convoluted. Although the main benefit of the thinking aloud protocol is to better understand the user's mental model and interaction with the product, there are other benefits as well. For example, the terminology the user uses to express an idea or function should be incorporated into the product design or at least into its documentation. However, the main drawbacks of Thinking Aloud Protocol are the non-natural environment of the testing process to the user and the inability to capture quantitative data.

Usability principles covered:

Effectiveness: Yes**, Efficiency:** No**, Satisfaction:** Yes

Retrospective Testing

Retrospective Testing or Retrospective Think Aloud is a form of Think Aloud Protocol that is performed after the user testing session activities instead of during them. Fairly often the retrospective protocol is stimulated by using a visual reminder such as a video replay. If a video replay of the usability test session is available, the observers can collect more information by reviewing the replay together with the user participants and asking them questions regarding their behavior during the test. Consequently, this technique should be used along with other techniques, especially those where the interaction between the observers and the participants is restricted. Moreover, both quantitative and qualitative data can be collected while in concurrent thinking aloud quantitative information gathering is not an option. However, in retrospective testing each test session lasts at least twice as long. Another obvious requirement for using this technique is that the user's interaction with the computer needs to be recorded and replayed.

Usability principles covered:

Effectiveness: Yes, **Efficiency:** Yes, **Satisfaction:** Yes

Co-discovery Learning

Co-discovery Learning is an adaptation of the most commonly used Think Aloud Protocol. In Co-discovery Learning, users are grouped in pairs and attempt to perform tasks together by talking aloud naturally to each other whilst being observed. They are to help each other in the same manner as they would if they were working together towards accomplishing a common goal using the product. They are encouraged to explain what they are thinking about while working on the tasks. Compared to Think Aloud Protocol, this technique makes it more natural for the test users to verbalize their thoughts during the test while retaining the great facilities of thinking aloud, pursuing of the users train of thought and notating erroneous assumptions about the system. It is also optimal to pair users who know each other so that they do not feel uncomfortable working together.

Co-discovery Learning is more realistic than a single user scenario, as people in work environments often work in teams. The users often find it easier and more natural to vocalize thoughts with a colleague present. The evaluators can also quantify the time taken for various tasks, the number of tasks competed correctly, the error frequency, numbers of times the users accessed the help system etc.[7]. These observations can form the ground to make more qualitative judgments such as the success or lack of the entire system, system sub-components, help system, effort required to achieve a particular result and quality of interface.

Usability principles covered:

Effectiveness: Yes, **Efficiency:** No, **Satisfaction:** Yes

Eye tracking

Eye tracking is the process of measuring either the point of gaze (where one is looking) or the motion of an eye relative to the head. Eye movement is typically divided into fixations and saccades – when the eye gaze pauses in a certain position, and when it moves to another position, respectively. The resulting series of fixations and saccades is called a scanpath. Scanpaths are useful for analyzing cognitive intent, interest, and salience while eye tracking in human-computer interaction (HCI) typically investigates the scanpaths for usability purposes.

There are numerous eye tracking techniques but the most popular and widely used are video-based eye trackers. A camera focuses on one or both eyes and records their movement as the viewer looks at some kind of stimulus. Most modern eye-trackers use the center of the pupil and infrared / near-infrared non-collimated light to create corneal reflections. The vector between the pupil center and the corneal reflections can be used to compute the point of regard on surface or the gaze direction. A simple calibration procedure of the individual is usually needed before using the eye tracker.

A wide variety of disciplines use eye tracking techniques, including cognitive science, psychology, HCI, marketing research, and medical research. Specific applications include the tracking eye movement in language reading, music reading, human activity recognition, the perception of advertising, and the playing of sports. More recently, eye tracking has become a key method to test usability of software. While traditional usability techniques are often quite powerful in providing information on clicking and scrolling patterns, eye tracking augments traditional usability methods by providing additional indisputable, objective and convincing data describing behavior and usability problems that the test participant cannot report and the researcher cannot observe. More specifically, it provides observers and testers with the ability to analyze user interaction between the clicks, how much time a user spends between clicks and unique information about first glance, search patterns and failed search. Eye tracking can be used together with a variety of research methods, including observations, interviews and Think Aloud Protocols. As a result it may yield valuable insight into which features are the most eye-catching, cause confusion or be ignored altogether as well as facilitate the assessment of navigation usability, distinctiveness, attractiveness and overall design.

Usability principles covered:

Effectiveness: No, **Efficiency:** Yes, **Satisfaction:** No

3.2.2.2 Inquiry

During usability test, evaluators need to obtain information about users' likes, dislikes, needs and understanding of the system by talking to them, observing them or letting them answer questions

verbally or in written form. The inquiry data collection methods can be divided into two categories, the traditional ones and the modern software-based ones.

3.2.2.3 Traditional approaches

Field Observation

Field Observation involves the visit of the usability evaluators to the users' workplace and observation of their work in order to understand how the users use the system to accomplish their tasks, if they use the system the way expected and what kind of mental model the users have about the system. However, field observation is time consuming, there is usually insufficient number of observations and the presence of observers may alter the behavior of the users and the working procedure in general.

Usability principles covered:

Effectiveness: Yes, **Efficiency:** No, **Satisfaction:** Yes

Focus Groups

This is a data collecting technique where about 6 to 9 users are brought together to discuss issues relating to the system. A usability evaluator plays the role of a moderator, who needs to prepare the list of issues to be discussed beforehand and seek to gather the needed information from the discussion. This can capture spontaneous user reactions and ideas that evolve in the dynamic group process. A serious consideration about Focus Groups technique is the skillfulness of the moderator who needs to be experienced in group facilitation and communication to make a focus group successful. It is not as simple as preparing questions since moderator needs to facilitate and guide discussion in real time. In addition, the data collected may possibly be biased, have low validity and be difficult to analyze because of their unstructured free-flowing nature and participants' inability to be candid.

Usability principles covered:

Effectiveness: Yes, **Efficiency:** No, **Satisfaction:** Yes

Questionnaires

Questionnaires have long been used to evaluate products since they provide answers to a variety of questions according to the needs. Moreover, questionnaires can be answered anonymously, allow time before responding, can be administered to many users at distant sites simultaneously and impose uniformity by asking all respondents the same questions. On the other hand, people can often express themselves better orally than in writing and informative questions take time to be developed and are not as flexible as interviews.

Questionnaires can be either "home grown" or measure against a benchmark of the use of standardized and publicly available surveys such as SUMI and WAMMI which are marked against a database of previous usability measurements. SUMI (University College Cork) is a brief questionnaire that is marked against a benchmark of responses to surveys of systems. WAMMI is an on-line survey administered as a page on the web site and users are asked to complete it before they leave the page. This gives ongoing feedback to continue monitoring how the web site is used. Each organization using the SUMI or WAMMI surveys send back their results to the Human Factor Research Group (HFRG) who provides statistical results from the database build of all SUMI/WAMMI

users. Other questionnaires specifically designed to access aspects of usability, the validity and/or reliability are the following: QUIS (Questionnaire for User Interface Satisfaction) developed by University of Maryland, PUEU (Perceived Usefulness and Ease of Use) developed by IBM, CSUQ (Computer System Usability Questionnaire) developed by IBM and PUTQ (Purdue Usability Testing Questionnaire) developed by Purdue University.

Usability principles covered:

Effectiveness: Yes, **Efficiency:** No, **Satisfaction:** Yes

Interviews

In this technique, usability observers formulate questions about the product based on the issues of interest. Then, they interview representative users to ask them these questions in order to gather the information desired. Interviews are flexible, suitable to get in-depth information for sensitive topics and allow the interviewer to pursue unanticipated lines of inquiry. On the contrary, interviews are time consuming and sometimes the interviewer can unduly influence the responses of the interviewee. The methods of interviewing include unstructured interviewing and structured interviewing. Unstructured interviewing methods are used during the earlier stages of usability testing. The objective of the investigator at this stage is to gather as much information as possible concerning the user's experience. The interviewer does not have a well-defined agenda and is not concerned with any specific aspects of the system. The primary objective is to obtain information on procedures adopted by users and on their expectations of the system. Structured interviewing has a specific, predetermined agenda with specific questions to guide and direct the interview. Structured interviewing is more of an interrogation than unstructured interviewing, which is closer to a conversation.

A useful technique to obtain further information after the original questions are answered is the use of probes. Probes are used to encourage the subjects to continue speaking, or to guide their response in a particular direction so a maximum amount of useful information is collected. Types of probes include:

- Addition probe encourages more information or clarifies certain responses from the test users. Either verbally or nonverbally the message is, "Go on, tell me more" or "Don't stop".
- Reflecting probe, by using a nondirective technique, encourages the test user to give more detailed information. The interviewer can reformulate the question or synthesize the previous response as a proposition.
- Directive probe specifies the direction in which a continuation of the reply should follow without suggesting any particular content. A directive probe may take the form of "Why is the (the case)?"
- Defining probe requires the subject to explain the meaning of a particular term or concept.

Usability principles covered:

Effectiveness: Yes, **Efficiency:** No, **Satisfaction:** Yes

3.2.2.4 Software based approaches

In recent years an increasing number of software tools involved in the usability evaluation process have emerged. These tools aim to automatically collect statistics about the detailed use of the

examined system, a process called logging. Logging is useful because it shows how users perform their actual work and enables effortless automatically collection of data from a large number of users working under different circumstances. Typically, a software product log will contain data about distance covered by mouse cursor, speed of cursor, use of keyboard, use of mouse button, total time of user activity, the frequency with which each user has used each feature in the product and the frequency with which various events of interest (such as error messages) have occurred. Moreover, same logging tools enable the capturing of screenshots and videos, storage of user activities in log files and creation, storage and implementation of macros. Such information can be used to optimize frequently used features and to identify the features that are rarely used or not used. Statistics showing the frequency of various error situations and the use of online help can be used to improve the usability of future releases of the system by redesigning the features causing the most errors and most access for online help. Some of the most popular logging software approaches are shown in the next table.

#	Software Name	Freeware
01	Mousotron Pro 5.0	YES
02	Mouse Off-road 2.15	YES
03	Mini-Input 2.0	NO
04	Mouse Odometer 4.0	YES
05	Mouse Meter 1.51	NO
06	My Mouse Meter 1.0.9	YES
07	Mouse Clocker 1.0	YES
08	Exact Mouse 2.0	NO
09	Usability Logger 2.3	YES
10	321 Soft Screen Video Recorder 1.05	NO
11	Screen VidShot 2.2.0.14	NO
12	ZD Soft Screen Recorder 2.6.4.0	NO
13	Screen Video Recorder 1.5	NO
14	Screen Tracker 2.0	NO
15	Advanced Key and Mouse Recorder 2.80	NO
16	Action Mouse Mover 1.0	NO
17	Adamant Key Mouse Pro 3.3	NO
18	Axife Mouse Recorder 5.0.1	NO
19	ECTI 1.73	NO
20	Mouse Tamer 2.0	NO
21	Smack 1.06	NO

22	Mouse Machine 1.1	YES
23	Jitbit Macro Recorder 3.82	NO
24	Mouse Master 2.1	NO
25	Macro Wizard 4.1	NO

3.3 Platform technical validation

To evaluate the ProsocialLearn platform it is necessary to perform a quantitative and qualitative analysis. The quantitative analysis includes tests of functional requirements and additional performance and stress tests.

On the other hand, the qualitative analysis concentrates on non-functional requirements like usability and portability. Examples of target non-functional requirements follow:

Usability:

- Ease of installation
- Ease of administration
- Comparison to existing solutions

Gaming provider and platform operators are interested in an easy-to-use solution. If the handling of ProsocialLearn is too complicated, this could impact of the level of acceptance of the solution.

Reliability: Once ProsocialLearn platform is operative, as specified, and delivered, the reliability characteristic defines the capability of the system to maintain its service provision under defined conditions for defined periods of time. One aspect of this characteristic is fault tolerance that is the ability of a system to withstand component failure.

Efficiency as the characteristic that is concerned with the resources consumed when providing the functionalities implemented by the ProsocialLearn platform.

Portability as the ability of the ProsocialLearn platform to run on different platforms.

3.3.1 Quantitative Analysis

The quantitative analysis mainly consists on the test of the functional and (some) non-functional requirements.

In addition to these requirements it is necessary to carry out stress and load tests for ProsocialLearn platform to show the performance in real environments.

3.3.2 Qualitative Analysis

Besides fulfilling the functional requirements it is essential for a later “market-ready” application of ProsocialLearn platform to meet non-functional requirements such as usability and portability.

For usability studies, for example, we directly observe how customers use technology (or not) to meet their needs. This provides the ability to ask questions, examine the behaviour and in case suggest changes to meet the objectives. In this case, differently from the quantitative analysis, the data analysis is usually not mathematical.

3.3.3 Methods

To check the criteria defined in the sub-sections 3.3.1 and 3.3.2, the following methods are adopted:

- Acceptance testing/Specification-based testing
- Load testing
- Performance/Scalability testing
- Stress testing
- Recovery testing
- Documentation testing
- Regression testing
- Long term testing
- Interoperability testing

An explanation of objectives and actions behind these methods is provided below:

Acceptance Testing/Specification Testing

Acceptance testing is usually an interactive test. Acceptance testing checks if the system meets the functional requirements as well as the non-functional requirements. A report is written specifying how close the system is to fulfil the requirements list and which changes are necessary to do so.

Load Testing

Load testing models the expected use of ProsocialLearn by simulating the simultaneous access from multiple users. During the load testing all actions and answers are monitored.

Performance/Scalability Testing

Performance of a system indicates the efficiency of the system while performing tasks. It includes total throughput of an operation as well as memory and disk space efficiency.

Stress Testing

Stress testing determines the behavior of the ProsocialLearn platform while the offered load is in excess of its designed capacity. The system is deliberately stressed by pushing it to and beyond its specified limits. Stress tests are targeted to bring out the problems associated with one or more of the following:

- Memory leaks.
- Buffer allocation and memory carving.

In terms of the project the stress testing will mainly focus to the ProsocialLearn components of the management service.

Recovery Testing

Recovery testing means the capacity to verify the recovery property of the ProsocialLearn platform during the failure of the software. It will be made in a variety of ways to verify that recovery is properly performed.

Documentation Testing

Documentation testing means verifying the technical accuracy and readability of the user manuals, including possible tutorials or online documentation. This test can be divided into two different sections

- Read Test: In this test documentation is reviewed for clarity, organization, flow, and accuracy without executing the documented instructions on the system.

- Functional Test: The instructions embodied in the documentation are followed to verify that the system works as it has been documented.

Long Term Testing

Long term testing is designed to ensure that the system remains stable for a long period of time under a high load. A system might function flawlessly when tested about some specific aspect, however, when a system runs for a long period of time without restarting, a number of problems are likely to occur: the system slows down, the system encounters functionality problems, the system silently fails over, and the systems crashes.

Interoperability/Portability Testing

Interoperability is the "ability to work with other systems"; in ProsocialLearn it means that we have to be able to guarantee component integration through the ProsocialLearn APIs on one hand, and the ability of ProsocialLearn platform to run on a different infrastructure.

3.4 Scientific effectiveness of prosocial games

The DoW states "Social exclusion is a key concept in Europe social policy, and both the Europe 2020 strategy and the Digital Agenda for Europe aim to ensure greater social cohesion and employment. Support for disengaged and disadvantaged learners, enhancing their employability and integration into society is a key. This includes helping people with learning disabilities, and young people to be more employable. **Children in danger of social exclusion, showing little to no signs of empathy and high levels of aggressive or anti-social behaviours should benefit from digital games tailored to teach prosocial skills that can help them achieve academically, appreciate team work and recognize the value of understanding other people's needs.**"

From this, we can identify three criteria for the selection of the participants in the longitudinal studies and three outcomes measures. In this section 3.4, we describe the criteria for selection, the outcome measures and we develop the methodology to be used to assess the scientific effectiveness of the prosocial games. We finish by summarizing the main points to follow to conduct the longitudinal studies.

3.4.1 Criteria for selection of our participants

The DoW identifies the following three criteria for selection: children in danger of (1) **social exclusion**, (2) showing little to no signs of **empathy** and (3) high levels of **aggressive or anti-social behaviours.**

3.4.1.1 Social Exclusion

The literature on social exclusion/inclusion is somewhat limited, particularly in children. In adults, a large variety of questionnaires can be used (See reviews in [17][19]) but they are limited to looking at employment status, social contact with workmates or community activism; none of these being translatable to a school setting. In children, there are a varieties of methods used to measure social inclusion but most of them are time consuming or not adaptable to our study. One research for instance reports recording 6 hours of video during play and meal times and coding these videos for negative interaction [21]. Such methodology is not feasible in the time frame delegated to testing in this project. Indeed, such methodology would be time consuming in terms of data collection and data analysis with the creation of new coding scheme etc. Another research reports using their own questionnaire that includes questions such as "I have many friends" or "I feel connected to my classmates" [22]. Although this questionnaire has not been tested for reliability, we could potentially

use some of the questions for our own evaluation. Another way of measuring social inclusion has been conducted by asking children to nominate who they like and dislike in their classroom or to write messages to their peers and later on coding the messages for presence of 'sincere compliments' or 'close relationship' [21]. However, this was conducted to see the inclusion of children with ADHD and it is not clear whether we would be able to see an improvement with typically developed children after playing our prosocial games.

However, a recent article [24] describing a framework for European action on child poverty and social exclusion opens new perspectives. This article defines children at risk of poverty and social exclusion (AROPE) if they experience one or more of three specific types of poverty: (1) monetary poverty, (2) material deprivation and (3) low work intensity.

(1) **Monetary poverty** is calculated as the family income. If it is lower than the national poverty threshold, then the children are at risk of monetary poverty.
(2) A recent article defined a new questionnaire to measure **material deprivation** [18] using 18 items such as 'the family cannot afford but would like to have : two pairs of properly fitting shoes; to have regular leisure activities; to have a computer and an Internet connection etc'
(3) **Low work intensity** is defined for children who live in a household where adults have worked less than 20% of their available work time in the previous years.

Although these criteria would help us define with precision the children at risk of social exclusion, asking hundreds of parents to fill in questionnaires might not be the best strategy for the time frame allocated to the longitudinal studies. Therefore, another approach would be to **select at the national level the poor areas and work within these schools**. Such information should be publicly available for each European countries where the testing will take place. The intervention could compare schools in high poverty areas vs. rich areas. The methodology described in 3.4.3 will develop this further.

3.4.1.2 Empathy

A questionnaire measuring empathy has been developed and tested in a variety of European countries (Germany, UK, Portugal) for the age group we are interested in (their questionnaire is for 8-14). It consists of 28 questions and measures affective and cognitive empathy [23]. We suggest that in each school, children willing to take part should fill in this questionnaire so we can identify the least empathetic children who should be targeted. Although these children should be targeted, we suggest that all children in the classroom participate in the intervention. The methodology described in 3.4.3 will develop this further.

Table 2: Scales of the Empathy Questionnaire.

Scale	*Item*
Cognitive Empathy	When I am angry or upset at someone, I usually try to imagine what he or she is thinking or feeling
	I can tell by looking at a person, whether they are happy
	I really like to watch people open presents, even when I don't get a present myself
	When I am arguing with my friends about what we are going to do, I think carefully about what they are saying before I decide whose idea is best
	I can tell what mood my parents are in by the look on their faces
	I notice straight away when something makes my best friend unhappy
	I can often guess the ending of other people's sentences because I know what they are about to say
	I often try to understand my friends better by seeing things from their point of view
	On the phone I can tell if the other person is happy or sad by the tone of their voice
	I often know the ending of movies or books before they have finished
	I think people can have different opinions on the same thing
	I can tell by the look on my parent's face whether it's a good time to ask them for something
Affective Empathy	It makes me sad to see a child who can't find anyone to play with
	Seeing a child who is crying makes me feel like crying
	Sometimes I cry when I watch TV
	It get upset when I see a child being hurt
	Some songs make me so sad I feel like crying
	When I see someone suffering, I feel bad too
	When I walk by a needy person I feel like giving them something
	It upsets me when another child is being shouted at
	When my parents get upset I feel bad
	I get upset when I see an animal being hurt

Figure 5 - Empathy questionnaire as in Zoll & Enz, 2005

3.4.1.3 Aggressive or anti-social behavior

A relatively recent questionnaire has been developed to measure aggressive or anti-social behavior in children [20]. It consists of 21 questions such as 'How often do you kick a classmate, say bad things to a classmate etc'. We suggest that in each school, children willing to take part should fill in this questionnaire so we can identify the most aggressive children who should be targeted. Although these children should be targeted, we suggest that all children in the classroom participate in the intervention. The methodology described in 3.4.3 will develop this further.

3.4.2 Outcome measures

As mentioned above, the DoW highlights three outcome measures: academic performances, team work and understanding other people's need.

Regarding **academic performances**, various methods could be used. One possibility to reduce the world load for the teachers and the children involved in the study would be to use tests (maths and reading tests) that are already part of the curriculum. This way, the children would not have to take any additional test and this would also not create any additional work for the teachers.

As a note, we do not have to have only maths and reading test. If the school is also using other test to measure academic achievement, we should include them if the teachers are happy to perform

these tests on their pupils. However, because it has to be consistent over countries, we suggest to have at least maths and reading test.

Regarding **team work**, various scenario could be used. We could develop a puzzle game where children play together and ask the teachers to rate the children's behavior. For instance, the scale could ask questions such as: 'Do they all participate? Y/N'; 'How many children are involved in solving the task?'; 'Do they shout or do they explain their ideas in turn?' However, this might be a lot of work for the teachers and we do not have the necessary resources to video record the children while they play and rate their behavior afterwards.

Finally, to measure the children's ability to **understand other people's needs**, different approaches could be used. First of all, we could ask the children to fill in the empathy questionnaire for the second time (the first time being used to select the children for the intervention). This way, we could see improvements that will be directly related to our intervention. Additionally, we could also measure the children's improvement on all the other core domains of prosociality by asking the teachers to rate this behaviour at the classroom level with some questions as detailed below. IT will be important to first debrief the teachers on to what we mean by each domain of prosociality.

Please rate all questions on a scale from 1 to 5; 1 being not true at all and 5 very true

1. I have the impression that the children in my classroom are more cooperative with each other
2. I have the impression that the children in my classroom are more trusting of each other
3. I have the impression that the children in my classroom are more fair to each other
4. I have the impression that the children in my classroom are more generous to each other
5. I have the impression that the children in my classroom are more compassionate to each other
6. I have the impression that the children in my classroom are more empathetic
7. I have the impression that the children in my classroom are showing less aggressiveness

Note:

We feel necessary to repeat a risk identified in D9.3 as it is relevant to the outcomes of the longitudinal studies. This risk concerns the ability to see an improvement in academic ability and social inclusion after only playing the game for a few months (each longitudinal study is 6 months). During each 6-month period, it is reasonable to think that the children will likely only play during 4 month (accounting for school holidays and actual testing phase). If they play 1 hour per week (which is what can be realistically expected), then we have a total of maximum 18 hours of sensitization to our prosocial objective. It is highly unlikely that this will be sufficient to see an increase in academic achievement or social inclusion. Report D2.1 suggests that increasing prosociality will in the long term increase academic achievement. Therefore, we hope to be able to see improvement in prosocial skills (even after such a short period) and from that we will infer improvement in academic achievement and social inclusion in the long term.

3.4.3 Methodology

3.4.3.1 Selection of the participants

As described above, the most accurate way of selecting the participants for the longitudinal studies would be to ask a large range of children and their families to complete a survey about their income, work intensity and material deprivation. This way, we could recruit the children who are the most at risk of social exclusion. However, such methodology would be extremely tedious and isn't feasible in

the time frame of this project. Therefore, we suggest to select two types of schools in poor and rich areas. This way, we can determine the type of context where our games are the most needed for future distribution.

Within each school, we also recommend to ask the children to complete the empathy questionnaire and the social aggression questionnaire. From these questionnaires, we should identify the children the more at risks within each classroom and see who benefit the most from our intervention. We recommend selecting 3 to 5 aggressive and low in empathy children and 3 to 5 non aggressive and high in empathy children. The teachers do not have to know how we selected these children if this can cause a problem with the ethics.

We also recommend testing children in different age group to cover the 7-12 age range. For instance, in the UK system, we recommend testing Year 3, 4, 5, 6 (primary school) and 7 (secondary school).

We also recommend testing schools in as many European countries as possible to see whether this has an influence. For this project, because of the variety of partners we have, the testing will most likely take place in Greece, Italy, Spain, FYROM, Lithuania, Bulgaria and UK.

Finally, we recommend having the games played in the whole classroom. We think that having the games as part of a lesson plan will help social cohesion and will help every students make the most of our intervention. Future reports should develop a methodology on how to include lessons plans, or teaching suggestions to accompany each game and help teachers plan a lesson around each games.

3.4.3.2 Intervention

For the intervention, we strongly suggest having a **control group** for each of the schools identified and each age group. This is because at these ages, children learn a lot in terms of academic abilities and socialization. Just by attending school, we would expect to find an increase in the variables measured (academic achievement, team work and understanding people's needs), whether the children play our games or not. Therefore, in order to demonstrate that our games increase these skills more so than by just attending school, we need to have a control group that will not receive any intervention. Not including a control group will prevent the generalizability of our data and might compromise our ability to sell these games into schools. However, getting teachers on board is already a tedious task and it will likely be even harder if we cannot even offer them a compensation with the games. We suggest a solution where half of the schools selected get the opportunity to use the games during the 6 month testing phase and to let the other schools use the games for the other half of the year (without investigating the effects; act as the control group). See Figure 5. However, such method has the disadvantage that we cannot guarantee that the testing will take place in the first 6 months of the academic year and the control school might not want to participate if they can only use our games for a few months (if we only start the testing in December for instance).

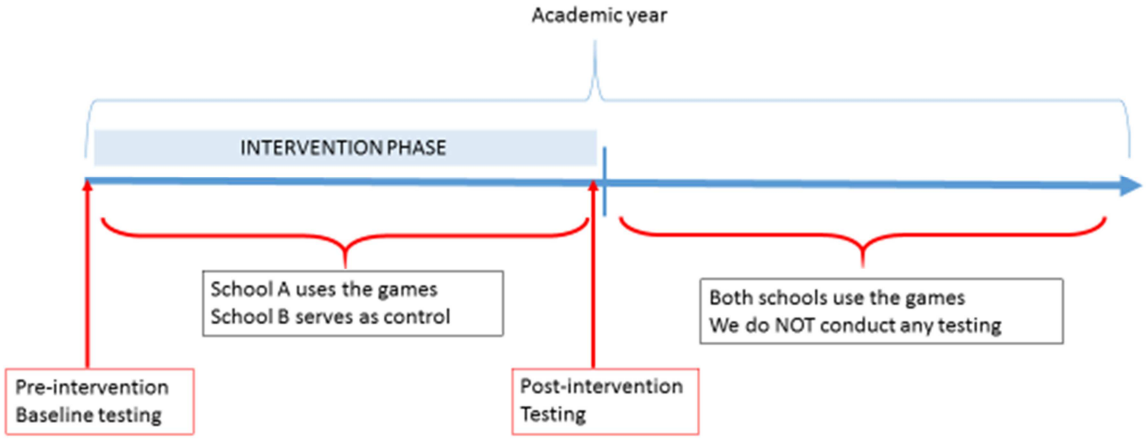


Figure 6 - Testing with a control group

An alternative for a control group could potentially be found by repeatedly testing the children/teachers on short 10-minutes questionnaires that would be filled in every week for the duration of the testing phase.

3.4.4 Summary

To summarise, below is the methodology that we recommend for the longitudinal studies. For each countries:

1. Get information at the national levels on rich and poor areas.
2. Select a few schools in these areas.
3. Select 5 age group within each schools (to test influence of age)
4. Ask the children to fill in the empathy and aggression questionnaires
5. Select 3 children high in aggression and low in empathy and 3 children low in aggression and high in empathy.
6. Pre-intervention testing. Ask the teachers to mark maths and reading tests. We might only want to look at the six children selected but we should keep a record of all children.
7. Intervention for only HALF of these schools. Give the schools access to the API platform to have access to our games. The other schools (controls) will not yet have access to the games.
8. Post-intervention testing. Ask the teachers to mark maths and reading tests. Also ask the children to fill in the empathy and aggression questionnaires. Ask the teachers if they saw improvements, at the classroom level, in the 6 domains of prosociality.
9. Give the game to ALL schools (so the control school gets something out of participating in this project).

With such a design, we would need to recruit:

2 (poor/rich areas) * 5 (age groups) * 6 (high/low empathy and aggression) * 2 (intervention/control) * 5 (countries) = 600 students; or 120 students per countries.

We might get more than 6 students in each classroom if we do this at the classroom level but we might only look at 6 of the children's data.

Note:

Additionally, we want to collect information about their:

- Gender
- Age
- Country of origin (culture)
- Personality (Agreeableness on the Big-5 in particular – 13 items)
- Attachment style (ASCQ Finzi-Dottan, 2012 -15 items)

Indeed, these variables might have an influence in the final model of prosociality we are designing. Document D3.2 describes the model and it will be made clear what additional information we want to collect during the longitudinal studies. This is not directly related to testing the effectiveness of the games but will test the effectiveness of our model so we thought it had to be included in the methodology of this section.

4 Assessment of Market Value Impact

4.1 Market Viability Tests

The ProsocialLearn D1.1 "Market and Competition Analysis" already contextualized a ProsocialLearn market. Besides, it analysed a number of relevant markets in order to draw insights for the development of the ProsocialLearn platform.

In order to develop the ProsocialLearn exploitation strategy, the next step is to find out the most appropriate business model for the ProsocialLearn solution. For this purpose, in the scope of the ProsocialLearn project we will implement a methodology to describe and assess business models through labeling value chain activities as revenue-generating / cost-generating activities.

Nevertheless, the ProsocialLearn business models have to be built upon one or more ***value propositions*** (be defined by the project month 18 in the deliverable D1.2), which define a specific value created by organizations using ProsocialLearn technology for the specific client-types, by defining where revenues are generated, what services are provided to whom and including appropriate definition of delivery and payment models.

In the context of this deliverable, as initial step, we have defined (listed below and described in the next sub-section) three possible exploitation routes and associated value chains to explore:

1. ProsocialLearn as a standalone and "exclusive" SaaS solution
2. ProsocialLearn as a SaaS Service that relies on a third party Marketplace functionalities
3. ProsocialLearn as a SaaS Service that relies on an own Marketplace as facilitator.

This initial version will be the baseline for an extensive work that will be delivered on month 18 in the deliverable D1.2. This analysis will also drive the definition of the most appropriate business model for the ProsocialLearn solution.

A further step of the ProsocialLearn exploitation strategy is to close analyze the business models of "who is going to pay for using the ProsocialLearn technology" (section 2.2.2 of the D1.1). In reality, these business models may be quite different among them due to the multiple stakeholders interested in the ProsocialLearn solution and the heterogeneity of their markets, i.e. in Europe, from country to country (but even in the same country) schools follow different purchase processes and strategies with different commercial routes.

This analysis will finally drive the ProsocialLearn exploitation route and the associated ProsocialLearn business model(s) (at this moment in the project, it is not excluded, ProsocialLearn may operate more than one business model).

4.1.1 Possible ProsocialLearn exploitation routes

Exploitation Route	Deployment scenarios description
ProsocialLearn as a standalone and "exclusive" SaaS solution	In this scenario the ProsocialLearn platform will be delivered as on-line SaaS solution in an exclusive and closed domain (i.e. Spanish Public Administration). The ProsocialLearn solution provider (solution operator) will be the market creator (who implement the ProsocialLearn market and operate the ProsocialLearn business model(s)): • It will offer to local public administrations a catalogue of prosocial games. The games are going to be off-line negotiated and integrated in the ProsocialLearn

	platform. • It will ensure the operability of the ProsocialLearn platform with a decent QoS/QoE monitoring the platform at runtime. **Strengths.** An exclusive-closed domain limits the "Privacy and Data Protection" concerns while theQoS/QoE could be easily maximized (using cloud technologies). The ProsocialLearn business model could be easily adapted(case by case) to different stakeholders (national public administrations, local public administrations, schools, etc) **Weaknesses.** The impact (on the gaming industry) will be certainly limited; the games catalogue will be limited as well. With a limited games catalogue the impact on the schools may be limited as well. The below picture show the value chain of this model.

Figure 7 - ProsocialLearn as a standalone and "exclusive" SaaS solution Value Chain

Exploitation Route	Deployment scenarios description
ProsocialLearn as a SaaS Service that relies on a third party Marketplace functionalities	As well as the first exploitation route, in this second scenario the ProsocialLearn solution will be delivered as on-line SaaS service in a closed domain. Unlike the previous case, the ProsocialLearn platform relies on "third party marketplace" (STEAM, Google App Market Place, Apple Market Place). In this context: • Games are physically offered by these third party marketplaces. • Games interact with the ProsocialLearn platform at runtime • The ProsocialLearn platform provider (market maker) delivers an operative ProsocialLearn platform (the added value prosocial functionalities) **Strengths.** As well as the previous example an exclusive-closed domain limits the "Privacy and Data Protection" concerns while the QoS/QoE could be easily maximized (using cloud technologies). This business model may fit

	with the School Business Model. **Weaknesses: The** ProsocialLearn operator/provider doesn't "control "the Market Place. ProsocialLearn is not the fully owner of the market. Performance issues at runtime. Issues with the integration of new Games (overhead with the integration).

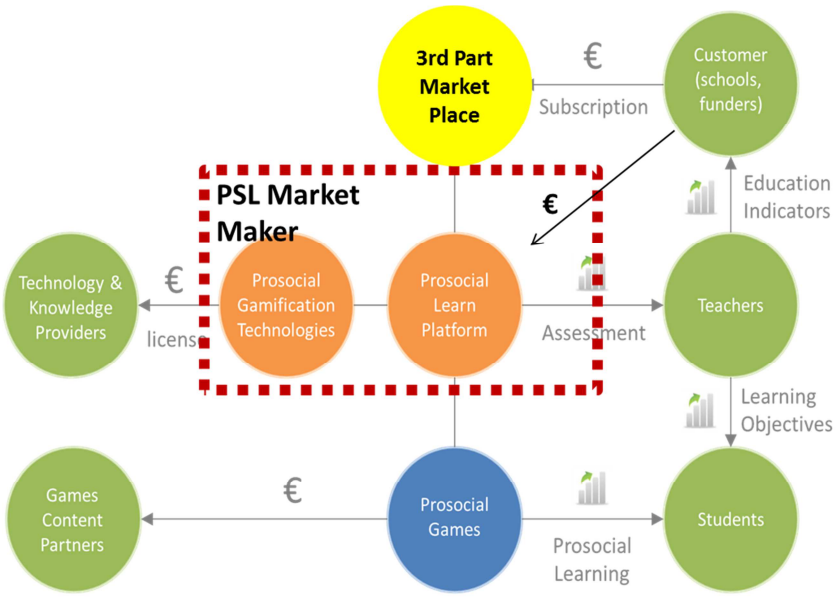


Figure 8 - ProsocialLearn as a SaaS Service that relies on a third party Marketplace functionalities Value Chain

Exploitation Routes	Deployment scenarios description
ProsocialLearn as a SaaS Service that relies on an own Marketplace as facilitator.	In this scenario the ProsocialLearn solution will be delivered as on-line PaaS/SaaS service to everyone is willing to consume prosocial games (the service is fully accessible via the web). • Gaming providers will consume the ProsocialLearn PaaS functionalities (through open APIs or a dashboard) to manage games in the system • Gaming providers may decide price/business models of their games will operate. • Schools/Local Administration can consume games registering and look for them in the ProsocialLearn Market Place. • Gaming providers are responsible (Liability) for the content of the games they offer. **Strengths:** this model may increase the impact of the ProsocialLearn technology on the gaming industry. layer to address Privacy and Data Protection. The solution should be full multitenant. The solution should provide a Service Level Agreement Management layer etc. *That Business Model may not be compatible with how schools acquire technological Educational Services.* **Weaknesses:** Without a gaming critical mass the marketplace may die. the ProsocialLearn solution Should be robust and secure as well as multitenant.

	The Business Model may not be fully compatible with how schools acquire technological Educational (school services. purchase processes).

Figure 9 - ProsocialLearn as a SaaS Service that relies on an own Marketplace as facilitator Value Chain

4.1.2 Testing viability of different market models

The appropriateness of these different models will depend on the attractiveness to games content partners and technology and knowledge partners on the supply side and schools/funders on the demand side.

There will be opportunities to test and compare the viability of these different market models through the work already planned on community management and evaluation. The importance of testing the overall market viability as well as the usability of individual games will be recognised in these plans. In this way, every opportunity is used to develop insights into the pros and cons of different business models and the factors that can determine which business model is preferable in a given context. Options include:

- Planned survey of schools in the Prosocial Learn community (as set out in D7.2) to cover procurement models and constraints e.g. whether existing software used in the school is purchased from exclusive-closed domain or marketplace, what marketplaces they use, any requirements a new provider must meet, at what level in their education system decisions about software purchase are made
- Induction events and workshops could explore initial preferences for different business models
- Small-scale experiments in schools to test feasibility could include interview with procurement lead to explore the perceived advantages and barriers to different business models.

4.2 Strategies to explore KPIs

The methodology to assess and control properly the quality and performance of the work carried out by the project was delivered in D9.3[2]. This section aims to quantitative and qualitative progress and refine the key Market Viability performance indicators related to the achievement of the project objectives, specifically:

Innovation: A ground-breaking gaming market for prosocial digital games targeting the education sector that relies on the innovation capacity of SMEs from the traditional gaming industry to produce engaging and exciting digital games for children.

Obj. 1

A new ecosystem must be established for student learning and skill acquisition based on Prosocial Gaming that channels creativity, innovation and technologies from the traditional gaming industry to the education sector. The traditional game industry is thriving with ideas and technical solutions that can directly compliment and benefit serious games, however, the financial risk to small game companies must be significantly reduced to incentivize new game productions by offering domain specific expertise, marketing and distribution channels for digital games. The perception that games are for entertainment must be overcome to increase acceptance of their use by teaching professionals in school curricula.

Name	Description	M12	M18	M24	M36	M36+
KPI 1	Commercial contacts established (according to ProsocialLearn data base)		25%	50%	100%	
KPI 2	Customers in the education sector willing to pay for prosocial games (according to contacts established KPI 1.1)			>3%	>10%	>20%
KPI 3	Distribution channels established towards the European education sector (pilots engagement)	Reach 2 schools			Reach 20 schools	Reach 50 schools
KPI 4	Number of games in catalogue	2		5	8	15
KPI 5	Size of the developer community including both leisure and serious games developer participation					5 developers
KPI 6	Public administration purchase's processes analyzed		5		10	
KP1 7	Elaboration of Business plan for the ProsocialLearn platform		YES		YES	

KPI 1.1: Commercial contacts established (according to ProsocialLearn data base). The consortium is working on the elaboration of a data base of public administration contacts (ministries of education, boards of education, schools board). The data base will be used to send commercial material about the project (factsheets, newsletter, links to demos etc)

[2]ProsocialLearn Consortium, D9.3 Risk Identification and Management and Quality Plan (2015)

KPI 1.2: Customers in the education sector willing to pay for prosocial games (according to contacts established KPI 1.1). A questionnaire will be sent to the data base contacts to gather information about their valorization regarding the outcomes of the project and also regarding their intention to pay for such service.

KPI 1.3: Distribution channels established towards the European education sector (pilots engagement): WP7 will conduct small experimental studies and longitudinal studies of prototype. The KPI measures the number of schools engaged in this process. KPI 1.4 Number of games in catalogue: number of games developed and ready to be included in the PSL platform

KPI 1.5: Size of the developer community including both leisure and serious games developer participation. This KPI will analyses the efforts to spread acknowledge, engagement and use of PSL platform by means of an active developers' community.

KPI 1.6: Public administration purchase's processes analyzed. As described in previous section there are relevant differences in how the schools can access to the PSL Platform. Administration purchase process differs from country to country and must be analyzed to be taken into account in the ProsocialLearn Business strategy development.

KPI 1.7: Elaboration of Business plan for the ProsocialLearn platform: two business and exploitation plan must be delivered on M18 and M36 to ensure the commercial viability and sustainability of the project's results.

5 Assessment of Ethics and Experiments Procedures

5.1 Evaluation protocols for experimental studies in schools

This section presents a concise but rather comprehensive draft of the experimental studies that are going to be conducted using the prosocial games developed within the framework of the project. The evaluation design of the studies will take into consideration the Runeson and Host approach [8], but will be suitably adapted to the needs of ProsocialLearn experimental studies. More specifically, the evaluation framework of the studies is the following:

- **Objectives**: The objectives of an experimental study are defined according to the evaluation of some major technical criteria, the usability/acceptability of the platform and prosocial games or the evaluation of the scientific effectiveness. In essence the objectives of the study should give a clear answer to the question: What to achieve?
- **Methodology:** The methodology engaged depends on the type of the experimental study, i.e., small scale or longitudinal, the PLO examined or the sensor technology supported by the game. The experimental study will include capturing and analysis of data using various sensor technologies and digital games with each one of them to support a different prosocial learning objective. The methodology should make clear what is studied in each experiment, where the researchers should seek the data and what methodology will be adopted for the evaluation of the collected data for technical and usability/acceptability validation as well as for the evaluation of the scientific effectiveness.
- **Data collection tools:** Critical technical information and usability/acceptability data will be acquired using some of the data collection techniques analyzed above. Here the evaluator should give answer to the question: How to collect data?
- **Type of actors:** Each experimental study will employ a sufficient number of users, students or teachers, in order to acquire ample information and elicit valid and meaningful conclusions.
- **Requirements examined and KPIs:** Experimental studies will be designed so as to examine and evaluate as many user and system requirements and KPIs as possible. In this direction, conducting all the small scale and longitudinal experimental studies entails the evaluation of the entire set of user and system requirements.
- **Evaluation Phase:** The evaluator should define here the evaluation phase and the objectives that will be fulfilled in each phase.

5.1.1 Small experimental studies

Small experimental studies will be conducted in operational or near operational school conditions during the preliminary and the first two evaluation phases. Depending on the evaluation phase, prototype prosocial games will be used to validate different functionalities e.g., player input modalities, multimodal data fusion, user modelling, adaptation algorithm, game mechanics etc. In addition, the prototype games will be used for the validation of user's acceptance and the optimization of ProsocialLearn platform.

5.1.1.1 Game 1: Path of Trust

Path of Trust (PoT) is a cooperative game where the objective is to collect treasure while navigating through a maze inside an Egyptian tomb, avoiding mummies and deadly traps. The player who assumes the role of wandering around (henceforth referred to as the Muscle) is attributed with Sensory Deprivation while the partner, unable to directly determine the course of movement, uses a top-down map view to navigate both of them safely through the maze, without being caught (henceforth referred to as the Guide). A sense of trust must be built between both players in order for the game to be completed; the Muscle player must trust their partner to provide guidance away from danger and the Guide must trust their partner to listen to directions. The two players are engaged in a multiplayer game where one is shown the 3D world, as shown in the Figure below, while the other is shown a top-down view of a 2D map. Both players have a treasure indicator on the right side of the screen which shows their individual progress in collecting treasure. Both start at 0 and have to reach the end goal. Whoever reaches the end goal first is declared winner of the game. Players are left to decide during gameplay if they shall work together to reap equal rewards or if they want to go out for themselves, endangering a spurious cooperation that might lead to both players' downfall.

The two players have to collaborate to collect treasure by avoiding traps and monsters lurking in the dark corridors of the tomb. They collect treasures (represented by diamonds) by having the Muscle touch them as he passes through the maze-like corridors. Unequal Pay is a game mechanic designed to introduce the element of competition and a desire to switch roles. It dictates that one player (e.g. the Muscle) is rewarded higher for accomplishing a task (i.e. collecting a treasure piece) than the other. Both players are meant to realize the benefits, as well as formulate a desire for re-routing resources. Hence, the mechanic of Switching Places, allows players to pass through a 3D Magic Portal, after which the character roles, gameplay, graphics and benefits are switched. As the weaker party at the end of the bargain (e.g. the Guide) is aware of when the opportunity to switch places presents itself, it's left up to the player to determine when to propose a bargain for the benefits to be exchanged. Likewise, it is up to the other player to evaluate the proposition and understand whether the offer was birthed out of a justified feeling of fairness or pure greed.

Figure 10 - The two screens of Path of Trust.

Game 1	**ID**	IG1	**Title**	Path of Trust
Objectives	1) Assessment of recognition accuracy of facial expression algorithm 2) Assessment of gaze analysis algorithm 3) Assessment of body motion analysis for emotion recognition			

	4) Assessment of recognition accuracy of gesture recognition algorithm for natural game interaction. 5) Assessment of data fusion algorithm 6) Assessment of adaptation algorithm 7) Platform technical validation
Methodology	The game aims to build up trustworthiness and teamwork among children aged 7-10. The evaluation consists of different steps. One step is related to the capture of users' motions through the natural game interaction module. These motions will be annotated and the recognition module will be tested based on these reference annotations. The second step concerns the evaluation of the input data modalities and data fusion algorithm. Again the data will be annotated by experts and the accuracy of the algorithms will be tested using the annotated dataset. For the evaluation of the adaptation algorithm, preference learning approaches will be applied. The final step of evaluation concerns the validation of the game in terms of usability, acceptability and effectiveness using traditional and software-based approaches.
Data Collection Tools	1) Metrics defined for player input data modalities, data fusion and adaptation algorithm (Objectives 1, 2 ,3, 4, 6, 7) 2) Use of data analytics collected by the game (Objective 5) 3) Use of Questionnaires (Objective 5 & 8)
Type of Actors (Number)	10-40 children aged 7-10 At least 3 researchers present
Requirements Examined (D2.3)	m.REQ1, m.REQ2, m.REQ3, p.REQ1, p.REQ2, p.REQ3, p.REQ4, eREQ1, eREQ4, eREQ5, uREQ1, uREQ7, uREQ11, uREQ12, uREQ18, uREQ20, uREQ27
KPI's (D9.3)	KPI 2.1, KPI 2.2, KPI3.2, KPI 4.1, KPI 4.2, KPI 5.1 KPI5.3
Evaluation Phase	• Preliminary evaluation phase: Objectives 1, 4 & 5. • First Evaluation Phase: Objectives 1-6 • Second Evaluation Phase: Objectives 1-7

5.1.1.2 Game 2: Kitty King's Candy Quest

Kitty King's Candy Quest (KKCQ) provides a set of scenarios that present specific moments for pairs of participants to make decisions of *Generosity* and *Fairness* in nature, as described in D2.2. KKCQ is a web-based, two-player game, focused on decision points that deal with prosocial concepts of fairness and generosity. There are four variations of the game, each one contained within the same game package.

A single gameplay cycle is broken down into the following player actions: at the start of the cycle, players complete a short round of collecting candy by clicking on a candy jar. One player is assigned the role of the *Giver*. This player gets all of the candy collected and has to decide how much to share with the other player, who takes on the role of the *Receiver*. The Receiver then decides if the sharing was done in a fair manner. A game consists of several cycles, involving different variants of the above situation with subtle variation that test different generosity and fairness attitudes and responses (i.e. a second variant allows both players to collect candy simultaneously, each player having his own

candy jar, while clicking contributes to a shared total). Each of the mini-games takes 1 or 2 minutes to complete.

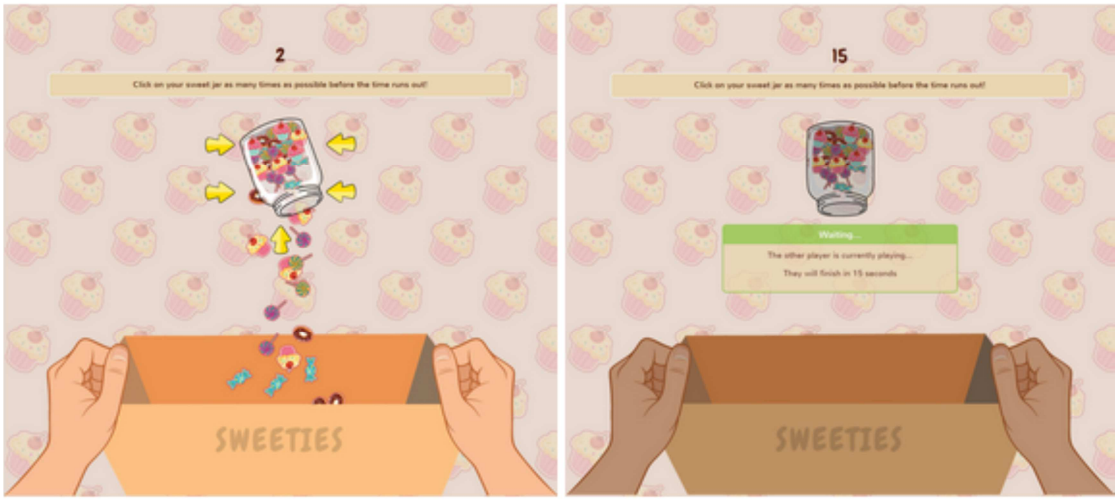


Figure 11 - Screenshots from the KKCQ game.

Game 2	**ID**	IG2	**Title**	Kitty King's Candy Quest
Objectives	1) Validation of the ProsocialLearn concept 2) Assessment of Voice Analysis Module 3) Platform technical validation			
Methodology	The goal of the study is to provide data for validating the central tenet of the ProsocialLearn concept. Namely to ascertain to what extent it may be possible to identify or measure prosocial intent or prosocial response through sensor data such as the webcam and microphones, as afforded by the project partners. For this purpose four inter-related minigames have been developed that provide specific instances of prosocial decisions and response actions. The so called Kitty King's Candy Quest game provides a set of scenarios that present specific moments for pairs of participants to make decisions of Generosity and Fairness in nature, as described in D2.2. These games are not intended to teach prosociality, rather they are experimental instruments designed to provide measurable moments where participants may exercise prosocial behaviour. By combining the sensor data with game data which elicit specific prosocial decision points and records the participants responses, it is hoped to establish a scientific basis for the project where there is currently little to no existing literature. The games must be played by the intended project audience (7 to 10 year olds). Ideally at least 30 pairs of participants would complete the experiment to provide a strong scientific basis for any usable generalizations. The researchers do not need to be present as the experimental instrument has been designed to take participants through the process, step by step, however the presence of at least one support member who is familiar with the system is highly advised. For more details about the experimental procedure see deliverable D7.2 "1st Experimental Planning and			

	Community Management".
Data Collection Tools	1) Use of data analytics collected by the game (Objective 1) 2) Video and audio record of each player's response at the end of each game round, about how they felt about their own or the other player's actions (Objective 1 & 3) 3) Metrics defined for player audio input data modality (Objectives 2)
Type of Actors (Number)	Minimum of 30 children (7 to 10 year olds) 1 researcher
Requirements Examined (D2.3)	m.REQ1, m.REQ2, m.REQ3, p.REQ1, p.REQ2, p.REQ3, p.REQ4, eREQ1, eREQ4, eREQ5, uREQ1, uREQ7, uREQ11, uREQ12, uREQ18, uREQ20, uREQ21,
KPI's (D9.3)	KPI 2.1, KPI 2.2, KPI3.2, KPI 5.1 KPI5.3
Evaluation Phase	• First Evaluation Phase: Objective 1, 2 • Second Evaluation Phase: Objectives 1-3

5.1.1.3 Game 3: Cooperative game

The Cooperative Game is based on cooperative mechanisms grounded on the theory of public goods game that considers costs/benefits of decisions associated with collective or individual action. The game aims to explore the definition of a "Cooperation" prosocial domain including how to measure cooperation and observe emotional affect. The goal of the game is for players to transfer the maximum amount of resource to an end point of a path where the resources are converted to private and collective benefits. Each player starts with resources and it is fixed that half of these resources will contribute to the personal good that will be translated in personal benefit at the end of the game and the other half will contribute to the collective goods that will be converted into the global benefit. Players must work together to avoid threats that reduce goods (both public and private). The game has four players. It is a turn based game with two dices rolled each turn. The result of the dice may move each player, may move a threat or both of them. On each turn, one player is in charge of deciding how to use the results of the dice. The decision may lead to three classes of movement: an individual movement, a collective movement (maximizing the collective benefit, for instance helping someone else), and a neutral movement. Each cooperative movement has a cost for a player. While the concept of cost is immediately clear to players, gaining an understanding that through cooperation the final benefit usually overcomes the cost will be part of the learning process. For instance the resource spent for performing a cooperative move may well be balanced by the fact that the move saves more resources belonging to another player, so globally preserving more Collective Goods.

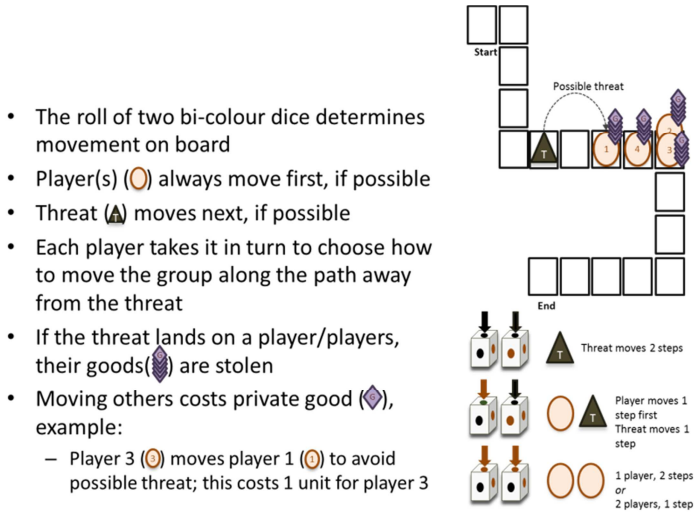


Figure 12 - Cooperation game logic

Game 3	**ID**	IG3	**Title**	Cooperative game
Objectives	1) Study the effect of voice interaction on collaborative behavior			
Methodology	The game is based on cooperative mechanisms grounded on the theory of public goods game that considers costs/benefits of decisions associated with collective or individual action. The game aims to explore the definition of a "Cooperation" prosocial domain including how to measure cooperation and observe emotional affect. The game can be played in two different modalities: with or without voice activated. Voice channel is designed as an independent channel that may stay always open or may be open/closed as necessary. Interaction through this channel is a way for sharing opinions across the players on the best strategy to follow. The voice will be used as a source of emotion observation. A game feature called mood feedback collector allows the students to provide feedback on their mood choosing from a finite set of options. A critical factor in the use of voice in a cooperation setting is the relative influence of group members. Without voice it is possible for individual decisions to be isolated, with voice some measure of influence would need to be measured to determine how much of the collective decision was related to a given player. For supporting experimentation and in particular removing the unpredictability of dice rolls results, it is possible to play the game with free dice rolls (i.e. each dice roll is unpredictable), or with fixed dice rolls (i.e. the sequences of dice rolls results along the game can be a-priory defined). This option is particularly welcome if/when multiple game runs with different configurations are used to perform tests. In this case it is possible to avoid the influence of random results, simply replicating them. Players, so they should remain uninformed of this specific aspect of the game.			

Data Collection Tools	1) Data will be collected through the game (Objective 1-2) 2) Metrics defined for player audio input data modality (Objectives 2)
Type of Actors (Number)	10-40 children aged 7-10 1 researcher present
Requirements Examined (D2.3)	m.REQ1, m.REQ2, m.REQ3, p.REQ1, p.REQ2, p.REQ3, p.REQ4, eREQ1, eREQ4, eREQ5, uREQ1, uREQ7, uREQ11, uREQ12, uREQ18, uREQ20, uREQ21,
KPI's (D9.3)	KPI 2.1, KPI 2.2, KPI3.2, KPI 5.1 KPI5.3
Evaluation Phase	• First Evaluation Phase: Objective 1 • Second Evaluation Phase: Objectives 1

5.1.1.4 Prototype Prosocial Games

PG and RK will develop two prototype prosocial games, which will support prosocial learning objectives associated with at least two prosocial skills. The games will build directly on the ProsocialLearn platform exploiting as many of the features as possible within gameplay. These games will be used in the second evaluation phase (M16-M24) during small experimental studies for the validation of the platform's performance as well as the optimization of WP3 and WP4 modules.

Game	**ID**	PS_SS	**Title**	Prosocial Game
Objectives	1) Validation/optimization of input modalities 2) Validation/optimization of data fusion and adaptation module 3) Evaluation of game's usability/acceptability 4) Platform technical validation			
Methodology	The games will support prosocial learning objectives associated with at least two prosocial skills. The experiments will run in operational or near operational conditions. Data will be collected to validate the performance of WP3 and WP4 modules, the platform and the acceptability of games.			
Data Collection Tools	1) Metrics defined for player input data modalities, data fusion and adaptation algorithm (Objectives 1, 2) 2) Use of data analytics collected by the game (Objective 2-3) 3) Use of Questionnaires (Objective 2 & 4)			
Type of Actors (Number)	30-40 pairs of children (7 to 10 year olds) per game			

Requirements Examined (D2.3)	m.REQ_X, p.REQ_X, e_REQ_X, u_REQ_X
KPI's (D9.3)	KPI 2.X, KPI 3.X, KPI 4.X, KPI5.1, KPI5.3
Evaluation Phase	• Second Evaluation Phase: Objectives 1-4

5.1.2 Longitudinal studies

Apart from the two prototype prosocial games that will be developed by PG and RK, three additional prosocial games will be developed by the SME game companies joining in the third year of the project. Each SME is expected to develop a game targeting a specific prosocial learning objective defined by teaching professionals. The games will be used for longitudinal studies in European schools.

The first set of studies will commence in M25 using games developed by PG and RK while the second set of studies will commence in M31 and will be based on games developed by partners involved in the third year. All studies will be conducted using a mature platform and tested in real conditions (schools).

	ID	PS_LS	**Title**	Prosocial Games
Objectives	1) Prosocial learning outcomes of students 2) Evaluation of improvement in social inclusion 3) Evaluation of improvement in academic performance			
Methodology	Five prototype prosocial games will be used during the longitudinal studies. The studies will start in the final year and each will be 6 months in duration. Experiments will be repeated in the same settings at frequent intervals of 8-12 weeks and reports will be gathered by the platform (QoE/S, affective/game-related cues), while feedback will be gathered from the tutors as well. Specifically, feedback will be received in the form of questionnaires before a student plays the game for the first time and about one month after the session, in order to evaluate whether results are sustained.			
Data Collection Tools	1) Use of data collected by the platform (Objective 1) 2) Use of Questionnaires (Objective 1)			
Type of Actors (Number)	20 - 30 children (7 to 10 year olds) per study (15 studies in total) 10-20 teachers in total			
Requirements Examined (D2.1)	m.REQ_X, p.REQ_X, e_REQ_X, u_REQ_X			
KPI's (D9.3)	KPI 1.X, KPI 2.X, KPI 3.X, KPI 4.X, KPI5.2, KPI5.3			

Evaluation Phase	Third Evaluation Phase: Objective 1, 2, 3

5.2 Assessment of ethical procedure

Deliverable D7.1 *ProsocialLearn ethical oversight procedures* provided a detailed description of the procedures to be used in ensuring appropriate ethical oversight of the various trials. In this section, we will summarise the relevant tenets of those procedures along with a brief update as this relates to an assessment of the ethical procedures described. Although the basic principles are well understood – respect for participants, the assumed beneficence of project results, and the equitable distribution of any beneficial outcomes [25] – the inclusion of minors of necessity presents specific concerns. For instance, how might the effects of attendant power relations influence consent and participation [26]? But at the same time, we should recognize that any ethics procedures should be flexible enough to balance participant and research interests [27]. In this context, ProsocialLearn has set out three constructs:

1. The *Ethics Management Board* is responsible for overall ethics oversight;
2. The inclusion of an *Ethics Advisory Board* made up of external experts in the field provides additional and *quasi*-independent checks; finally
3. Since much of the debate around ethical conduct in research relates to participant data handling, a *Privacy Impact Assessment checklist* has also been provided as a quick reference list of the main considerations in reviewing proposed trials

These are summarized in the following sections. In the final section of this Chapter, the validation process for these structures is described. These outcomes should provide additional checks for the structures put in place and described in the other sections of the Chapter.

5.2.1 Ethics management board

The role of the Ethics Management Board (EMB) is the overall oversight of the ethical execution of the proposed research activities in ProsocialLearn. Members include:

- A chairperson, rotated bi-annually between the partners
- Each work package leader
- An external, advisory board of experts (Section **¡Error! No se encuentra el origen de la referencia.**)

The EMB is responsible to ensure the overall compliance of the project with legal and ethical guidelines, as well as to review the internal (is it being properly run?) and external (will it deliver societally beneficial outcomes?) validity of the trials. A set of general ethical principles have been defined, including consent, confidentiality and the traditional data controller and data processor roles; as well as an overview of the processes they will follow (meetings, responsibilities, etc.).

5.2.1.1 Ethics Advisory Board and external experts

Not least because of potential conflicts of interest, as well as an inherently vulnerable population [28], the EMB includes a semi-independent advisory board of three experts; although paid *pro rata* by the project, they are not main beneficiaries of the project or its outcomes. They provide specific expertise for research involving minors, education and security & privacy. It is their responsibility primarily to advise the EMB when specific concerns or questions arise.

5.2.2 Privacy Impact Assessment checklist

In addition to the general ethical principles, the EMB has also defined a set of specific items which should be used as a framework against which the management and execution of trials should be verified. These are summarized in **¡Error! No se encuentra el origen de la referencia.** below.

AREA	#	DESCRIPTION
General ethical issues	9	These include general topics which should form part of explicit *participant information* provided to potential participants in support of their decision to take part or not. In addition, the checklist would provide a solid basis for any external ethical approval if required and sought[3].
Location data issues	11	These cover the main data management procedures, including access to the data, pseudonymisation, and curation.
Profiling issues	6	These constrain the extent of and information included as part of any profiling activity.
Tracking issues	2	These relate to the specific limitations and provisions around tracking or identifying location, especially in respect of persistence across different games.
Consent issues	4	These cover the management of informed consent, including the mechanisms to record it and the form it should take.
Anonymisation issues	2	Anonymisation relates to the process to obscure identification as well as any related issues of storage location.

As well as conforming with the relevant provisions of Directive 29/46/EC4 and the Data Protection Working Party, as stated, these provide a guided structure for the trials and how they should be handled.

5.2.3 Ethics assessment and validation

Checks and balances are therefore in place within ProsocialLearn to ensure overall management and guidance on ethics (Section 5.2.1), including suitable semi-independent expertise in an advisory capacity given the nature of the work being carried out (Section **¡Error! No se encuentra el origen de la referencia.**), and a set of guiding principles to help structure the trials themselves (Section **¡Error! No se encuentra el origen de la referencia.**). Thus, the project works with reference to the checklist and guidance provided, under the oversight of the EMB, who may call upon specific independent expertise for advice and arbitration if required. This process is already in operation and helping support the trials as they are developed.

[3] See, for instance, http://www.research-integrity.admin.cam.ac.uk/research-ethics/ethical-review and http://www.southampton.ac.uk/ris/policies/ethics.html

[4] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML

Above and beyond this, there is a need for project-wide validation that this process is both fit for purpose and working as outlined. To this end, one partner will take responsibility for reviewing all the workings and decisions of the EMB annually, at M12, M24 and M36. They will produce a summary of activities, along with a comparison against the relevant checks and procedures outlined in D7.1. The summary will be cross-checked by each of the External advisory board for each of the separate reports (i.e., they will validate one summary each). Each summary will highlight any concerns, along with the agreed mitigation for them. Finally, the summaries will be included as an Appendix in each of the iterations of D7.3 for external review.

6 Conclusions

This deliverable presented a detailed assessment plan, for the evaluation of the system's independent modules, the integrated platform and the prosocial games. In particular, it defines the evaluation strategy for the game effectiveness, market value impact and ethics procedures to drive detailed planning of technical validation, short and longitudinal studies and market viability tests. The methodology adopted in this deliverable includes a preliminary evaluation phase during the development phase of WP3 and WP4, as well as three successive evaluation phases aiming to provide a multilateral assessment process covering the technical validation of the platform and the proposed technology, as well as the overall scientific effectiveness of the games. More details on the experiment planning will be described in D.7.2 "1st Experiment planning and community management" (Mont 9), D.7.3 "2nd Experiment planning and community management" (Month 15) and D.7.4 "3rd Experiment planning and community management" (Month 27), while the evaluation results will be presented in deliverables D7.8 "1st Results of small experimental studies", D7.9 "2nd Results of small experimental studies", D7.10 "1st Validation activities in operating school conditions" and D7.11 "2nd Validation activities in operating school conditions".

7 References

[1] M. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, "The first Facial Expression Recognition and Analysis Challenge," Proc. Int'l Conf. Automatic Face and Gesture Recognition, 2011.

[2] G. T. Papadopoulos, K. Apostolakis, P. Daras, "Gaze-based Relevance Feedback for Realizing Region-based Image Retrieval", IEEE Transactions on Multimedia, Vol.16, No.2, pp.440,454, Feb. 2014.

[3] ISO 9241-11:1998, Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability, Retrieved from http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=16883.

[4] ISO/IEC 9126:1991. Information Technology - Software Product Evaluation - Quality Characteristics and Guidelines for the User.

[5] Lewis, C. H. (1982). Using the "Thinking Aloud" Method In Cognitive Interface Design (Technical report). IBM. RC-9265.

[6] Ericsson, K. & Simon, H. (1993). Protocol Analysis: Verbal Reports as Data (2nd ed.). Boston: MIT Press.

[7] Doubleday, A., Ryan, M., Springett, M., & Sutcliffe, A. (1997). A comparison of usability techniques for evaluating design. Proceedings of DIS 97. 101-110. New York.

[8] Runeson, P. &Höst, M. (2008). Guidelines for Conducting and Reporting Case Study Research in Software Engineering. Empirical Software Engineering, 14(2): 131-164.

[9] MSRC-12 Dataset: http://research.microsoft.com/en-us/um/cambridge/projects/msrc12/

[10] G3D Dataset: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6239175

[11] MSR Action3D Dataset: http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/

[12] D. Keltner, A. Kogan, P.K. Piff, S.R. Saturn (2014). The sociocultural appraisals, values, and emotions (SAVE) framework of prosociality: Core processes from gene to meme. Annual Review of Psychology, 65: 425-460

[13] M. Stirrat, D.I. Perret (2010). Valid facial cues to cooperation and trust: Male facial width and trustworthiness, Psychological Science, 21(3): 349-354

[14] M. Dyck, M. Winbeck, S. Leiberg, Y. Chen, R.C. Gur, K. Mathiak (2008). Recognition profile of emotions in natural and virtual faces, PLoS ONE 3(11): e3628

[15] M. McRorie, I. Sneddon, G. McKeown, E. Bevacqua, E. de Sevin and C. Pelachaud (2010). Evaluation of four designed virtual agent personalities, IEEE Transactions on Affective Computing 3(3): 311-322

[16] A. Kleinsmith, P.R. De Silva, and N. Bianci-Berthouze (2006). Cross-cultural differences in recognizing affect from body posture. Interacting with Computers, 18(6): 1371–1389

[17] J. N. Baumgartner, & J. K. Burns, (2014). Measuring social inclusion-a key outcome in global mental health. International Journal of Epidemiology, 43(December 2013), 354–364. doi:10.1093/ije/dyt224

[18] Commission European, & Eurostat. (2012). Measuring material deprivation in the EU. Working Paper. doi:10.2785/33598

[19] T. Coombs, A. Nicholas, & J. Pirkis, (2013). A review of social inclusion measures. The Australian and New Zealand Journal of Psychiatry, 47(10), 906–19. doi:10.1177/0004867413491161

[20] P. Gremigni, B.F. Damasio, & J.C. Bors, (2013). Development and validation of a questionnaire to evaluate overt aggression. Psicologia: Reflexao E Critica, 26(2), 311–318.

[21] A. Y. Mikami, M.S. Griggs, M.D. Lerner, C.C. Emeh, M.M. Reuland, A. Jack, & M.R. Anthony, (2013). A randomized trial of a classroom intervention to increase peers' social inclusion of children with attention-deficit/hyperactivity disorder. Journal of Consulting and Clinical Psychology, 81(1), 100–112. doi:10.1055/s-0029-1237430.Imprinting

[22] G. F. Welch, E. Himonides, J. Saunders, I. Papageorgi, & M. Sarazin, (2014). Singing and social inclusion. Frontiers in Psychology, 5(July), 1–12. doi:10.3389/fpsyg.2014.00803

[23] C. Zoll, & S. Enz, (2005). A Questionnaire to Assess Affective and Cognitive Empathy in Children. Journal of Child Psychology, 15, 165–174.

[24] R. Davis. Child poverty and social exclusion: A framework for European action. Library Briefing. Library of the European Parliament. June 14th 2013.

[25] The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (n.d.) Ethical Principles and Guidelines for the Protection of Human Subjects of Research "The Belmont Report". Retrieved from: http://www.hhs.gov/ohrp/policy/belmont.html; accessed on 25.ix.15.

[26] Morrow, V., & Richards, M. (1996). The ethics of social research with children: An overview1. *Children & society, 10*(2), 90-105. doi:10.1111/j.1099-0860.1996.tb00461.x

[27] Wiles, R., Crow, G., Charles, V., & Heath, S. (2007). Informed consent and the research process: Following rules or striking balances? *Sociological Research Online, 12*(2). doi:10.5153/sro.1208

[28] London, L. (2002). Ethical oversight of public health research: can rules and IRBs make a difference in developing countries? *American Journal of Public Health, 92*(7), 1079-1084. doi:10.2105/AJPH.92.7.1079

Appendix 1 - Questionnaire for Social Inclusion

SOCIAL INCLUSION

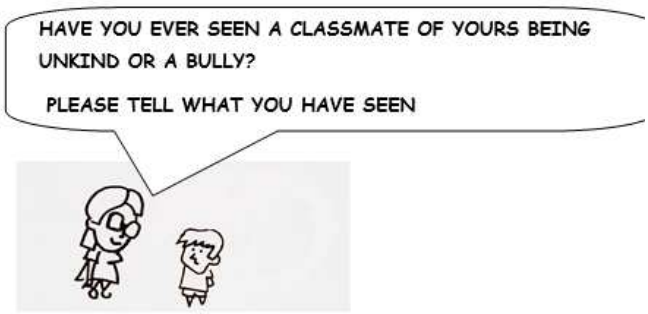
Children

Three options: True, I don't know, Not true

1. I have many friends
2. I feel connected to my classmates
3. I know how to be with other people

The questions below are from a Daphne project, namely Prosave http://www.era-edu.com/csfvm/ProSAVE

4. SOMETIMES, AT SCHOOL, SOME CHILDREN ARE UNKIND TO OTHER CHILDREN …



5 How do you feel in your class?

My class is:

1. A place where I feel protected

2. A place where I do not feel protected

3. A place where I meet my friends

4. A place I necessarily have to go to

5. A really nice place!

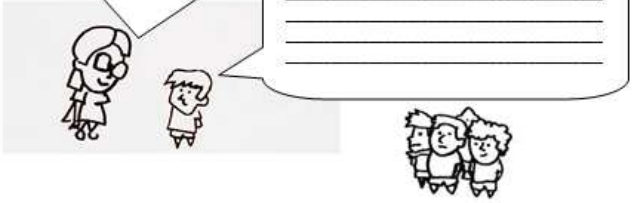
6. A place not for me

6

When you are at school, do you ever happen to remain on your own during recess time ?

_____
_____
_____
_____
_____

7 What are those unkind things you wouldn't want your classmates to do to you?



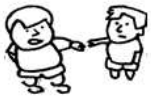
_____
_____
_____
_____

Please tell us what you think of these stories:

Scene 1. A child is mocked by some classmates …

1. The child isn't offended and laughs …

2. The child gets angry and mocks one of the children …



3. The child is sad and doesn't say anything …

Scene 2. A little girl invites everybody to her birthday party except a little boy/girl …

1. The child thinks "What a luck I'm not going, I've got sports! …"

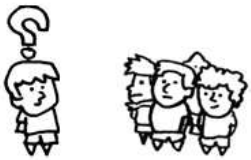
2. The child asks the girl why he/she wasn't invited …

3. The child is sad and doesn't say anything …

Scene 3. A group of children are talking together in a low voice. They're gossiping about a classmate of theirs … who is standing aside with his head lowered …

1. The child goes to his/her classmates and happily starts talking with them …

2. The child asks his/her classmates what are they talking about …



3. The child is sad and doesn't say anything …

Scene 4. A real bully notices the pen of another child …

1. The child gives him his/her pen but asks for another one in exchange …

2. The child doesn't give him the pen and answers back in a rude manner …

3. The child is sad and doesn't say anything …

Teachers

On a scale of 1 to 10, rate pupil X

1. X has many friends
2. X is generally excluded from the other children during playtime
3. X is happy to be on his/her own
4. X can ask his/her classmate for help
5. X looks like he is connected to his/her classmates
6. X knows how to be with other people
7. X is generally socially competent
8. X can solve relationship problems on his/her own without the intervention of a teacher
9. X heavily relies on teacher or adults to solve conflicts

ACADEMIC ACHIEVEMENTS

Teachers

1. Do you think X improved in reading skills?
2. Do you think X improved in mathematical skills?

Additionally, we should ask the teachers to provide us with 'test/exam' results from before and after the intervention. They should not do additional tests, just use the one they would do to measure academic achievement in their classroom. This will likely be different between countries and even between class room but we will just measure percentage of improvement between before and after the intervention.

PROSOCIAL LEARNING OBJECTIVE

Children

These questions should be open ended and presented with pictures like it is the case in the Daphne project above.

1. Can you tell me about the last time you worked with one of your classmate? What did you like/dislike about it?
2. Can you tell me about the last time you trusted a classmate with your things? Like asked a classmate to hold a precious marble while you do your shoelaces? What did you like/dislike about it?
3. Can you tell me about the last time you shared something with a classmate/friend? For instance if you shared a cookie or part of your lunch? What did you like/dislike about it?
4. Can you tell me about the last time you helped a friend who was feeling a bit sad? Did you realize he/she was feeling sad? What did you do? What did you like/dislike about it?

Teacher

On a scale of 1 to 10, rate pupil X

1. X demonstrates awareness and understanding of cooperation
2. X demonstrates awareness and understanding of trust

3. X demonstrates awareness and understanding of fairness
4. X demonstrates awareness and understanding of generosity
5. X demonstrates awareness and understanding of other's emotional state
6. X demonstrates awareness and understanding of compassion

Ask the children to play a puzzle game and ask the teacher to rate their behaviour using this scale:

1. Do they all participate or is there one (or more) children excluded?
2. Do they shout or do they explain their ideas in turn?

Appendix 2 - Technical Assessment Report Template

Technical Report ID	
Date	
Module/Entity Tested	
Test Leader	

Summary

Summarize what item was tested, what features or combination of features were tested, how the item was tested, what was the approach, what were the main things that happened, what resources were used (tools, people, time)

Variances

If any test items differed from their specifications, describe that. If the testing process didn't go as planned, describe that. Say why things were different

Results

Assessment Indices results

Assessment Category	**Assessment Indices**			
	ID	**Description**	**Desirable Value or Fail/Pass criteria**	**Value**

Other results

Evaluation of Results

How good are the test items? What's the risk that they might fail?

Appendix 3 - Experimental Study Evaluation Report Template

Experimental Study Report ID	
Date	
Prosocial Game Tested	
Experimental Study Responsible	

Introduction

Provide a brief introduction of the experiment

Objectives

Describe the objectives of the experiment

Methodology

Describe the experiment, how it was conducted, how were the targets evaluated, how were the data collected, what resources were used (tools, people, time)

Variances

In case of unmet objectives describe the reasons that led to this variance.

Results

Describe the results of the experiment

Evaluation of Results

Provide comments on the results